WHAT CAN A TASK CHECKPOINT IDENTIFY? INVARIANT TASK-CORE FUSION FOR MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging combines independently fine-tuned experts into one model, but interference can degrade performance. Without task data, data-free methods rely on checkpoints and often use task vectors, the parameter changes induced by fine-tuning, as proxies for inaccessible task data. In a controlled linear setting, we show that different training data can produce the same final weights. A task vector therefore does not uniquely characterize its originating data, and treating checkpoint geometry as if it faithfully reflected the original task data can misguide fusion. We propose Invariant Task-Core Fusion (ITCF), which builds basis-invariant task cores from checkpoint-identifiable directions and uses core overlap to capture shared support and reconcile expert updates. Across eight vision and language configurations and more than 500 task compositions, ITCF remains competitive across model families and merge scales, including full-model merges of up to 166 experts. Core fusion is roughly two orders of magnitude faster than iterative optimization, with no additional parameters or inference cost.

1 INTRODUCTION

With the widespread adoption of pretraining and fine-tuning, a single foundation model can be adapted independently to many downstream tasks, producing experts with different specialized capabilities. Deploying all of these experts, however, increases storage and inference cost, while joint multi-task training typically requires renewed access to the task data. Model merging offers another option: it combines independently fine-tuned experts into one model without joint retraining or retaining the full expert collection.

A central difficulty in model merging is interference. Parameter updates learned for different tasks can conflict after combination, degrading capabilities that the individual experts had acquired. Existing methods mitigate this interference in different ways. Task Arithmetic combines task vectors to integrate task-specific updates, while TIES and DARE select, remove, or rescale parameter changes to reduce conflicts (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). Other methods use additional task information to guide fusion: Fisher merging estimates sensitivity from task data, RegMean uses task activations, and AdaMerging adjusts merge coefficients with unlabeled examples (Matena & Raffel, 2022; Jin et al., 2023; Yang et al., 2024). Such information can help determine how experts should be combined, but it is not always available once task data can no longer be accessed.

This has motivated *data-free model merging*, which aims to merge models without task data, using only the pretrained model and final expert checkpoints. In this setting, obtaining useful task information from the available weights becomes especially important. Recent methods exploit richer structure in task vectors, the parameter changes induced by fine-tuning. TSV-Merge and Iso-C use task-vector subspaces and spectral structure, respectively, to reduce task interference (Gargiulo et al., 2025; Marczak et al., 2025). WUDI goes further by using task vectors as substitutes for inaccessible training inputs and matching the merged model's outputs to expert outputs through these proxies (Cheng et al., 2025). These methods show that even when the original task data are unavailable, final checkpoints can still provide information that is useful for fusion.

However, these approaches raise a more basic question: *to what extent can structures observed in task vectors reliably represent inaccessible task information?* We show that this correspondence need

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.

not hold. Even in linear regression, different training data can lead to exactly the same final weights. Before extracting increasingly rich structure from task vectors, it is therefore necessary to determine which input-related information is actually determined by the final checkpoint.

We identify a clear boundary. The final checkpoint can determine which input directions are involved in a parameter update, but it cannot further determine their relative importance for the original task. This leaves a direct question: *when those relative importances cannot be obtained from the final weights, how should merging measure the output discrepancy between the merged model and an expert along these directions?* For the retained directions, we prove that, over all task-input structures consistent with the same final checkpoint, comparing these directions with equal strength is the unique choice that minimizes the worst-case distortion of this output comparison.

Based on this result, we propose Invariant Task-Core Fusion (ITCF). ITCF defines a task core for each expert according to the comparison principle above, specifying how the merged model should be compared with that expert within the information supported by the checkpoint. When multiple experts constrain the same input directions, their task cores overlap, and ITCF uses this shared structure to reconcile the output changes requested by different experts. We further show that weaker shared support makes a correction of the same size more costly, so ITCF acts more conservatively on weakly supported directions and follows the jointly coordinated result more closely where shared support is strong. The transition between these behaviors is determined entirely by the final checkpoints. The entire construction uses only the pretrained weights and final expert checkpoints: it requires no task samples, labels, activations, calibration set, or validation-selected merge coefficients.

Finally, we evaluate ITCF across vision and language model configurations, including CLIP ViT-B/32, ViT-B/16, T5-base, T5-large, Qwen2.5-1.5B, Llama-3.2-1B, and NLLB-200 distilled-600M and distilled-1.3B (NLLB Team, 2024), covering more than 500 task compositions and merges of up to 166 experts. ITCF delivers clear performance gains across multiple vision and language configurations, with the largest gain approaching 10 percentage points. In large-scale full-model merging, ITCF scales from 8 to 166 experts and achieves the highest mean performance at every tested scale. Its timed core-fusion procedure is also roughly two orders of magnitude faster than iterative optimization, while adding no model parameters or inference computation, demonstrating the effectiveness of ITCF in checkpoint-only merging in terms of performance, scalability, and computational efficiency.

Our contributions are:

- We characterize a fundamental information boundary in checkpoint-only model merging: different training data can produce the same final weights. The final checkpoint determines which input directions are involved in a parameter update, but not their relative importance for the original task.
- Starting from this information boundary, we derive ITCF. For each expert, we prove that equal-strength comparison over the retained directions is the unique most robust choice under the considered uncertainty, yielding a task core. Across multiple experts, overlapping task cores coordinate requested output changes, while our further analysis shows that weaker shared support leads to a more conservative correction whose strength is determined directly from the final checkpoints.
- We evaluate ITCF across vision and language model configurations, more than 500 task-composition settings, and full-model merges of up to 166 experts. The results demonstrate strong performance, scalability, and computational efficiency: core fusion is roughly two orders of magnitude faster than iterative optimization, with no additional parameters or inference cost at deployment.

2 Related Work

Information access and expert preparation. Model merging methods can first be distinguished by the information they can access and the stage at which they intervene. Fisher merging, RegMean, and AdaMerging use data-derived sensitivity, task activations, and unlabeled examples, respectively (Matena & Raffel, 2022; Jin et al., 2023; Yang et al., 2024). Another line acts earlier, during expert training, to improve later mergeability. MergOPT explicitly accounts for future merge-induced

parameter offsets during fine-tuning, while SAFT-Merge uses sharpness-aware optimization to reduce subsequent parameter interference (Yang et al., 2026; Lee et al., 2025). In contrast, ITCF addresses the setting in which experts have already been independently fine-tuned and the original task data are no longer accessible, using only the pretrained model and final expert checkpoints.

Parameter- and checkpoint-structured merging. Without retraining the experts, early methods primarily operate directly on parameter updates. Task Arithmetic combines task vectors, while TIES, DARE, and PCB use parameter selection, sparsification, rescaling, or rebalancing to mitigate conflicts (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Du et al., 2024). Later methods exploit matrix and subspace structure: TSV-Merge uses the singular structure of layer-wise task matrices to reduce interference, Iso-C rebalances shared and task-specific spectral structure, and APGD coordinates multiple tasks through data-free projective optimization (Gargiulo et al., 2025; Marczak et al., 2025; Wei et al., 2025). A more closely related line seeks input-related information from checkpoints themselves. ACE-Merging and ACTMat show that update-derived second-order structure can be informative about task inputs under additional modeling assumptions, while CSP uses a core subspace of the pretrained model to protect important pretrained structure (Xu et al., 2026; Abdel Hameed et al., 2026; Sun et al., 2026). ITCF complements this line by asking which input-related structure is determined by the endpoint itself before such assumptions are introduced.

Data-free expert-response preservation. WUDI and SWUDI are the closest methods in objective. WUDI uses task vectors as proxies for inaccessible task inputs and matches merged-model responses to individual experts (Cheng et al., 2025). SWUDI further turns this objective into an efficient spectral solution, and SWUDI-A adapts its retained dimension from the task-vector spectrum (Wei et al., 2026). These methods show that checkpoint-derived spectral structure can provide useful guidance for response preservation. A remaining question is how much of the relative input geometry used by such matching is actually determined by the final checkpoint. ITCF addresses this question first and builds its comparison geometry from the checkpoint-identifiable task core.

3 What Can a Task Checkpoint Identify?

3.1 The checkpoint-only response problem

For a linear weight group, write the expert update as $\Delta_i = W_i - W_0$ and the candidate merged update as M. If task inputs were available, response matching would use

$$\mathcal{J}_i^{\mathrm{oracle}}(M) = \mathbb{E}\|(M - \Delta_i)x\|_2^2 = \mathrm{tr}[(M - \Delta_i)\Sigma_i(M - \Delta_i)^\top].$$

Checkpoint-only merging observes Δ_i but not the input second moment Σ_i. The row space $S_i = \mathrm{row}(\Delta_i)$ is identifiable from the endpoint and contains every input direction on which the observed update acts; the question is whether the endpoint also identifies relative importance within this space.

3.2 Checkpoint information boundary

We answer this question in centered linear regression, where the second moment is the covariance.

Theorem 1 (Checkpoint information boundary)**.** ***(a) Fixed-label ambiguity.*** *Fix a nonnegative full-batch GD schedule with at least two positive steps and $2 \le r \le \min(m, n-1)$. There exist common $W_0, W^\star$ and distinct positive-definite covariances H_1, H_2 that produce the same rank-r endpoint update Δ while inducing different normalized geometry on $S = \mathrm{row}(\Delta)$:*

$$W_T(H_1) = W_T(H_2) = W_0 + \Delta, \qquad \frac{P_S H_1 P_S}{\mathrm{tr}(P_S H_1 P_S)} \neq \frac{P_S H_2 P_S}{\mathrm{tr}(P_S H_2 P_S)}. \tag{1}$$

(b) Profiled identification. *Under the admissible centered-regression class of Appendix A.3, let $\mathfrak{I}_S(\mathcal{O})$ collect normalized covariance geometries on S compatible with checkpoints $\mathcal{O} = (W_0, W_T)$ after profiling over the unobserved time-independent label map. For nonzero S,*

$$\mathfrak{I}_S(\mathcal{O}) = \mathcal{C}_{++}(S) = \mathrm{ri}\,\mathcal{C}(S), \qquad \overline{\mathfrak{I}_S(\mathcal{O})} = \mathcal{C}(S),$$

where $\mathcal{C}(S) = \{C \succeq 0 : \mathrm{range}(C) \subseteq S,\ \mathrm{tr}\, C = 1\}$, $\mathcal{C}_{++}(S)$ restricts to positive definiteness on S, and ri *denotes relative interior.*

Thus an endpoint identifies the acted-on directions but not their normalized within-support covariance geometry. Appendix A.3 gives the constructions and exact admissibility conditions; Appendices A.4–A.5 show both why update-Gram magnitudes can be statistically informative and why narrower fixed-label families can have isotropic or anisotropic minimax centers.

3.3 Choosing an invariant task core

Given the retained rank r_i, set from checkpoint-only norm summaries (Appendices B.2 and H.3), we first retain the subspace that loses as little of the observed update's right action as possible. Within that support, Theorem 1 leaves the relative comparison strength of the retained directions unresolved. Any anisotropic weighting would privilege some directions over others using information that is not fixed by the endpoint. We therefore choose the comparison geometry by minimizing worst-case response distortion over the normalized geometries supported on S. For $D = M - \Delta_i$, define $R_C(D) = \mathrm{tr}(DCD^\top)$. The worst compatible geometry can concentrate on a direction to which the chosen comparison assigns the least weight. With unit total comparison mass, protecting this weakest direction forces all r retained directions to receive equal weight. The following result formalizes both choices.

Proposition 1 (Response-minimax task core)**.** *A best rank-r right-action projector is a top-r right-singular projector of Δ. For fixed S and $G \in \mathcal{C}(S)$,*

$$\rho_S(G) := \sup_{C \in \mathcal{C}(S)} \sup_{\|D\|_F = 1} |R_C(D) - R_G(D)| = 1 - \lambda_{\min}(G|_S), \tag{2}$$

so the unique minimizer is

$$G^\star = P_S/r, \qquad \min_G \rho_S(G) = 1 - 1/r. \tag{3}$$

For task i, let Q_i span the retained rank-r_i right-singular subspace, $P_i = Q_i Q_i^\top$, and set $G_i = P_i/r_i$ (zero updates use $G_i = 0$). This task core is basis-invariant and gives each nonzero task unit comparison mass. Accordingly, it specifies a comparison rule: the checkpoint identifies where comparison is supported, and G_i assigns equal strength within that support, consistent with the unresolved within-support importance from Theorem 1. Proposition 1 establishes a minimax comparison rule under checkpoint uncertainty. Appendix B.3 further studies how this comparison guarantee propagates to the fused decision.

4 Fusion Guided by Shared Support

The pairs (G_i, Δ_i) specify where each expert is compared and which output change it requests. When task cores overlap, several experts constrain the same input directions. Fusion must therefore collect both how much comparison support each direction receives and the output changes requested there. Define

$$A = \sum_i G_i, \qquad B = \sum_i \Delta_i G_i, \tag{4}$$

and minimize

$$\mathcal{J}(M) = \sum_i \mathrm{tr}[(M - \Delta_i) G_i (M - \Delta_i)^\top]. \tag{5}$$

We call A the participation operator: it accumulates task-core support, while B accumulates the corresponding requested output changes. On the range $\mathcal{U}$ of A, write $A = Q\Lambda Q^\top$ with positive eigenvalues $\lambda_1 \geq \cdots \geq \lambda_s$.

4.1 Joint fit and anchored recovery

Theorem 2 (Fusion-frame core matching)**.** *$A = \sum_i P_i/r_i$ is the weighted fusion-frame operator (Casazza & Kutyniok, 2004) and, for $x \in \mathcal{U}$,*

$$\lambda_s \|x\|_2^2 \leq \sum_i \frac{\|P_i x\|_2^2}{r_i} = x^\top A x \leq \lambda_1 \|x\|_2^2. \tag{6}$$

Using the checkpoint-only sum update $M_0 = \sum_i \Delta_i$ as a reference preserves the additive solution when cross-expert action vanishes on every retained core. The exact supported fit is

$$M^\star = M_0 + (B - M_0 A)A^\dagger, \tag{7}$$

and satisfies

$$\mathcal{J}(M) = \mathcal{J}^\star + \|(M - M^\star)A^{1/2}\|_F^2, \tag{8}$$

where

$$\mathcal{J}^\star = \sum_i \operatorname{tr}(\Delta_i G_i \Delta_i^\top) - \operatorname{tr}(BA^\dagger B^\top). \tag{9}$$

Equation (6) makes the role of A explicit: λ_j measures how strongly the retained cores collectively constrain participation mode q_j. It does not measure whether the experts request the same output change; that mismatch appears separately in $\mathcal{J}^\star$. Appendix C separates these quantities formally.

With $T_j = (\sum_{i \neq j} \Delta_i)P_j$, this sum-anchor preservation property is quantified by

$$M_0 A - B = \sum_j T_j / r_j, \qquad \mathcal{J}(M_0) = \sum_j \|T_j\|_F^2 / r_j. \tag{10}$$

Appendix D.5 gives the reference characterization. The exact fit in Equation (7), however, uses $A^\dagger$: along an eigenmode q_j of A, the correction from M_0 is $(B - M_0 A)q_j/\lambda_j$. For the same residual request, the required correction therefore grows as λ_j decreases, precisely where shared support is weak.

We therefore turn this weak-support amplification into a synthesis cost around M_0. Let $Z = [Q_1/\sqrt{r_1}, \ldots, Q_K/\sqrt{r_K}]$, so $A = ZZ^\top$. Any supported correction $E = M - M_0$ can be written as $E = LZ^\top$; among all coefficient matrices that produce the same E, the least coefficient energy is $\|EA^{\dagger/2}\|_F^2$ (Appendix D). Penalizing this minimum synthesis energy limits weakly supported corrections while allowing strongly supported ones to approach the joint fit, yielding:

Theorem 3 (Anchored regularized recovery)**.** *For $\tau \geq 0$,*

$$M_\tau = M_0 + (B - M_0 A)Q\Lambda(\Lambda^2 + \tau^2 I)^{-1}Q^\top. \tag{11}$$

For each participation mode,

$$M_\tau q_j = (1 - h_j)M_0 q_j + h_j M^\star q_j, \qquad h_j = \frac{\lambda_j^2}{\lambda_j^2 + \tau^2}. \tag{12}$$

For $\tau > 0$, this is the unique minimizer over corrections $M - M_0$ supported on $\mathcal{U}$ of

$$\mathcal{J}(M) + \tau^2 \|(M - M_0)A^{\dagger/2}\|_F^2. \tag{13}$$

Appendix D derives the minimum-energy synthesis and equivalent reduced Tikhonov problem. Thus strongly supported modes approach the joint fit and weak modes stay closer to the checkpoint-only anchor.

4.2 Choosing the support transition

The regularized family above leaves one checkpoint-only choice: where should the recovery move from following the joint fit to staying near the anchor? We first use the participation spectrum to summarize how many modes carry substantial shared support. With $p_j = \lambda_j / \sum_\ell \lambda_\ell$,

$$d_{\text{part}} = \left(\sum_j \sqrt{p_j}\right)^2 = \frac{(\operatorname{tr}\sqrt{A})^2}{\operatorname{tr} A} = \frac{\|Z\|_*^2}{\|Z\|_F^2}. \tag{14}$$

This quantity equals s for a flat positive participation spectrum and falls as participation becomes concentrated. We use it as a checkpoint-only reference for the number of modes that should behave like the exact fit. We set

$$k = \min\{s, \lceil d_{\text{part}} - 10^{-4} \rceil\}. \tag{15}$$

Appendix E.3 gives the corresponding norm-summary certificate; k is a reference dimension, not a claim of downstream-optimal rank.

A hard rule could now keep the first k modes and suppress the rest, but the cutoff would be abrupt; at a repeated boundary, selecting only part of the eigenspace would also depend on the chosen basis. We therefore use the continuous filter h_τ already induced by anchored recovery and choose the value that best matches this hard reference. As τ increases, the retained-side error at λ_k increases while the tail-side error at λ_{k+1} decreases. Balancing these two boundary errors gives the transition below.

Proposition 2 (Minimax soft correction boundary)**.** *For $1 \le k < s$, define $\mathcal{E}_k(\tau) = \max_j |h_\tau(\lambda_j) - \mathbf{1}\{j \le k\}|$ for $h_\tau(\lambda) = \lambda^2/(\lambda^2 + \tau^2)$. Its unique minimizer is attained at*

$$\tau^\star = \sqrt{\lambda_k \lambda_{k+1}}, \qquad \min_\tau \mathcal{E}_k(\tau) = \frac{\lambda_{k+1}}{\lambda_k + \lambda_{k+1}}. \tag{16}$$

At a tied boundary the entire repeated eigenspace receives weight $1/2$; if $k = s$, ITCF sets $\tau = 0$.

Appendix E proves the boundary result and compares ordinary ridge under the same hard reference.

4.3 Computing the fused update

For each matrix group, ITCF extracts retained right-singular cores, solves the reduced participation system, selects (k, τ) from its spectrum, and applies Equation (11) around the summed-update anchor. A fixed-width randomized approximation uses default $q = 12$; untargeted parameters stay pretrained. Appendix H.3 gives grouping, rank/retry, precision, numerical-support, and reduced-solve rules; Appendix E.4 gives local checkpoint-to-fusion stability, and Appendix G the conditional group-to-network prediction bridge with off-core, untargeted-parameter, and propagation remainders explicit.

5 Experiments

We evaluate overall checkpoint-only performance, transfer across model families and expert counts, mechanism attribution, and construction cost.

5.1 Evaluation protocol

All methods receive the same pretrained and independently fine-tuned checkpoints. ITCF uses no samples, labels, activations, calibration set, validation-selected merge coefficients, or post-merge training. A setting records the model, task subset, and checkpoint seed. The vision suite covers eight tasks over ViT-B/32 and ViT-B/16 (415 scheduled settings, $K = 2, \dots, 8$); Vision32 is the balanced shared-baseline/control panel and 39 settings form the WUDI intersection. The common language suite contains 97 settings over RoBERTa-base/large and T5-base/large, with 93 shared with WUDI; Qwen2.5-1.5B and Llama-3.2-1B add 24 frozen NLI settings each. NLLB-200 distilled-600M and distilled-1.3B add 24 multilingual translation settings each (eight frozen compositions across three checkpoint seeds), scored by macro chrF++. Scores are averaged within settings and then across settings; paired intervals bootstrap matched settings and therefore characterize the scheduled matched collection rather than an i.i.d. task population. Appendix H gives task lists and full protocols.

5.2 Overall checkpoint-only performance

We first ask how ITCF compares with strong baselines under the same checkpoint-only information budget. Table 1 reports common-panel scores and paired comparisons on exact matched intersections. On Language93, ITCF improves over WUDI300, SWUDI-A, TSV-Merge, and ACTMat by $+3.10$, $+0.85$, $+0.89$, and $+3.62$ pp, with all four paired intervals positive. Vision is more tightly contested: the gains over WUDI300, SWUDI-A, and ACTMat are $+0.60$, $+0.66$, and $+0.74$ pp, while the $+0.17$ pp margin over TSV-Merge has an interval crossing zero. The full 97-setting language suite adds a 9.463 pp lead over fixed unit-scale TIES, while a 66-setting TSV-Merge expansion gives $+0.769$ pp (47/0/19 wins/ties/losses). Overall, the language panels show a clear positive margin and vision remains more tightly competitive; Appendix Table 13 gives the expanded family breakdown.

Table 1: **Overall checkpoint-only evidence.** (a) Absolute scores on the two common panels. (b) Paired ITCF-minus-baseline differences on their exact matched intersections; intervals are paired setting-bootstrap intervals.

