# GRAM: Group-wise Rank-Aware Model Merging via Subspace Alignment

**Di Wu***
JPMorgan Chase & Co.
di.x5.wu@chase.com

**Shuaidong Pan***
JPMorgan Chase & Co.
shuaidong.pan@chase.com

## Abstract

Merging different models into a single model is highly desirable for multi-task deployment, particularly in regulated domains like healthcare. In such settings, models are often locally fine-tuned due to strict data privacy policies, necessitating that aggregation operates exclusively on the fine-tuned weights. Existing methods apply a uniform merge across all tasks, ignoring the geometric relationships among their learned subspaces. When task updates occupy misaligned subspaces, projecting them onto a shared low-rank basis discards task-relevant directions, a failure mode named *subspace interference*. Through this study, we show that the per-task projection error of a merged model is governed by the principal-angle alignment between each task and the shared subspace, so that grouping geometrically compatible tasks provably reduces this error within each group. Motivated by this observation, we propose **Group-wise Rank-Aware Model Merging (GRAM)**, which uses the principal-angle geometry of task vectors to decide which tasks should be merged together and then performs a standard SVD-based merge within each coalition; the pipeline requires only the fine-tuned weights at merge time. With only one additional model ($K$=2), GRAM significantly mitigates subspace interference and consistently outperforms prior merging methods across language understanding (GLUE with Flan-T5), 8-task CLIP vision, and medical instruction tuning (Qwen2-7B-Instruct with four English/Chinese medical LoRAs).

## 1 Introduction

Model merging [Wortsman et al., 2022, Ilharco et al., 2022, Tang et al., 2023] aggregates multiple fine-tuned, task-specific models into a single deployable model without the need for additional training. This paradigm is highly attractive for multi-task deployment, as it circumvents the substantial costs of storing and serving independent models. The structured nature of weight updates—particularly in parameter-efficient fine-tuning (e.g., LoRA [Hu et al., 2022])—makes model merging highly effective. This approach is especially critical in privacy-constrained domains such as healthcare, where institutions independently fine-tune a shared backbone on local clinical data. Because data sharing across sites is strictly restricted, any unified multi-institution model must be constructed solely from the fine-tuned weights, precluding the use of the original training data.

Prior work on model merging spans parameter-space arithmetic with conflict resolution [Ilharco et al., 2022, Yadav et al., 2023], auxiliary signal coefficient learning [Yang et al., 2023, Tang et al., 2023], and structured approximations exploiting low-rank or layer-wise properties [Gargiulo et al., 2025, Stoica et al., 2024, Marczak et al., 2025, Wei et al., 2025, Sun et al., 2025, Chen et al., 2025a]. A pervasive limitation of these methods is their reliance on a uniform global merge: all tasks are compressed into a single shared space. Consequently, tasks whose learned parameter directions are least compatible with the majority suffer from disproportionate performance degradation—a

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

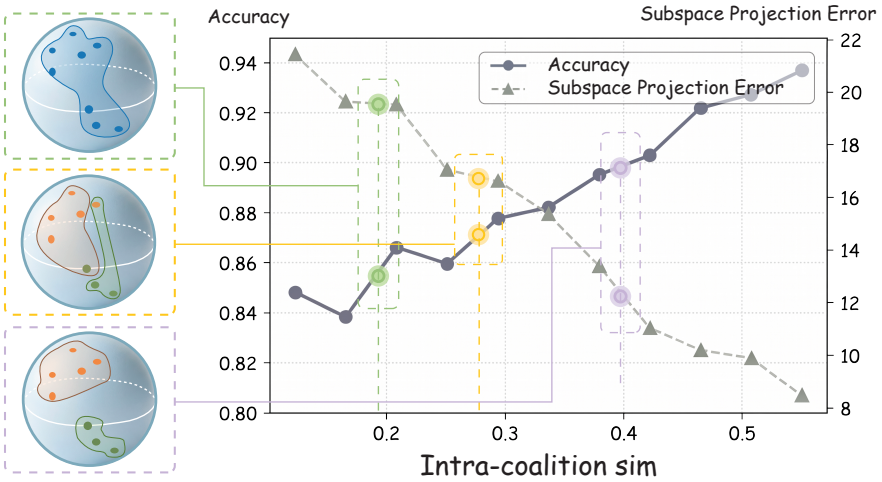



Figure 1: How tasks are grouped on the Grassmannian governs merging quality. *Insets:* three partitions of the per-task low-rank subspaces visualised on the Grassmannian—all tasks merged into a single global coalition (top, low intra-coalition similarity), three groups (middle), and the two-group GRAM partition (bottom, high intra-coalition similarity). *Curve:* as intra-coalition similarity increases, merged-model accuracy rises (left axis, solid) and the within-group subspace projection error $\sum_{\text{pairs}\in G}(1-\cos^2\theta)$ falls (right axis, dashed). Highlighted markers tie each partition to its location on the trade-off curve: GRAM places geometrically compatible tasks in the same coalition, reducing the alignment penalty that global merging incurs.

phenomenon we term subspace interference. Furthermore, many existing techniques require access to task-specific calibration data or activations, rendering them inapplicable in strict privacy-preserving deployments where data cannot leave its source institution.

A critical signal overlooked by current methods is the inter-task geometry of the learned parameter subspaces. Existing techniques process updates individually or globally, systematically neglecting the pairwise geometric relationships between tasks. As illustrated in Figure 1, subspace similarity is highly predictive of merging compatibility: grouping tasks with high intra-coalition subspace similarity simultaneously increases merged-model accuracy and minimizes projection error. We observe distinct geometric signatures—such as cluster-outlier structures in GLUE tasks and block-diagonal affinities in vision tasks (section D). Crucially, the outlier tasks are precisely those that disproportionately degrade standard global SVD merging, and thus benefit the most from group-wise isolation (section 5). Subspace geometry therefore carries predictive information that current methods fail to exploit.

Motivated by this geometric insight, we propose Group-wise Rank-Aware Model Merging (GRAM). GRAM targets the task-aware deployment setting, where task identity is known at inference. We reframe the merging objective from compressing $T$ models into a single model, to compressing $T$ models into $K \ll T$ models, trading a marginal increase in storage for a targeted, significant reduction in subspace interference. Crucially, GRAM is entirely data-free: the pipeline operates exclusively on the fine-tuned weights, requiring no source data, activations, or calibration sets at merge time. Our primary contributions are as follows:

1. **A Geometry-Aware Merging Algorithm:** We introduce GRAM, which partitions tasks using Grassmannian $k$-means++ on chordal distances. Within each coalition, it dynamically merges models via a greedy adaptive rank allocator that pools singular values across members, achieving provable optimality under a global rank budget.

2. **Theoretical Guarantees:** We rigorously bound the projection error of our clustering objective. We prove that the resulting partition yields a merge error within $O(\log K)$ of the optimal partition, and its expected error is strictly bounded below that of a standard global merge.

3. **State-of-the-Art Data-Free Performance:** Extensive evaluations across language understanding, vision, and medical instruction-tuning demonstrate that GRAM consistently outperforms existing data-free baselines. Notably, on the medical benchmark, GRAM surpasses even data-dependent methods (e.g., RegMean, Fisher-Merge) that rely on per-task calibration data.

## 2 Related Work

**Model Merging.** Model merging combines independently fine-tuned parameters into a single deployable model without retraining. Parameter-level methods average fine-tuned weights or arithmetically combine task vectors, sometimes with sparsification or sign/magnitude-based conflict resolution [Wortsman et al., 2022, Ilharco et al., 2022, Yadav et al., 2023, Yu et al., 2024]. Coefficient-learning methods learn per-task or per-layer mixing weights from auxiliary signals or coordinate-wise masks [Yang et al., 2023, Tang et al., 2023]. Structured-update methods exploit the low-rank or layered structure of fine-tuning updates, operating in the singular-value or per-layer domain to reduce inter-task interference [Gargiulo et al., 2025, Stoica et al., 2024, Marczak et al., 2025, Chen et al., 2025b, Wei et al., 2025, Wang et al., 2024b, Sun et al., 2025, Fu et al., 2025]. Multi-model approaches retain some per-task structure rather than collapsing to a single shared model [Huang et al., 2024]. A common thread across these single-model approaches is that all tasks participate in the same global merge, implicitly assuming sufficient subspace alignment—an assumption that fails when tasks span diverse domains. We depart from it by measuring inter-task subspace geometry directly and merging within geometrically compatible groups.

**Task Grouping in Multi-Task Learning.** Deciding which tasks should be trained together has a long history in multi-task learning. Zamir et al. [2018] computationally map transfer relationships across 26 visual tasks, and Standley et al. [2020] formalize the grouping problem and search over partitions to trade accuracy against compute. Task2Vec [Achille et al., 2019] represents each task by a Fisher-information vector, while branched networks [Vandenhende et al., 2019] and mixture-of-experts models [Chen et al., 2023] realize grouping through the architecture itself. Affinity-based methods avoid combinatorial search: TAG [Fifty et al., 2021] estimates inter-task affinity from gradient signals in a single training run, MTG-Net [Song et al., 2022] meta-learns a predictor of per-task gains across task combinations, Grad-TAG [Li et al., 2024] scales affinity estimation via gradient linearization, and Wang et al. [2024a] cast grouping as a mathematical program with theoretical guarantees. Complementary work on dynamic loss weighting [Liu et al., 2022] modulates task competition during training rather than partitioning tasks explicitly. Dynamic client cooperation [Yang et al., 2026] mitigates interference through adaptive coalition formation rather than explicitly partitioning tasks. All of these approaches require training-time data and gradients, making them inapplicable to the post-hoc merging setting where only frozen weights are available. Our grouping criterion is data-free: it uses only the left singular subspaces of the task-specific weight updates.

## 3 Preliminaries

**Task vectors.** We consider a pre-trained model with $L$ layers and weights $W_0 = \{W_0^{(\ell)}\}_{\ell=1}^{L}$. After adapting the model to task $i \in \{1, \dots, T\}$, we obtain fine-tuned weights $W_i$. Following the task-vector view of model merging [Ilharco et al., 2022, Yadav et al., 2023, Yu et al., 2024], the adaptation is summarized by the per-layer *task vector*

$$\Delta W_i^{(\ell)} \;=\; W_i^{(\ell)} - W_0^{(\ell)} \;\in\; \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \tag{1}$$

i.e., the difference between the task-adapted and pre-trained weights at layer $\ell$. The matrix formulation in (1) applies to *linear* layers — MLP blocks and attention projection matrices — which covers all merged parameters in our experiments: the LoRA settings adapt only linear projections by construction, and in the fully fine-tuned vision setting SVD-based merging is applied to all 2-D weight matrices while the remaining non-matrix parameters (LayerNorm gains, biases) are merged by simple averaging (section D.6). For models tuned with Low-Rank Adaptation (LoRA) [Hu et al., 2022], the task vector inherits a rank-$r$ factorization

$$\tau_i^{(\ell)} \;=\; B_i^{(\ell)} A_i^{(\ell)}, \qquad B_i^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times r},\; A_i^{(\ell)} \in \mathbb{R}^{r \times d_{\text{in}}},\; r \ll \min(d_{\text{out}}, d_{\text{in}}). \tag{2}$$

**SVD of task updates.** The compact SVD of each update is

$$\Delta W_i^{(\ell)} = U_i^{(\ell)} \, \Sigma_i^{(\ell)} \, V_i^{(\ell)\top}, \tag{3}$$

where $U_i^{(\ell)} \in \mathbb{R}^{d_{\text{out}} \times r}$ contains the left singular vectors (output subspace), $\Sigma_i^{(\ell)} \in \mathbb{R}^{r \times r}$ is a diagonal matrix of singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and $V_i^{(\ell)} \in \mathbb{R}^{d_{\text{in}} \times r}$ contains the right singular vectors (input subspace). We drop the layer index $(\ell)$ when the context is unambiguous.