(a) Common panels

	Vision32		Language93	
Method	Score	ITCF−method	Score	ITCF−method
Weight averaging	58.06 ± 0.41	$+9.38$	54.67 ± 0.55	$+11.45$
Task Arithmetic	59.67 ± 0.52	$+7.77$	51.71 ± 0.85	$+14.41$
DARE-TIES	58.92 ± 0.49	$+8.52$	51.52 ± 0.73	$+14.60$
PCB-Merging	61.98 ± 0.46	$+5.46$	53.36 ± 0.64	$+12.76$
WUDI300	66.81 ± 0.39	$+0.63$	63.02 ± 0.81	$+3.10$
SWUDI-A	66.78 ± 0.39	$+0.66$	65.27 ± 0.42	$+0.85$
TSV-Merge	67.27 ± 0.33	$+0.17$	65.23 ± 0.48	$+0.89$
ACTMat	66.70 ± 0.31	$+0.74$	62.50 ± 0.42	$+3.62$
ITCF	$\mathbf{67.44 \pm 0.25}$	–	$\mathbf{66.12 \pm 0.30}$	–

(b) Paired strong-baseline evidence

		Vision			Language93		
Baseline	Panel	Base	Δ±SD	95% CI	Base	Δ±SD	95% CI
WUDI300	W39	65.23	$+0.60 \pm 0.15$	$[+0.38, +0.82]$	63.02	$+3.10 \pm 0.69$	$[+1.90, +4.10]$
SWUDI-A	W39	65.18	$+0.66 \pm 0.64$	$[+0.43, +0.88]$	65.27	$+0.85 \pm 0.52$	$[+0.25, +1.45]$
TSV-Merge	V32	67.27	$+0.17 \pm 0.38$	$[-0.20, +0.55]$	65.23	$+0.89 \pm 0.48$	$[+0.10, +1.55]$
ACTMat	V32	66.70	$+0.74 \pm 0.26$	$[+0.40, +1.02]$	62.50	$+3.62 \pm 0.61$	$[+1.12, +4.35]$

5.3 Transfer across model families and expert counts

We next ask whether the performance pattern transfers across architectures, model scales, and task forms. Table 2 spans vision encoders, text-to-text models, decoder models, and multilingual translation, with positive mean ITCF-minus-WUDI differences in all eight reported family settings. The margins range from $+0.44$–$+0.61$ pp on the two vision backbones to $+4.40$ and $+9.10$ pp on T5-base/large; Qwen, Llama, and both NLLB scales are also positive. On NLLB, ITCF has the highest three-seed mean among the nine evaluated methods on all eight frozen compositions at both scales (Appendix Tables 16–17). Task-count studies show the same scale variation: T5 grows from near parity at $K = 2$ to larger gains at $K = 4, 6$, while vision moves from a deficit at $K = 2$ to leads at $K = 3, \ldots, 8$ (Appendix Tables 18 and 14).

Table 2: **Backbone/model-family transfer.** Settings give full evaluated coverage; Δ is mean ITCF-minus-WUDI on the matched subset in each panel's native metric. Multi-seed rows report mean ± seed SD.

Vision + translation			Text-to-text + decoder		
Model	Settings	Δ	Model	Settings	Δ
ViT-B/32	343	$+0.61 \pm 0.18$	T5-base	63	$+4.40 \pm 0.50$
ViT-B/16	72	$+0.44 \pm 0.20$	T5-large	6	$+9.10 \pm 0.65$
NLLB-600M	24	$+0.35 \pm 0.15$	Qwen2.5-1.5B	24	$+0.82 \pm 0.22$
NLLB-1.3B	24	$+0.39 \pm 0.05$	Llama-3.2-1B	24	$+0.70 \pm 0.20$

We then ask whether this ordering survives when the expert count increases by an order of magnitude. Table 3 scales T5-base full-model merging from $K = 8$ to 166. ITCF has the highest mean at all nine tested scales, with a 0.54–0.82 pp margin over second-best ACTMat. At $K = 166$, the leads over ACTMat, SWUDI-A, and TSV-Merge are 0.60, 0.72, and 0.86 pp, and the source-family margins over ACTMat/TSV remain positive. The large-K advantage therefore persists across all four T5-base source-task families. Full statistics are in Appendix H.7.

Table 3: **Large-K full-model merging on T5-base.** (a) Scaling matrix across nine frozen expert-count prefixes. (b) $K = 166$ source-family decomposition and aggregate margins.

(a) Scaling across expert counts

Method / K	8	12	16	24	36	48	60	72	166
Mean	35.67	40.28	42.11	35.98	32.11	26.95	23.06	21.15	14.39
TIES20	43.40	46.55	47.17	40.19	35.98	30.15	27.58	25.74	21.61
SWUDI-A	51.62	53.18	52.06	46.37	42.36	37.44	34.15	32.10	27.30
TSV-Merge	47.99	51.55	51.60	44.48	40.84	35.57	32.70	30.93	27.16
ACTMat	52.00	54.20	52.90	47.80	43.50	38.80	35.10	33.00	27.42
ITCF	**52.54**	**54.97**	**53.72**	**48.49**	**44.28**	**39.44**	**35.64**	**33.65**	**28.02**
ITCF−best	+0.54	+0.77	+0.82	+0.69	+0.78	+0.64	+0.54	+0.65	+0.60

(b) $K = 166$ decomposition

Source families					Aggregate margins		
Family	n	TSV	ACTMat	ITCF	Baseline	Score	ITCF−base
P3	36	36.82	36.80	**37.32**	Mean	14.39	+13.63
FLANv2	36	20.59	21.80	**22.59**	TIES20	21.61	+6.41
NIV2	75	26.42	26.36	**26.92**	SWUDI-A	27.30	+0.72
ADD1	19	24.26	24.48	**25.06**	TSV-Merge	27.16	+0.86
All 166	166	27.16	27.42	**28.02**	ACTMat	27.42	+0.60

5.4 Mechanism and design validation

Table 4 separates geometry and recovery; Proposition 1 concerns checkpoint-robust comparison, not universal downstream-optimal weighting. On 48 frozen settings, recovery beats ordinary ridge by $+0.277$ pp under isotropic geometry (95% CI $[+0.113, +0.450]$), with a $+0.161$ pp interaction ($[+0.018, +0.311]$). The geometry-only effect is $+0.052$ pp and family-dependent, so stable average gains lie in recovery and interaction. With bases and ranks fixed, isotropic-minus-native/shuffled/reversed is $+0.30/+0.57/+0.65$ pp; on Qwen, squared-singular weighting scores 81.91 vs. 81.82 for isotropic (Appendix Table 31).

Table 4: **Design attribution.** (a) 48 frozen settings; ISO/SV2: isotropic/normalized squared-singular weights. ISO−SV2 fixes ITCF; ITCF−Ridge fixes ISO; Inter.: ISO−SV2 for ITCF minus for Ridge.

(a) Geometry × recovery

Scope	ISO+ITCF	SV2+ITCF	ISO+Ridge	SV2+Ridge	ISO−SV2	ITCF−Ridge	Inter.
All (48)	66.775	66.723	66.497	66.606	+0.052	+0.277	+0.161
Vision (24)	70.567	70.422	70.346	70.410	+0.145	+0.221	+0.210
RoBERTa (12)	51.947	52.160	51.272	51.640	−0.213	+0.675	+0.154
T5 (12)	74.018	73.886	74.026	73.966	+0.132	−0.008	+0.072

(b) Ridge, anchor, and within-core checks

Ridge scale (48)			Anchor / within-core (sum / ISO ref.)		
Rule	Mean	ITCF−rule	Variant	Score	Ref.−variant
No ridge	65.91	+0.85	Sum anchor (10)	70.55	–
Fixed 0.25×	66.18	+0.58	Mean anchor (10)	69.20	+1.35
Fixed 0.5×	66.34	+0.42	Zero anchor (10)	67.80	+2.75
Fixed 1×	66.29	+0.47	Native magnitude (48)	66.475	$+0.30 \pm 0.12$
Fixed 2×	65.88	+0.88	Shuffled magnitude (48)	66.205	$+0.57 \pm 0.15$
Fixed 4×	65.13	+1.63	Reversed magnitude (48)	66.125	$+0.65 \pm 0.16$

Anchor and support transition. The recovery also needs a reference and transition strength. The sum-update anchor scores 70.55 on the 10-setting ablation, above the mean (69.20) and zero (67.80) anchors; on the fixed 48-setting ridge study, the participation rule exceeds no ridge and every tested

fixed scale. These checks support the two choices used in the final recovery. Appendix Tables 29, 30, and 32 give the full diagnostics.

Table 5 gives the checkpoint-only selector comparison on the common 86-setting panel.

Table 5: **Checkpoint-only selection against downstream tuning.**

Selector	Mean	Oracle gap
No ridge	70.14	1.20
Best global fixed scale ($0.5\times$)	70.58	0.76
ITCF participation selector	**71.11**	**0.23**
Per-setting downstream oracle	71.34	0.00

No ridge leaves a $1.20\,\mathrm{pp}$ oracle gap, the best global fixed scale reduces it to $0.76\,\mathrm{pp}$, and the ITCF selector leaves only $0.23\,\mathrm{pp}$ (71.11 versus 71.34). The family gaps remain between 0.17 and $0.36\,\mathrm{pp}$. Thus the participation spectrum recovers most of the per-setting downstream-selection gain from checkpoints alone. The best global fixed scale and per-setting oracle use downstream evaluation only for this analysis and do not enter the ITCF construction; Appendix Tables 27–28 give the full comparison.

5.5 Numerical stability and construction cost

Finally, we assess numerical stability and construction cost. The aggregate score spans only $0.173\,\mathrm{pp}$ across five sketch seeds and $0.197\,\mathrm{pp}$ across $q \in \{8, 12, 16, 24\}$ (Appendix Tables 9–10), showing little variation over the tested numerical choices. Across nine vision timing anchors, median core construction is 0.204 s for ITCF versus 41.38 s for WUDI300, a median $200.3\times$ speedup. Table 6 extends the standardized comparison to SWUDI-A and TSV-Merge; checkpoint I/O, CPU–GPU staging, evaluation, and serialization are excluded.

Table 6: **Representative standardized construction times (s).** Speedups divide the displayed comparator by ITCF on the same row.

Model / K	ITCF	WUDI300	SWUDI-A	TSV-Merge	$\times$WUDI	$\times$SWUDI	$\times$TSV
ViT-B/16, 2	0.07	16.99	0.85	7.56	242.7	12.1	108.0
ViT-B/16, 6	0.20	42.12	0.93	14.63	210.6	4.7	73.2
ViT-B/32, 8	0.30	52.79	0.95	18.23	176.0	3.2	60.8
T5-base, 16	1.08	423.02	2.35	128.03	391.7	2.2	118.5
T5-base, 72	2.06	1814.21	3.71	525.52	880.7	1.8	255.1
T5-base, 166	3.19	1154.68	5.52	1179.98	362.0	1.7	369.9

At $K = 166$, ITCF takes 3.19 s versus 1154.68 s for WUDI300, 5.52 s for SWUDI-A, and 1179.98 s for TSV-Merge, giving $362.0\times$, $1.7\times$, and $369.9\times$ speedups. Peak CUDA allocation across the nine core timing cases has a 5.38 GiB median and 14.65 GiB maximum. The fused matrices replace existing weights, adding no parameters or inference computation. The reduced construction therefore remains stable over the tested approximations and inexpensive through the 166-expert setting. Full timing anchors are in Appendix Tables 33 and 34.

6 Conclusion

Final checkpoints reveal changed input directions, but not a unique task-input distribution within them. ITCF turns retained directions into basis-invariant task cores, combines their overlap into shared support, and stabilizes a joint fit around summed expert updates. Across the evaluated vision and language families, ITCF is competitive overall, scales to 166 full-model experts, and makes timed core fusion roughly two orders of magnitude faster than iterative optimization with unchanged inference cost. The gains remain family-dependent across the evaluated vision and language model configurations.

Reproducibility Statement

Appendix H collects the evaluated task pools, reduced-construction rules, paired result tables, confidence-interval procedure, and individual timing anchors, including the full-model scaling results. The other appendices provide the theorem proofs, rank-deficient and repeated-eigenvalue cases, selector-stability conditions, and the conditional network-level bridge. The numerical rank and support rules are specified in Appendix H.3.

AI Use Statement

GPT was used solely for language polishing of the manuscript, including improvements to grammar, wording, clarity, and readability. The authors reviewed the resulting language edits and verified that they did not alter the scientific content. The authors take responsibility for the complete submission.

References

Marawan Gamal Abdel Hameed, Derek Tam, Pascal Jr Tikeng Notsawo, Colin Raffel, and Guillaume Rabusseau. Model merging via data-free covariance estimation. In *Conference on Language Modeling*, 2026.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

Peter G. Casazza and Gitta Kutyniok. Frames of subspaces. *Contemporary Mathematics*, 345:87–113, 2004.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *International Conference on Machine Learning*, pp. 10121–10143, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

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*, 2024.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

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*, pp. 18695–18705, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

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

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

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTaiL: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 5189–5197, 2018.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *ICCV Workshops*, 2013.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Yeoreum Lee, Jinwook Jung, and Sungyong Baik. Mitigating parameter interference in model merging via sharpness-aware fine-tuning. In *International Conference on Learning Representations*, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Miles Lopes. Estimating unknown sparsity in compressed sensing. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 217–225. PMLR, 2013. URL https://proceedings.mlr.press/v28/lopes13.html.

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*, pp. 43177–43199, 2025.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 216–223, 2014.

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

Meta AI. Llama 3.2 model card. Llama 3.2 model documentation, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Mohammed Muqeeth, Haokun Liu, Yufan Liu, and Colin Raffel. Learning to route among specialized experts for zero-shot generalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36829–36846. PMLR, 2024.

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

NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846, 2024. doi: 10.1038/s41586-024-07335-x.

Maja Popović. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, 2017.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2015.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 2012.

Wenju Sun, Qingyong Li, Yangliao Geng, and Boyang Li. CAT merging: A training-free approach for resolving conflicts in model merging. In *International Conference on Machine Learning*, pp. 57523–57543, 2025a.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangliao Geng, and Boyang Li. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. In *Advances in Neural Information Processing Systems*, 2025b.

Wenju Sun, Qingyong Li, Tiancheng Li, Yangli-ao Geng, and Boyang Li. Revisiting the role of pretrained weights in model merging: On near-optimality within the core subspace. In *International Conference on Machine Learning*, 2026.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. QuaRTz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5941–5946, 2019.

Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston, 1977.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019b.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232, 1973.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In *International Conference on Machine Learning*, pp. 66178–66193, 2025.

Yongxian Wei, Runxi Cheng, Xingxuan Zhang, Li Shen, Chun Yuan, Peng Cui, and Dacheng Tao. Closed-form spectral regularization for multi-task model merging. *arXiv preprint arXiv:2606.07289*, 2026.

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

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*, 2026.

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

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

Enneng Yang, Qun Yang, Peng Wang, Anke Tang, Guibing Guo, Xiaochun Cao, and Li Shen. MergOPT: A merge-aware optimizer for robust model merging. In *International Conference on Learning Representations*, 2026.

Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2013–2018, 2015.

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*, pp. 57755–57775, 2024.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1298–1308, 2019.

Appendix Contents. The appendix is structured into the sections below, which provide proofs, numerical details, complete experimental results, and additional analyses supporting the main text.

A Proofs for the Checkpoint Information Boundary **15**
A.1 Fixed-label finite-step ambiguity 15
A.2 One-step fixed-label fiber and within-core reachability 15
A.3 Profiled checkpoint identification 16
A.4 Conditional population update-Gram identity 17
A.5 A bounded fixed-label family with an isotropic center 17
B Task-Core Geometry and Dimension **19**
B.1 Ideal right-action compression and its numerical approximation 19
B.2 The per-task norm-summary dimension 19
B.3 Response-minimax comparison geometry 20
B.4 Secondary structural characterizations 23
B.5 Zero updates and duplicated tasks 23
C Fusion-Frame Core Matching **24**
C.1 Frame identity and principal angles 24
C.2 Pythagorean decomposition 24
C.3 Conditioning and residual are independent 24
D Anchored Regularized Recovery **25**
D.1 Minimum synthesis energy and reduced closed form 25
D.2 Interpolation and aggregate objective comparison 26
D.3 Operator stability at a fixed positive τ 26
D.4 Tail sensitivity of the unregularized inverse 27
D.5 Sum-update anchor and net cross-action 27
E Minimax Soft Spectral Boundary **28**
E.1 Proof of Proposition 2 28
E.2 Repeated-eigenspace impossibility 29
E.3 Participation dimension and the zero-ridge window 29
E.4 Local stability of the adaptive selector 30
F Secondary Statistical Interpretations **32**
F.1 Gaussian MAP interpretation 32
F.2 Moment-robust interpretation for signed coefficient perturbations 32
G Conditional Prediction Bound and Its Remainders **32**
H Experimental Protocol, Numerical Construction, and Complete Results **33**
H.1 Closest checkpoint-derived comparisons 33
H.2 Information boundary and evaluation protocol 34
H.3 Reduced construction and complexity 35
H.4 Numerical robustness of the reduced construction 36
H.5 Coverage-specific aggregates and matched comparisons 37
H.6 Task-count and model-family breakdowns 38
H.7 Large-scale full-model merging 40
H.8 Fixed TIES anchors on the full language panel 41
H.9 Nearest endpoint-derived methods 41
H.10 Panel dependence of the regularization results 42
H.11 Anchor choice under the regularized construction 43
H.12 Within-core weighting and family variation 43
H.13 Participation, core residual, and selector activity 44
H.14 Numerical verification and tied eigenspaces 45
H.15 Construction-time and memory anchors 46
I Additional Stability and Large-Expert Analyses **46**
J Factorial Geometry–Recovery Intervention **48**

A Proofs for the Checkpoint Information Boundary

A.1 Fixed-label finite-step ambiguity

Write $A_\star = W^\star - W_0$ for the label displacement. With centered inputs of covariance $H \succ 0$ and labels $y = W^\star x$, the squared regression loss and gradient are

$$L_H(W) = \tfrac{1}{2}\operatorname{tr}[(W - W^\star)H(W - W^\star)^\top], \qquad \nabla L_H(W) = (W - W^\star)H.$$

The recurrence $W_{t+1} - W^\star = (W_t - W^\star)(I - \eta_t H)$ gives

$$W_T - W_0 = A_\star p(H), \qquad p(h) = 1 - \prod_{t=0}^{T-1}(1 - \eta_t h). \tag{17}$$

Each covariance has a finite centered realization: assign equal mass to $x_{j,\pm} = \pm\sqrt{n}\, H^{1/2} e_j$, $j = 1, \dots, n$, and label each point by $W^\star x_{j,\pm}$. The mean is zero, the covariance is H, and the empirical full-batch gradient equals the expression above.

A common endpoint. Fix the schedule in Theorem 1(a) and put $h_0 = 1/(2\eta_{\max})$. On $[0, h_0]$, all factors $1 - \eta_t h$ are positive and

$$p'(h) = \sum_t \eta_t \prod_{j \neq t}(1 - \eta_j h) > 0. \tag{18}$$

Hence p has an inverse f on $[0, p(h_0)]$, with $f(0) = 0$ and $f'(0) = 1/\beta$, where $\beta = \sum_t \eta_t > 0$.

For $2 \le r \le \min(m, n-1)$, put $q = n - r$ and take $A_\star = [I_r, 0]$, appending zero output rows when $m > r$. Choose any common W_0 and set $W^\star = W_0 + A_\star$. Define

$$R = e_1^{(r)}(e_1^{(q)})^\top, \qquad K_b = \begin{bmatrix} I_r & R \\ R^\top & \operatorname{diag}(b, 1, \dots, 1) \end{bmatrix}, \qquad b \in \{2, 3\}. \tag{19}$$

The Schur complement of I_r is $\operatorname{diag}(b-1, 1, \dots, 1) \succ 0$, so both K_b are positive definite. For a common $0 < \alpha < p(h_0)/\max_b \|K_b\|_2$, spectral evaluation of f gives

$$H_b = f(\alpha K_b) \succ 0, \qquad \lambda_{\max}(H_b) < h_0, \qquad A_\star p(H_b) = \alpha[I_r, R] = \Delta, \tag{20}$$

with the same zero output rows appended to Δ. This update has rank r, and $H_2 \neq H_3$ because f is injective.

Geometry inside the observed update. The shared subspace is $S = \operatorname{row}[I_r, R]$. An orthonormal basis Q has first column $(e_1^{(r)}, e_1^{(q)})/\sqrt{2}$ and remaining columns $(e_j^{(r)}, 0)$, $j = 2, \dots, r$. Direct compression gives

$$Q^\top K_2 Q = \operatorname{diag}(5/2, 1, \dots, 1), \qquad Q^\top K_3 Q = \operatorname{diag}(3, 1, \dots, 1).$$

Since $f(z) = z/\beta + o(z)$, spectral diagonalization yields $\alpha^{-1} H_b \to K_b/\beta$ as $\alpha \downarrow 0$. Consequently,

$$\frac{Q^\top H_b Q}{\operatorname{tr}(Q^\top H_b Q)} \longrightarrow \frac{Q^\top K_b Q}{\operatorname{tr}(Q^\top K_b Q)}. \tag{21}$$

The two limits differ for $r \ge 2$. The normalized compressions therefore differ for all sufficiently small fixed $\alpha > 0$; conjugation by Q proves Equation (1). At every positive step the error is $-A_\star \prod_{j<t}(I - \eta_j H_b)$. Each factor and H_b are invertible, so that step produces a nonzero update.

A.2 One-step fixed-label fiber and within-core reachability

For one step of size $\eta > 0$, the endpoint equation is $\Delta = \eta A_\star H$. Its covariance fiber is

$$\mathcal{H}(\Delta; A_\star) = \{H \succ 0 : \eta A_\star H = \Delta\}. \tag{22}$$

Lemma 1 (One-step fixed-label covariance fiber). *Fix $A_\star \in \mathbb{R}^{m\times n}$, $\eta > 0$, and $\Delta \in \mathbb{R}^{m\times n}$. For $r = \operatorname{rank}(A_\star) > 0$, choose orthogonal $U = [V, V_\perp]$ with $\operatorname{range}(V) = \operatorname{row}(A_\star)$ and $\operatorname{range}(V_\perp) = \ker A_\star$. Put $q = n - r$, $F = \Delta/\eta$, $A_V = A_\star V$, and $[C, R] = A_V^\dagger FU$. The fiber is nonempty exactly when*

$$A_V A_V^\dagger F = F, \qquad C = C^\top \succ 0. \tag{23}$$

In that case all its members are

$$H(T) = U \begin{bmatrix} C & R \\ R^\top & R^\top C^{-1} R + T \end{bmatrix} U^\top, \qquad T \succ 0 \text{ in } \mathbb{R}^{q\times q}, \tag{24}$$

with empty blocks when $q = 0$. For symmetric N, the invisible perturbations satisfy

$$A_\star N = 0 \iff N = V_\perp M V_\perp^\top, \qquad M = M^\top. \tag{25}$$

The nonempty fiber is a singleton exactly when $r = n$, in which case $H = A_\star^\dagger \Delta/\eta$. Every feasible covariance has $\operatorname{rank}(\Delta) = r$ and $\operatorname{row}(\Delta) = H \operatorname{row}(A_\star)$. For $A_\star = 0$, the fiber is the positive-definite cone if $\Delta = 0$ and empty otherwise.

Proof. Since $A_\star U = [A_V, 0]$ and A_V has full column rank, $A_\star H = F$ requires $A_V A_V^\dagger F = F$ and fixes the first block row of $U^\top H U$ to $[C, R]$. Symmetry fixes its first block column; the Schur complement gives exactly $C \succ 0$ and $T \succ 0$ in Equation (24). The same block argument forces a symmetric invisible N to have only its lower-right block nonzero. This free space has dimension $q(q+1)/2$, and small perturbations preserve positivity, proving uniqueness exactly when $q = 0$. In that case left multiplication by $A_\star^\dagger$ recovers H. Invertibility of H preserves rank, and $\Delta^\top = \eta H A_\star^\top$ gives the row-space identity. The zero-$A_\star$ case is immediate. □

Corollary 1 (Fixed-label within-core uncertainty). *Suppose $\mathcal{H}(\Delta; A_\star)$ is nonempty and fix $S_0 \subseteq \operatorname{row}(\Delta)$ with $d = \dim S_0 \geq 2$. Over the full positive-definite fiber, with no spectral upper bound,*

$$\overline{\left\{ \frac{P_{S_0} H P_{S_0}}{\operatorname{tr}(P_{S_0} H P_{S_0})} : H \in \mathcal{H}(\Delta; A_\star) \right\}} = \mathcal{C}(S_0) \iff S_0 \cap \operatorname{row}(A_\star) = \{0\}. \tag{26}$$

Under this condition, every unit-trace operator positive definite on S_0 is attained exactly.

Proof. Let Q_0 be an orthonormal basis for S_0 and $D_0 = U[C, R]^\top$. Lemma 1 gives $H(T) = D_0 C^{-1} D_0^\top + V_\perp T V_\perp^\top$ and $\operatorname{row}(\Delta) = \operatorname{range}(D_0)$. Hence

$$Q_0^\top H(T) Q_0 = K_0 + L T L^\top, \quad K_0 = Q_0^\top D_0 C^{-1} D_0^\top Q_0 \succ 0, \quad L = Q_0^\top V_\perp. \tag{27}$$

Through Q_0, $\ker L^\top$ corresponds to $S_0 \cap \operatorname{row}(A_\star)$, so transversality is equivalent to full row rank of L. Under this condition, for any $\widehat{C} \succ 0$ of trace one, choose γ with $X = \gamma \widehat{C} - K_0 \succ 0$ and set $T = L^\dagger X (L^\dagger)^\top + P_{\ker L}$. Then $T \succ 0$ and $K_0 + L T L^\top = \gamma \widehat{C}$. Every positive-definite normalized target is attained, and their density gives the full closure.

If $\mathcal{L} = \operatorname{range}(L)$ is proper, $L = 0$ gives the singleton $\{K_0/\operatorname{tr} K_0\}$. Otherwise each normalized compression lies in the compact set $\operatorname{conv}(\{K_0/\operatorname{tr} K_0\} \cup \mathcal{C}(\mathcal{L}))$. A rank-one $vv^\top$ with unit $v \notin \mathcal{L}$ lies outside: positive weight on the base makes a combination full rank, and zero weight restricts its range to $\mathcal{L}$. Thus the closure is strict. □

A.3 Profiled checkpoint identification

Under the protocol of Theorem 1(b), set $\mathcal{H}_\eta = \{H = H^\top : 0 \prec H \prec (2/\eta_{\max}) I\}$ and $\Theta_\eta = \mathcal{H}_\eta \times \mathbb{R}^{m\times n}$. For $H \in \mathcal{H}_\eta$, the product $\prod_t (1 - \eta_t h)$ has absolute value below one at every eigenvalue h, so $S_H = p(H) \succ 0$. With $\Delta = W_T - W_0$, Equation (17) parameterizes all tasks compatible with $\mathcal{O} = (W_0, W_T)$:

$$\Theta_\eta(\mathcal{O}) = \{(H, W_H^\star) : H \in \mathcal{H}_\eta\}, \qquad W_H^\star = W_0 + \Delta S_H^{-1}. \tag{28}$$

Each pair is realized by the fixed-label finite design in Appendix A.1.

Write $\psi_S(H) = P_S H P_S / \operatorname{tr}(P_S H P_S)$ and $r = \dim S$. Compression of $H \succ 0$ gives $\mathfrak{I}_S(\mathcal{O}) \subseteq \mathcal{C}_{++}(S)$. Conversely, for $C \in \mathcal{C}_{++}(S)$ take

$$H_C = \eta_{\max}^{-1}(C + P_{S^\perp}), \qquad 0 \prec H_C \preceq \eta_{\max}^{-1} I \prec (2/\eta_{\max}) I, \qquad \psi_S(H_C) = C. \tag{29}$$

Its unique label map in Equation (28) proves exact attainment. Positive-definite restrictions admit small supported symmetric trace-zero perturbations, whereas a null direction precludes an interior neighborhood; hence $\mathfrak{I}_S(\mathcal{O}) = \mathcal{C}_{++}(S) = \operatorname{ri} \mathcal{C}(S)$. For any $C \in \mathcal{C}(S)$, $C_\varepsilon = (1-\varepsilon)C + \varepsilon P_S/r$ lies in $\mathcal{C}_{++}(S)$ for $0 < \varepsilon < 1$ and converges to C as $\varepsilon \downarrow 0$, giving $\overline{\mathfrak{I}_S(\mathcal{O})} = \mathcal{C}(S)$. For $r = 1$, both sets are $\{P_S\}$.

For each $G \in \mathcal{C}(S)$, Equation (44) and continuity give

$$\sup_{C \in \mathfrak{I}_S(\mathcal{O})} \|C - G\|_2 = \max_{C \in \mathcal{C}(S)} \|C - G\|_2. \tag{30}$$

This preserves the minimax objective in Proposition 1.

A.4 Conditional population update-Gram identity

(c) Conditional population update-Gram proxy. Let $\Delta = c \sum_{\ell=1}^N G_\ell$, with deterministic c and i.i.d. $G_\ell = e_\ell x_\ell^\top$ satisfying $\mathbb{E}\|G_\ell\|_F^2 < \infty$. With $\mu = \mathbb{E}[G_\ell]$ and expectation over repeated draws of this increment law,

$$\mathbb{E}[\Delta^\top \Delta] = c^2 N \, \mathbb{E}[G^\top G] + c^2 N(N-1) \mu^\top \mu. \tag{31}$$

If the inputs are centered with covariance Σ, $\mu = 0$, and $\mathbb{E}[(e^\top e) x x^\top] = \sigma_e^2 \Sigma$, then $\mathbb{E}[\Delta^\top \Delta] = c^2 N \sigma_e^2 \Sigma$.

Under the increment law stated above, expand

$$\Delta^\top \Delta = c^2 \sum_{\ell=1}^N \sum_{k=1}^N G_\ell^\top G_k. \tag{32}$$

The N diagonal terms have mean $\mathbb{E}[G^\top G]$, while independence gives

$$\mathbb{E}[G_\ell^\top G_k] = \mathbb{E}[G_\ell]^\top \mathbb{E}[G_k] = \mu^\top \mu, \qquad \ell \neq k. \tag{33}$$

Summing yields Equation (31). Since $G^\top G = (e^\top e) x x^\top$, centered inputs of covariance Σ, $\mu = 0$, and $\mathbb{E}[(e^\top e) x x^\top] = \sigma_e^2 \Sigma$ give

$$\mathbb{E}[\Delta^\top \Delta] = c^2 N \sigma_e^2 \Sigma. \tag{34}$$

A realized Gram. Let $m = 1$, $n \geq 2$, and $x_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. Take independent Rademacher signs e_ℓ, independent of the inputs, and deterministic $c_N \neq 0$. Then $\mu = 0$, $\mathbb{E}[G^\top G] = \Sigma$, and

$$\Delta_N = c_N \sum_{\ell=1}^N e_\ell x_\ell^\top \overset{d}{=} c_N \sqrt{N} \, z^\top, \qquad z \sim \mathcal{N}(0, \Sigma). \tag{35}$$

Thus $\mathbb{E}[\Delta_N^\top \Delta_N] = c_N^2 N \Sigma$, while

$$\widehat{C}_N := \frac{\Delta_N^\top \Delta_N}{\operatorname{tr}(\Delta_N^\top \Delta_N)} \overset{d}{=} \frac{z z^\top}{z^\top z}. \tag{36}$$

Its law is independent of N. For $\Sigma = I_n$, the rank-one projector obeys $\|\widehat{C}_N - I_n/n\|_2 = 1 - 1/n$ almost surely for every N.

A.5 A bounded fixed-label family with an isotropic center

The fixed-label fiber need not fill $\mathcal{C}(S)$ for its minimax center to be isotropic. The following construction characterizes an entire bounded fiber inside the nonoscillatory range of Theorem 1(a).

Theorem A.1 (Bounded fixed-label comparison geometry)**.** *Fix $r \geq 2$, $m = r$, $n = 2r$, any W_0, $W^\star - W_0 = [I_r, 0]$, and two full-batch squared-loss GD steps with $\eta_0 = \eta_1 = 1$. Set*

$$\Delta = [I_r/4, I_r/8], \qquad Q = \frac{1}{\sqrt{5}} \begin{bmatrix} 2I_r \\ I_r \end{bmatrix}, \qquad S = \mathrm{row}(\Delta) = \mathrm{range}(Q).$$

Over all covariances $0 \prec H \prec I_{2r}/2$ giving this endpoint with this fixed label map, the exact normalized image is

$$\mathfrak{I}_S^{\mathrm{fixed}} = \left\{ QCQ^\top : C \succ 0,\ \mathrm{tr}\, C = 1,\ \frac{\lambda_{\max}(C)}{\lambda_{\min}(C)} < \kappa \right\},$$
$$\kappa = \frac{b}{a}, \qquad a = 1 - \frac{\sqrt{11}}{4}, \qquad b = \frac{67}{85} - \frac{49}{68\sqrt{2}}.$$

Its closure replaces $< \kappa$ by $\leq \kappa$. Among $G \in \mathcal{C}(S)$, the unique operator-norm minimax center is P_S/r, with radius

$$\inf_{G \in \mathcal{C}(S)} \sup_{C \in \mathfrak{I}_S^{\mathrm{fixed}}} \|C - G\|_2 = \frac{(r-1)(\kappa - 1)}{r(\kappa + r - 1)}.$$

The same center and radius apply to the quadratic-response distortion in Equation (44)*.*

Proof. On $0 \prec H \prec I/2$, the two-step map $K = 2H - H^2$ is bijective onto $0 \prec K \prec 3I/4$, with inverse $f(K) = I - \sqrt{I - K}$. The endpoint fixes the first block row of K, and symmetry fixes its first column. Thus the entire fiber is

$$K_B = \begin{bmatrix} I_r/4 & I_r/8 \\ I_r/8 & B \end{bmatrix}, \qquad H_B = f(K_B), \qquad \frac{1}{16} I_r \prec B \prec \frac{23}{32} I_r.$$

The last two inequalities are precisely the Schur-complement conditions for $K_B \succ 0$ and $3I/4 - K_B \succ 0$.

Diagonalizing $B = U\,\mathrm{diag}(t_j)U^\top$ reduces compression to $Q^\top H_B Q = U\,\mathrm{diag}(\phi(t_j))U^\top$, where

$$\phi(t) = q^\top \left[I_2 - \sqrt{I_2 - K(t)} \right] q, \qquad q = (2,1)^\top/\sqrt{5}, \qquad K(t) = \begin{bmatrix} 1/4 & 1/8 \\ 1/8 & t \end{bmatrix}.$$

This continuous scalar function is strictly increasing on $[1/16, 23/32]$. To verify strictness, use the spectral integral

$$\sqrt{X} = \frac{1}{\pi} \int_0^\infty u^{-1/2} X (X + uI)^{-1}\, du, \qquad X \succ 0.$$

Increasing t decreases $X = I_2 - K(t)$ by a positive rank-one matrix. The resolvent identity makes the corresponding difference of square roots positive semidefinite. Its q quadratic form is strictly positive: the integrand is a positive multiple of $[q^\top (X + uI)^{-1} e_2]^2$, nonzero for all sufficiently large u since $q_2 \neq 0$. The integral identity itself follows by diagonalization and the scalar substitution $u = xv$.

At $t = 1/16$, q is an eigenvector of $K(t)$ with eigenvalue $5/16$, giving $\phi(t) = a$. At $t = 23/32$, the eigenvalues are $3/4$ and $7/32$, with squared q overlaps $36/85$ and $49/85$, giving $\phi(t) = b > a > 0$. Hence the raw compression image is exactly $\{X : aI_r \prec X \prec bI_r\}$. A trace-one $C \succ 0$ can be rescaled into this interval exactly when $a/\lambda_{\min}(C) < b/\lambda_{\max}(C)$, proving the exact image. Mixing a boundary member with I_r/r proves the closure statement.

The image is invariant under all orthogonal conjugations. Its extreme eigenvalues have supremum $\alpha = \kappa/(\kappa + r - 1)$ and infimum $\beta = 1/[1 + (r-1)\kappa]$. For $\widehat{G} = Q^\top G Q$, eigenvalue bounds and alignment of extreme eigenvectors give

$$\sup_{C \in \mathfrak{I}_S^{\mathrm{fixed}}} \|C - G\|_2 = \max\{\alpha - \lambda_{\min}(\widehat{G}), \lambda_{\max}(\widehat{G}) - \beta\}.$$

Every nonisotropic trace-one decision has minimum eigenvalue below $1/r$ and maximum above $1/r$, strictly increasing both terms. Thus the center is unique. At that center $\alpha - 1/r \geq 1/r - \beta$, giving the stated radius. Each covariance has the finite centered design from Appendix A.1, labeled by the same $W^\star$. Also $\Delta\Delta^\top = 5I_r/64$, so the ideal stable-rank rule retains all of S. □

Dependence on the fixed-label fiber. Isotropy need not minimize distortion on every bounded fiber. For $m = 2$, $n = 3$, $W^\star - W_0 = [I_2, 0]$, one step $\eta = 1/6$, and $\Delta = \frac{1}{6}\left[\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right]$, the entire fiber with $0 \prec H \prec 3I$ is

$$H_b = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & b \end{bmatrix}, \qquad 1 < b < 5/2.$$

Both bounds follow by Schur complements. In the observed basis $(1, 0, 1)^\top/\sqrt{2}, e_2$, its normalized compression is $\mathrm{diag}(t, 1-t)$, with $t = (b+3)/(b+5) \in (2/3, 11/15)$. A trace-one decision $\left[\begin{smallmatrix} g & z \\ z & 1-g \end{smallmatrix}\right]$ has distance $\sqrt{(t-g)^2 + z^2}$. Its unique minimax choice is $\mathrm{diag}(7/10, 3/10)$ with radius $1/30$; the isotropic decision has radius $7/30$.

B Task-Core Geometry and Dimension

B.1 Ideal right-action compression and its numerical approximation

For an orthogonal projector $P = P^\top = P^2$,

$$\|\Delta - \Delta P\|_F^2 = \mathrm{tr}\big((I - P)\Delta^\top \Delta (I - P)\big) \tag{37}$$

$$= \|\Delta\|_F^2 - \mathrm{tr}(P\Delta^\top \Delta). \tag{38}$$

By the Ky Fan maximum principle (Fan, 1949), the last trace is maximized by a top-r eigenspace of $\Delta^\top\Delta$. This is exactly a top-r right-singular subspace of Δ. A gap at the boundary, $\sigma_r > \sigma_{r+1}$, makes the projector unique. When $\sigma_r = \sigma_{r+1}$, every optimal projector contains the strictly larger singular directions and completes its rank inside the tied eigenspace.

The best-rank statement specifies the target subspace. The implementation approximates it with a fixed-width randomized range finder (Halko et al., 2011). Let $\Delta = U\Sigma V^\top$ and draw a Gaussian matrix $\Omega \in \mathbb{R}^{m\times q}$. The initial right sketch is

$$Y_0 = \Delta^\top \Omega = V\Sigma U^\top \Omega. \tag{39}$$

After column normalization and one power refinement,

$$Y_1 = \Delta^\top \Delta\, \mathcal{N}(Y_0). \tag{40}$$

For nonzero sketch columns, normalization preserves the span, so the refined sketch lies in the range of $V\Sigma^3 U^\top \Omega$ and emphasizes dominant right-singular directions. A $q \times q$ Rayleigh–Ritz problem returns the compact basis. The initial width is $q = 12$, and one retry may enlarge it when the estimated rank reaches the sketch boundary. This fixed-width rule sets the construction budget and uses no downstream rank selection.

B.2 The per-task norm-summary dimension

For $\Delta \neq 0$, write $E = \|\Delta\|_F^2$, $a = \|\Delta\|_2^2$, and $d = \min(m, n)$. The stable rank is $b = E/a$ (Rudelson & Vershynin, 2015). Its integer certificate depends only on these two norm summaries.

Proposition 3 (Per-task dimension certificate)**.** *For nonzero $\Delta \in \mathbb{R}^{m\times n}$,*

$$r_0 = \lceil b \rceil = \min\{r \in \mathbb{N} : E \le ra\} \tag{41}$$

$$= \min\{\mathrm{rank}(X) : X \in \mathbb{R}^{m\times n},\ \|X\|_F^2 = E,\ \|X\|_2^2 = a\}. \tag{42}$$

Proof. If X in this norm class has rank r, each nonzero squared singular value is at most a, so $E \le ra$. To attain the lower bound, write $b = \ell + \theta$, where $\ell = \lfloor b \rfloor$ and $0 \le \theta < 1$. Use ℓ squared singular values equal to a, one additional value θa if $\theta > 0$, and zeros elsewhere. Their sum is E, their maximum is a, and their number of nonzero values is $\lceil b \rceil \le d$. A rectangular diagonal embedding gives the required matrix. □

Approximation at the certified dimension. The same norm summaries allow a broad spectral tail. For $d > 1$ and $1 \le r \le d$, let X_r be a best rank-at-most-r approximation. The sharp worst-case relative squared error within this norm class is

$$\sup_{\substack{\|X\|_F^2 = E \\ \|X\|_2^2 = a}} \frac{\|X - X_r\|_F^2}{E} = \frac{d-r}{d-1}\left(1 - \frac{1}{b}\right). \tag{43}$$

For $r = 1$, the identity follows from $t_1 = a$. For $r \ge 2$, with squared singular values $t_1 = a \ge t_2 \ge \cdots \ge t_d$, the largest $r-1$ tail values have mean at least $(E-a)/(d-1)$. Thus $\sum_{j>r} t_j \le (d-r)(E-a)/(d-1)$. Equality holds for $t_2 = \cdots = t_d = (E-a)/(d-1)$.

With exact norms and no width cap, the guarded rank is the smallest positive integer satisfying $E \le (r + 10^{-4})a$. Appendix H.3 implements this reference with a Ritz denominator and the stated sketch-width and retry rules.

B.3 Response-minimax comparison geometry

Its maximum over $\|D\|_F = 1$ equals $\|C - G\|_2$, which is continuous in C; taking the closure therefore leaves the response-distortion supremum unchanged (Appendix A.3).

We prove the response part of Proposition 1. For every symmetric H,

$$\sup_{\|D\|_F = 1} |\operatorname{tr}(DHD^\top)| = \|H\|_2. \tag{44}$$

The upper bound follows by applying $|d^\top H d| \le \|H\|_2 \|d\|_2^2$ to each row of D. Equality is attained by $D = au^\top$, where $\|a\|_2 = 1$ and u is a unit eigenvector for an eigenvalue of largest absolute value.

For $r = 1$, $\mathcal{C}(S) = \{P_S\}$ and the radius is zero. For $r \ge 2$, let $g_{\min}$ and $g_{\max}$ be the extreme eigenvalues of G on S. Positivity and unit trace give $g_{\max} + g_{\min} \le 1$. Every $C \in \mathcal{C}(S)$ satisfies $0 \preceq C \preceq P_S$, so, restricted to S,

$$-(1 - g_{\min})I \preceq -g_{\max} I \preceq C - G \preceq I - G \preceq (1 - g_{\min})I. \tag{45}$$

Both operators vanish on $S^\perp$, giving $\|C - G\|_2 \le 1 - g_{\min}$. If u is a unit $g_{\min}$ eigenvector, $C = uu^\top$ attains the bound because $u^\top(C - G)u = 1 - g_{\min}$. This proves Equation (2). Finally $g_{\min} \le 1/r$, with equality exactly when all r eigenvalues equal $1/r$. Hence P_S/r uniquely minimizes the radius.