**Merging objective.** Given $T$ task vectors, each of rank $r$ across $L$ layers, the goal is to produce $K < T$ merged task vectors that minimize the total reconstruction error across all tasks subject to a per-merge rank budget:

$$\min_{\text{partition, } \{\widetilde{\Delta W}_k\}_{k=1}^K} \sum_{i=1}^{T} \left\| \Delta W_i - \widetilde{\Delta W}_{\pi(i)} \right\|_{\text{F}}^2 \quad \text{s.t.} \quad \text{rank}(\widetilde{\Delta W}_k) \leq r \;\; \text{for all } k \in \{1, \ldots, K\}, \tag{4}$$

where $\pi(i) \in \{1, \ldots, K\}$ assigns task $i$ to its group $G_{\pi(i)}$ and $\widetilde{\Delta W}_k \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the merged update for group $G_k$. The model deployed for group $G_k$ is, per layer,

$$W_{G_k}^{(\ell)} \;=\; W_0^{(\ell)} + \widetilde{\Delta W}_k^{(\ell)}, \tag{5}$$

i.e., each of the $K$ deployed models is the pre-trained backbone plus its group's merged task vector, and task $i$ is served by the model of group $G_{\pi(i)}$.

**Subspace interference.** Let $B \in \mathbb{R}^{d_{\text{out}} \times r}$ be a shared output basis with orthonormal columns ($B^\top B = I_r$). Substituting the compact SVD $\Delta W_i = U_i \Sigma_i V_i^\top$, the projection error of task $i$ onto $\text{col}(B)$ is

$$\mathcal{E}_i \;=\; \left\| (I - BB^\top) \, \Delta W_i \right\|_{\text{F}}^2 \;=\; \left\| (I - BB^\top) \, U_i \Sigma_i \right\|_{\text{F}}^2, \tag{6}$$

We adopt $\mathcal{E}_i$ as a principled measure of inter-task interference as it directly measures the task-relevant information lost when updates are forced into a shared subspace. Crucially, $\mathcal{E}_i$ depends only on $U_i$ and $\Sigma_i$ — not on $V_i$ — so partitioning by output-subspace alignment captures the interference signal a shared basis can control.

To isolate the geometric contribution of subspace misalignment from the magnitude of the task update, we drop the singular-value weighting and analyze $\left\| (I - BB^\top) \, U_i \right\|_{\text{F}}^2$. Every column of $U_i$ admits a unique orthogonal decomposition into a component in $\text{col}(B)$ and a component in $\text{col}(B)^\perp$. Applied column-wise,

$$U_i \;=\; \underbrace{BB^\top U_i}_{\in \text{col}(B)} + \underbrace{(I - BB^\top) \, U_i}_{\in \text{col}(B)^\perp}.$$

The two summands live in orthogonal subspaces, so by the Pythagorean theorem (in the Frobenius inner product), $\|U_i\|_{\text{F}}^2 \;=\; \left\| BB^\top U_i \right\|_{\text{F}}^2 + \left\| (I - BB^\top) \, U_i \right\|_{\text{F}}^2$. Since $B$ has orthonormal columns, $\left\| BB^\top U_i \right\|_{\text{F}}^2 = \left\| B^\top U_i \right\|_{\text{F}}^2$, and $\|U_i\|_{\text{F}}^2 = r$. Rearranging,

$$\left\| (I - BB^\top) \, U_i \right\|_{\text{F}}^2 \;=\; r - \left\| B^\top U_i \right\|_{\text{F}}^2. \tag{7}$$

Let $\theta_1, \ldots, \theta_r$ denote the principal angles between $\text{col}(U_i)$ and $\text{col}(B)$; throughout the paper we reserve $\theta$ for principal angles and $k$ for merge groups (model weights are denoted $W$), and index principal angles by $j$. A standard result [Golub and Van Loan, 1996] states that the singular values of $B^\top U_i$ are exactly $\{\cos\theta_j\}_{j=1}^r$, so $\left\| B^\top U_i \right\|_{\text{F}}^2 = \sum_{j=1}^r \cos^2\theta_j$. Substituting into (7) and using $\sin^2\theta_j + \cos^2\theta_j = 1$,

$$\left\| (I - BB^\top) \, U_i \right\|_{\text{F}}^2 \;=\; r - \sum_{j=1}^{r} \cos^2\theta_j \;=\; \sum_{j=1}^{r} \sin^2\theta_j, \tag{8}$$

which is the squared chordal (projection F-) distance between the two subspaces. The expression depends only on the relative geometry of $\text{col}(U_i)$ and $\text{col}(B)$ and is invariant to the choice of basis. When the subspaces are well-aligned ($\theta_j \approx 0$ for all $j$), the projection error vanishes. When they are misaligned, important task directions are discarded by the merge. This motivates grouping tasks by (singular-value-weighted) subspace compatibility before merging.

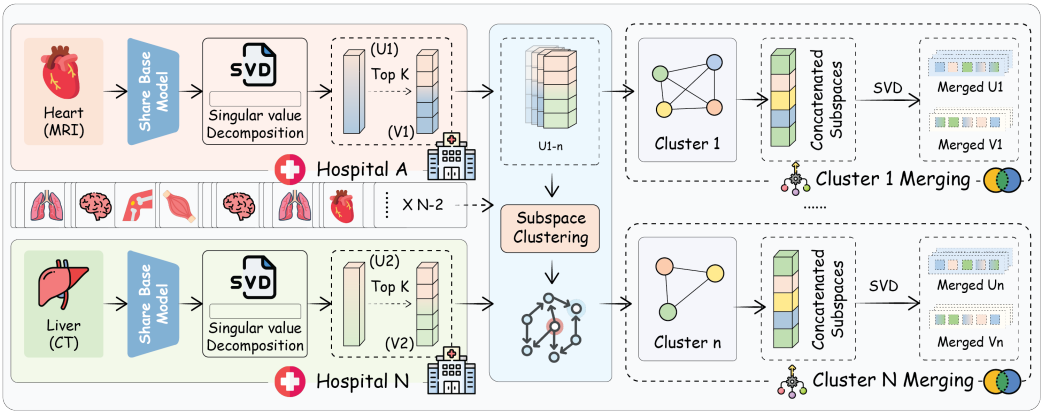



Figure 2: The GRAM pipeline. Each task vector $\Delta W_i$ is decomposed via SVD to extract its top-$r$ left- and right-singular subspaces $(U_i, V_i)$ (left). The collection of $U_i$'s is fed into a subspace clustering step that partitions the $T$ tasks into $K$ coalitions by chordal distance (middle). Within each coalition, the selected top-$k_i$ singular components from member tasks are concatenated and re-orthonormalized via thin SVD to yield a single rank-$r$ merged update (right).

## 4 Method: Group-wise Rank-Aware Model Merging

Our goal is to merge $T$ task vectors into $K \ll T$ per-group updates in a way that protects geometrically incompatible tasks from mutual interference. To accomplish this, we decompose the pipeline (fig. 2) into three stages. **(1) Quantifying pairwise subspace geometry.** We compute the pairwise chordal distance between task output subspaces. This quantity directly governs the projection error a shared basis would incur (Theorem B.3), so it is the natural input for the clustering step rather than coordinate-wise distances on flattened parameters. **(2) Partitioning tasks into geometrically coherent coalitions.** Treating each task's projection matrix $P_i = U_i U_i^\top$ as a point in Euclidean space, we run Grassmannian $k$-means++ on the chordal-distance matrix to obtain $K$ coalitions with low within-group and high between-group subspace distance, importing the $O(\log K)$ approximation guarantee of $k$-means++ into the Grassmannian setting. **(3) Per-coalition low-rank merge.** Within each coalition we concatenate the per-task top-$k_i$ singular triples under an adaptive rank allocator that distributes the rank budget to minimize the reconstruction error, then orthonormalize via a thin SVD to produce a single rank-$r$ merged update per coalition.

### 4.1 Chordal Subspace Distance

We measure the geometric compatibility between two task subspaces via the *squared chordal distance*. At layer $\ell$, the chordal distance between the column spaces of $U_i^{(\ell)}$ and $U_j^{(\ell)}$ is

$$d_c^2\big(U_i^{(\ell)}, U_j^{(\ell)}\big) \;=\; r \;-\; \Big\| U_i^{(\ell)\top} U_j^{(\ell)} \Big\|_{\mathrm{F}}^2 \;=\; \sum_{s=1}^{r} \sin^2 \theta_s, \tag{9}$$

where $\theta_1, \ldots, \theta_r$ are the principal angles between $\mathrm{col}(U_i^{(\ell)})$ and $\mathrm{col}(U_j^{(\ell)})$ (indexed by $s$ here since $j$ denotes a task). By eq. (8), this quantity directly governs the projection error incurred when two tasks share a merged basis, making it the natural metric for deciding which tasks to group. Furthermore, we show that the chordal distance on the Grassmannian is equivalent to the squared Frobenius distance between the orthogonal projectors $P_i^{(\ell)} := U_i^{(\ell)} U_i^{(\ell)\top}$ up to a factor of 2 (section C.4), so off-the-shelf $k$-means++ on $\{P_i^{(\ell)}\}$ inherits its $O(\log K)$ approximation guarantee in the chordal geometry (section B).

### 4.2 Grassmannian $k$-Means++ Clustering

To obtain a single distance per task pair, we average the per-layer chordal distance across layers:

$$\bar{d}_c^2(i,j) \;=\; \frac{1}{L}\sum_{\ell=1}^{L} d_c^2\big(U_i^{(\ell)},\, U_j^{(\ell)}\big). \tag{10}$$

We then run $k$-means++ [Arthur et al., 2007] on $\bar{d}_c^2$ to partition the $T$ tasks into $K$ groups $\{G_1, \ldots, G_K\}$. The number of groups $K$ is a user-specified *compression budget* that trades deployment cost ($K$ deployed models) against merging fidelity. In our experiments (section 5), $K{=}2$ provides the best cost–accuracy trade-off on all benchmarks.

### 4.3 Per-Group Low-Rank Merge

For each group $G_k$ and layer $\ell$, we merge the constituent task vectors into a single rank-$r$ update. The procedure operates on all three SVD components—left singular vectors $U_i$, singular values $\sigma_i$, and right singular vectors $V_i$—and allocates rank budget adaptively rather than uniformly across tasks.

**Adaptive rank allocation.** A naïve approach gives each of the $|G_k|$ tasks an equal share $r/|G_k|$ of the rank budget. Since intrinsic rank varies sharply across tasks and layers in practice (fig. 4, section D.3), this uniform allocation wastes capacity on compact tasks and starves complex ones.

We instead formulate rank allocation as a discrete optimization problem. Let $\Delta W_i = U_i \Sigma_i V_i^\top$ be the SVD of task $i$'s update, and let $r_i$ denote the number of singular directions allocated to task $i$. The per-task reconstruction error after keeping the top $r_i$ components is

$$\mathcal{L}_i(r_i) = \big\|(I - B_i^{(r_i)} {B_i^{(r_i)}}^\top)\,\Delta W_i\big\|_{\mathrm{F}}^2 = \sum_{j>r_i} \sigma_{i,j}^2, \tag{11}$$

where $B_i^{(r_i)} \in \mathbb{R}^{d_{\text{out}} \times r_i}$ contains the top-$r_i$ left singular vectors of $\Delta W_i$, and the second equality follows from the Eckart–Young theorem. We seek the allocation that minimizes total reconstruction error under a global budget:

$$\min_{\{r_i\}} \sum_{i \in G_k} \sum_{j>r_i} \sigma_{i,j}^2 \qquad \text{s.t.} \quad \sum_{i \in G_k} r_i \le r. \tag{12}$$

The marginal gain of assigning one additional rank slot to task $i$ is $\mathcal{L}_i(r_i) - \mathcal{L}_i(r_i+1) = \sigma_{i,\,r_i+1}^2$. Since singular values are sorted in descending order within each task, the optimal strategy is *greedy*: pool all candidate singular values $\{\sigma_{i,j}^2\}$ across tasks in the group, select the top-$r$ entries, and count how many are drawn from each task. This yields optimal per-task allocations $r_1, \ldots, r_{|G_k|}$ with $\sum_i r_i = r$, ensuring that rank budget flows to the directions that most reduce reconstruction error.

**Coalition merge.** Given the per-task allocations $\{r_i\}$, we follow TSV-Merge [Gargiulo et al., 2025]: concatenate the selected top-$r_i$ SVD components across tasks into $\bar{U} \in \mathbb{R}^{d_{\text{out}} \times r}$, $\bar{s} \in \mathbb{R}^r$, and $\bar{V} \in \mathbb{R}^{r \times d_{\text{in}}}$, sort globally by singular value, replace $\bar{U}$ and $\bar{V}^\top$ with their orthogonal polar factors $\widehat{U}, \widehat{V}$ (rather than computing an SVD of the joint operator), and reconstruct $\widetilde{\Delta W}_k^{(\ell)} = \widehat{U}\ \mathrm{diag}(\bar{s})\ \widehat{V} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ as the task vector for group $G_k$ at layer $\ell$.