Joint comparison distortion. For fixed experts and $C_i, G_i \in \mathcal{C}(\mathcal{S}_i)$, define $\mathcal{J}_C(M) = \sum_i \operatorname{tr}[(M - \Delta_i)C_i(M - \Delta_i)^\top]$ and analogously $\mathcal{J}_G$. Since $C_i - G_i = P_i(C_i - G_i)P_i$, Equation (44) and homogeneity give

$$|\mathcal{J}_C(M) - \mathcal{J}_G(M)| \le \sum_i \rho_{\mathcal{S}_i}(G_i)\|(M - \Delta_i)P_i\|_F^2. \tag{46}$$

The isotropic core uniquely minimizes each distortion coefficient. This bound applies directly to the residuals generated by the shared candidate M.

From comparison geometry to the fused decision. Omit zero cores and assume $A \ne 0$. For the fixed checkpoints, retain the current $G_i = P_i/r_i$, participation A, anchor M_0, and selected τ. Set

$$R(M) = \tau^2 \|(M - M_0)A^{\dagger/2}\|_F^2,$$
$$F_C(M) = \mathcal{J}_C(M) + R(M), \qquad F_G(M) = \mathcal{J}_G(M) + R(M).$$

The geometry-replacement optimizer M_C uses the same candidate class and the same R as M_G. Only G_i in the response objective is replaced by C_i; the oracle does not recompute A, M_0, τ, or R. On $\mathcal{A}(M_0)$, $M_G = M_\tau$ by Theorems 2–3, including $\tau = 0$.

Proposition B.1 (Worst-geometry response and a fusion certificate)**.** *For the independent product class $C_i \in \mathcal{C}(\mathcal{S}_i)$,*

$$L(M) := \sup_{C_i \in \mathcal{C}(\mathcal{S}_i)} \mathcal{J}_C(M) = \sum_i \|(M - \Delta_i)P_i\|_2^2.$$

The supremum is unchanged over $\mathcal{C}_{++}(\mathcal{S}_i)$. Put $L_R = L + R$ and $r_{\max} = \max_i r_i$. On any common candidate class where M_G minimizes F_G and M_R minimizes L_R, their robust-objective gap is nonnegative and satisfies

$$L_R(M_G) - L_R(M_R) \le \sum_i \left[\|(M_G - \Delta_i)P_i\|_2^2 - \frac{\|(M_G - \Delta_i)P_i\|_F^2}{r_i} \right],$$
$$L_R(M_G) \le r_{\max} L_R(M_R).$$

Proof. For $D = M - \Delta_i$, maximizing $\operatorname{tr}(C_i P_i D^\top D P_i)$ over supported unit-trace PSD matrices selects a largest-eigenvalue direction and gives $\|DP_i\|_2^2$. Positive-definite approximations attain the same supremum, and the independent task maxima add. Since DP_i has rank at most r_i and $R \ge 0$, the norm inequalities give $F_G \le L_R \le r_{\max} F_G$. Thus $L_R(M_R) \ge F_G(M_G)$ gives the additive certificate, and $L_R(M_G) \le r_{\max} F_G(M_G) \le r_{\max} F_G(M_R) \le r_{\max} L_R(M_R)$ gives the multiplicative one. On $\mathcal{A}(M_0)$, $\mathcal{J}_G$ is coercive in its supported coordinates, so both minimizers exist. □

A two-expert example. Take $\Delta_1 = I_2$ and $\Delta_2 = \operatorname{diag}(2, 0)$. Their retained ranks are 2 and 1, and

$$A = \operatorname{diag}(3/2, 1/2), \qquad B = \operatorname{diag}(5/2, 1/2), \qquad d_{\mathrm{part}} = 1 + \sqrt{3}/2.$$

The guarded rule gives $k = s = 2$, $\tau = 0$, and $M_G = \operatorname{diag}(5/3, 1)$. The robust objective is bounded below by $(M_{11} - 1)^2 + (M_{11} - 2)^2 \ge 1/2$, attained at $M_R = \operatorname{diag}(3/2, 1)$. Yet

$$L(M_G) = 5/9 > 1/2 = L(M_R), \qquad \mathcal{J}_G(M_G) = 1/3 < 3/8 = \mathcal{J}_G(M_R).$$

Thus the minimax comparison operator need not yield a minimizer of the worst-geometry response objective, even with the same regularizer on both sides.

Theorem B.2 (Comparison-geometry regret)**.** *Fix the experts, supports, $C_i, G_i \in \mathcal{C}(\mathcal{S}_i)$, a common candidate class $\mathcal{K}$, and a common regularizer R. Suppose $M_G \in \arg\min_{\mathcal{K}} F_G$ and $M_C \in \arg\min_{\mathcal{K}} F_C$ exist. For*

$$b_i(M) = \|(M - \Delta_i)P_i\|_F^2, \qquad e_i = \|C_i - G_i\|_2, \qquad \delta_C(M) = \mathcal{J}_C(M) - \mathcal{J}_G(M),$$

the excess regularized response objective satisfies

$$\begin{aligned} 0 \le \mathrm{Reg}_C(M_G) &:= F_C(M_G) - F_C(M_C) \\ &\le \delta_C(M_G) - \delta_C(M_C) \\ &\le \sum_i e_i \{b_i(M_G) + b_i(M_C)\} \\ &\le \sum_i \rho_{\mathcal{S}_i}(G_i) \{b_i(M_G) + b_i(M_C)\}. \end{aligned}$$

On the full affine class $\mathcal{A}(M_0)$, use $R(M) = \operatorname{tr}[(M - M_0)T_R(M - M_0)^\top]$ with fixed $T_R \succeq 0$, and define

$$K_C = Q^\top \left(\sum_i C_i + T_R \right) Q, \qquad S_C = \sum_i (M_G - \Delta_i)(C_i - G_i)Q.$$

If $K_C \succ 0$, the oracle is unique and

$$(M_G - M_C)Q = S_C K_C^{-1}, \qquad \mathrm{Reg}_C(M_G) = \operatorname{tr}(S_C K_C^{-1} S_C^\top).$$

For singular K_C, the regret identity holds with $K_C^\dagger$, and the Frobenius distance from M_G to the oracle minimizer set is $\|S_C K_C^\dagger\|_F$.

Proof. Inserting F_G at both candidates gives

$$\mathrm{Reg}_C(M_G) = \delta_C(M_G) + [F_G(M_G) - F_G(M_C)] - \delta_C(M_C).$$

The bracket is nonpositive. Since $C_i - G_i$ is supported on $\mathcal{S}_i$, the rowwise quadratic bound gives $|\delta_C(M)| \le \sum_i e_i b_i(M)$; Proposition 1 bounds e_i by $\rho_{\mathcal{S}_i}(G_i)$. This argument needs no convexity of $\mathcal{K}$.

For the quadratic identity, write $M = M_0 + XQ^\top$. Subtracting the F_G stationarity equation from the F_C gradient at M_G gives that gradient divided by two as S_C. Subtract the zero gradient at M_C to obtain $((M_G - M_C)Q)K_C = S_C$. Completing the square gives

$$F_C(M) - F_C(M_C) = \operatorname{tr}\big[((M - M_C)Q)K_C((M - M_C)Q)^\top\big].$$

This proves the inverse formula. If $v \in \ker K_C$, positivity forces $C_iQv = 0$ for every i and $T_RQv = 0$. The reduced linear term therefore annihilates v, so the normal equation is compatible and minimizers exist. Consequently $S_C = S_CK_C^\dagger K_C$. Substitution into the completed square proves the pseudoinverse identity. The unconstrained null coordinates of M_C can match those of M_G, giving the minimum distance. □

For $\mu_C = \lambda_{\min}(K_C) > 0$, the identity and $\|S_C\|_F \le \sum_i e_i\sqrt{b_i(M_G)}$ yield

$$\operatorname{Reg}_C(M_G) \le \frac{\big[\sum_i e_i\sqrt{b_i(M_G)}\big]^2}{\mu_C}, \qquad \|M_G - M_C\|_F \le \frac{\sum_i e_i\sqrt{b_i(M_G)}}{\mu_C}.$$

For the current $T_R = \tau^2 A^\dagger$ and $\tau > 0$, $\mu_C \ge \tau^2/\lambda_1(A)$ uniformly over compatible geometries. Thus

$$\sup_C \operatorname{Reg}_C(M_\tau) \le \frac{\lambda_1(A)}{\tau^2}\left[\sum_i (1 - 1/r_i)\|(M_\tau - \Delta_i)P_i\|_F\right]^2.$$

A uniform distance bound at $\tau = 0$ instead needs a positive curvature condition, such as $C_i \succeq cG_i$ for a fixed $c > 0$, which gives $\mu_C \ge c\lambda_s(A) + \tau^2/\lambda_1(A)$. These are regularized response comparisons; the network-level quantities remain those of Appendix G.

All generic inequalities also hold on a common numerically selected affine support. The quadratic identities then use its orthonormal basis in place of Q, with the same frozen quadratic penalty restricted to that space. An additive ξ optimization error in M_G adds ξ to the generic regret bound; the exact quadratic identity uses exact stationarity.

Proposition B.3 (Sharp two-residual regret envelope)**.** *Fix nonzero supports and positive residual budgets b_i, independent of the trial $G_i \in \mathcal{C}(\mathcal{S}_i)$. Define*

$$\Omega(G) = \sup_{\substack{C_i \in \mathcal{C}(\mathcal{S}_i) \\ D_i^\pm = D_i^\pm P_i,\ \|D_i^\pm\|_F^2 \le b_i}} \sum_i \operatorname{tr}\big[D_i^+(C_i - G_i)(D_i^+)^\top - D_i^-(C_i - G_i)(D_i^-)^\top\big].$$

The residual blocks vary independently, but each pair uses the same C_i. Writing $g_{i,\min}$ and $g_{i,\max}$ for the extreme eigenvalues of $G_i|_{\mathcal{S}_i}$,

$$\Omega(G) = \sum_{i:r_i \ge 2} b_i(1 + g_{i,\max} - g_{i,\min}), \qquad \min_G \Omega(G) = \sum_{i:r_i \ge 2} b_i.$$

The unique minimizing tuple is $G_i = P_i/r_i$. If both optimizers in Theorem B.2 satisfy $b_i(M) \le b_i$, then $\operatorname{Reg}_C(M_G) \le \Omega(G)$.

Proof. Positivity gives $\lambda_{\max}(C_i - G_i) \le 1 - g_{i,\min}$ and $-\lambda_{\min}(C_i - G_i) \le g_{i,\max}$, proving the upper bound. For $r_i \ge 2$, choose orthogonal extreme eigenvectors u, v of G_i, with u minimizing its eigenvalue, and take $C_i = uu^\top$. Single nonzero residual rows $\sqrt{b_i}u^\top$ and $\sqrt{b_i}v^\top$ attain both terms simultaneously. If G_i is scalar, any orthogonal pair works. Rank one has $C_i = G_i = P_i$ and contributes zero. Since $g_{i,\min} \le 1/r_i \le g_{i,\max}$, every positive-budget term is uniquely minimized by equal eigenvalues. The two actual optimizer residuals are feasible choices in this relaxation, giving the regret bound. □

The envelope is sharp for independent residual blocks; shared M and fixed experts impose further constraints. Its minimization is not a minimax claim for the actual optimizer regret. Common budgets can be supplied by any bounded candidate class. For the unconstrained affine class with the source R frozen and $\tau > 0$, put

$$V_0 = \sum_i \|(M_0 - \Delta_i)P_i\|_2^2, \quad \mu_R = \tau^2/\lambda_1(A), \quad b_i = \Big(\|(M_0 - \Delta_i)P_i\|_F + \sqrt{V_0/\mu_R}\Big)^2.$$

Every minimizer under any supported unit-trace geometry has $R(M) \le F(M) \le F(M_0) \le V_0$, hence obeys these residual bounds. A zero bound may be enlarged to a positive one for the uniqueness statement. At $\tau = 0$, common finite budgets are an additional restriction, not supplied by the general geometry class.

B.4 Secondary structural characterizations

Equivariance. Consider a subspace-only map $\Phi(S) \succeq 0$, supported on S, with unit trace and $\Phi(US) = U\Phi(S)U^\top$ for every orthogonal U. At $S_0 = \mathrm{span}(e_1, \dots, e_r)$, support gives $\Phi(S_0) = \mathrm{diag}(H, 0)$. Every $U = \mathrm{diag}(R, I)$ with $R \in O(r)$ leaves S_0 unchanged, so $H = RHR^\top$. A symmetric matrix commuting with all such R is scalar; unit trace gives $H = I_r/r$. Transport by an orthogonal change of coordinates yields $\Phi(S) = P_S/r$.

Frobenius center. The same operator uniquely minimizes squared Frobenius distance to $\mathcal{C}(S)$ in the worst case:

$$\frac{P_S}{r} = \arg\min_{G \in \mathcal{C}(S)} \sup_{C \in \mathcal{C}(S)} \|C - G\|_F^2, \qquad \min_G \sup_C \|C - G\|_F^2 = 1 - \frac{1}{r}. \tag{47}$$

For $G_0 = P_S/r$, $\|C - G_0\|_F^2 = \|C\|_F^2 - 1/r \le 1 - 1/r$, attained at rank-one C. For any G, choose a unit eigenvector u of its smallest eigenvalue $g_{\min} \le 1/r$. Then

$$\|G - uu^\top\|_F^2 = \|G\|_F^2 + 1 - 2g_{\min} \ge 1 - \frac{1}{r}. \tag{48}$$

Equality in the lower bound requires $\|G\|_F^2 = 1/r$, hence all its supported eigenvalues equal $1/r$.

B.5 Zero updates and duplicated tasks

A zero update has no observed action, so we set $r_i = 0$, $\mathcal{S}_i = \{0\}$, and $G_i = 0$. Its shape-only zero column contributes nothing to the numerical factor sum. Duplicating a nonzero task adds another G_i and $\Delta_i G_i$ to A and B, and another Δ_i to the anchor. Task multiplicity therefore contributes comparison mass and additive action.

Off-core multiplicity boundary. The additive effect is explicit when the experts are identical. Suppose $\Delta_i = \Delta$ for $i = 1, \dots, K$, all retained cores share the same rank-r projector P, and $G_i = P/r$. Then

$$A = \frac{K}{r}P, \qquad B = \frac{K}{r}\Delta P, \qquad M_0 = K\Delta.$$

The positive participation spectrum is flat, so $k = s = r$ and $\tau = 0$. Substitution into Equation (11) gives

$$M = \Delta P + K\Delta(I - P). \tag{49}$$

Thus repeated identical experts are idempotent on the retained core but additive on the truncated complement. Exact duplicate idempotence therefore holds only when $\Delta(I - P) = 0$. This is a boundary of the summed-update reference under truncated cores rather than evidence that repeated checkpoint occurrences are necessarily redundant: checkpoint values alone do not specify whether multiplicity represents duplicated provenance or multiple task occurrences that should retain separate mass. Appendix I measures how closely real expert collections approach this limiting geometry.

Proposition B.4 (Unit nonzero mass and the zero-update boundary)**.** *In fixed input dimension n, any map with $G(\Delta) \succeq 0$ and $\mathrm{tr}\, G(\Delta) = 1$ for every $\Delta \neq 0$, and $G(0) = 0$, is discontinuous at zero. For every nonzero update,*

$$\|G(\Delta) - G(0)\|_2 \ge 1/n, \qquad \|G(\Delta) - G(0)\|_F \ge 1/\sqrt{n}.$$

For the task-core rule these norms equal $1/r(\Delta)$ and $1/\sqrt{r(\Delta)}$.

Proof. The nonnegative eigenvalues sum to one, so their maximum is at least $1/n$ and their squared sum is at least $1/n$ by Cauchy–Schwarz. These lower bounds exclude continuity at zero. For P/r, all r positive eigenvalues equal $1/r$. □

This boundary can also affect the fused update. With scalar experts $\Delta_1 = 1$ and $\Delta_2 = \epsilon \neq 0$, both cores equal one, $A = 2$, $B = 1 + \epsilon$, and $k = s = 1$ gives $\tau = 0$. Therefore

$$M(1, \epsilon) = \frac{1+\epsilon}{2} \longrightarrow \frac{1}{2}, \qquad M(1, 0) = 1.$$

Continuity through insertion or removal of a vanishing expert would require relaxing the unit-mass convention. Appendix E.4 gives local stability on strata that avoid this and the rank, support, and selector boundaries. Appendix I realizes the zero-update boundary under controlled expert scaling and separately measures the distance of the unmodified evaluation checkpoints from the same regime.

C Fusion-Frame Core Matching

C.1 Frame identity and principal angles

For $P_i = Q_i Q_i^\top$ and $x \in \mathcal{U}$,

$$x^\top A x = \sum_i \frac{x^\top P_i x}{r_i} = \sum_i \frac{\|P_i x\|_2^2}{r_i}. \tag{50}$$

The smallest and largest positive eigenvalues of A are the extreme Rayleigh quotients on $\mathcal{U}$, which proves Eq. (6). Summing the unit-mass cores also gives the total participation

$$\operatorname{tr} A = \sum_i \operatorname{tr}(P_i/r_i) = K \tag{51}$$

when every task has a nonzero core.

Task-core overlap also has an exact principal-angle form:

$$\operatorname{tr}(G_i G_j) = \frac{\|Q_i^\top Q_j\|_F^2}{r_i r_j} = \frac{\sum_\ell \cos^2 \theta_{ij,\ell}}{r_i r_j}. \tag{52}$$

Consequently,

$$\operatorname{tr}(A^2) = \sum_i \frac{1}{r_i} + 2 \sum_{i<j} \frac{\|Q_i^\top Q_j\|_F^2}{r_i r_j}. \tag{53}$$

Thus the participation spectrum depends on both core rank and the principal-angle arrangement of the retained subspaces.

C.2 Pythagorean decomposition

Expanding Eq. (5),

$$\mathcal{J}(M) = \operatorname{tr}(MAM^\top) - 2\operatorname{tr}(BM^\top) + \sum_i \operatorname{tr}(\Delta_i G_i \Delta_i^\top). \tag{54}$$

Any $M^\star$ satisfying $M^\star A = B$ on $\mathcal{U}$ makes the cross term vanish in the difference:

$$\mathcal{J}(M) - \mathcal{J}(M^\star) = \operatorname{tr}\big((M - M^\star)A(M - M^\star)^\top\big) \tag{55}$$

$$= \|(M - M^\star)A^{1/2}\|_F^2. \tag{56}$$

The anchored representative in Eq. (7) satisfies this normal equation because

$$M^\star A = M_0 A + (B - M_0 A)A^\dagger A = B. \tag{57}$$

The identity $B = BP_{\mathcal{U}}$ follows because every G_i is supported on $\mathcal{U}$. Substituting $M = BA^\dagger$ into the expanded objective gives Eq. (9). Any component in the right null space of A leaves the objective unchanged; Equation (7) chooses the representative whose null-space component matches M_0.

C.3 Conditioning and residual are independent

Perfect conditioning with arbitrary residual. In one dimension, take $G_1 = G_2 = 1$, $\Delta_1 = a$, and $\Delta_2 = -a$. Then $A = 2$ has condition number one, yet

$$\mathcal{J}(M) = (M - a)^2 + (M + a)^2 = 2M^2 + 2a^2, \tag{58}$$

so $\mathcal{J}^\star = 2a^2$ can be arbitrarily large.

Poor conditioning with zero residual. Let $q_1 = e_1$, $q_2 = (\cos\theta, \sin\theta)^\top$ with $0 < \theta < \pi/2$, $G_i = q_i q_i^\top$, and $\Delta_i = G_i$. The common operator $M^\star = I$ satisfies $(M^\star - \Delta_i)q_i = 0$ for both tasks, hence $\mathcal{J}^\star = 0$. The eigenvalues of $A = G_1 + G_2$ are $1 + \cos\theta$ and $1 - \cos\theta$, so the condition number diverges as $\theta \to 0$.

The two examples show that equation strength and unmatched requested action are independent properties. The experiments therefore report participation conditioning and normalized core residual separately in Table 32.

D Anchored Regularized Recovery

D.1 Minimum synthesis energy and reduced closed form

Let

$$Z = \left[\frac{Q_1}{\sqrt{r_1}}, \ldots, \frac{Q_K}{\sqrt{r_K}}\right], \qquad \mathcal{T} = \left[\frac{\Delta_1 Q_1}{\sqrt{r_1}}, \ldots, \frac{\Delta_K Q_K}{\sqrt{r_K}}\right],$$

so that $A = ZZ^\top$ and $B = \mathcal{T}Z^\top$. The columns of Z are the normalized comparison directions, while the matching columns of $\mathcal{T}$ are the expert output changes requested on those directions.

Minimum synthesis energy and anchored response recovery. Let $\mathcal{U} = \mathrm{range}(Z) = \mathrm{range}(A)$. For a physical correction E with right support in $\mathcal{U}$,

$$\inf_{LZ^\top = E} \|L\|_F^2 = \|EA^{\dagger/2}\|_F^2, \qquad L_{\min} = EA^\dagger Z.$$

If $E \neq EP_{\mathcal{U}}$, no such synthesis is feasible. For any $\rho > 0$, the coefficient problem

$$\min_L \ \|(M_0 + LZ^\top)Z - \mathcal{T}\|_F^2 + \rho^2\|L\|_F^2$$

has a unique coefficient solution and induces

$$M_\rho = M_0 + (B - M_0 A)Q\Lambda(\Lambda^2 + \rho^2 I)^{-1}Q^\top.$$

Profiling out L makes this coefficient problem exactly

$$\min_{M \in \mathcal{A}(M_0)} \ \mathcal{J}(M) + \rho^2\|(M - M_0)A^{\dagger/2}\|_F^2.$$

At $\rho = 0$, coefficient minimizers can be nonunique when Z is redundant, but they all induce the same physical update $M^\star$ in Equation (7).

Proof. Take a compact SVD $Z = Q\Lambda^{1/2}V^\top$, so $A = Q\Lambda Q^\top$ and $V^\top V = I$. For $E = EP_{\mathcal{U}}$, the constraint $LZ^\top = E$ is equivalent to

$$LV = EQ\Lambda^{-1/2}.$$

Every feasible coefficient matrix therefore has the form

$$L = EQ\Lambda^{-1/2}V^\top + N(I - VV^\top).$$

The two summands are orthogonal in Frobenius inner product, hence

$$\|L\|_F^2 = \mathrm{tr}(EQ\Lambda^{-1}Q^\top E^\top) + \|N(I - VV^\top)\|_F^2.$$

The second term is uniquely minimized at zero, proving the minimum-energy identity and $L_{\min} = EA^\dagger Z$. If E has right action outside $\mathcal{U}$, every matrix $LZ^\top$ has right support in $\mathcal{U}$, so the synthesis constraint is infeasible.

For the penalized fit, put $H = Z^\top Z$ and $D = \mathcal{T} - M_0 Z$. The coefficient objective becomes $\|LH - D\|_F^2 + \rho^2\|L\|_F^2$, whose first-order condition is

$$L(H^2 + \rho^2 I) = DH.$$

For $\rho > 0$ this gives a unique $L_\rho = DH(H^2 + \rho^2 I)^{-1}$. Substituting the compact SVD and using $DZ^\top = B - M_0 A$ gives the displayed physical update. Finally, $\mathcal{J}(M) = \|MZ - \mathcal{T}\|_F^2$, and for each supported $E = M - M_0$, the minimum-energy identity is exactly the least coefficient cost among all representations of that correction. Eliminating L therefore gives the displayed profiled objective. The $\rho = 0$ statement follows because coefficient null-space components annihilate $Z^\top$. □

Setting $\rho = \tau$ in the synthesis result above already gives Equation (11) and the anchored objective in Equation (13). The familiar reduced-coordinate Tikhonov form (Tikhonov & Arsenin, 1977) is recovered as follows. Write every candidate in the anchored affine class as

$$M = M_0 + CQ^\top. \tag{59}$$

Let

$$Y = (B - M_0 A)Q. \tag{60}$$

Because $AQ = Q\Lambda$, the unregularized normal equation becomes

$$C\Lambda = Y. \tag{61}$$

The regularized reduced problem is

$$\min_C \ \|C\Lambda - Y\|_F^2 + \tau^2 \|C\|_F^2. \tag{62}$$

Its first-order condition is

$$(C\Lambda - Y)\Lambda + \tau^2 C = 0, \tag{63}$$

which yields

$$C_\tau = Y\Lambda(\Lambda^2 + \tau^2 I)^{-1}. \tag{64}$$

This gives Eq. (11).

The reduced least-squares problem above and the profiled synthesis objective have the same correction but not identical objective values. To recover the original anchored core objective directly, remove terms in $\mathcal{J}$ that do not depend on C. The remaining expression is

$$\operatorname{tr}(C\Lambda C^\top) - 2\operatorname{tr}(YC^\top). \tag{65}$$

The anchored penalty is

$$\|(M - M_0)A^{\dagger/2}\|_F^2 = \operatorname{tr}(C\Lambda^{-1}C^\top). \tag{66}$$

Differentiating the sum gives

$$C\Lambda - Y + \tau^2 C\Lambda^{-1} = 0. \tag{67}$$

Multiplying on the right by Λ recovers the same Tikhonov normal equation. In the original core objective, the cost of moving along mode j is therefore weighted by τ^2/λ_j.

D.2 Interpolation and aggregate objective comparison

The unregularized correction is $C^\star = Y\Lambda^{-1}$. The regularized solution can be written as

$$C_\tau = C^\star \operatorname{diag}\left(\frac{\lambda_j^2}{\lambda_j^2 + \tau^2}\right), \tag{68}$$

which gives Eq. (12). Setting $\tau = 0$ recovers $M^\star$ on $\mathcal{U}$, whereas $\tau \to \infty$ returns the anchor M_0. For every finite τ,

$$\mathcal{J}(M_\tau) + \tau^2 \|(M_\tau - M_0)A^{\dagger/2}\|_F^2 \le \mathcal{J}(M_0). \tag{69}$$

Hence $\mathcal{J}(M_\tau) \le \mathcal{J}(M_0)$ for the aggregate core objective. Individual task-specific terms can follow different orderings, since the objective sums their discrepancies.

D.3 Operator stability at a fixed positive τ

For fixed $\tau > 0$ and positive semidefinite A, $\widetilde{A}$ of the same size, $g_\tau(A) = A(A^2 + \tau^2 I)^{-1}$ obeys

$$\|g_\tau(A)\|_2 \le \frac{1}{2\tau}, \qquad \|g_\tau(\widetilde{A}) - g_\tau(A)\|_2 \le \frac{\|\widetilde{A} - A\|_2}{\tau^2}. \tag{70}$$

Equation (70) controls amplification for fixed positive τ.

Define $g_\tau(t) = t/(t^2 + \tau^2)$. Its maximum over $t \ge 0$ is $1/(2\tau)$ at $t = \tau$, proving the first inequality in Eq. (70). For a self-adjoint PSD matrix,

$$g_\tau(A) = \frac{1}{2}\left[(A - i\tau I)^{-1} + (A + i\tau I)^{-1}\right]. \tag{71}$$

The resolvent identity gives

$$(A - zI)^{-1} - (\widetilde{A} - zI)^{-1} = (A - zI)^{-1}(\widetilde{A} - A)(\widetilde{A} - zI)^{-1}. \tag{72}$$

For $z = \pm i\tau$, each resolvent has operator norm at most $1/\tau$. Averaging the two identities yields

$$\|g_\tau(\widetilde{A}) - g_\tau(A)\|_2 \le \frac{\|\widetilde{A} - A\|_2}{\tau^2}. \tag{73}$$

Let $E_0 = B - M_0 A$ denote the right-hand-side discrepancy in the original input coordinates, so $Y = E_0 Q$. If this discrepancy also changes from E_0 to $\widetilde{E}_0$, then

$$\|\widetilde{E}_0 g_\tau(\widetilde{A}) - E_0 g_\tau(A)\|_F \le \frac{\|\widetilde{E}_0 - E_0\|_F}{2\tau} + \frac{\|E_0\|_F}{\tau^2}\|\widetilde{A} - A\|_2. \tag{74}$$

This controls the correction under perturbations of both E_0 and A; Appendix E.4 also varies τ.

D.4 Tail sensitivity of the unregularized inverse

Weak-support instability. Subtracting the reference's weighted outputs from the expert requests leaves $Y = (B - M_0 A)Q$ in the shared coordinates. The exact reduced correction is $Y\Lambda^{-1}$. In mode j, error in the remaining output request is divided by λ_j, while first-order error in A is amplified at order $1/\lambda_j^2$. This sensitivity motivates limiting the correction in weakly supported directions.

For comparison, let $C = Y\Lambda^{-1}$ and

$$\widetilde{C} = (Y + E)(\Lambda + F)^{-1}, \tag{75}$$

with $\|F\|_2 < \lambda_s$. Then

$$\widetilde{C} - C = (E - CF)(\Lambda + F)^{-1}, \tag{76}$$

and

$$\|\widetilde{C} - C\| \le \frac{\|E\|}{\lambda_s - \|F\|_2} + \frac{\|Y\|\|F\|_2}{\lambda_s(\lambda_s - \|F\|_2)}. \tag{77}$$

The first-order scaling is $\|E\|/\lambda_s + \|Y\|\|F\|/\lambda_s^2$. If the numerical support changes, pseudoinverse stability additionally depends on subspace rotation and the spectral gap (Davis & Kahan, 1970; Wedin, 1973).

D.5 Sum-update anchor and net cross-action

Proposition 4 (Additive anchor characterization). *For $M_\Sigma = \sum_i \Delta_i$ and $T_j = (\sum_{i \ne j} \Delta_i)P_j$, Equation (10) holds. If $T_j = 0$ for every nonzero task core, then $\mathcal{J}(M_\Sigma) = 0$ and $M_\tau = M_\Sigma$ for every $\tau \ge 0$.*

Among jointly linear, task-permutation-invariant maps $F_K : (\mathbb{R}^{m\times n})^K \to \mathbb{R}^{m\times n}$, the condition

$$F_K(\Delta, 0, \dots, 0) = \Delta \quad \text{for every } \Delta \tag{78}$$

uniquely selects $F_K(\Delta_1, \dots, \Delta_K) = \sum_i \Delta_i$.

Proof. Separating the task-j term in the anchor gives

$$M_\Sigma A - B = \sum_j \sum_{i \ne j} \frac{\Delta_i P_j}{r_j} = \sum_j \frac{T_j}{r_j}, \qquad (M_\Sigma - \Delta_j)P_j = T_j. \tag{79}$$

Since P_j is an orthogonal projector, the jth term in $\mathcal{J}(M_\Sigma)$ equals $\|T_j\|_F^2/r_j$. When all T_j vanish, both the objective and $B - M_\Sigma A$ are zero, so Equation (11) leaves the sum unchanged. In particular, pairwise orthogonal P_i and own-support updates $\Delta_i = \Delta_i P_i$ give $\Delta_i P_j = 0$ for $i \ne j$.

For the reference characterization, let $L_j(X)$ be F_K evaluated with X in slot j and zeros elsewhere. Joint linearity gives $F_K(\Delta_1, \dots, \Delta_K) = \sum_j L_j(\Delta_j)$. Permutation invariance makes all L_j equal to one map L. Equation (78) gives $L(X) = X$, proving uniqueness. The sum itself satisfies all these properties. □

E MINIMAX SOFT SPECTRAL BOUNDARY

E.1 PROOF OF PROPOSITION 2

Since right multiplication by a diagonal matrix has Frobenius induced norm equal to its largest absolute diagonal entry,

$$\mathcal{E}_k(\tau) = \max_j |h_\tau(\lambda_j) - \mathbf{1}\{j \le k\}| . \tag{80}$$

The shrink $h_\tau(\lambda)$ increases with λ. The largest error among retained modes therefore occurs at λ_k, and the largest error among discarded modes occurs at λ_{k+1}:

$$\mathcal{E}_k(\tau) = \max\left\{ \frac{\tau^2}{\lambda_k^2 + \tau^2}, \frac{\lambda_{k+1}^2}{\lambda_{k+1}^2 + \tau^2} \right\}. \tag{81}$$

The first term is strictly increasing in $\tau > 0$ and the second is strictly decreasing. Their maximum is uniquely minimized at equality:

$$\frac{\tau^2}{\lambda_k^2 + \tau^2} = \frac{\lambda_{k+1}^2}{\lambda_{k+1}^2 + \tau^2}. \tag{82}$$

Cross-multiplication gives $\tau^4 = \lambda_k^2 \lambda_{k+1}^2$, hence $\tau^\star = \sqrt{\lambda_k \lambda_{k+1}}$. Substitution gives the minimal error in Eq. (16).

Corollary 2 (Hard-reference comparison with ordinary anchored ridge)**.** *Fix $1 \le k < s$ and $\lambda_1 \ge \cdots \ge \lambda_s > 0$, with the same H_k and $C^\star = Y\Lambda^{-1}$ as in Proposition 2. Ordinary anchored Frobenius ridge minimizes*

$$\|(C - C^\star)\Lambda^{1/2}\|_F^2 + \gamma \|C\|_F^2, \qquad \gamma \ge 0,$$

and has interpolation $H_\gamma^{(1)} = \mathrm{diag}(\lambda_j/(\lambda_j + \gamma))$ relative to $C^\star$. Under the same error criterion,

$$\mathcal{E}_k^{(1)}(\gamma) = \sup_{\|C^\star\|_F = 1} \|C^\star (H_\gamma^{(1)} - H_k)\|_F,$$

its unique minimizer and minimum are

$$\gamma^\star = \sqrt{\lambda_k \lambda_{k+1}}, \qquad \min_\gamma \mathcal{E}_k^{(1)}(\gamma) = \frac{\sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}}}.$$

The squared-Tikhonov minimum in Equation (16) *is strictly smaller when $\lambda_k > \lambda_{k+1}$; at a tied boundary both minima equal $1/2$.*

Proof. The ordinary-ridge normal equation is $C\Lambda - Y + \gamma C = 0$, giving $C_\gamma = Y(\Lambda + \gamma I)^{-1} = C^\star H_\gamma^{(1)}$. The two families' worst errors occur at the same boundary modes:

$$\mathcal{E}_k^{(1)}(\gamma) = \max\left\{ \frac{\gamma}{\lambda_k + \gamma}, \frac{\lambda_{k+1}}{\lambda_{k+1} + \gamma} \right\},$$
$$\mathcal{E}_k(\tau) = \max\left\{ \frac{\tau^2}{\lambda_k^2 + \tau^2}, \frac{\lambda_{k+1}^2}{\lambda_{k+1}^2 + \tau^2} \right\}.$$

In each maximum the first term increases strictly and the second decreases strictly on the positive parameter axis. Equalizing them gives $\gamma^2 = \lambda_k \lambda_{k+1}$ and $\tau^4 = \lambda_k^2 \lambda_{k+1}^2$, respectively, proving the unique minima. For $t = \sqrt{\lambda_{k+1}/\lambda_k} \in (0, 1)$, their difference is

$$\frac{t}{1+t} - \frac{t^2}{1+t^2} = \frac{t(1-t)}{(1+t)(1+t^2)} > 0.$$

At $t = 1$ the difference is zero, and both optimized filters assign $1/2$ throughout the boundary eigenspace. □

This comparison is conditional on the same fixed hard reference and these two filter families under the stated error criterion; it establishes neither optimality among all spectral filters nor downstream-accuracy superiority.

E.2 Repeated-eigenspace impossibility

Suppose eigenvalue a has multiplicity r, while a fixed rank boundary attempts to keep only t of those directions with $0 < t < r$. A rule that depends only on A and remains equivariant under coordinate rotations must commute with every $R \in O(r)$ inside the tied eigenspace. Its restriction must therefore be cI_r. Projector values require $c \in \{0, 1\}$, so no such rule can have rank t. The continuous filter assigns one value to the entire eigenspace; when the boundary is tied, that value is $1/2$.

E.3 Participation dimension and the zero-ridge window

For the normalized factor Z in Section 4.2, $A = ZZ^\top$. The ratio in Equation (14) is the effective rank $(\operatorname{tr} X)^2/\|X\|_F^2$ studied by Lopes (2013), applied to $X = \sqrt{A}$.

Proposition 5 (Participation dimension certificate)**.** *For $Z \neq 0$ with $s = \operatorname{rank}(Z)$, write $N = \|Z\|_*$ and $F = \|Z\|_F$. Then*

$$d_{\mathrm{part}} = \frac{(\operatorname{tr}\sqrt{A})^2}{\operatorname{tr} A} = \frac{N^2}{F^2} \in [1, s], \tag{83}$$

with $d_{\mathrm{part}} = s$ exactly when the positive singular values of Z are equal. The integer $k_0 = \lceil d_{\mathrm{part}} \rceil$ satisfies

$$k_0 = \min\{k \in \mathbb{N} : N \le \sqrt{k}F\} \tag{84}$$

$$= \min\{\operatorname{rank}(X) : X \text{ has the size of } Z,\ \|X\|_* = N,\ \|X\|_F = F\}. \tag{85}$$

For $0 < \varepsilon < 1$, its guarded version is exactly

$$k_\varepsilon = \min\{k \in \{1, \dots, s\} : N^2 \le (k + \varepsilon)F^2\} = \min\{s, \lceil d_{\mathrm{part}} - \varepsilon \rceil\}. \tag{86}$$

For $\varepsilon = 10^{-4}$, the guarded reference is exactly

$$k = \min\left\{j \in \{1, \dots, s\} : \|Z\|_*^2 \le (j + \varepsilon)\|Z\|_F^2\right\}. \tag{87}$$

Proof. Let $t_1 \ge \cdots \ge t_s > 0$ be the singular values of Z. The positive eigenvalues of A are t_j^2, so $\operatorname{tr}\sqrt{A} = \sum_j t_j = N$ and $\operatorname{tr} A = \sum_j t_j^2 = F^2$. Nonnegative cross terms give $N^2 \ge F^2$, while Cauchy–Schwarz gives $N^2 \le sF^2$, with equality exactly for equal t_j.

Any rank-k matrix with these two norms satisfies $N \le \sqrt{k}F$. To attain the smallest integer $k = \lceil N^2/F^2 \rceil$, first suppose $k = 1$. Then $N = F$ and one singular value N suffices. For $k \ge 2$, choose

$$a_1 = \frac{N + \sqrt{(k-1)(kF^2 - N^2)}}{k}, \qquad a_2 = \cdots = a_k = \frac{N - \sqrt{(kF^2 - N^2)/(k-1)}}{k}. \tag{88}$$

The radicands are nonnegative because $N^2 \le kF^2$, and $a_2 > 0$ because $N^2 > F^2$. Direct substitution gives $\sum_j a_j = N$ and $\sum_j a_j^2 = F^2$. Padding with zeros gives a matrix of the required size and rank $k \le s$.

Finally, $N^2 \le (k + \varepsilon)F^2$ is equivalent to $k \ge d_{\mathrm{part}} - \varepsilon$. Since $1 \le d_{\mathrm{part}} \le s$ and $0 < \varepsilon < 1$, its smallest positive integer solution is Equation (86). Taking $\varepsilon = 10^{-4}$ proves Equation (87). □

Participation mass beyond the reference dimension. For $1 \le k < s$, put $a = t_{k+1}$ and $T_{\mathrm{tail}} = \sum_{j>k} t_j$. Each tail value is at most a, while the head sum is at least ka. Hence

$$\sum_{j>k} t_j^2 \le aT_{\mathrm{tail}} \le \frac{(N - T_{\mathrm{tail}})T_{\mathrm{tail}}}{k} \le \frac{N^2}{4k}. \tag{89}$$

Dividing by F^2 gives the conservative participation-mass bound

$$\frac{\sum_{j>k} \lambda_j}{\sum_j \lambda_j} \le \frac{d_{\mathrm{part}}}{4k}. \tag{90}$$

For $k = k_\varepsilon < s$, the right-hand side is at most $1/4 + \varepsilon/(4k_\varepsilon)$.

Spectrum shape and the zero-ridge window. With $p_j = \lambda_j / \sum_\ell \lambda_\ell$ and $u_j = 1/s$, $d_{\mathrm{part}}/s = (\sum_j \sqrt{p_j u_j})^2$ is the squared Bhattacharyya coefficient between p and the uniform spectrum. The symmetric concavity of $\sum_j \sqrt{p_j}$ makes d_{part} decrease under increasing concentration in the majorization order.

For $s \geq 2$, $d_{\mathrm{part}} - 10^{-4} > s - 1$ selects $k = s$ and $\tau = 0$. To describe this window, take a spectrum with total mass K, $s-1$ equal leading eigenvalues, and one tail value ϵ. Direct substitution gives

$$d_{\mathrm{part}} - (s-1) = \frac{2\sqrt{(s-1)(K-\epsilon)\epsilon} - (s-2)\epsilon}{K}. \tag{91}$$

As $\epsilon \to 0$, this difference tends to zero and the guarded rank drops below s; for $s = 1$, the rule always gives $k = 1$ and $\tau = 0$. Table 32 reports zero-ridge activity in the numerical construction.

E.4 Local stability of the adaptive selector

Consider a positive-ridge neighborhood with unchanged numerical support, a fixed selected index $1 \leq k < s$, $d_{\mathrm{part}} - 10^{-4}$ separated from the integers, and positive lower bounds on the boundary eigenvalues. The adaptive rule changes both A and τ, so the local estimate combines spectral perturbation with the fixed-parameter bound in Appendix D.3.

If $\|\widetilde{A} - A\|_2 \leq \epsilon$ under these conditions, Weyl's inequality gives $|\widetilde{\lambda}_j - \lambda_j| \leq \epsilon$. For $\tau = \sqrt{\lambda_k \lambda_{k+1}}$,

$$\frac{|\widetilde{\tau} - \tau|}{\tau} \leq \frac{\epsilon}{2}\left(\frac{1}{\lambda_k} + \frac{1}{\lambda_{k+1}}\right) + O(\epsilon^2). \tag{92}$$

The scalar derivative satisfies $\sup_{t \geq 0} |\partial_\tau g_\tau(t)| = 3\sqrt{3}/(8\tau^2)$, attained at $t = \tau/\sqrt{3}$. Integrating this bound between the two positive thresholds and combining it with the fixed-parameter estimate yields, with $\tau_{\min} = \min\{\tau, \widetilde{\tau}\} > 0$,

$$\|g_{\widetilde{\tau}}(\widetilde{A}) - g_\tau(A)\|_2 \leq \frac{\|\widetilde{A} - A\|_2}{\tau_{\min}^2} + \frac{3\sqrt{3}}{8}\frac{|\widetilde{\tau} - \tau|}{\tau_{\min}^2}. \tag{93}$$

Crossing an integer effective-rank boundary can change k discretely. Appendix H.4 measures this numerical sensitivity directly across fixed sketch seeds and sketch widths.

Checkpoint-to-fusion perturbations. The operator estimate also controls the complete construction when core ranks and selector branches remain fixed. The next result propagates perturbations from the expert updates, including changes in the anchor, and treats zero ridge separately.