**Theoretical guarantee for grouping.** The grouping step is backed by a deterministic projection-error guarantee: with optimal per-group bases, any partition into $K$ groups achieves total reconstruction error no larger than a single global rank-$r$ basis fit to all tasks (Theorem B.1 in section B; the proof is a direct optimality argument independent of the clustering rule). The guarantee explains *why* the empirical gains in section 5 concentrate on tasks geometrically furthest from the bulk—exactly the tasks for which a partition-aware basis improves most over the global optimum, making the inequality strict and large in magnitude—while the alignment-penalty machinery (Theorem B.3) and the $k$-means++ partition error chain in section B quantify *how much* grouping helps.

Table 1: GLUE results on Flan-T5-Base and Flan-T5-Large (LoRA $r$=16): validation accuracy, Spearman $\rho\times100$ for STS-B. Each cell reports Base / Large in that order. **Bold**: best merged result per column for each variant. Individual: per-task fine-tuned model (oracle upper bound).

| Method | CoLA | MNLI | MRPC | QNLI | QQP | RTE | SST-2 | STS-B | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Individual (oracle) | 69.1 / 80.1 | 82.7 / 88.5 | 85.5 / 89.2 | 90.9 / 94.4 | 84.0 / 87.1 | 84.4 / 91.3 | 92.9 / 95.2 | 87.4 / 90.9 | 84.6 / 89.6 |
| Weight Averaging | **69.7** / 74.6 | 59.7 / 84.3 | 78.9 / 84.1 | 90.1 / 92.8 | 83.8 / 86.2 | 80.5 / 85.9 | 91.2 / 94.8 | 72.0 / 88.0 | 78.2 / 86.3 |
| Task Arithmetic | 69.5 / 76.9 | 59.0 / 85.4 | 78.7 / 85.3 | 90.1 / **93.9** | 83.9 / 85.8 | 79.4 / 88.1 | 91.5 / 95.2 | 72.9 / 87.8 | 78.1 / 87.3 |
| Concrete TA | 69.1 / 76.6 | 58.1 / 86.3 | 78.4 / 86.3 | 89.9 / **93.9** | 83.5 / 85.9 | 79.4 / **88.4** | 91.6 / 95.2 | 73.4 / 87.9 | 78.0 / 87.5 |
| DOGE TA | 69.1 / 78.4 | 71.9 / 88.1 | 80.9 / 86.5 | **90.3** / 93.8 | 83.5 / **86.3** | 79.8 / 87.7 | 92.5 / 95.1 | 71.1 / 87.7 | 79.9 / 88.0 |
| TIES-Merging | 68.3 / 77.1 | 56.3 / 85.1 | 79.4 / 86.3 | 89.8 / **93.9** | 83.7 / 86.0 | 79.4 / 87.7 | 91.6 / 95.1 | 71.2 / 88.0 | 77.5 / 87.4 |
| AdaMerging++ | 69.1 / 76.7 | 60.3 / 87.6 | 78.4 / 84.8 | 90.0 / 93.8 | 83.6 / 85.9 | 79.1 / 88.1 | 91.6 / 95.2 | 74.1 / 88.6 | 78.3 / 87.6 |
| Concrete AM | 69.0 / 76.1 | 59.4 / 87.7 | 80.1 / 85.5 | 89.9 / 93.8 | 82.9 / 85.9 | 79.1 / 88.1 | 91.7 / 95.4 | 75.4 / 87.1 | 78.5 / 87.8 |
| TSV-Merge | 69.2 / 79.2 | 74.7 / 86.0 | 80.9 / 87.3 | 90.2 / 93.6 | 83.6 / 85.6 | 78.7 / 87.0 | 92.7 / 95.3 | 83.2 / 88.7 | 81.7 / 87.8 |
| Iso-C | 69.1 / 73.8 | 59.9 / 64.2 | 77.9 / 82.4 | 88.9 / 92.2 | 82.9 / 86.0 | **80.9** / 82.7 | 91.5 / 92.8 | 66.1 / 86.4 | 77.2 / 82.5 |
| Iso-CTS | 69.1 / 73.8 | 58.3 / 59.1 | 77.0 / 82.1 | 88.5 / 91.8 | 82.2 / 85.8 | 80.5 / 82.3 | 91.4 / 92.4 | 64.0 / 86.0 | 76.4 / 81.7 |
| **GRAM** (ours) | 69.5 / **79.3** | **82.8 / 89.0** | **84.3 / 88.7** | 90.2 / 93.8 | **84.0** / 85.7 | 79.8 / 88.1 | **93.2 / 95.4** | **86.7 / 89.9** | **83.8 / 88.7** |

# 5 Experiments

## 5.1 Setup

We evaluate GRAM on three benchmarks: GLUE [Wang et al., 2018] with Flan-T5-Base/Large [Chung et al., 2024], the 8-task vision benchmark with fully fine-tuned CLIP ViT-B/32 and ViT-L/14, and medical instruction tuning on Qwen2-7B-Instruct with four LoRA experts (section 5.4). All methods operate on the same pre-existing fine-tuned models—from FusionBench [Tang et al., 2025] for GLUE and from the TSV-Merge release [Gargiulo et al., 2025] for vision—and are re-run under a unified protocol (section D), so cross-method differences come only from merging. We report validation accuracy on GLUE (Spearman $\rho\times100$ for STS-B), test accuracy on vision, and the metrics described in section 5.4 for the medical benchmark. All experiments use task-aware inference: task identity is known at test time and selects the group-specific adapter. This is a preference rather than a requirement (section 5.5).

For GLUE and vision, we compare against seven baselines: Weight Averaging [Wortsman et al., 2022], Task Arithmetic [Ilharco et al., 2022], TIES-Merging [Yadav et al., 2023], AdaMerging++ [Yang et al., 2023], Iso-C and Iso-CTS [Marczak et al., 2025], and TSV-Merge [Gargiulo et al., 2025]. The GLUE results in table 1 additionally include Concrete TA, Concrete AM [Tang et al., 2023], and DOGE TA [Wei et al., 2025]. The CLIP ViT-B/16 results in section D.5 additionally include LiNeS-Merge [Wang et al., 2024b] and LOT-Merging [Sun et al., 2025] for completeness across recent structured-update baselines. The medical benchmark (section 5.4) additionally compares against Fisher-Merge [Matena and Raffel, 2022] and RegMean [Jin et al., 2022] as data-using references. Ablation includes random-equipartition controls under the same chordal metric.

## 5.2 Language Understanding

Table 1 presents GLUE results on Flan-T5-Base and Flan-T5-Large (LoRA $r$=16). We fix $K$=2 throughout to keep the budget overhead minimal—a single extra deployed model relative to the $K$=1 baselines—and ask whether this small extra resource yields meaningful gains. On Flan-T5-Base, GRAM reaches 83.8% average, $+2.1$ over the strongest baseline TSV-Merge (81.7%). On Flan-T5-Large, the gap between methods narrows in the higher-capacity regime, but GRAM still leads the field at 88.7%, $+0.9$ over TSV-Merge (87.8%)—all with only 2 deployed models and no per-task calibration data.

The per-task profile is consistent with the alignment-penalty mechanism of Theorem B.3: tasks at large principal angles from the bulk cluster bear the largest projection error under a global merge and stand to recover the most under per-coalition merging. MNLI and STS-B sit furthest from the {MRPC, QNLI, QQP, RTE} cluster in subspace geometry (section D), and they are the two tasks where GRAM recovers the most relative to TSV-Merge at both scales (MNLI $74.7\rightarrow82.8$ base, $86.0\rightarrow89.0$ large; STS-B $83.2\rightarrow86.7$ base, $88.7\rightarrow89.9$ large). On QQP, GRAM attains the highest base accuracy ($84.0$); on SST-2 it is within $0.1$ of the strongest baseline.

Table 2: Eight-task vision results on fully fine-tuned CLIP ViT-B/32 and ViT-L/14: test accuracy (%). Each cell reports B/32 / L/14 in that order. **Bold**: best merged result per column for each variant. Individual: per-task fine-tuned model (oracle upper bound).

| Method | MNIST | SVHN | GTSRB | DTD | Cars | RESISC45 | EuroSAT | SUN397 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Individual (oracle) | 99.7 / 99.8 | 97.6 / 98.1 | 99.1 / 99.2 | 97.7 / 98.5 | 78.2 / 92.2 | 95.9 / 97.3 | 99.8 / 99.8 | 75.3 / 82.0 | 92.9 / 95.9 |
| Weight Averaging | 86.5 / 96.8 | 62.7 / 77.7 | 53.3 / 68.9 | 54.7 / 66.8 | 62.6 / 81.7 | 71.9 / 82.3 | 71.0 / 88.7 | 65.1 / 71.5 | 66.0 / 79.3 |
| Task Arithmetic | 96.9 / 99.0 | 78.1 / 86.0 | 68.2 / 84.9 | 59.6 / 72.2 | 56.2 / 81.9 | 68.1 / 86.3 | 76.0 / 91.9 | 57.3 / 72.9 | 70.1 / 84.4 |
| TIES-Merging | 97.0 / 99.0 | 80.5 / 87.6 | 69.3 / 83.7 | 66.2 / 77.5 | 61.1 / 83.3 | 73.2 / 87.7 | 79.4 / 93.8 | 62.7 / 74.0 | 73.7 / 85.8 |
| AdaMerging++ | 98.2 / 99.0 | 90.1 / 90.2 | 73.2 / 95.7 | 81.3 / 93.7 | 69.2 / 88.6 | 78.2 / 91.5 | 93.9 / 97.2 | 65.7 / 74.7 | 81.2 / 91.3 |
| TSV-Merge | 99.3 / **99.6** | 91.5 / 95.1 | 91.8 / 96.3 | 83.8 / 93.4 | 69.5 / 89.3 | 84.8 / 93.2 | 94.1 / 98.5 | 65.6 / 76.8 | 85.0 / 92.8 |
| Iso-C | 98.9 / 99.3 | 87.0 / 90.5 | 91.0 / 95.8 | 88.1 / 96.1 | 71.0 / 91.4 | 87.0 / 94.7 | 92.4 / 98.8 | 69.1 / 79.7 | 85.6 / 93.3 |
| Iso-CTS | 98.9 / 99.4 | 84.9 / 91.2 | 91.6 / 96.6 | 89.0 / **97.0** | 71.4 / **91.9** | 88.4 / **95.4** | 92.5 / 99.1 | **69.4** / **80.2** | 85.8 / 93.9 |
| **GRAM** (ours) | **99.4 / 99.6** | **94.2 / 96.4** | **94.6 / 98.3** | **90.4** / 95.5 | **73.1** / 90.6 | **89.4** / 94.4 | **99.1 / 99.3** | 69.1 / 78.2 | **88.7 / 94.0** |

Table 3: Medical merging on Qwen2-7B-Instruct with four LoRA experts. MedQA / CMExam: accuracy; HealthcareMagic / cMedQA2: BLEU-4 / ROUGE-1/2/L. *Avg Impr.*: mean relative improvement over Pretrained across 10 columns; *Avg Rank.*: mean per-column rank (1 = best). **Bold** / underline: best / second-best per column among merging methods. Individual: per-task fine-tuned oracle.