Theorem E.1 (Local checkpoint-to-fusion stability)**.** *Let $\widetilde{\Delta}_i = \Delta_i + E_i$. Assume the same active updates remain nonzero, the retained ranks r_i are unchanged, and $P_i, \widetilde{P}_i$ are the respective exact top-r_i right-singular projectors. For $r_i < n$, assume*

$$\gamma_i = \sigma_{r_i}(\Delta_i)^2 - \sigma_{r_i+1}(\Delta_i)^2 > 0, \qquad q_i = (2\|\Delta_i\|_2 + \|E_i\|_2)\|E_i\|_2 < \gamma_i/2, \qquad \delta_i = \frac{q_i}{\gamma_i - q_i},$$

where singular values are padded with zeros. For $r_i = n$, put $\delta_i = 0$. Form both sets of cores, participation operators, requested-response operators, and sum anchors by the current definitions. Define

$$a = \sum_i \delta_i / r_i, \qquad b = \sum_i \frac{\|E_i\|_F + \|\Delta_i\|_F \delta_i}{r_i}, \qquad m_{\mathrm{err}} = \sum_i \|E_i\|_F,$$

$$E_0 = B - M_0 A, \quad e_0 = b + m_{\mathrm{err}}\|\widetilde{A}\|_2 + \|M_0\|_F a.$$

Then

$$\|\widetilde{P}_i - P_i\|_2 \leq \delta_i, \qquad \|\widetilde{G}_i - G_i\|_2 \leq \delta_i / r_i,$$

$$\|\widetilde{A} - A\|_2 \leq a, \quad \|\widetilde{B} - B\|_F \leq b, \quad \|\widetilde{M}_0 - M_0\|_F \leq m_{\mathrm{err}}, \quad \|\widetilde{E}_0 - E_0\|_F \leq e_0.$$

Assume $A, \widetilde{A}$ have the same positive rank $s \geq 1$ and all their positive eigenvalues are at least $\ell > 0$; their ranges may rotate. Write

$$u = \sum_{j=1}^{s} \sqrt{\lambda_j}, \quad \widetilde{u} = \sum_{j=1}^{s} \sqrt{\widetilde{\lambda}_j}, \quad v = \sum_{j=1}^{s} \lambda_j, \quad \widetilde{v} = \sum_{j=1}^{s} \widetilde{\lambda}_j, \quad v_{\min} = \min(v, \widetilde{v}), \quad d = u^2 / v.$$

A sufficient condition for the same selected k is

$$|\widetilde{d} - d| \le d_{\mathrm{err}} := \frac{sa}{v_{\min}} \left(\frac{u + \widetilde{u}}{2\sqrt{\ell}} + d \right) < \mathrm{dist}(d - 10^{-4}, \mathbb{Z}).$$

For the full unit-mass support, $v = \widetilde{v}$ is the active-task count and the d term in this bound can be omitted. If $k < s$, set

$$t_{\mathrm{err}} = \frac{(\lambda_k + \lambda_{k+1})a + a^2}{\tau + \widetilde{\tau}}, \qquad \tau_{\min} = \min(\tau, \widetilde{\tau}) > 0.$$

Then $|\widetilde{\tau} - \tau| \le t_{\mathrm{err}}$ and

$$\|\widetilde{M} - M\|_F \le m_{\mathrm{err}} + \frac{e_0}{2\tau_{\min}} + \frac{\|E_0\|_F}{\tau_{\min}^2} \left(a + \frac{3\sqrt{3}}{8} t_{\mathrm{err}} \right).$$

If $k = s$ on both sides, $\widetilde{\tau} = \tau = 0$ and instead

$$\|\widetilde{M} - M\|_F \le m_{\mathrm{err}} + \frac{e_0}{\ell} + \frac{3\|E_0\|_F}{\ell^2} a.$$

Proof. Expanding $\widetilde{\Delta}_i^\top \widetilde{\Delta}_i - \Delta_i^\top \Delta_i$ bounds its norm by q_i. The ordered-cluster Davis–Kahan argument gives the projector bound (Davis & Kahan, 1970). Explicitly, let $H = \Delta_i^\top \Delta_i$, $\widetilde{U}$ span its perturbed top cluster, and $U_\perp$ span the old complement. For $X = U_\perp^\top \widetilde{U}$,

$$X\widetilde{\Lambda} - H_\perp X = U_\perp^\top (\widetilde{H} - H)\widetilde{U}.$$

Weyl's inequality separates these two spectral blocks by at least $\gamma_i - q_i$. After shifting both blocks by $\lambda_{r_i+1}(H)$, the exponential-integral solution of this Sylvester equation has norm at most $q_i/(\gamma_i - q_i)$. The principal-angle identity for equal-dimensional subspaces gives $\|X\|_2 = \|\widetilde{P}_i - P_i\|_2$. Division by r_i proves the core bound. Next, $\widetilde{\Delta}_i \widetilde{P}_i - \Delta_i P_i = E_i \widetilde{P}_i + \Delta_i(\widetilde{P}_i - P_i)$ gives the B bound; summation gives the bounds for A and M_0. Expanding $\widetilde{E}_0 - E_0 = (\widetilde{B} - B) - (\widetilde{M}_0 - M_0)\widetilde{A} - M_0(\widetilde{A} - A)$ gives e_0.

Weyl's inequality yields $|\widetilde{\lambda}_j - \lambda_j| \le a$. Since positive eigenvalues are at least ℓ, $|\widetilde{u} - u| \le sa/(2\sqrt{\ell})$ and $|\widetilde{v} - v| \le sa$. Expanding

$$\widetilde{d} - d = \frac{\widetilde{u}^2 - u^2}{\widetilde{v}} + u^2 \left(\frac{1}{\widetilde{v}} - \frac{1}{v} \right)$$

proves d_{err} and the equal-trace refinement. The strict margin keeps the ceiling, and hence k, unchanged. For $k < s$, the difference of the two products defining τ^2 is bounded by $(\lambda_k + \lambda_{k+1})a + a^2$, proving the threshold bound.

For positive thresholds, combine the fixed-parameter resolvent estimate in Appendix D.3 with the scalar derivative estimate above. The exact decomposition is

$$\widetilde{M} - M = (\widetilde{M}_0 - M_0) + (\widetilde{E}_0 - E_0) g_{\widetilde{\tau}}(\widetilde{A}) + E_0 \big[g_{\widetilde{\tau}}(\widetilde{A}) - g_\tau(A) \big].$$

Using $\|g_{\widetilde{\tau}}(\widetilde{A})\|_2 \le 1/(2\tau_{\min})$ gives the positive-ridge inequality. For zero ridge, put $E = \widetilde{A} - A$ and use

$$\begin{aligned} \widetilde{A}^\dagger - A^\dagger = & -\widetilde{A}^\dagger E A^\dagger + (\widetilde{A}^\dagger)^2 E (I - AA^\dagger) \\ & + (I - \widetilde{A}^\dagger \widetilde{A}) E (A^\dagger)^2. \end{aligned}$$

Expansion verifies this pseudoinverse identity directly (Wedin, 1973). Each inverse has norm at most $1/\ell$, so the difference is bounded by $3a/\ell^2$. The same decomposition with $g_0(A) = A^\dagger$ proves the zero-ridge inequality. No gap between λ_k and λ_{k+1} is required for the soft filter. □

Numerically selected support. The theorem uses the full positive support. If the construction instead retains the top c modes with projector Π, define $A_c = \Pi A \Pi$ and $B_c = B\Pi$. Assume c and the support-threshold branch are unchanged, $\gamma_\Pi = \lambda_c(A) - \lambda_{c+1}(A) > 0$, and $a < \gamma_\Pi/2$. The same projector argument and product expansion give

$$\|\widetilde{\Pi} - \Pi\|_2 \le p_{\mathrm{err}} := \frac{a}{\gamma_\Pi - a}, \qquad \|\widetilde{A}_c - A_c\|_2 \le a + 2\|A\|_2 p_{\mathrm{err}}, \qquad \|\widetilde{B}_c - B_c\|_F \le b + \|B\|_F p_{\mathrm{err}}.$$

When the retained eigenvalues and selector also satisfy the theorem's lower-bound and margin conditions, substituting these operator bounds into the two recovery inequalities, with $s = c$ and the trace-varying d_{err}, gives the selected-support result. The full-space case uses $\Pi = I$ and $p_{\mathrm{err}} = 0$. For sketched cores, the downstream inequalities use bounds on the actual projector errors in place of δ_i. Fixed sketch draws, stable factorization and Ritz gaps, rank/retry margins, and arithmetic errors must then be treated according to Appendix H.3; an exact singular-value gap alone does not certify the numerical projector error.

F Secondary Statistical Interpretations

F.1 Gaussian MAP interpretation

For each participation mode, take independent c_j and ε_j with $\sigma, \alpha > 0$:

$$y_j = \lambda_j c_j + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I), \qquad c_j \sim \mathcal{N}(0, \alpha^2 I). \tag{94}$$

The posterior mean and MAP estimator are

$$\widehat{c}_j = \frac{\lambda_j}{\lambda_j^2 + \sigma^2/\alpha^2} y_j. \tag{95}$$

Setting $\tau^2 = \sigma^2/\alpha^2$ gives the ITCF filter. Even with homoscedastic noise, a smaller λ_j lowers the observation signal-to-noise ratio and induces stronger shrinkage. If the noise-to-prior ratio varies by mode, the estimator instead uses mode-specific τ_j values.

F.2 Moment-robust interpretation for signed coefficient perturbations

A second interpretation places uncertainty in the reduced fitting coefficients. Let F be a signed reduced coefficient perturbation with

$$\mathbb{E}[F] = 0, \qquad \mathbb{E}[FF^\top] \preceq \tau^2 I. \tag{96}$$

Then

$$\sup_{\mathbb{P}} \mathbb{E}\|C(\Lambda + F) - Y\|_F^2 = \|C\Lambda - Y\|_F^2 + \tau^2 \|C\|_F^2. \tag{97}$$

Expanding the loss, $\mathbb{E}[F] = 0$ removes the cross term. The upper bound follows from

$$\mathbb{E}\|CF\|_F^2 = \mathrm{tr}(C\mathbb{E}[FF^\top]C^\top) \le \tau^2 \|C\|_F^2, \tag{98}$$

and equality is attained by $F = \pm\tau I$ with equal probability. This gives the Tikhonov objective exactly for signed coefficient perturbations, with $\Lambda + F$ allowed to be indefinite.

G Conditional Prediction Bound and Its Remainders

Fix task i, an input x, and a common output space for the expert f_i and merged predictor f_M. Let g index the targeted linear weight groups, with $D_{g,i} = M_g - \Delta_{g,i}$, $P_{g,i} = Q_{g,i} Q_{g,i}^\top$, and

$$\mathcal{J}_{g,i}(M_g) = \mathrm{tr}(D_{g,i} G_{g,i} D_{g,i}^\top), \qquad \mathcal{J}_i(M) = \sum_g \mathcal{J}_{g,i}(M_g), \qquad \mathcal{J}(M) = \sum_i \mathcal{J}_i(M). \tag{99}$$

Here the network-level objective sums the layerwise comparisons.

Expand the network into its finite computation graph and replace targeted linear occurrences in topological order, starting from the expert. A grouped projection is one joint intervention; reused

matrices have several occurrences of the same group. Let o index these occurrences and $g(o)$ its weight group. Adjacent hybrids share the input activation $U_{o,i}(x; M)$ to the replaced occurrence. Let $L_{o,i}(x; M)$ be a valid downstream Lipschitz bound from its output perturbation in Frobenius norm to the chosen predictor norm. After all targeted replacements, denote the hybrid by $\bar{f}_{M,i}$; its untargeted parameters still equal the expert's.

For one occurrence with group g, suppressing arguments on U,

$$D_{g,i}U = D_{g,i}P_{g,i}U + D_{g,i}(I - P_{g,i})U, \tag{100}$$

$$\|D_{g,i}P_{g,i}U\|_F \le \|D_{g,i}Q_{g,i}\|_F \|Q_{g,i}^\top U\|_2 \le \sqrt{r_{g,i}\mathcal{J}_{g,i}(M_g)}\,\|Q_{g,i}^\top U\|_F. \tag{101}$$

To collect repeated occurrences before applying Cauchy–Schwarz, define

$$a_{g,i}(x; M) = \sqrt{r_{g,i}} \sum_{o:g(o)=g} L_{o,i}(x; M)\|Q_{g,i}^\top U_{o,i}(x; M)\|_F, \tag{102}$$

$$C_i^{\mathrm{hyb}}(x; M) = \left(\sum_g a_{g,i}(x; M)^2\right)^{1/2}, \tag{103}$$

$$R_i^{\perp,\mathrm{hyb}}(x; M) = \sum_o L_{o,i}(x; M)\|D_{g(o),i}(I - P_{g(o),i})U_{o,i}(x; M)\|_F, \tag{104}$$

$$R_i^{\mathrm{nt,hyb}}(x; M) = \|f_M(x) - \bar{f}_{M,i}(x)\|. \tag{105}$$

A zero core contributes $a_{g,i} = 0$ and $P_{g,i} = 0$. Telescoping the targeted interventions gives a controlled term $\sum_g a_{g,i}\sqrt{\mathcal{J}_{g,i}}$ plus $R_i^{\perp,\mathrm{hyb}}$. Cauchy–Schwarz over groups and the remaining non-target difference yield

$$\|f_M(x) - f_i(x)\| \le C_i^{\mathrm{hyb}}(x; M)\sqrt{\mathcal{J}_i(M)} + R_i^{\perp,\mathrm{hyb}}(x; M) + R_i^{\mathrm{nt,hyb}}(x; M). \tag{106}$$

For a comparison at fixed inputs, take common valid bounds $\bar{C}_i$ on the hybrid constants of both M_0 and M_τ. Then

$$\sum_i \bar{C}_i\sqrt{\mathcal{J}_i(M)} \le \left(\sum_i \bar{C}_i^2\right)^{1/2} \sqrt{\mathcal{J}(M)}. \tag{107}$$

Summing the layerwise objective comparisons gives $\mathcal{J}(M_\tau) \le \mathcal{J}(M_0)$, so this controlled upper envelope is no larger at M_τ.

For logits, use the ℓ_∞ predictor norm above. Let $\hat{y}_i = \arg\max_c z_{i,c}(x)$ and

$$\gamma_i(x) = z_{i,\hat{y}_i}(x) - \max_{c \ne \hat{y}_i} z_{i,c}(x). \tag{108}$$

If the complete right-hand side of Equation (106) is below $\gamma_i(x)/2$, the merged logits retain the expert's predicted class: each competing logit gap can shrink by at most twice that bound. Equation (106) isolates the task-core-controlled component while retaining off-core responses, untargeted-parameter differences, and hybrid-network constants explicitly; their uniform smallness in arbitrary deep networks is neither assumed nor established, so the bound remains conditional rather than an unconditional accuracy guarantee.

H Experimental Protocol, Numerical Construction, and Complete Results

H.1 Closest checkpoint-derived comparisons

Endpoint-derived operators. We next compare the complete ITCF construction with ACE-Merging and CSP, which also derive their merging operators from checkpoints. With identical checkpoints and fixed configurations, ITCF exceeds ACE-Merging by $9.192\,\mathrm{pp}$ across 12 settings (95% interval $[4.329, 13.524]$; 11 wins). It also scores higher than the released CSP $k = 1$ operator on the eight vision settings (Table 7). These comparisons favor ITCF on the tested shared task combinations.

Projection-based checkpoint-only comparison. APGD provides a complementary data-free projection baseline under the same information budget (Wei et al., 2025). On a broader 48-setting matched panel, ITCF exceeds the checkpoint-only public APGD adapter by $17.971\,\mathrm{pp}$ on average (median $+5.179\,\mathrm{pp}$; 41/0/7 wins/ties/losses; 95% interval $[+10.193, +26.617]$). Setting-level differences range from -34.64 to $+94.67\,\mathrm{pp}$, so the average advantage coexists with substantial variation across settings (Appendix Table 24).

Information-budget boundary. The CSP comparison uses the released $k = 1$ operator, labeled CSP-public; Appendix Table 25 gives its public $k = 1, \ldots, 10$ sweep. CAT and LOT use task samples or statistics computed from samples (Sun et al., 2025a;b). They therefore receive additional information beyond the checkpoints used in this ranking.

H.2 Information boundary and evaluation protocol

Construction access. Merge construction uses the pretrained checkpoint, independently fine-tuned checkpoints, and algebra derived from them. Task samples, labels, logits, activation caches, calibration sets, and validation scores are excluded from this stage. Downstream evaluation follows construction, with coefficients, task subsets, ranks, sketch width, and model-specific branches fixed independently of the evaluation scores.

Vision and language coverage. The vision tasks are Cars (Krause et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), GTSRB (Stallkamp et al., 2012), MNIST (LeCun et al., 1998), RESISC45 (Cheng et al., 2017), SUN397 (Xiao et al., 2010), and SVHN (Netzer et al., 2011). The full vision lattice contains 415 scheduled compositions over ViT-B/32 and ViT-B/16 (Radford et al., 2021). The 39 settings shared by ITCF and WUDI form the WUDI-matched vision panel. Vision32 is a separate balanced 32-setting panel for common-baseline summaries and controlled interventions.

The encoder-language tasks are CoLA, SST-2, MRPC, STS-B, QQP, MNLI, QNLI, and RTE. The text-to-text tasks are QASC, WikiQA, QuaRTz, PAWS, WinoGrande, and WSC (Khot et al., 2020; Yang et al., 2015; Tafjord et al., 2019; Zhang et al., 2019; Sakaguchi et al., 2020). Together they produce 97 compositions over RoBERTa-base, RoBERTa-large, T5-base, and T5-large (Liu et al., 2019; Raffel et al., 2020; Wang et al., 2019b;a). The 93 settings shared with WUDI form Language93. Each composition receives the mean of its task metrics in percentage units.

Qwen and Llama NLI settings. Qwen2.5-1.5B and Llama-3.2-1B (Qwen Team, 2024; Grattafiori et al., 2024; Meta AI, 2024) share 24 frozen settings over six full-parameter NLI experts: SNLI, MNLI, SICK, QNLI, RTE, and SciTail (Bowman et al., 2015; Marelli et al., 2014; Khot et al., 2018), with $K = 2, \ldots, 6$ compositions fixed by combinatorial coverage rather than downstream accuracy. Both use seeds $\{0, 1, 2\}$, fixed $q = 48$, and macro constituent-task accuracy as the setting score.

NLLB translation settings. NLLB-200 distilled-600M and distilled-1.3B (NLLB Team, 2024) use six full-parameter OPUS-100 experts (Zhang et al., 2020) for English-to-German, French, Spanish, Russian, Chinese, and Japanese translation. Both scales use seeds $\{0, 1, 2\}$ and the same eight frozen compositions per seed: two each at $K = 2, 3, 4$, one at $K = 5$, and one at $K = 6$. The setting score is the macro chrF++ score (Popović, 2017) across its constituent translation directions. Merge construction uses the same attention/feed-forward matrix scope for every method; embedding, norm, bias, and other untargeted tensors stay at the pretrained value.

Table 7: **Nearest endpoint-derived methods.** Scores use identical checkpoints and setting schedules. CSP-public denotes the released $k = 1$ operator; the public $k = 1, \ldots, 10$ sweep appears in the appendix.

Panel	n	ITCF	ACE	CSP-public
All	12	67.24	58.05	–
Vision	8	68.29	59.22	58.84
RoBERTa	4	65.14	55.70	–

Expert preparation. The T5-base scaling and Qwen/Llama NLI experts use full-parameter fine-tuning for 1,000 steps with AdamW (learning rate 10^{-5}, zero weight decay), micro-batch 16, gradient accumulation 1, gradient clipping 1.0, BF16, and 6% constant-with-warmup scheduling. The NLLB experts also use 1,000 AdamW steps at learning rate 10^{-5}, zero weight decay, effective batch 16, BF16 autocast with FP32 parameters/checkpoints, gradient clipping 1.0, and maximum source/target length 128; the 1.3B run uses gradient checkpointing. We use the final checkpoint without validation-based selection.

Paired inference. Baseline comparisons join rows by the exact setting identifier. Confidence intervals resample matched settings with replacement. Win counts use the sign of the setting-level difference; ties are exact at the stored precision.

H.3 Reduced construction and complexity

The ideal task core is defined spectrally; this subsection records how the main implementation approximates it and how the reduced solve avoids materializing the full participation matrix.

Randomized core extraction. The following construction describes the main CLIP/RoBERTa/T5/NLLB implementation. The Qwen2.5-1.5B and Llama-3.2-1B NLI panels use a separately fixed sketch width of $q = 48$; NLLB uses $q = 12$. For each task and matrix group, the randomized range finder in Appendix B.1 approximates the ideal right-singular core with initial width $q = \min(12, m, n)$ and one power refinement. Its small Rayleigh–Ritz eigensystem orders the candidate directions by decreasing nonnegative eigenvalue.

Retained task rank and width guard. Let $e_i = \max(\|\Delta_i\|_F^2, \epsilon_{\text{tiny}})$ and $\widehat{\lambda}_{i,1} = \max(\lambda_{i,1}^{\text{Ritz}}, \epsilon_{\text{tiny}})$, with ϵ_{tiny} the smallest positive normal FP32 value. The implementation computes $a_i = \min\{\min(m,n), \max(1, e_i/\widehat{\lambda}_{i,1})\}$, then sets $r_i = \min\{q, \max(1, \lceil a_i - 10^{-4} \rceil)\}$. The first r_i Ritz directions form the numerical core factor. If any task core in an equal-shape bucket reaches $r_i \geq \max(1, q-1)$ and $q < \min(m,n)$, one retry uses $q \leftarrow \min\{\min(m,n), \max(q+2, 2q)\}$ and recomputes the factors and ranks. Zero updates retain a zero column for tensor shape, contributing zero support as described in Appendix B. Large sketch products use BF16; the small eigensolves and rank decisions use FP32.

Replacing the Ritz denominator by $\|\Delta_i\|_2^2$ gives the ideal stable-rank reference in Proposition 3; the estimated, finite-width rule above is its numerical realization.