| | MedQA | CMExam | HealthcareMagic | | | | cMedQA2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Acc. | Acc. | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU-4 | ROUGE-1 | ROUGE-2 | ROUGE-L | Avg Impr. | Avg Rank. |
| Pretrained | 51.92 | 74.77 | 1.05 | 25.90 | 3.27 | 12.65 | 9.77 | 40.30 | 19.05 | 20.94 | – | – |
| Individual (oracle) | 60.25 | 88.63 | 5.43 | 30.60 | 8.07 | 18.48 | 11.17 | 41.71 | 19.23 | 24.02 | +69.62% | 4.55 |
| Weight Averaging | 56.64 | 85.63 | 4.49 | 30.71 | 7.13 | 17.70 | 10.78 | 41.26 | 20.08 | 22.28 | +55.23% | 7.65 |
| Task Arithmetic | 57.97 | 87.69 | 5.00 | 30.33 | 7.72 | 18.11 | 5.55 | 32.70 | 12.02 | 19.15 | +49.40% | 7.05 |
| TIES-Merging | 58.05 | 87.69 | 4.23 | 30.72 | 7.03 | 17.58 | 11.30 | 42.06 | 20.60 | 22.85 | +54.18% | 6.10 |
| Fisher-Merge | 60.25 | 85.65 | 2.30 | 27.87 | 4.58 | 14.59 | 12.01 | **44.06** | 21.36 | 24.04 | +27.18% | 6.25 |
| RegMean | **60.80** | **88.21** | 5.51 | **31.19** | 8.21 | 18.58 | 11.39 | 41.75 | 20.25 | 23.05 | +71.48% | 3.40 |
| Iso-C | 58.05 | 86.68 | 4.52 | 29.58 | 7.12 | 17.60 | 9.78 | 39.60 | 17.93 | 23.65 | +53.47% | 7.85 |
| Iso-CTS | 57.82 | 86.74 | 4.54 | 29.73 | 7.20 | 17.77 | 11.97 | 42.84 | 20.28 | **25.26** | +59.11% | 5.40 |
| TSV-Merge | 56.64 | 86.72 | 5.12 | 30.61 | 7.71 | 18.10 | 12.55 | 43.84 | **21.84** | 24.40 | +67.81% | 4.35 |
| **GRAM** (ours) | 58.52 | 87.60 | **5.62** | 30.70 | **8.30** | **18.68** | **12.67** | 43.68 | 21.49 | 25.01 | **+75.54%** | **2.40** |

### 5.3 Vision

Table 2 reports per-task accuracy on fully fine-tuned CLIP ViT-B/32 and ViT-L/14. GRAM attains the highest average accuracy on both architectures (88.7% B/32, 94.0% L/14), surpassing Iso-CTS (85.8% / 93.9%) and TSV-Merge (85.0% / 92.8%). The pairwise projection-error matrix $\sum_s \sigma_{i,s}^2 \sin^2 \theta_s$ on the 8 vision tasks (App. D.4) shows a clear block-diagonal structure under $K{=}2$, with high-chordal-distance pairs routed into different groups. The accuracy gains track this structure: GRAM is strongest on GTSRB (94.6 / 98.3) and EuroSAT (99.1 / 99.3), which sit furthest from the visual-recognition cluster (DTD/Cars/RESISC45/SUN397) and pay the largest off-diagonal penalty under $K{=}1$. The vision gap to the next-best baseline ($+2.9$ on B/32) exceeds the NLU gap ($+2.1$ on Flan-T5-Base), suggesting vision tasks present more heterogeneous subspaces where partitioning offers more benefit.

### 5.4 Medical Instruction Tuning

We set up a case study of the medical scenario from section 1: four publicly available LoRA experts on Qwen2-7B-Instruct, fine-tuned for distinct medical sub-domains across two task families and two languages: MedQA (English MCQA), CMExam (Chinese MCQA), HealthcareMagic (English doctor-patient dialogue), and cMedQA2 (Chinese doctor-patient dialogue). MCQA tasks are scored by accuracy; open-ended tasks by BLEU-4 and ROUGE-1/2/L, yielding 10 heterogeneous metric columns. Like the data-free baselines (Weight Averaging, Task Arithmetic, TIES-Merging, TSV-Merge, Iso-C, Iso-CTS), GRAM requires no calibration data. We additionally include Fisher-Merge [Matena and Raffel, 2022] and RegMean [Jin et al., 2022] as data-using references that consume a standard 64-sample-per-task calibration set.

Table 3 reports per-metric results. With experts spanning two languages and two task families (MCQA / dialogue), plain averaging visibly interferes across the gap—Task Arithmetic regresses on cMedQA2 BLEU-4 (5.55 vs Pretrained 9.77) and on its ROUGE columns. GRAM's $K{=}2$ coalitions route subspace-distant pairs into separate groups and attain the best Avg Impr ($+75.54\%$) and Avg Rank

Table 5: Clustering-algorithm ablation (Flan-T5-Base / GLUE): four ways of producing the $K=2$ partition from the *same* subspace geometry, plus the random control, under identical merging and evaluation.

| Partitioning of the same subspace geometry | $K=2$ partition | Avg. |
|---|---|---|
| GRAM | {CoLA, MRPC, QNLI, QQP, RTE, SST-2} \| {MNLI, STS-B} | **83.8** |
| Agglomerative (average linkage) | identical partition | 83.8 |
| Max-modularity communities (exact, all 127 bipartitions) | {CoLA, MNLI, SST-2, STS-B} \| {MRPC, QNLI, QQP, RTE} | 82.9 |
| Random equipartition | random $4/4$ | 82.4 |
| No grouping ($K=1$, re-merged) | all tasks | 81.9 |

(2.40) of any merging method, ahead of even the data-using RegMean ($+71.48\%$, 3.40) with its 64-sample-per-task calibration set. Gains concentrate on the open-ended generation columns (both BLEU-4 fields; HealthcareMagic ROUGE-2 / ROUGE-L), where preserving task-specific singular directions outweighs calibrated reweighting; on the two MCQA accuracy columns—the regime where calibration-based methods are strongest—GRAM trails the leader by under one point. The summary-metric edge also extends past the per-task Individual oracle ($+69.62\%$, 4.55), consistent with recent findings that LLM merging can match or exceed fine-tuned baselines on this benchmark family [Fu et al., 2025].

### 5.5 Ablation Studies

**Effects of each module.** Table 4 evaluates the contribution of each GRAM component to overall performance. Starting from full GRAM, we remove one component at a time and report the benchmark average for both PEFT merging (Flan-T5-Base on GLUE) and full-model merging (CLIP ViT-B/32 on the 8-task vision benchmark). Removing **PA-Sim** replaces the chordal subspace metric with an alternative that ignores cross-task geometry, breaking the alignment-penalty signal that drives the partition. Removing **Coalition** collapses GRAM to $K=1$ and recovers the global TSV-Merge baseline. Removing **Adapt-Rank** substitutes a uniform $r/|G_k|$ budget for the adaptive allocator, wasting capacity on sharp-spectrum tasks and starving long-tail ones. Each component is crucial, but their relative importance flips across modalities: on GLUE, PA-Sim drives the largest drop ($-3.0$), followed by Coalition ($-2.1$) and Adapt-Rank ($-1.6$); on ViT-B/32, Coalition dominates ($-3.4$) while PA-Sim ($-0.5$) and Adapt-Rank ($-0.1$) are small — the heterogeneous LoRA spectra make adaptive allocation vital for PEFT but largely cosmetic for the more uniform full-fine-tuned vision deltas. With all components combined, GRAM boosts TSV-Merge by $+2.1$ on GLUE and $+3.8$ on ViT-B/32, demonstrating the complementarity of these components.

Table 4: Effects of the proposed modules. Subscripts (red) report the change relative to full GRAM.

| Model | GLUE | ViT-B/32 |
|---|---|---|
| **GRAM** | **83.8** | **88.8** |
| − PA-Sim | $80.8_{-3.0}$ | $88.3_{-0.5}$ |
| − Coalition | $81.7_{-2.1}$ | $85.4_{-3.4}$ |
| − Adapt-Rank | $82.2_{-1.6}$ | $88.7_{-0.1}$ |

**Sensitivity to the number of groups $K$.** Table 6 evaluates GRAM for every $K$ from 1 to $T=8$ on Flan-T5-Base and Flan-T5-Large under one unified protocol (caption). At both scales the cost–accuracy frontier has a sharp knee at $K=2$: moving from Weight Averaging—the simplest single-model baseline—to GRAM at $K=2$ gains $+5.6$ (Base) and $+2.4$ (Large) points, capturing $87\%$ and $73\%$ of the respective gaps to the no-merging oracle, for a single extra rank-16 adapter ($0.7\%$ / $0.6\%$ of backbone parameters). Raising $K$ from 2 all the way to 7 adds at most $+0.1$ (Base) and $+0.6$ (Large) on top—on Base the average is not even monotone, dipping at $K=4$. Most of the achievable gain is therefore already available at small $K$: GRAM's geometry-aware clustering identifies compatible coalitions efficiently enough that a large $K$ is not needed to capture most of the gains in this setting. The practical rule is budget-first—deploy the smallest $K$ the serving budget allows; $K=2$, our default, captures the bulk of the gain at the tightest possible budget.

**Robustness of the partition.** Table 5 compares four ways of producing the $K=2$ partition from the *same* subspace geometry. The partition is stable: it is identical across 10 initialization seeds at both $K=2$ and $K=3$ ($10/10$), and an independent hierarchy-based algorithm (agglomerative, average linkage) reproduces GRAM's partitions exactly at $K \in \{2, 3, 5, 6, 7\}$, with the exact max-modularity communities coinciding for every $K \geq 3$. The partition structure is thus a property of the geometry,

not of a particular clustering algorithm—the gain comes from the subspace signal itself. The one disagreement is informative: at $K{=}2$ the degree-corrected modularity objective prefers a $4/4$ split that pulls MNLI into the low-similarity cluster and scores $0.9$ lower ($82.9$), whereas the solution used by GRAM respects the raw block structure of the similarity matrix and keeps the stronger $6/2$ partition.

**Deployment without oracle task identity.** Task-aware inference is a preference, not a requirement: a lightweight Mahalanobis task mapper [Lee et al., 2018], as deployed in SKILL [Ge et al., 2023]—class-conditional Gaussians in the frozen base encoder's feature space, $5$ training samples per class, no gradient training, no test-time task labels—routes $100.0\%$ of validation examples to the correct group on Flan-T5-Base / GLUE at $K{=}2$, so task-agnostic GRAM *equals* oracle GRAM ($83.8$ on every task); group-routing accuracy stays $\geq 99.98\%$ for every $K$. Its only errors (QQP$\leftrightarrow$MRPC, $0.18\%$ of QQP) are absorbed by the coalition structure, since GRAM groups together precisely the similar tasks a router confuses (section D.2).

**Scalability of the grouping step.** A natural concern is whether grouping becomes a bottleneck when the number of tasks is large. Figure 3 measures the end-to-end cost of grouping on a single RTX 3080 as $T$ grows from 8 to 1024. No benchmark offers thousands of adapters, so beyond our real tasks we use perturbed copies of them; grouping cost depends only on adapter shapes, not on their values, so these copies are a faithful proxy. Grouping is cheap: the real 8-task setting takes about one second, and 1024 tasks finish within a few minutes. Both time and memory grow roughly linearly with $T$, because decomposing each adapter dominates the cost and the pipeline processes one layer at a time. Llama-3-8B appears cheaper than Flan-T5 only because its LoRA factors are used directly rather than dense per-layer deltas.

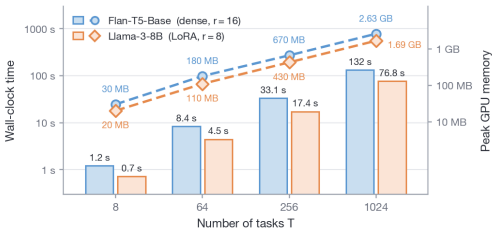



Figure 3: Grouping cost vs. number of tasks $T$ on one RTX 3080 (bars: wall-clock time; lines: peak GPU memory). $T{=}8$ is the real task set; larger $T$ uses synthetic tasks with the same layer shapes and rank.

## 6 Conclusion

GRAM partitions tasks by chordal subspace distance via Grassmannian $k$-means++ and merges within each group with a standard SVD-based procedure, avoiding the projection loss of global merging on misaligned subspaces. On GLUE with Flan-T5-Base/Large, GRAM ($K{=}2$) reaches 83.8%/88.7% average, matching the strongest baseline at the cost of one extra deployed model. On 8-task vision (CLIP ViT-B/32, ViT-L/14), GRAM reaches 88.7%/94.0%, the highest among compared methods. On the medical benchmark (Qwen2-7B-Instruct with four LoRA experts), GRAM attains the best Avg Rank without using any per-task calibration data. Gains concentrate on tasks whose subspaces are most dissimilar from the majority—precisely the tasks that the alignment penalty of Theorem B.3 predicts will benefit most from grouping.

**Limitations.** Our results have several scope limits: (i) single-seed reporting, so small Flan-T5-Large margins should be read with caution; (ii) one model family per modality (Flan-T5 for NLU, CLIP for vision, Qwen2 for medical), and the medical evaluation is a case study on public benchmarks rather than a real federated experiment from multiple sites; (iii) the theory analyzes idealized objects (optimal basis, Grassmannian $k$-means objective) rather than GRAM's exact column-stacked merge with adaptive allocation, so it should be read as geometric motivation rather than end-to-end optimality guarantee.

**Future work.** Scaling the task-agnostic router of section D.2 beyond small task collections—where group routing approaches full task identification [Ge et al., 2023]—and extending the medical evaluation to a real multi-site federated setup would broaden the deployment story.

## References


Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439, 2019.

David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, volume 7, pages 1027–1035, 2007.

Zhikang Chen, Sen Cui, Deheng Ye, Min Zhang, Gang Niu, Yu Zhang, Masashi Sugiyama, and Tingting Zhu. From coefficients to directions: Rethinking model merging with directional alignment. *arXiv preprint arXiv:2512.00391*, 2025a.

Zhikang Chen, Abudukelimu Wuerkaixi, Sen Cui, Haoxuan Li, Ding Li, Jingfeng Zhang, Bo Han, Gang Niu, Houfang Liu, Yi Yang, et al. Learning without isolation: Pathway protection for continual learning. *arXiv preprint arXiv:2505.18568*, 2025b.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34: 27503–27516, 2021.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. Training-free llm merging for multi-task learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33111–33124, 2025.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025.

Yunhao Ge, Yuecheng Li, Di Wu, Ao Xu, Adam M. Jones, Amanda Sofie Rios, Iordanis Fostiropoulos, Shixian Wen, Po-Hsuan Huang, Zachary William Murdock, Gozde Sahin, Shuo Ni, Kiran Lekkala, Sumedh Anand Sontakke, and Laurent Itti. Lightweight learner for shared knowledge lifelong learning. *Transactions on Machine Learning Research*, 2023.

Gene H Golub and Charles F Van Loan. Matrix computations johns hopkins university press. *Baltimore and London*, 41(62):2, 1996.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Scalable multitask learning using gradient-based estimation of task affinity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1542–1553, 2024.

Shikun Liu, Stephen James, Andrew J Davison, and Edward Johns. Auto-lambda: Disentangling dynamic task relationships. *arXiv preprint arXiv:2202.03091*, 2022.

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. *arXiv preprint arXiv:2502.04959*, 2025.

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

Xiaozhuang Song, Shun Zheng, Wei Cao, James Yu, and Jiang Bian. Efficient and effective multi-task grouping via meta learning on task combinations. *Advances in Neural Information Processing Systems*, 35:37647–37659, 2022.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. *arXiv preprint arXiv:2410.19735*, 2024.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangli-ao Geng, and Boyang Li. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. *arXiv preprint arXiv:2505.23859*, 2025.

Anke Tang, Li Shen, Yong Luo, Liang Ding, Han Hu, Bo Du, and Dacheng Tao. Concrete subspace learning based interference elimination for multi-task model fusion. *arXiv preprint arXiv:2312.06173*, 2023.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. FusionBench: A unified library and comprehensive benchmark for deep model fusion. *Journal of Machine Learning Research*, 26(307):1–38, 2025. URL `https://www.jmlr.org/papers/v26/25-1243.html`.

Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched multi-task networks: deciding what layers to share. *arXiv preprint arXiv:1904.02920*, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355, 2018.

Chenguang Wang, Xuanhao Pan, and Tianshu Yu. Towards principled task grouping for multi-task learning. *arXiv preprint arXiv:2402.15328*, 2024a.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. Lines: Post-training layer scaling prevents forgetting and enhances model merging. *arXiv preprint arXiv:2410.17146*, 2024b.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. *arXiv preprint arXiv:2501.01230*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Danni Yang, Zhikang Chen, Sen Cui, Mengyue Yang, Ding Li, Abudukelimu Wuerkaixi, Haoxuan Li, Jinke Ren, and Mingming Gong. Decentralized dynamic cooperation of personalized models for federated continual learning. *Advances in Neural Information Processing Systems*, 38:111897–111926, 2026.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*, 2023.

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 *Forty-first International Conference on Machine Learning*, 2024.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

# A Broader Impact

GRAM provides a way to combine independently fine-tuned task vectors into a single deployable model without retraining or pooling raw training data, and the geometry-aware grouping step is what makes this scale beyond a handful of tasks. We highlight two settings where this property carries concrete societal value.

**Healthcare and other privacy-constrained domains.** Clinical AI is trained on data that cannot legally or ethically be centralized: electronic health records, radiology images, pathology slides, and clinical-note adapters live behind institutional firewalls and HIPAA/GDPR boundaries. Today, a hospital that fine-tunes a foundation model on its own patient cohort cannot easily benefit from another hospital's adapter for a related task (a different imaging modality, a different patient population, or a different downstream label) without either sharing the underlying data or running an expensive joint training run. GRAM operates entirely on *released task vectors*—the parameter deltas, never the data—and merges them by clustering subspaces that are geometrically compatible. This lets a coalition of institutions pool model expertise across complementary specialties (e.g., radiology + pathology + clinical-note triage) while keeping each institution's data on-premises, and without forcing dissimilar tasks into a single averaged model where high-stakes minority tasks would otherwise be drowned out by majority gradients. The same property applies to other privacy-sensitive verticals: legal, financial, and on-device personalization.

**Decentralized and community-driven model merging.** The open-model ecosystem now hosts hundreds of thousands of community-released LoRA adapters and full fine-tunes per base model, but practitioners have no principled way to combine more than a few at a time. Naive averaging across many adapters destroys task-specific structure; per-adapter routing keeps a linear deployment cost. GRAM's coalition structure is a middle path: a small number $K$ of merged checkpoints that respect subspace geometry, with the compression budget $K$ left as an explicit knob. This matches the way decentralized model development actually proceeds—asynchronous, uncoordinated, contributors who never share data—and gives downstream users a way to consume many adapters as a few well-chosen merged models, lowering the deployment cost of community fine-tuning without forcing centralization.

**Risks and limitations.** Two risks are worth flagging. First, merging task vectors does not erase the biases or harmful behaviors of any constituent model; coalitions of biased adapters can yield biased merged checkpoints, and the chordal-distance grouping is geometric, not semantic, so it offers no protection against value misalignment in the inputs. Practitioners should audit each merged coalition with the same fairness and safety scrutiny they would apply to a single fine-tune. Second, GRAM lowers the cost of redistributing capabilities encoded in fine-tuned adapters, including capabilities the original fine-tuner did not intend to release. Standard adapter-licensing and provenance practices—adapter cards, licensing terms, watermarking—apply unchanged, and we encourage their use when GRAM is deployed in shared settings.

# B Deferred Theoretical Analysis

We collect the geometric and complexity results that motivate GRAM but are deferred from the main text for space. Theorem B.1 below is the headline guarantee that grouping reduces total projection error (referenced from section 5.5); the results that follow extend the picture to the GRAM column-stacked basis, the partition-level error chain that ties the chordal clustering rule to a $k$-means++ approximation guarantee, the computational cost, and a Rademacher-complexity bound. Equation (8) in section 3 already establishes that, for any orthonormal $B \in \mathbb{R}^{d_{\text{out}} \times r}$, $\left\|(I - BB^\top)U_i\right\|_{\text{F}}^2 = \sum_{j=1}^{r} \sin^2 \theta_j$ — the squared chordal distance between $\text{col}(U_i)$ and $\text{col}(B)$. All proofs are in section C.

**Theorem B.1** (Grouping reduces total error)**.** *Let $T$ tasks be partitioned into $K$ groups $\{G_1, \ldots, G_K\}$. For each group $G_k$, let $B^{(k)} \in \mathbb{R}^{d_{\text{out}} \times r}$ be the optimal rank-$r$ basis minimizing $\sum_{i \in G_k} \left\|(I - BB^\top)\Delta W_i\right\|_{\text{F}}^2$ over orthonormal $B$, and let $B^* \in \mathbb{R}^{d_{\text{out}} \times r}$ be the analogous*

*global optimum for all $T$ tasks. Then*

$$\sum_{k=1}^{K}\sum_{i\in G_k}\left\|(I-B^{(k)}B^{(k)\top})\Delta W_i\right\|_{\mathrm{F}}^2 \;\le\; \sum_{i=1}^{T}\left\|(I-B^{*}B^{*\top})\Delta W_i\right\|_{\mathrm{F}}^2, \tag{13}$$

*with strict inequality whenever some per-group optimal basis $B^{(k)}$ improves on the global optimum $B^*$ within group $G_k$. The proof (section C.1) is a direct optimality argument and does not depend on the specific clustering rule.*

**Theorem B.2** (Direct projection-error bound for the GRAM per-group basis)**.** *Let $G_k = \{1,\dots,m\}$ be a coalition with rank allocation $(k_1,\dots,k_m)$ satisfying $0 \le k_i \le r$ and $\sum_{i=1}^{m} k_i \le r$ (the budget enforced by the GRAM allocator of section 4.3), and let*

$$M^{(k)} \;:=\; \left[\,U_1^{(k_1)} \;\middle|\; \cdots \;\middle|\; U_m^{(k_m)}\,\right] \;\in\; \mathbb{R}^{d_{\mathrm{out}}\times R}, \qquad R=\sum_i k_i \le r,$$

*be the column-wise concatenation of the per-task top-$k_i$ left singular vectors of $\Delta W_i$, with the convention that $U_i^{(0)}$ is an empty matrix (no contribution). Let $\widehat{U}^{(k)} \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ be any orthonormal basis whose column space contains $\mathrm{col}(M^{(k)})$ (e.g. the GRAM merge basis, padded if $R<r$). Then for every task $i \in G_k$,*

$$\left\|(I-\widehat{U}^{(k)}\widehat{U}^{(k)\top})\,\Delta W_i\right\|_F^2 \;\le\; \sum_{\ell=k_i+1}^{r}\sigma_{i,\ell}^2, \tag{14}$$

*where the right-hand side reduces to $\|\Delta W_i\|_{\mathrm{F}}^2$ when $k_i = 0$.*

### B.1 Grassmannian $k$-Means++ Error Bound

**Lemma B.3** (Alignment penalty for two equally weighted subspaces)**.** *Let $U, V \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ have orthonormal columns and principal angles $0 \le \theta_1 \le \cdots \le \theta_r \le \frac{\pi}{2}$. Fix $\sigma > 0$. Then for every matrix $B \in \mathbb{R}^{d_{\mathrm{out}}\times r}$ with $B^\top B = I_r$,*

$$\left\|(I-BB^\top)\sigma U\right\|_F^2 + \left\|(I-BB^\top)\sigma V\right\|_F^2 \;\ge\; \sigma^2\sum_{j=1}^{r}\big(1-\cos\theta_j\big), \tag{15}$$

*with equality when the columns of $B$ are the top-$r$ left singular vectors of $[\,U \mid V\,]$.*

**Deterministic latent-group model.** There are $G \ge 2$ latent groups with prototype subspaces $Q_1,\dots,Q_G \in \mathbb{R}^{d_{\mathrm{out}}\times r}$, each with orthonormal columns, satisfying $\theta_j(Q_g,Q_h) \ge \phi_{\min} > 0$ for all $g \ne h$ and $j = 1,\dots,r$. There are $T = mG$ tasks with $m \ge 2$ per group; task $i$ in group $g(i)$ has $\Sigma_i = \sigma I_r$ and $d_c^2(U_i, Q_{g(i)}) \le \epsilon$. For a partition $\pi = \{G_1,\dots,G_K\}$, define the *optimal-basis merge error*

$$\mathrm{Err}(\pi) \;:=\; \sum_{k=1}^{K}\min_{B\in\mathbb{R}^{d_{\mathrm{out}}\times r},\, B^\top B=I_r}\sum_{i\in G_k}\left\|(I-BB^\top)\Delta W_i\right\|_F^2. \tag{16}$$

Under $\Sigma_i = \sigma I_r$, the optimal per-group basis is the chordal Fréchet mean on $\mathrm{Gr}(r, d_{\mathrm{out}})$ (Theorem C.5), and $\mathrm{Err}(\pi) = \sigma^2\Phi(\pi)$ with the Grassmannian $k$-means cost $\Phi(\pi) := \sum_k \min_{C\in\mathrm{Gr}(r,d_{\mathrm{out}})}\sum_{i\in G_k} d_c^2(U_i, C)$. The bounds below compare partitions through the projection error of the *optimal* per-group basis ($\mathrm{Err}(\pi)$); they should be read as guarantees on the partition objective and how grouping shapes it, not as guarantees on GRAM's actual column-stacked basis. Theorem B.2 provides the algorithmic-side bound on the GRAM basis itself, and section C.5 discusses when the column-stacked basis coincides with the Fréchet mean and when it merely upper-bounds it. Let $\pi_{++}$ denote the partition from $k$-means++ on $\{P_i = U_iU_i^\top\}$, $\pi_{\mathrm{rand}}$ a uniformly random equipartition into $G$ groups of size $m$, $\pi^* := \arg\min_\pi \Phi(\pi)$, and $\pi_{\mathrm{global}}$ the trivial single-block partition ($K = 1$).