Targeted groups and output template. CLIP ViT targets each residual block's attention input/output projections and feed-forward input/output matrices. RoBERTa concatenates attention query, key, and value matrices by rows; attention output and the two feed-forward matrices form separate groups. T5 uses the same query/key/value grouping for encoder and decoder self-attention and decoder cross-attention, with output projections and each feed-forward input/output matrix grouped separately. NLLB uses the analogous self-attention, cross-attention, and feed-forward matrices. The corresponding key suffixes are:

Table 8: Matrix-key suffixes for the targeted model families.

Family	Matrix-key suffixes within the corresponding block
CLIP ViT	`attn.in_proj_weight`, `attn.out_proj.weight`, `mlp.c_fc.weight`, `mlp.c_proj.weight`
RoBERTa	`attention.self.{query,key,value}.weight`; `attention.output.dense.weight`, `intermediate.dense.weight`, `output.dense.weight`
T5	`SelfAttention.{q,k,v,o}.weight`, decoder `EncDecAttention.{q,k,v,o}.weight`; `DenseReluDense.wi*.weight`, `DenseReluDense.wo.weight`
NLLB	encoder/decoder `self_attn.{q,k,v}_proj.weight`, `self_attn.out_proj.weight`; decoder `encoder_attn.{q,k,v}_proj.weight`, `encoder_attn.out_proj.weight`; `fc1.weight`, `fc2.weight`

Grouped matrices share an input dimension and are split by their original row counts after fusion. The sparse patch is added to the base template; all untargeted parameters retain base values. Task-specific evaluation heads follow the evaluation protocol. The grouping utility also has a two-dimensional-tensor fallback for unrecognized layouts; the named families above use their explicit group definitions.

Reduced solve and numerical support. Fusion pools the normalized factors into one reduced problem. Let $r_\Sigma = \sum_i r_i$ and concatenate the normalized factors as $Z = [Q_1/\sqrt{r_1}, \ldots, Q_K/\sqrt{r_K}]$. Because $A = ZZ^\top$, its nonzero eigensystem can be recovered from the factor-space matrix $Z^\top Z \in \mathbb{R}^{r_\Sigma \times r_\Sigma}$ instead of materializing $A \in \mathbb{R}^{n\times n}$; this matrix is smaller when $r_\Sigma < n$. For batched computation, let R_{pad} be the largest r_Σ in the bucket. If all task updates in a targeted group are exactly zero, the group returns a zero fused update and skips the participation-dimension, k, and τ computations. Numerical support keeps eigenvalues above $\lambda_1 \max(n, R_{\text{pad}})\epsilon_{\text{FP32}}$. An empty support mask retains its first slot for tensor-shape consistency. This numerical support defines s; the reference dimension k is then computed by Equation (15) on the retained eigenvalues. For this selected spectrum, the factor-norm identity in Equation (14) uses the projection of Z onto the selected eigenspace. The solve applies the continuous gain $\lambda/(\lambda^2 + \tau^2)$ on supported modes. Reciprocal-square-root and gain denominators use floors 10^{-20} and 10^{-30}, respectively. Construction is dominated by checkpoint-sketch products and small factor algebra. The deployed model retains the pretrained architecture, parameter count, and inference path.

H.4 Numerical robustness of the reduced construction

The main implementation uses randomized range finding with a fixed default sketch width. The following tests vary only those numerical choices while leaving the checkpoints and fusion rule unchanged.

Sketch seed. Table 9 reports five fixed sketch seeds on 10 settings spanning ViT-B/32, RoBERTa-base, and T5-base. Scores remain close across the randomized numerical realizations.

Table 9: Sketch-seed stability on 10 cross-family settings. Rank and k agreement are measured against the reference seed 20260821 over the corresponding matrix groups.

Seed	Mean score	Rank match	k match
20260821	70.5926	100.00%	100.00%
20260838	70.5627	99.50%	88.75%
20260855	70.5025	99.61%	89.17%
20260872	70.4369	99.55%	88.96%
20260889	70.4195	99.49%	89.42%

Across these seeds, the aggregate mean spans 0.1731 percentage points. The mean within-setting score standard deviation is 0.1310 points, and the largest within-setting range is 1.1840 points. Rank agreement remains high while k varies more often, and the aggregate downstream score stays stable on this panel.

Sketch width. Table 10 varies q without choosing a value from downstream scores. The default $q = 12$ was fixed before this study.

Table 10: Sketch-width sensitivity on the same 10 settings. Saturation is the fraction of initial sketches reaching the width guard; expansion is the fraction widened beyond the initial q.

q	Mean score	Mean core s	Initial saturation	Expanded
8	70.3956	0.2442	28.52%	23.02%
12	70.5926	0.2033	13.94%	11.73%
16	70.5273	0.1937	8.17%	6.82%
24	70.4989	0.3019	3.36%	3.20%

The full score span is 0.1970 points. Larger q reduces saturation and retry frequency, and the timing minimum on this panel occurs at $q = 16$. The default $q = 12$ lies on the same shallow performance plateau and serves as the fixed numerical budget used by the main experiments.

H.5 Coverage-specific aggregates and matched comparisons

The main text distinguishes fixed-coverage summaries from paired claims. We first report the standard panels, then give the additional TSV-Merge matched expansion that tests the same ordering on a separate set of settings.

Aggregate coverage and standard matched panels.

Table 11: Aggregate results at each method's fixed coverage. ITCF spans the full 415-setting vision lattice and 97-setting language suite; pairwise intersections are reported in Table 12. TIES uses a global unit coefficient on all 97 language settings.

Method	Vision n	Vision mean	Language n	Language mean
Weight averaging	39	57.48	93	54.67
Task arithmetic	39	58.84	93	51.71
TIES, fixed unit scale	39	61.35	97	57.07
DARE-TIES	39	58.13	93	51.52
PCB-Merging	39	61.07	93	53.36
WUDI300	39	65.23	93	63.02
SWUDI-A	39	65.18	93	65.27
TSV-Merge	32	67.27	93	65.23
ITCF	415	68.87	97	66.53

Table 12: Paired ITCF differences on matched baseline intersections. Values are mean $\Delta\pm$ seed standard deviation across seeds within the listed panel. Confidence intervals are paired setting-bootstrap intervals.

Baseline	Panel	Mean $\Delta\pm$ seed SD	95% setting-bootstrap CI
WUDI300	WUDI-matched vision	$+0.60 \pm 0.15$	$[+0.38, +0.82]$
SWUDI-A	WUDI-matched vision	$+0.66 \pm 0.64$	$[+0.43, +0.88]$
TSV-Merge	Vision32	$+0.17 \pm 0.38$	$[-0.20, +0.55]$
ACTMat	Vision32	$+0.74 \pm 0.26$	$[+0.40, +1.02]$
WUDI300	Language93	$+3.10 \pm 0.69$	$[+1.90, +4.10]$
SWUDI-A	Language93	$+0.85 \pm 0.52$	$[+0.25, +1.45]$
TSV-Merge	Language93	$+0.89 \pm 0.48$	$[+0.10, +1.55]$
ACTMat	Language93	$+3.62 \pm 0.61$	$[+1.12, +4.35]$

Additional TSV-Merge matched expansion. The standard Vision32 and Language93 comparisons remain as reported above. Table 13 adds a separate 66-setting matched panel; its family rows are descriptive decompositions of the same combined result.

Table 13: Additional matched ITCF–TSV-Merge evidence. Confidence intervals are paired bootstrap intervals for the combined and source panels; family rows show descriptive decompositions.

Group	n	Mean Δ	Median Δ	W/T/L	95% interval
Combined additional panel	66	$+0.769$	$+1.046$	47/0/19	$[+0.057, +1.455]$
Cross-family panel	48	$+0.694$	$+0.926$	32/0/16	$[-0.311, +1.632]$
Independent K-scaling panel	18	$+0.966$	$+1.196$	15/0/3	$[+0.379, +1.515]$
Vision	42	$+0.917$	$+0.926$	32/0/10	–
T5	12	$+1.466$	$+1.296$	10/0/2	–
RoBERTa	12	-0.449	-2.489	5/0/7	–

H.6 Task-count and model-family breakdowns

The following tables localize the aggregate results by task count and model family. The held-out RoBERTa/T5 contrast is retained as a boundary case; the NLLB tables report the three-seed multilingual translation transfer used in Table 2.

Table 14: Every task count in the full 415-setting vision lattice. WUDI coverage is the number of settings in the pairwise intersection.

K	ITCF settings	ITCF mean	WUDI matches	ITCF−WUDI
2	56	68.22	6	-0.798
3	84	69.18	6	$+0.991$
4	98	68.86	6	$+0.844$
5	84	68.90	5	$+1.147$
6	56	68.85	11	$+0.705$
7	36	69.21	4	$+0.738$
8	1	67.57	1	$+0.737$

Table 15: Held-out model-family contrast.

Panel	n	ITCF	WUDI300	ITCF−WUDI
RoBERTa-base combinations	6	53.85	56.45	-2.594
T5-base combinations	6	71.54	61.57	$+9.969$

Table 16: NLLB-200 distilled-600M on eight frozen multilingual translation compositions. Scores are macro chrF++ across the constituent translation directions and are reported as mean ± seed SD over seeds $\{0, 1, 2\}$. DE/FR, ZH/JA, and analogous labels denote the target-language subsets.

Method	K2 DE/FR	K2 ZH/JA	K3 DE/FR/ES	K3 RU/ZH/JA
Weight averaging	51.16±0.02	17.54±0.02	53.76±0.02	27.76±0.03
Task Arithmetic	51.00±0.03	17.43±0.02	53.72±0.01	27.75±0.05
DARE-TIES	50.98±0.03	17.41±0.07	53.69±0.03	27.71±0.02
PCB-Merging	51.14±0.03	17.38±0.05	53.67±0.02	27.60±0.05
WUDI300	51.33±0.04	17.78±0.03	53.90±0.00	28.06±0.02
SWUDI-A	51.31±0.01	17.76±0.03	53.87±0.05	28.01±0.02
TSV-Merge	51.20±0.07	17.67±0.03	53.81±0.02	27.93±0.04
ACTMat	51.36±0.05	17.74±0.07	53.89±0.06	27.98±0.03
ITCF	**51.98±0.01**	**17.90±0.06**	**54.06±0.20**	**28.47±0.36**

Method	K4 DE/FR/ES/RU	K4 ES/RU/ZH/JA	K5 DE/FR/ES/RU/ZH	K6 All	Mean
Weight averaging	52.41±0.02	35.56±0.02	46.31±0.01	40.67±0.01	40.65±0.00
Task Arithmetic	52.44±0.05	35.60±0.02	46.49±0.01	40.78±0.04	40.65±0.02
DARE-TIES	52.43±0.01	35.59±0.03	46.43±0.01	40.78±0.03	40.63±0.02
PCB-Merging	52.45±0.01	35.53±0.06	46.49±0.01	40.71±0.03	40.62±0.02
WUDI300	52.63±0.04	35.82±0.02	46.62±0.02	40.83±0.03	40.87±0.01
SWUDI-A	52.58±0.06	35.77±0.02	46.56±0.03	40.89±0.01	40.84±0.02
TSV-Merge	52.52±0.03	35.69±0.02	46.57±0.01	40.89±0.03	40.78±0.01
ACTMat	52.61±0.03	35.75±0.05	46.53±0.03	40.90±0.02	40.84±0.02
ITCF	**52.87±0.08**	**36.12±0.30**	**46.98±0.31**	**41.40±0.42**	**41.22±0.14**

Table 17: NLLB-200 distilled-1.3B on the same eight frozen multilingual translation compositions. Scores are macro chrF++ across the constituent translation directions and are reported as mean ± seed SD over seeds $\{0, 1, 2\}$.

Method	K2 DE/FR	K2 ZH/JA	K3 DE/FR/ES	K3 RU/ZH/JA
Weight averaging	52.83±0.04	18.84±0.05	55.22±0.03	29.24±0.04
Task Arithmetic	52.37±0.08	18.63±0.07	55.09±0.01	29.23±0.03
DARE-TIES	52.26±0.10	18.56±0.04	55.08±0.01	29.20±0.03
PCB-Merging	52.75±0.07	18.78±0.11	55.28±0.02	29.43±0.04
WUDI300	52.91±0.03	19.56±0.11	55.33±0.03	29.89±0.05
SWUDI-A	52.89±0.05	19.50±0.08	55.36±0.08	29.79±0.11
TSV-Merge	52.90±0.01	19.15±0.05	55.31±0.04	29.62±0.03
ACTMat	52.89±0.02	19.42±0.19	55.36±0.05	29.80±0.09
ITCF	**53.09±0.15**	**19.81±0.17**	**55.88±0.07**	**29.94±0.02**

Method	K4 DE/FR/ES/RU	K4 ES/RU/ZH/JA	K5 DE/FR/ES/RU/ZH	K6 All	Mean
Weight averaging	54.01±0.03	36.90±0.02	47.90±0.04	42.03±0.01	42.12±0.01
Task Arithmetic	54.10±0.03	36.99±0.05	48.23±0.02	42.39±0.03	42.13±0.00
DARE-TIES	54.01±0.02	36.94±0.02	48.11±0.04	42.30±0.03	42.06±0.01
PCB-Merging	54.15±0.02	37.14±0.07	48.30±0.01	42.41±0.04	42.28±0.01
WUDI300	54.19±0.02	37.43±0.08	48.45±0.04	42.50±0.05	42.53±0.02
SWUDI-A	54.21±0.03	37.38±0.06	48.43±0.02	42.55±0.03	42.51±0.02
TSV-Merge	54.21±0.04	37.27±0.06	48.33±0.02	42.50±0.03	42.41±0.02
ACTMat	54.22±0.03	37.36±0.06	48.39±0.03	42.56±0.03	42.50±0.03
ITCF	**54.83±0.21**	**37.84±0.10**	**48.88±0.08**	**43.14±0.47**	**42.93±0.06**

Table 18: Independent three-seed T5-base replication. Values are mean ± standard deviation across seeds.

Method	$K = 2$	$K = 4$	$K = 6$
ITCF	66.08 ± 1.47	70.19 ± 0.08	64.11 ± 1.13
WUDI300	65.93 ± 1.61	64.20 ± 3.01	44.11 ± 1.59
TSV-Merge	64.62 ± 1.01	67.71 ± 0.43	64.77 ± 0.26

H.7 LARGE-SCALE FULL-MODEL MERGING

The standard panels merge at most a small number of experts. This subsection tests whether the ordering persists when independently fine-tuned full-model experts scale from $K = 8$ to $K = 166$.

Task setting and expert parameterization. The scaling study follows the 166-expert FLAN task collection and canonical ordering of Muqeeth et al. (2024): 36 P3, 36 FLANv2, 75 NIV2, and 19 ADD1 tasks, using frozen prefixes at the reported K. PHATGOOSE uses routed LoRA experts; we inherit only the tasks and ordering, using independently full-parameter-fine-tuned T5-base experts without its routing.

Frozen large-K baseline configurations. Task Arithmetic uses $\alpha = 1$; TIES20 a global top-20% mask with unit scaling; DARE90 drop rate 0.9 with $1/(1 - 0.9)$ rescaling, mask seed 0, and $\alpha = 1$; DARE-TIES90 then applies TIES20. TSV-Merge uses `sv_reduction=1/K`, leading per-task singular slots, and mean non-matrix delta. WUDI300 uses Adam for 300 steps at 2×10^{-5}; WUDI300† in Table 21 is matrix-scope.

Table 19: **Full-model scaling comparisons.** Macro source-task scores (%) are reported as mean ± seed standard deviation. Each row uses the same K independently fine-tuned T5-base full-model experts.

K	SWUDI-A	TSV-Merge	ACTMat	ITCF
8	51.62 ± 0.338	47.99 ± 0.431	52.00 ± 0.322	$\mathbf{52.54 \pm 0.251}$
12	53.18 ± 0.312	51.55 ± 0.468	54.20 ± 0.358	$\mathbf{54.97 \pm 0.247}$
16	52.06 ± 0.379	51.60 ± 0.348	52.90 ± 0.329	$\mathbf{53.72 \pm 0.253}$
24	46.37 ± 0.318	44.48 ± 0.409	47.80 ± 0.391	$\mathbf{48.49 \pm 0.249}$
36	42.36 ± 0.362	40.84 ± 0.452	43.50 ± 0.313	$\mathbf{44.28 \pm 0.254}$
48	37.44 ± 0.301	35.57 ± 0.367	38.80 ± 0.354	$\mathbf{39.44 \pm 0.218}$
60	34.15 ± 0.247	32.70 ± 0.343	35.10 ± 0.253	$\mathbf{35.64 \pm 0.223}$
72	32.10 ± 0.252	30.93 ± 0.307	33.00 ± 0.249	$\mathbf{33.65 \pm 0.216}$
166	27.30 ± 0.181	27.16 ± 0.241	27.42 ± 0.177	$\mathbf{28.02 \pm 0.183}$

Across the complete scaling panel, ITCF has the highest mean among the four reported methods at every tested expert count, with ACTMat second; the ITCF-minus-ACTMat margin ranges from $+0.54$ pp to $+0.82$ pp. At $K = 166$, ITCF scores 28.02, compared with 27.42 for ACTMat, 27.30 for SWUDI-A, and 27.16 for TSV-Merge. It leads these baselines by 0.60 pp, 0.72 pp, and 0.86 pp, respectively. The family decomposition shows ITCF ahead of both ACTMat and TSV-Merge in all four source families (Table 20). In paired comparisons over all 166 constituent tasks, ITCF records 101 wins, 3 ties, and 62 losses against SWUDI-A, and 96 wins, 4 ties, and 66 losses against TSV-Merge (Table 21).

Table 20: **Task-family decomposition at $K = 166$.** Family macro scores (%) use n constituent source tasks and are reported as mean ± seed standard deviation.

Source family	n	TSV-Merge	ACTMat	ITCF
P3	36	36.82 ± 0.251	36.80 ± 0.219	$\mathbf{37.32 \pm 0.197}$
FLANv2	36	20.59 ± 0.299	21.80 ± 0.281	$\mathbf{22.59 \pm 0.247}$
NIV2	75	26.42 ± 0.202	26.36 ± 0.178	$\mathbf{26.92 \pm 0.183}$
ADD1	19	24.26 ± 0.246	24.48 ± 0.223	$\mathbf{25.06 \pm 0.201}$
All 166	**166**	$\mathbf{27.16 \pm 0.241}$	$\mathbf{27.42 \pm 0.177}$	$\mathbf{28.02 \pm 0.183}$

Table 21: **Paired task-level comparisons at $K = 166$.** W/T/L counts are from ITCF's perspective over the 166 constituent source tasks. Macro scores are percentages; differences are percentage points. WUDI300† uses the matrix scope of the same full-model scaling experiment.

Baseline	Baseline macro	ITCF−baseline	W/T/L
Mean	14.39	+13.63	137/2/27
Task arithmetic	5.88	+22.14	161/1/4
TIES20	21.61	+6.41	126/4/36
DARE90	5.97	+22.05	161/1/4
DARE-TIES90	6.49	+21.53	160/1/5
SWUDI-A	27.30	+0.72	101/3/62
TSV-Merge	27.16	+0.86	96/4/66
WUDI300†	5.88	+22.14	160/1/5

H.8 Fixed TIES anchors on the full language panel

Because a fixed TIES coefficient is used in the main language comparison, this subsection shows the surrounding fixed anchors rather than selecting a coefficient from ITCF outcomes.

TIES uses one global top-20% mask across all 97 language settings. Table 22 reports several fixed coefficient anchors; the main comparison uses coefficient 1.0.

Table 22: TIES with a global top-20% mask and fixed coefficient on 97 language settings.

α	0.3	0.8	1.0	1.5	2.0	2.5
Mean	16.70	56.49	57.07	51.34	43.46	37.14

H.9 Nearest endpoint-derived methods

This comparison asks whether the checkpoint-derived geometry used by ITCF is competitive with alternative endpoint-only constructions, first on the 12-setting ACE/CSP panel and then on a broader APGD panel.

ACE and CSP.

Table 23: Nearest endpoint-derived methods on 12 shared settings. CSP is available on the eight vision settings.

Panel	n	ITCF	ACE	CSP-$k1$	ITCF−ACE
All	12	67.24	58.05	–	+9.192
Vision	8	68.29	59.22	58.84	+9.071
RoBERTa	4	65.14	55.70	–	+9.436

APGD. APGD is evaluated through the checkpoint-only public DOGE/APGD adapter with the public hyperparameters fixed before downstream evaluation. Table 24 reports the matched difference on 48 settings.

Table 24: Matched ITCF–APGD comparison on 48 checkpoint-only settings. The final column shows the range of setting-level differences.

n	Mean Δ	Median Δ	W/T/L	95% interval	Range
48	$+17.971$	$+5.179$	41/0/7	$[+10.193, +26.617]$	$[-34.64, +94.67]$

The mean and median favor ITCF strongly, while the wide setting-level range shows substantial heterogeneity across matched settings.

CSP sweep and access boundary. Table 25 gives the public CSP sweep on the eight vision settings. The main text uses its released $k = 1$ operator under the label CSP-public.

Table 25: Public CSP operator grid on the fixed eight-setting vision panel.

k	1	2	3	4	5	6	7	8	9
Mean	58.84	58.76	58.67	58.59	58.51	58.45	58.37	58.29	58.23

k	10	ITCF
Mean	58.17	68.29

H.10 Panel dependence of the regularization results

Panel dependence and positive floors. A separate 12-setting subset favors ITCF over no ridge by $1.323\,\mathrm{pp}$, compared with $0.85\,\mathrm{pp}$ on the broad 48-setting panel. On that smaller subset, the deterministic fixed-ridge rule scores $0.439\,\mathrm{pp}$ higher than ITCF. Small positive floors applied only to the $k = s$ branch are effectively tied with ITCF on the broad panel: $c = 0.01, 0.05, 0.10$ change the mean by 0.00, $+0.03$, and $+0.01\,\mathrm{pp}$ relative to ITCF, respectively, with all paired intervals spanning zero (Appendix Table 26). The downstream ordering thus depends on the panel, separately from the objective, interpolation, and fixed-positive-τ stability properties in Appendix D.3.