**Theorem B.4** ($k$-means++ partition error bounds)**.** *Under the deterministic latent-group model with $K = G$:*

*1.* ($k$-means++ upper bound on the optimal-basis error.)

$$\mathbb{E}\big[\mathrm{Err}(\pi_{++})\big] \;\le\; 8(\ln K+2)\cdot\mathrm{Err}(\pi^*) \;\le\; 8(\ln K+2)\,\sigma^2 T\epsilon. \tag{17}$$

*2.* (Random partition lower bound, valid for $\epsilon \le r(1-\cos\phi_{\min})/4$.)

$$\mathbb{E}\big[\mathrm{Err}(\pi_{\mathrm{rand}})\big] \;\ge\; G\left(1-\frac{G}{\binom{T}{m}}\right)\sigma^2\big[r(1-\cos\phi_{\min})-4\sqrt{r\epsilon}\,\big]. \tag{18}$$

*3.* (Global merge lower bound, valid for $\epsilon \le r(1-\cos\phi_{\min})/4$.)

$$\mathrm{Err}(\pi_{\mathrm{global}}) \;\ge\; \lfloor T/2\rfloor\,\sigma^2\big[r(1-\cos\phi_{\min})-4\sqrt{r\epsilon}\,\big]. \tag{19}$$

**Corollary B.5** (Projection error chain)**.** *Under the deterministic latent-group model with $K = G$, in the regime $\epsilon \le c\,r(1-\cos\phi_{\min})^2/(m\ln K)$ for a small absolute constant $c > 0$,*

$$\mathbb{E}\big[\mathrm{Err}(\pi_{++})\big] \;\le\; \mathbb{E}\big[\mathrm{Err}(\pi_{\mathrm{rand}})\big] \;\le\; \mathrm{Err}(\pi_{\mathrm{global}}). \tag{20}$$

*The right inequality is deterministic and follows from groupwise optimality (Theorem B.1); the left inequality follows from comparing parts (1) and (2) of Theorem B.4, with the regime chosen so that the random lower bound exceeds the GRAM upper bound.*

Theorem B.5 captures the headline message: refining the trivial single-block merge into $K$ groups via $k$-means++ on subspace similarities is never worse than a random equipartition (in expectation, in the small-$\epsilon$ regime), and an arbitrary equipartition is in turn never worse than not grouping at all (deterministically). A non-uniform singular-value extension is in section C.9.

### B.2 Computational Complexity

Let $d = \max(d_{\mathrm{out}}, d_{\mathrm{in}})$. The GRAM pipeline incurs, once per model:

- *Per-task SVD:* For LoRA deltas $\Delta W_i = B_i A_i$ (Flan-T5/GLUE, Qwen medical), thin QR of $B_i$ and $A_i^\top$ followed by an $r \times r$ SVD costs $O((d_{\mathrm{out}} + d_{\mathrm{in}})r^2 + r^3)$ per task per layer. For full-rank deltas (vision), a top-$r$ truncated SVD costs $O(d_{\mathrm{out}} d_{\mathrm{in}} r)$ per task per layer.
- *Pairwise chordal distance:* $d_c^2(U_i, U_j) = r - \big\|U_i^\top U_j\big\|_{\mathrm{F}}^2$ is $O(d_{\mathrm{out}} r^2)$ per pair; the full $T \times T$ matrix is $O(T^2 d_{\mathrm{out}} r^2)$ — no need to materialize the projection matrices $P_i$.
- *Clustering:* $k$-means++ on the precomputed pairwise-distance matrix; negligible when $T$ is small.
- *Per-group merge:* $O(KLd_{\mathrm{out}}r^2)$ for the polar-factor orthogonalization plus $O(|G_k|r\log r)$ per layer for rank allocation.

The dominant term is $O(TL(d_{\mathrm{out}} + d_{\mathrm{in}})r^2)$ for LoRA deltas and $O(TLd_{\mathrm{out}}d_{\mathrm{in}}r)$ for full-rank deltas, both strictly cheaper than one epoch of fine-tuning ($O(n_{\text{train}}Ld^2)$ per task). At inference, each group's rank-$r$ merged update costs $O(dr)$ per layer—identical to individual or global merge on a per-query basis. GRAM stores $K$ merged updates instead of $T$ per-task models, a factor $T/K$ reduction in storage with no additional per-query FLOPs.

### B.3 Generalization via Rademacher Complexity

**Definition 1** (Linear-update hypothesis class)**.** *Fix $W_0 \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, $a \in \mathbb{R}^{d_{\mathrm{out}}}$ with $\|a\|_2 = 1$, and $\beta > 0$. Let $\mathcal{H}_a(\beta) = \big\{x \mapsto a^\top(W_0 + \Delta W)\,x : \|\Delta W\|_{\mathrm{F}} \le \beta\big\}$.*

**Theorem B.6** (Rademacher bound for linear updates)**.** *Let $\{x_1, \dots, x_n\}$ be an i.i.d. sample with $\|x_i\|_2 \le R$. Then*

$$\hat{\mathfrak{R}}_n\big(\mathcal{H}_a(\beta)\big) \;\le\; \beta R/\sqrt{n}. \tag{21}$$

**Corollary B.7** (Per-group capacity is no larger than the global class)**.** *If the GRAM per-group merged updates $\{\widetilde{\Delta W}^{(k)}\}_{k=1}^K$ all satisfy $\left\|\widetilde{\Delta W}^{(k)}\right\|_{\mathrm{F}} \le \beta$, each per-group update belongs to $\mathcal{H}_a(\beta)$, so the per-group Rademacher complexity is bounded by $\beta R/\sqrt{n}$, matching the global class.*

## C Proof Details

### C.1 Proof of Theorem B.1

*Proof of Theorem B.1.* For the global merge, let $B^* \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ be the optimal rank-$r$ basis minimizing $\sum_{i=1}^T \big\|(I - B^* B^{*\top})\Delta W_i\big\|_{\mathrm{F}}^2$ over all tasks.

For a partition $\{G_1, \ldots, G_K\}$, let $B^{(k)}$ be the optimal rank-$r$ basis for group $G_k$, i.e.,

$$B^{(k)} = \underset{B \in \mathbb{R}^{d_{\text{out}} \times r},\, B^\top B = I}{\arg\min} \sum_{i \in G_k} \left\| (I - BB^\top) \Delta W_i \right\|_{\text{F}}^2 . \tag{22}$$

By optimality of $B^{(k)}$ within group $G_k$, for any orthonormal $B \in \mathbb{R}^{d_{\text{out}} \times r}$:

$$\sum_{i \in G_k} \left\| (I - B^{(k)} B^{(k)\top}) \Delta W_i \right\|_{\text{F}}^2 \leq \sum_{i \in G_k} \left\| (I - BB^\top) \Delta W_i \right\|_{\text{F}}^2 . \tag{23}$$

In particular, choosing $B = B^*$ (the global optimum) for each group:

$$\sum_{k=1}^{K} \sum_{i \in G_k} \left\| (I - B^{(k)} B^{(k)\top}) \Delta W_i \right\|_{\text{F}}^2 \leq \sum_{k=1}^{K} \sum_{i \in G_k} \left\| (I - B^* B^{*\top}) \Delta W_i \right\|_{\text{F}}^2 = \sum_{i=1}^{T} \left\| (I - B^* B^{*\top}) \Delta W_i \right\|_{\text{F}}^2 , \tag{24}$$

which establishes that group-wise merging with per-group optimal bases always achieves error no greater than global merging.

Strictness of the inequality requires a uniqueness/non-degeneracy assumption: if some $B^{(k)}$ strictly improves on $B^*$ within group $G_k$ (which generically holds when the within-group leading singular subspace differs from the global one), the inequality is strict; otherwise it can be tight. □

### C.2 Proof of Theorem B.2

We restate the theorem for convenience and then prove it via three elementary lemmas.

**Lemma C.1** (Pythagorean identity for orthogonal projectors)**.** *For any orthogonal projector $P \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ ($P^2 = P = P^\top$) and any matrix $A \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $\|(I - P)A\|_{\text{F}}^2 = \|A\|_{\text{F}}^2 - \|PA\|_{\text{F}}^2$.*

*Proof.* $\|A\|_{\text{F}}^2 = \|PA + (I - P)A\|_{\text{F}}^2 = \|PA\|_{\text{F}}^2 + 2\langle PA, (I - P)A \rangle_F + \|(I - P)A\|_{\text{F}}^2$. The cross term vanishes because $\langle PA, (I - P)A \rangle_F = \operatorname{tr}(A^\top P^\top (I - P)A) = \operatorname{tr}(A^\top P(I - P)A) = 0$, using $P = P^\top$ and $P(I - P) = 0$. □

**Lemma C.2** (Subspace monotonicity of projected energy)**.** *Let $P, Q \in \mathbb{R}^{d_{\text{out}} \times d_{\text{out}}}$ be orthogonal projectors with $\operatorname{col}(Q) \subseteq \operatorname{col}(P)$. Then $P \succeq Q$ in the Loewner order, and for every $A \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, $\|PA\|_{\text{F}}^2 \geq \|QA\|_{\text{F}}^2$.*

*Proof.* For orthogonal projectors, $\operatorname{col}(Q) \subseteq \operatorname{col}(P)$ is equivalent to $PQ = Q$ (since $Q$ projects into a subspace of $\operatorname{col}(P)$ and $P$ acts as the identity there). Taking transposes and using $P = P^\top$, $Q = Q^\top$ also gives $QP = Q$, so $P - Q$ is symmetric and idempotent: $(P - Q)^2 = P^2 - PQ - QP + Q^2 = P - Q - Q + Q = P - Q$. Hence $P - Q$ is itself an orthogonal projector, and for any $x$, $x^\top (P - Q)x = \|(P - Q)x\|_2^2 = \|(I - Q)Px\|_2^2 \geq 0$, where the last equality uses $(P - Q)x = Px - Qx = Px - QPx = (I - Q)Px$. Thus $P \succeq Q$, and applying this with $x = Av$ gives $A^\top (P - Q)A \succeq 0$, so $\operatorname{tr}(A^\top PA) \geq \operatorname{tr}(A^\top QA)$, i.e. $\|PA\|_{\text{F}}^2 \geq \|QA\|_{\text{F}}^2$. □

**Lemma C.3** (Self-projection recovers the truncation energy)**.** *For any task $i$ with thin SVD $\Delta W_i = U_i \Sigma_i V_i^\top$ and any $0 \leq k_i \leq r$, letting $P_i := U_i^{(k_i)} U_i^{(k_i)\top}$ (with $P_i = 0$ when $k_i = 0$), $\|P_i \, \Delta W_i\|_{\text{F}}^2 = \sum_{\ell=1}^{k_i} \sigma_{i,\ell}^2$, where the sum is empty (and equals 0) when $k_i = 0$.*

*Proof.* The case $k_i = 0$ is immediate from $P_i = 0$. For $k_i \geq 1$, $U_i^{(k_i)\top} U_i = [\, I_{k_i} \,|\, 0_{k_i \times (r - k_i)} \,]$, so $U_i^{(k_i)\top} \Delta W_i = [\, I_{k_i} \,|\, 0 \,] \Sigma_i V_i^\top = [\, \Sigma_i^{(k_i)} \,|\, 0 \,] V_i^\top = \Sigma_i^{(k_i)} V_i^{(k_i)\top}$, where $\Sigma_i^{(k_i)} = \operatorname{diag}(\sigma_{i,1}, \ldots, \sigma_{i,k_i})$ and $V_i^{(k_i)} \in \mathbb{R}^{d_{\text{in}} \times k_i}$. Since $U_i^{(k_i)}$ has orthonormal columns, $\|P_i \Delta W_i\|_{\text{F}}^2 = \left\| U_i^{(k_i)\top} \Delta W_i \right\|_{\text{F}}^2 = \left\| \Sigma_i^{(k_i)} V_i^{(k_i)\top} \right\|_{\text{F}}^2 = \left\| \Sigma_i^{(k_i)} \right\|_{\text{F}}^2 = \sum_{\ell=1}^{k_i} \sigma_{i,\ell}^2$ (using $\left\| XY^\top \right\|_{\text{F}} = \|X\|_{\text{F}}$ when $Y$ has orthonormal columns). □

*Proof of Theorem B.2.* Let $P^{(k)} := \widehat{U}^{(k)}\widehat{U}^{(k)\top}$ be the orthogonal projector onto $\mathrm{col}(\widehat{U}^{(k)})$. By the construction of $\widehat{U}^{(k)}$, $\mathrm{col}(\widehat{U}^{(k)}) = \mathrm{col}(M^{(k)})$. (This holds in particular for the orthogonal polar factor of $M^{(k)}$ used in GRAM: if $M^{(k)} = u\,s\,v^\top$ is the SVD, then $uv^\top$ has the same column space as $u$, which equals $\mathrm{col}(M^{(k)})$.)

Since $U_i^{(k_i)}$ is one of the column blocks of $M^{(k)}$ by construction, every column of $U_i^{(k_i)}$ lies in $\mathrm{col}(M^{(k)}) = \mathrm{col}(\widehat{U}^{(k)})$, so

$$\mathrm{col}(P_i) \;=\; \mathrm{col}(U_i^{(k_i)}) \;\subseteq\; \mathrm{col}(\widehat{U}^{(k)}) \;=\; \mathrm{col}(P^{(k)}). \tag{25}$$

By Theorem C.2, $\left\|P^{(k)}\Delta W_i\right\|_{\mathrm{F}}^2 \geq \|P_i\Delta W_i\|_{\mathrm{F}}^2$, and by Theorem C.3, $\|P_i\Delta W_i\|_{\mathrm{F}}^2 = \sum_{\ell=1}^{k_i}\sigma_{i,\ell}^2$. Combining,

$$\left\|P^{(k)}\Delta W_i\right\|_{\mathrm{F}}^2 \;\geq\; \sum_{\ell=1}^{k_i}\sigma_{i,\ell}^2. \tag{26}$$

Applying Theorem C.1 to $P = P^{(k)}$ and $A = \Delta W_i$ gives $\left\|(I-P^{(k)})\Delta W_i\right\|_{\mathrm{F}}^2 = \|\Delta W_i\|_{\mathrm{F}}^2 - \left\|P^{(k)}\Delta W_i\right\|_{\mathrm{F}}^2$. Substituting the lower bound and using $\|\Delta W_i\|_{\mathrm{F}}^2 = \sum_{\ell=1}^{r}\sigma_{i,\ell}^2$,

$$\left\|(I-P^{(k)})\Delta W_i\right\|_{\mathrm{F}}^2 \;\leq\; \sum_{\ell=1}^{r}\sigma_{i,\ell}^2 - \sum_{\ell=1}^{k_i}\sigma_{i,\ell}^2 \;=\; \sum_{\ell=k_i+1}^{r}\sigma_{i,\ell}^2, \tag{27}$$

which is (14).

**Equality case ($m = 1$).** If the coalition has a single task, $M^{(k)} = U_i^{(k_i)}$ and $\mathrm{col}(P^{(k)}) = \mathrm{col}(P_i)$, so the inequality in Theorem C.2 is an equality, and the bound is attained. □

**What this proof does and does not give.** The proof bounds the projection error onto the GRAM column basis, not the reconstruction error $\|\Delta W_i - \widetilde{\Delta W}^{(k)}\|_F^2$ of the merged delta itself. The latter decomposes orthogonally as

$$\|\Delta W_i - \widetilde{\Delta W}^{(k)}\|_F^2 \;=\; \|(I-P^{(k)})\Delta W_i\|_F^2 + \|P^{(k)}\Delta W_i - \widetilde{\Delta W}^{(k)}\|_F^2, \tag{28}$$

and Theorem B.2 controls only the first term. The second term — the in-subspace fitting error of the single merged delta against task $i$'s in-subspace component — is not analyzed here and would require a separate argument tracking how $\widetilde{\Delta W}^{(k)} = \widehat{U}^{(k)}\,\mathrm{diag}(\bar{s})\,\widehat{V}^{(k)\top}$ is constructed from the per-task singular values and right-singular vectors. The bound in Theorem B.2 is also loose precisely when other tasks in the coalition contribute new directions that align with $\mathrm{col}(U_i) \setminus \mathrm{col}(U_i^{(k_i)})$ — the regime in which grouping is most useful — because the proof discards all inter-task overlap by lower-bounding $\left\|P^{(k)}\Delta W_i\right\|_{\mathrm{F}}^2$ via the smaller projector $P_i$ alone.

### C.3 GRAM vs. global merge under uniform allocation (statement and proof)

**Corollary C.4** (Grouping reduces the projection-error upper bound for every task)**.** *Fix a LoRA rank budget $\rho$ per merged adapter. Compare the global merge over all $T$ tasks under uniform allocation $k_i^{\mathrm{glob}} := \lfloor \rho/T \rfloor$ to a GRAM coalition $G_k$ of size $m < T$ under uniform allocation $k_i := \lfloor \rho/m \rfloor$. Then $k_i^{\mathrm{glob}} \leq k_i$, and for every task $i \in G_k$ the projection-error upper bounds from Theorem B.2 satisfy*

$$\underbrace{\sum_{\ell=k_i+1}^{r}\sigma_{i,\ell}^2}_{\textit{GRAM bound}} \;\leq\; \underbrace{\sum_{\ell=k_i^{\mathrm{glob}}+1}^{r}\sigma_{i,\ell}^2}_{\textit{global bound}}, \qquad \textit{improvement} \;\geq\; \sum_{\ell=k_i^{\mathrm{glob}}+1}^{k_i}\sigma_{i,\ell}^2 \;\geq\; 0. \tag{29}$$

*The improvement is strictly positive whenever any of the singular values $\sigma_{i,k_i^{\mathrm{glob}}+1}, \ldots, \sigma_{i,k_i}$ is non-zero.*

*Proof.* $T > m$ implies $\lfloor \rho/T \rfloor \le \lfloor \rho/m \rfloor$, hence $k_i^{\mathrm{glob}} \le k_i$. The tail sum $\sum_{\ell=k+1}^{r} \sigma_{i,\ell}^2$ is non-increasing in $k$, so the GRAM bound is no larger than the global bound; the difference is exactly $\sum_{\ell=k_i^{\mathrm{glob}}+1}^{k_i} \sigma_{i,\ell}^2$. Applying Theorem B.2 to both bases yields (29). □

**Remark 1** (Tightness)**.** *Theorem C.4 states a worst-case guarantee on the gap between GRAM and global merging at the level of upper bounds, not a comparison of actual projection errors; the looseness of Theorem B.2 discussed above carries through here. The multiplicative gap between alignment-aware and random grouping is captured by Theorem B.4, not this corollary.*

### C.4 Equivalence of chordal and Euclidean distances on projectors

To run $k$-means++ off the shelf we need a Euclidean embedding of the Grassmannian. Each subspace $\mathrm{col}(U_i^{(\ell)})$ has a unique *orthogonal projector* $P_i^{(\ell)} := U_i^{(\ell)} U_i^{(\ell)\top} \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{out}}}$, which is invariant to the choice of orthonormal basis for the subspace and serves as that embedding. On these projectors,

$$\left\| P_i^{(\ell)} - P_j^{(\ell)} \right\|_{\mathrm{F}}^2 \;=\; \left\| P_i^{(\ell)} \right\|_{\mathrm{F}}^2 + \left\| P_j^{(\ell)} \right\|_{\mathrm{F}}^2 - 2 \left\langle P_i^{(\ell)}, P_j^{(\ell)} \right\rangle \;=\; 2\, d_c^2\big(U_i^{(\ell)},\, U_j^{(\ell)}\big), \tag{30}$$

using $\left\| P_i^{(\ell)} \right\|_{\mathrm{F}}^2 = r$ (rank-$r$ idempotent) and $\langle P_i^{(\ell)}, P_j^{(\ell)} \rangle = \left\| U_i^{(\ell)\top} U_j^{(\ell)} \right\|_{\mathrm{F}}^2$. The chordal distance on the Grassmannian is therefore Euclidean (up to a factor of 2) on the projector embedding.

### C.5 Chordal Fréchet mean and the GRAM merge basis

**Proposition C.5** (Fréchet mean under chordal distance)**.** *Let $\{U_1, \dots, U_m\} \subset \mathbb{R}^{d_{\mathrm{out}} \times r}$ have orthonormal columns and let $M := \sum_{i=1}^m U_i U_i^\top$. Any minimizer of $\sum_{i=1}^m d_c^2(U_i, C)$ over $C \in \mathrm{Gr}(r, d_{\mathrm{out}})$ is a top-$r$ invariant subspace of $M$; the minimizer is unique whenever $\lambda_r(M) > \lambda_{r+1}(M)$, and otherwise any choice of $r$-dimensional subspace within the top eigenspace is optimal.*

*Proof.* Write $d_c^2(U_i, C) = r - \left\| C^\top U_i \right\|_{\mathrm{F}}^2 = r - \mathrm{tr}(C^\top U_i U_i^\top C)$ for any orthonormal $C \in \mathbb{R}^{d_{\mathrm{out}} \times r}$. Therefore

$$\sum_{i=1}^{m} d_c^2(U_i, C) \;=\; mr - \mathrm{tr}\big(C^\top M C\big).$$

Minimizing over orthonormal $C$ is equivalent to maximizing $\mathrm{tr}(C^\top M C)$ where $M \succeq 0$. By the Ky Fan max-trace theorem, the maximum is $\sum_{j=1}^r \lambda_j(M)$ and is achieved when $C$ spans the top-$r$ eigenspace of $M$. □

**Relation to the GRAM column-stacked basis.** Theorem B.2 concerns the column space of $M^{(k)} = [U_1^{(k_1)} \mid \cdots \mid U_m^{(k_m)}]$, which has dimension at most $R = \sum_i k_i$ (and is set to $r$ by the GRAM allocation rule of section 4.3, which exhausts the rank budget). The Fréchet-mean basis above is the top-$r$ eigenspace of $\sum_i U_i U_i^\top$, and the two coincide only when $\mathrm{col}(M^{(k)})$ equals that eigenspace; in general they differ. For example, with $m = r = 2$, $\Sigma_i = \sigma I_2$, $U_1 = U_2 = [e_1, e_2]$, and allocation $k_1 = k_2 = 1$, the selected column $e_1$ from each task gives $\mathrm{col}(M^{(k)}) = \mathrm{span}(e_1)$ while the Fréchet mean is $\mathrm{span}(e_1, e_2)$. We therefore distinguish two regimes in the analysis below: (i) the per-task projection-error *upper* bound eq. (14), which holds for *any* orthonormal basis of $\mathrm{col}(M^{(k)})$ — including the GRAM output — without invoking the Fréchet-mean identification; and (ii) the latent-group analysis ($k$-means++ gap, error chain), in which $\mathrm{Err}(\pi)$ is defined via the optimal-basis projection error and the Fréchet mean is invoked as the optimum, not as GRAM's actual output. This separation makes the theory bridge orthogonal to whether GRAM's column-stacked basis exactly matches the Fréchet mean.