The broad ridge stress asks whether the participation-selected regularization is better than no ridge or fixed scales across a mixed panel; the smaller 12-setting subset shows how that conclusion changes with panel composition.

Table 26: Broad 48-setting ridge stress. Each fixed scale multiplies the same deterministic reference scale within a setting. Positive-floor variants modify only the $k = s$ branch, replacing $\tau = 0$ with $\tau = c\bar{\lambda}$, where $\bar{\lambda}$ is the mean positive participation eigenvalue.

Rule	Mean	ITCF$-$rule	95% interval
No ridge	65.91	$+0.85$	$[+0.31, +1.38]$
Fixed $0.25\times$	66.18	$+0.58$	$[+0.20, +0.96]$
Fixed $0.5\times$	66.34	$+0.42$	$[+0.12, +0.72]$
Fixed $1\times$	66.29	$+0.47$	$[+0.19, +0.76]$
Fixed $2\times$	65.88	$+0.88$	$[+0.54, +1.21]$
Fixed $4\times$	65.13	$+1.63$	$[+1.17, +2.06]$
Positive floor, $0.01\bar{\lambda}$	66.76	$+0.00$	$[-0.08, +0.08]$
Positive floor, $0.05\bar{\lambda}$	66.79	-0.03	$[-0.11, +0.05]$
Positive floor, $0.10\bar{\lambda}$	66.77	-0.01	$[-0.10, +0.08]$
ITCF	66.7630	–	–

The no-ridge contrast varies with the evaluation panel. On the 12-setting subset, ITCF leads no ridge by $1.323\,\mathrm{pp}$ with interval $[+0.663, +1.951]$; the deterministic fixed-ridge rule is $0.439\,\mathrm{pp}$ higher than ITCF. On the broader 48-setting panel, the no-ridge difference is $0.85\,\mathrm{pp}$. The strongest no-ridge separation is concentrated in the smaller subset. Replacing only the zero-ridge $k = s$ branch with small positive floors changes the broad-panel mean by at most $0.03\,\mathrm{pp}$ relative to ITCF, and all three paired intervals include zero.

Comparison with downstream ridge selection. We also compare the participation selector with ridge scales selected from downstream scores. Table 27 reports the comparison by model family, and Table 28 compares the selectors on the same 86-setting panel.

Table 27: Participation-based selector against downstream oracle ridge selection. Oracle selects the best ridge scale separately for each composition using downstream evaluation and is used only for analysis. Gap is Oracle minus ITCF.

Model family	Settings	ITCF	Oracle	Gap
Vision	32	67.44	67.61	0.17
RoBERTa	12	61.38	61.74	0.36
T5	18	69.82	70.05	0.23
Qwen	24	81.82	82.07	0.25
All	86	71.11	71.34	0.23

Table 28: Ridge selectors on 86 settings. The best global fixed scale uses one scale across all compositions. The per-setting oracle selects a scale separately for each composition. Both are selected using downstream evaluation for analysis. Oracle gap is the oracle mean minus the selector mean.

Selector	Mean	Oracle gap
No ridge	70.14	1.20
Best global fixed scale ($0.5\times$)	70.58	0.76
ITCF participation selector	71.11	0.23
Oracle per-setting	71.34	0.00

H.11 Anchor choice under the regularized construction

The regularized solution falls back toward M_0 when participation is weak, so the reference itself can affect downstream behavior. We therefore hold the task cores, ranks, participation system, and selector fixed and replace only the anchor.

Table 29: Anchor ablation on 10 settings under the regularized ITCF construction. Mean scores are reported as mean $\pm$ seed standard deviation.

Anchor	Mean score
Sum update	$\mathbf{70.55 \pm 0.20}$
Mean update	69.20 ± 0.22
Zero	67.80 ± 0.25

H.12 Within-core weighting and family variation

The task core fixes which directions are retained; these experiments test how their internal magnitudes affect the comparison weights across model families.

Table 30: Within-core construction comparisons on common retained bases and ranks. Differences are isotropic minus each variant, in percentage points, reported as mean ± seed standard deviation.

Within-core weighting variant	Isotropic − variant (pp)
Native singular magnitude	$+0.30 \pm 0.12$
Shuffled magnitude	$+0.57 \pm 0.15$
Reversed magnitude	$+0.65 \pm 0.16$

The seed-repeated comparison preserves the same ordering across the three tested weighting variants: the isotropic-minus-variant mean difference is smallest for native singular magnitude and larger for shuffled and reversed magnitude. Isotropic weighting gives the retained directions equal weight, whereas native weighting uses their singular magnitudes to assign direction-dependent weights.

Table 31: Qwen2.5-1.5B fixed-$q = 48$ probe. The merge construction reuses the extracted bases and retained ranks. Each normalized weighting recomputes the fusion system and ridge using the same rules.

Variant	Mean	Worst task
Isotropic	81.82	67.14
Singular-value weights	81.66	68.02
Squared singular-value weights	81.91	68.73
WUDI300 reference	81.4571	64.1772

H.13 Participation, core residual, and selector activity

Irreducible joint-fit residual. Theorem 2 separates the strength of the comparisons from the mismatch remaining at their best joint fit. We measure the latter relative to the anchor's fitting error as $\rho_{\mathrm{core}} = \mathcal{J}^\star/(\mathcal{J}(M_0) + 10^{-12})$ (Appendix H.13). Across 4,176 matrix groups, its setting-average value falls from 4.62×10^{-4} at $K = 2$ to 3.0×10^{-6} at $K = 8$. This trend differs from selector activity: ρ_{core} measures relative disagreement between requested changes, whereas the zero-ridge rate records how often regularization is applied.

To connect performance ablations with mechanism behavior, we report checkpoint-only geometry diagnostics for directional agreement and selector activity alongside the downstream scores.

The normalized irreducible residual is

$$\rho_{\mathrm{core}} = \frac{\mathcal{J}^\star}{\mathcal{J}(M_0) + 10^{-12}}. \tag{109}$$

This diagnostic uses checkpoint quantities only. Table 32 averages it by setting over 4,176 matrix groups.

Table 32: Checkpoint-only geometry by task count. Conditioning and residual are reported separately. Panel (b) restricts to groups selecting $\tau = 0$ and reports their condition-number distribution and relative correction magnitude $r_{\mathrm{corr}} = \|M_\tau - M_0\|_F / (\|M_0\|_F + 10^{-12})$.

(a) All matrix groups

K	Settings	ρ_{core}	$\tau = 0$	Median cond.
2	10	4.62×10^{-4}	67.25%	7.73
3	10	1.82×10^{-4}	21.15%	15.14
4	8	1.04×10^{-4}	3.91%	24.93
5	8	6.02×10^{-5}	1.04%	35.15
6	6	2.18×10^{-5}	0.00%	44.73
7	2	1.84×10^{-5}	0.00%	77.40
8	4	3.00×10^{-6}	0.00%	54.39

(b) Conditional diagnostics for $\tau = 0$ groups

K	Cond. med.	Cond. p95	Cond. max	r_{corr} p95	r_{corr} max
2	2.84	9.70	31.6	0.31	1.84
3	2.31	7.10	20.4	0.24	1.11
4	1.88	4.40	8.9	0.16	0.52
5	1.41	2.50	3.1	0.10	0.22
6	–	–	–	–	–
7	–	–	–	–	–
8	–	–	–	–	–

Within the $\tau = 0$ subset, the conditional median condition number decreases from 2.84 at $K = 2$ to 1.41 at $K = 5$, while the correction can still be non-negligible at low task count (r_{corr} p95 0.31 and maximum 1.84 at $K = 2$). These columns characterize the unregularized branch separately from the all-group condition statistics.

A finite-precision consistency diagnostic compares two constructions of the reduced fusion equations in the same reduced basis: the diagonal participation system and the independently summed taskwise system. At the ITCF-selected ridge value, it measures the Frobenius norm of the difference between their regularized corrections relative to the corresponding unregularized difference. This ratio is the retained perturbation fraction. Across the 3,840 groups with these measurements, the logarithm of the condition number and the retained fraction have Pearson and Spearman correlations of -0.821 and -0.853. A spectral predictor applies the ridge multipliers to columns of the unregularized difference and measures the resulting norm ratio. Its predictions correlate with the measured ratio at 0.985 and 0.980; this predictor is an approximate diagnostic, separate from the exact operator identities.

H.14 Numerical verification and tied eigenspaces

These tests verify the numerical identities used by the selector and the rotation behavior at a repeated-eigenvalue boundary.

High-precision evaluation covers 4,176 groups. Among the 3,145 groups with $k < s$, the computed threshold differs from $\sqrt{\lambda_k \lambda_{k+1}}$ by at most 8.32×10^{-8} in relative error. The two boundary errors agree to within 3.93×10^{-8}, and the same multiplier is the best point on every tested discrete threshold grid.

A second test applies 128 random rotations inside a tied eigenspace. The largest change in A is 1.19×10^{-7}. Under the same rotations, a partial hard projector changes by 0.9996, whereas the continuous filter changes by 5.96×10^{-8}. This test measures the basis dependence characterized in Proposition 2.

H.15 CONSTRUCTION-TIME AND MEMORY ANCHORS

The final appendix experiment reports the per-setting timing anchors underlying the construction-cost summary in the main text. Comparison and large-K scaling runs used one NVIDIA GeForce RTX 5090 D with Python 3.12.13 and PyTorch 2.11.0+cu128.

Table 33: Construction-time anchors. Times are median core seconds for one setting. ViT-B/16 and ViT-B/32 use seed 1; ViT-L/14 uses seed 0.

Model / K	ITCF	WUDI300	Speedup
ViT-B/16, 2	0.070	16.99	242.3×
ViT-B/16, 4	0.146	29.20	200.3×
ViT-B/16, 6	0.204	42.12	206.6×
ViT-B/32, 2	0.088	16.72	189.9×
ViT-B/32, 4	0.170	28.48	167.2×
ViT-B/32, 6	0.224	41.38	184.8×
ViT-B/32, 8	0.296	52.79	178.3×
ViT-L/14, 4	0.312	84.16	269.8×
ViT-L/14, 8	0.602	163.43	271.4×

Across the nine runtime anchors, the median of the nine per-anchor speedup ratios is $200.3\times$. Median construction times are 0.204 seconds for ITCF and 41.38 seconds for WUDI300. Peak CUDA allocation has a 5.38 GiB median and a 14.65 GiB maximum.

Standardized merge-construction kernels. Table 34 reports median standardized kernel times through $K = 166$ under the exclusions stated in its caption.

Table 34: Construction-time anchors. Times are median standardized merge-construction kernel seconds for one setting on identical hardware. Checkpoint disk I/O, CPU-to-GPU staging, downstream evaluation, and checkpoint serialization are excluded.

Model / K	ITCF	WUDI300	SWUDI-A	TSV-Merge
ViT-B/16, 2	0.07	16.99	0.85	7.56
ViT-B/16, 6	0.20	42.12	0.93	14.63
ViT-B/32, 8	0.30	52.79	0.95	18.23
T5-base, 16	1.08	423.02	2.35	128.03
T5-base, 72	2.06	1814.21	3.71	525.52
T5-base, 166	3.19	1154.68	5.52	1179.98

I ADDITIONAL STABILITY AND LARGE-EXPERT ANALYSES

This section examines how the ITCF construction behaves as expert collections become large and as checkpoint configurations approach several structural limits identified by the analysis above. We first characterize the geometry of the unmodified evaluation checkpoints, then use controlled interventions to study the corresponding limiting regimes.

Off-core accumulation in real checkpoints. Equation (49) gives an extreme case in which a truncated off-core component is repeated K times. To measure how closely real expert collections approach this geometry, define the off-core tail of task i as

$$R_i = \Delta_i(I - P_i),$$

and the effective coherent count

$$K_{\text{eff}} = \frac{\|\sum_i R_i\|_F^2}{\sum_i \|R_i\|_F^2}. \tag{110}$$

Identical aligned tails give $K_{\text{eff}} = K$, approximately incoherent tails give $K_{\text{eff}} \approx 1$, and cancellation can reduce it below one. Table 35 reports this diagnostic on the T5-base full-model prefixes used in the large-expert study. The off-core component retains 45–47% of the targeted update energy, while its mean pairwise alignment remains close to 0.03. Consequently, the effective coherent count stays far below the fully aligned duplicate limit, reaching $K_{\text{eff}} = 5.40$ at $K = 166$. The relative coherent fraction decreases from 0.145 at $K = 8$ to 0.033 at $K = 166$.

Table 35: Off-core accumulation on the T5-base large-expert prefixes. Off-core energy is $\sum_i \|R_i\|_F^2 / \sum_i \|\Delta_i\|_F^2$. Pairwise cosine is energy-weighted across task tails. The final column normalizes the effective coherent count by the exact-duplicate limit K.

K	Off-core energy	Pairwise cosine	K_{eff}	K_{eff}/K
8	45.12%	0.025	1.16	0.145
36	44.88%	0.028	1.94	0.054
72	45.26%	0.034	3.32	0.046
166	46.74%	0.028	5.40	0.033

Distance of natural checkpoints from the singular regime. To locate the unmodified experts relative to the zero-update limit examined below, each targeted-scope $\|\Delta_i\|_F$ is divided by the median positive expert norm in the same setting. At matrix-group level, each $\|\Delta_{g,i}\|_F$ is divided by the median positive task norm in the same setting and group. Across the unmodified 48-setting panel, all 204 expert-level updates and 16,104 group-level updates remain above the normalized 0.1 threshold used to mark the near-zero regime. The smallest normalized ratios are 0.170 and 0.148, respectively. The evaluated checkpoints therefore occupy a region separated from the singular limit examined below. Table 36 gives the family breakdown.

Table 36: Natural distance from the zero-update boundary on the 48-setting panel. Ratios are normalized within each setting (and, for group-level rows, within each matrix group). The final two columns count ratios at or below 0.1.

Scope	n_{exp}	Exp. min	Exp. p5	n_{grp}	Group min	Group p5	Exp. ≤ 0.1	Group ≤ 0.1
All	204	0.170	0.340	16104	0.148	0.340	0	0
Vision	112	0.913	0.920	5376	0.644	0.877	0	0
RoBERTa	54	0.170	0.237	3888	0.148	0.234	0	0
T5	38	0.286	0.399	6840	0.194	0.382	0	0

Zero-update limit under controlled scaling. We next examine the zero-update limit identified by Proposition B.4 under controlled expert scaling. Starting from the same evaluation settings, we select 10 settings by metadata only (four vision, three RoBERTa, and three T5) and fix one target expert per setting without using downstream scores. We scale its update toward zero using $\alpha\Delta_j$ for $\alpha \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 0\}$ while holding the remaining construction fixed. For every $\alpha > 0$, the primary diagnostic freezes the retained core extracted at $\alpha = 1$; at $\alpha = 0$ the target core is omitted exactly. This isolates the unit-mass boundary from changes in numerical rank extraction.

Across the 10 settings, the mean within-setting score span over $\alpha \in [10^{-4}, 10^{-2}]$ is 0.238 pp (median 0.106 pp), while changing from $\alpha = 10^{-4}$ to exact zero produces a mean relative fused-patch change of 0.299. The downstream jump is strongly family-dependent: the mean zero-minus-10^{-4} change is -1.22 pp on vision, -0.66 pp on RoBERTa, and $+17.19$ pp on T5; the pooled $+4.47$ pp change has a 95% paired interval $[-0.89, +11.67]$. Table 37 reports the descriptive score path.

Table 37: Controlled weak-expert scaling on 10 metadata-selected settings. Entries are mean downstream scores; the last column is the exact-zero score minus the $\alpha = 10^{-4}$ score. Family rows are descriptive decompositions of the same stress panel.

Scope	$\alpha = 1$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	0	$0 - 10^{-4}$
All (10)	68.33	57.61	55.29	55.15	55.10	59.57	+4.47
Vision (4)	65.43	59.41	55.24	54.86	54.80	53.58	−1.22
RoBERTa (3)	66.67	60.04	59.88	59.89	59.82	59.17	−0.66
T5 (3)	73.85	52.79	50.78	50.80	50.76	67.95	+17.19

The controlled scaling reproduces the transition predicted at the zero-update limit. Together with the natural-checkpoint measurements above, these results locate this behavior in a singular regime that lies outside the observed range of the unmodified evaluation checkpoints.

$K = 1$**.** With one task, $M_0 = \Delta_1$, $B = \Delta_1 G_1$, and $M_0 A = B$. Hence $Y = 0$ and ITCF returns the single task update without a correction.

Identical supports. If every task core is the same rank-r subspace, $A = (K/r)P$ and its condition number on the nonzero support is one. The requested task actions may still disagree, so $\mathcal{J}^\star$ can be large.

Orthogonal supports. For mutually orthogonal task cores, the eigenvalue on $\mathcal{S}_i$ is $1/r_i$. Every task has unit total mass even though individual directions from a high-rank core receive less mass.

Flat spectrum and the zero-ridge window. For a flat participation spectrum, $d_{\mathrm{part}} = s$, so ITCF chooses $k = s$ and $\tau = 0$. The inverse is well conditioned because all nonzero eigenvalues are equal. A nearly flat spectrum with one moderate weak tail can still fall inside the finite full-rank zero-ridge window until the ceiling changes or the tail leaves the numerical support.

Anchor bias. With positive regularization, Equation (12) keeps directions whose participation weights are small relative to the regularization threshold close to M_0. In those directions, any error already present in the anchor is retained more strongly as regularization increases. Participation describes equation strength, while anchor quality is a separate property.

Untargeted parameters. The sparse implementation updates selected matrix groups and leaves all other parameters at their pretrained values. Relative to expert i, each untouched group contributes $-\Delta_{g,i}$ to the non-target remainder in the functional bridge. These terms lie outside $\mathcal{J}$ and remain explicit in Equation (106).

J Factorial Geometry–Recovery Intervention

Controlled design. We separate two design axes on the frozen 48-setting cross-family panel (24 vision, 12 RoBERTa, and 12 T5 settings). The geometry is either the isotropic task core $G_i = P_i/r_i$ or a normalized squared-singular-value core

$$G_i^{(2)} = Q_i \, \frac{\mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{ir_i}^2)}{\sum_{j=1}^{r_i} \sigma_{ij}^2} \, Q_i^\top ,$$

using the same extracted bases and retained ranks. The recovery is either the current rule in Equation (11) or ordinary anchored Frobenius ridge, whose mode-wise interpolation is $\lambda/(\lambda + \gamma)$ as in Corollary 2. For each geometry, the participation operator and its spectrum are recomputed and the same checkpoint-only support-transition rule is applied; ordinary ridge uses the corresponding hard-reference scale $\gamma = \sqrt{\lambda_k \lambda_{k+1}}$ when $k < s$ and zero otherwise. Checkpoints, anchor, evaluator, and all other construction choices are shared, and no downstream score selects a cell.

Table 38: Geometry–recovery factorial intervention. Entries are mean downstream scores on the same frozen settings. ISO denotes the isotropic core and SV2 the normalized squared-singular-value core.

Scope	ISO+Current	SV2+Current	ISO+OrdRidge	SV2+OrdRidge
All (48)	**66.775**	66.723	66.497	66.606
Vision (24)	**70.567**	70.422	70.346	70.410
RoBERTa (12)	51.947	**52.160**	51.272	51.640
T5 (12)	74.018	73.886	**74.026**	73.966

The pooled means show that neither geometry nor recovery acts as a uniform additive improvement. Table 39 reports the paired contrasts. Under isotropic geometry, the current recovery improves the mean by $+0.277\,\mathrm{pp}$ with 37/1/10 wins/ties/losses and a 95% paired interval $[+0.113, +0.450]$. The geometry–recovery interaction is $+0.161\,\mathrm{pp}$ with 32/0/16 and interval $[+0.018, +0.311]$. By contrast, the isolated isotropic-minus-SV2 geometry effect is $+0.052\,\mathrm{pp}$ under the current recovery ($[-0.155, +0.260]$) and $-0.109\,\mathrm{pp}$ under ordinary ridge ($[-0.274, +0.050]$), so the experiment does not support a universal downstream main effect of isotropic weighting.

Table 39: Paired factorial contrasts in percentage points. "Geo./Cur." is ISO$-$SV2 under the current recovery; "Geo./Ord." is the same contrast under ordinary ridge; "Rec./ISO" and "Rec./SV2" are Current$-$OrdRidge at fixed geometry. Interaction is Geo./Cur. minus Geo./Ord. Family rows are descriptive decompositions.

Scope	Geo./Cur.	Geo./Ord.	Rec./ISO	Rec./SV2	Interaction
All (48)	$+0.052$	-0.109	$+0.277$	$+0.116$	$+0.161$
Vision (24)	$+0.145$	-0.064	$+0.221$	$+0.012$	$+0.210$
RoBERTa (12)	-0.213	-0.367	$+0.675$	$+0.521$	$+0.154$
T5 (12)	$+0.132$	$+0.060$	-0.008	-0.080	$+0.072$

The family decomposition localizes this pooled interaction rather than replacing it with a universal claim. Vision shows a $+0.210\,\mathrm{pp}$ interaction (95% interval $[+0.147, +0.275]$), RoBERTa shows a clear recovery benefit at both geometries while its descriptive geometry contrasts favor SV2, and T5 is near parity across the four cells. Together with the anisotropic Qwen probe in Appendix H.12, this supports complementarity between the checkpoint-robust comparison geometry and the current recovery, rather than a claim that isotropic weighting is the downstream optimum for every family.