### C.6 Proof of Theorem B.3

*Proof of Theorem B.3.* Let $P := BB^\top$ and $S := UU^\top + VV^\top$. Using $\|(I-P)\sigma X\|_{\mathrm{F}}^2 = \sigma^2\big(\|X\|_{\mathrm{F}}^2 - \left\|B^\top X\right\|_{\mathrm{F}}^2\big)$ with $\|U\|_{\mathrm{F}}^2 = \|V\|_{\mathrm{F}}^2 = r$ and $\left\|B^\top X\right\|_{\mathrm{F}}^2 = \mathrm{tr}(B^\top X X^\top B)$,

$$\|(I-P)\sigma U\|_{\mathrm{F}}^2 + \|(I-P)\sigma V\|_{\mathrm{F}}^2 \;=\; \sigma^2\big(2r - \mathrm{tr}(B^\top S\, B)\big). \tag{31}$$

Since $S = MM^\top$ with $M := [\,U \mid V\,] \in \mathbb{R}^{d_{\text{out}} \times 2r}$, the nonzero eigenvalues of $S$ equal the nonzero eigenvalues of

$$M^\top M \;=\; \begin{pmatrix} I_r & U^\top V \\ V^\top U & I_r \end{pmatrix}.$$

The singular values of $U^\top V$ are $\cos\theta_1, \ldots, \cos\theta_r$ (Golub–Van Loan), so $M^\top M$ is unitarily similar to $\bigoplus_{j=1}^{r} \left(\begin{smallmatrix} 1 & \cos\theta_j \\ \cos\theta_j & 1 \end{smallmatrix}\right)$ and its eigenvalues are $\{1 + \cos\theta_j,\, 1 - \cos\theta_j\}_{j=1}^{r}$. Since $\cos\theta_j \geq 0$, the top-$r$ eigenvalues of $S$ are $\{1 + \cos\theta_j\}_{j=1}^{r}$. By the Ky Fan max-trace theorem,

$$\operatorname{tr}(B^\top S\, B) \;\leq\; \sum_{j=1}^{r} (1 + \cos\theta_j), \tag{32}$$

with equality when $\operatorname{col}(B)$ is a top-$r$ eigenspace of $S$, equivalently the top-$r$ left singular subspace of $[\,U \mid V\,]$. Substituting (32) into (31) gives

$$\|(I-P)\sigma U\|_{\text{F}}^2 + \|(I-P)\sigma V\|_{\text{F}}^2 \;\geq\; \sigma^2 \Big(2r - \sum_{j=1}^{r} (1 + \cos\theta_j)\Big) \;=\; \sigma^2 \sum_{j=1}^{r} (1 - \cos\theta_j),$$

which is (15). □

### C.7 Proof of Theorem B.4

We prove each part of the theorem under the deterministic latent-group model: $G$ groups, $m$ tasks per group ($T = mG$), prototype subspaces $Q_1, \ldots, Q_G$ with pairwise principal angles $\geq \phi_{\min}$, and within-group spread $d_c^2(U_i, Q_{g(i)}) \leq \epsilon$ for all $i$.

*Proof of part (1): upper bound.* Define $P_i = U_i U_i^\top$ and recall $\|P_i - P_j\|_{\text{F}}^2 = 2\, d_c^2(U_i, U_j)$.

*Step 1 (seeding guarantee).* $k$-means++ selects $K$ seeds $s_1, \ldots, s_K$ from $\{P_i\}$ via $D^2$-sampling and assigns each $P_i$ to its nearest seed, yielding partition $\pi_{++}$. By Arthur et al. [2007],

$$\mathbb{E}\left[\sum_{k=1}^{K} \sum_{i \in G_k^{++}} \|P_i - s_k\|_{\text{F}}^2\right] \;\leq\; 8(\ln K + 2) \cdot \text{OPT}_E, \tag{33}$$

where $\text{OPT}_E := \min_{\pi, \{\mu_k\}} \sum_k \sum_{i \in G_k} \|P_i - \mu_k\|_{\text{F}}^2$.

*Step 2 (seeds are projections).* Each seed $s_k = P_{j_k}$ is a rank-$r$ projection matrix, hence a valid point on $\text{Gr}(r, d_{\text{out}})$. The chordal Fréchet mean minimizes cost over rank-$r$ projections and can only reduce cost:

$$\Phi(\pi_{++}) \;\leq\; \sum_k \sum_{i \in G_k^{++}} d_c^2(U_i, U_{j_k}) \;=\; \tfrac{1}{2} \sum_k \sum_{i \in G_k^{++}} \|P_i - s_k\|_{\text{F}}^2 .$$

*Step 3 (Euclidean $\leq$ Grassmannian).* For any partition $\pi$ and group $G_k$, the Euclidean mean $\bar{P}_k$ minimizes $\sum_i \|P_i - \mu\|_{\text{F}}^2$ over all matrices $\mu$, while the Fréchet mean $P_{C_k}$ is constrained to rank-$r$ projections. Therefore $\sum_i \|P_i - \bar{P}_k\|_{\text{F}}^2 \leq \sum_i \|P_i - P_{C_k}\|_{\text{F}}^2 = 2 \sum_i d_c^2(U_i, C_k)$. Minimizing over $C_k$ and summing: $\text{OPT}_E \leq 2\, \Phi(\pi^*)$.

*Combining.* $\mathbb{E}[\Phi(\pi_{++})] \leq \frac{1}{2} \cdot 8(\ln K + 2) \cdot 2\, \Phi(\pi^*) = 8(\ln K + 2)\, \Phi(\pi^*)$. Multiplying by $\sigma^2$ gives $\mathbb{E}[\text{Err}(\pi_{++})] \leq 8(\ln K + 2)\, \text{Err}(\pi^*)$.

*Absolute bound.* The latent partition $\pi_{\text{lat}}$ is feasible: $\Phi(\pi^*) \leq \Phi(\pi_{\text{lat}}) = \sum_{g=1}^{G} \sum_{i:g(i)=g} d_c^2(U_i, C_g^{\text{lat}}) \leq \sum_{i=1}^{T} d_c^2(U_i, Q_{g(i)}) \leq T\epsilon$, where the second inequality uses $d_c^2(U_i, C_g^{\text{lat}}) \leq d_c^2(U_i, Q_{g(i)})$ since $C_g^{\text{lat}}$ is the cost-minimizing center. □

*Proof of part (2): lower bound.* Call $G_k$ *mixed* if it contains tasks $i, j$ with $g(i) \neq g(j)$, and pick any such pair $(i, j)$ in each mixed $G_k$. Let $C_k \in \mathbb{R}^{d_{\text{out}} \times r}$ be the chordal Fréchet mean of $G_k$, which is the minimizer in (16) for $\Sigma_l = \sigma I_r$ by Theorem C.5. Theorem B.3 holds for *every* orthonormal $B$,

so applying it with $B = C_k$ to the pair $(U_i, U_j)$ and dropping the non-negative contributions from the remaining $|G_k| - 2$ tasks,

$$\sum_{l \in G_k} \big\|(I - C_k C_k^\top)\Delta W_l\big\|_F^2 \;\geq\; \sigma^2 \sum_{s=1}^{r} \big(1 - \cos\theta_s(U_i, U_j)\big). \tag{34}$$

The left side is exactly the contribution of $G_k$ to $\mathrm{Err}(\pi)$ (since $C_k$ achieves the minimum in (16)), so this is a per-mixed-group lower bound on $\mathrm{Err}(\pi)$.

*Procrustes metric and triangle inequality.* Define $\delta_p(U, V) := \big[2\sum_{s=1}^{r}\big(1 - \cos\theta_s(U, V)\big)\big]^{1/2} = \min_{R \in O(r)} \|U - VR\|_{\mathrm{F}}$, the Procrustes (chordal Fubini–Study) distance, which is a metric on $\mathrm{Gr}(r, d_{\mathrm{out}})$ [Edelman et al., 1998]; $\delta_p^2$ denotes its square. We use two facts: *(a)* since $1 - \cos t \leq \sin^2 t$ for $t \in [0, \pi/2]$, $\delta_p^2(U, V) \leq 2\sum_s \sin^2\theta_s = 2d_c^2(U, V)$; and *(b)* since $1 - \cos t \geq (1 - \cos t) \cdot 1$ trivially with $\theta_s(Q_g, Q_h) \geq \phi_{\min}$, $\delta_p^2(Q_g, Q_h) \geq 2r(1 - \cos\phi_{\min})$. Combining (a) with $d_c^2(U_i, Q_g) \leq \epsilon$ gives $\delta_p(U_i, Q_g) \leq \sqrt{2\epsilon}$, and likewise for $j$. By the triangle inequality for $\delta_p$,

$$\delta_p(U_i, U_j) \;\geq\; \delta_p(Q_g, Q_h) - \delta_p(U_i, Q_g) - \delta_p(U_j, Q_h) \;\geq\; \sqrt{2r(1 - \cos\phi_{\min})} - 2\sqrt{2\epsilon}. \tag{35}$$

The right side is non-negative iff $\epsilon \leq r(1 - \cos\phi_{\min})/4$, which is the hypothesis of part (2). Squaring and dropping the non-negative cross term,

$$\begin{aligned}
\delta_p^2(U_i, U_j) \;&\geq\; 2r(1 - \cos\phi_{\min}) - 2 \cdot 2\sqrt{2}\sqrt{r(1 - \cos\phi_{\min})} \cdot \sqrt{2\epsilon} \\
&=\; 2r(1 - \cos\phi_{\min}) - 8\sqrt{r\epsilon(1 - \cos\phi_{\min})} \\
&\geq\; 2r(1 - \cos\phi_{\min}) - 8\sqrt{r\epsilon},
\end{aligned}$$

using $1 - \cos\phi_{\min} \leq 1$. Halving, $\sum_{s=1}^{r}(1 - \cos\theta_s(U_i, U_j)) \geq r(1 - \cos\phi_{\min}) - 4\sqrt{r\epsilon}$. Substituting into (34), the per-mixed-group contribution to $\mathrm{Err}(\pi_{\mathrm{rand}})$ is at least $\sigma^2[r(1 - \cos\phi_{\min}) - 4\sqrt{r\epsilon}]$.

*Expected number of mixed groups.* Under a uniformly random equipartition into $G$ groups of size $m$, each group is a random $m$-subset; exactly $G$ of the $\binom{T}{m}$ such subsets are monochromatic (one per latent group). Hence $\mathbb{E}[N_{\mathrm{mix}}] = G\big(1 - G/\binom{T}{m}\big)$. Linearity of expectation across mixed groups gives (18). □

*Proof of part (3): global merge lower bound.* Let $B^* \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ be the global merge basis ($\pi_{\mathrm{global}}$ has a single block containing all $T = mG$ tasks). We construct $\lfloor T/2 \rfloor$ pairwise-disjoint inter-group pairs and apply Theorem B.3 to each.

*Existence of the matching.* With $G \geq 2$ latent groups of size $m$, no group contains more than $T/2$ tasks (since $m \leq T/2$ for $G \geq 2$). A standard graph-theoretic result then gives a matching of size $\lfloor T/2 \rfloor$ in the inter-group graph (the complete multipartite graph $K_{m,m,\ldots,m}$): order tasks as $\tau_1, \ldots, \tau_T$ by interleaving across groups (one task from each group, then a second from each, etc.), and pair $(\tau_{2k-1}, \tau_{2k})$ for $k = 1, \ldots, \lfloor T/2 \rfloor$; consecutive tasks in this ordering belong to different groups whenever $G \geq 2$.

*Per-pair lower bound.* For each disjoint inter-group pair $(i, j)$ with $g(i) \neq g(j)$, applying Theorem B.3 with $B = B^*$ and the Procrustes argument from the proof of part (2),

$$\big\|(I - B^*B^{*\top})\Delta W_i\big\|_{\mathrm{F}}^2 + \big\|(I - B^*B^{*\top})\Delta W_j\big\|_{\mathrm{F}}^2 \;\geq\; \sigma^2\big[r(1 - \cos\phi_{\min}) - 4\sqrt{r\epsilon}\big], \tag{36}$$

valid for $\epsilon \leq r(1 - \cos\phi_{\min})/4$.

*Summation.* Summing (36) over the $\lfloor T/2 \rfloor$ disjoint pairs and adding the non-negative contribution of any unmatched task (when $T$ is odd),

$$\mathrm{Err}(\pi_{\mathrm{global}}) \;=\; \sum_{i=1}^{T} \big\|(I - B^*B^{*\top})\Delta W_i\big\|_{\mathrm{F}}^2 \;\geq\; \lfloor T/2 \rfloor\, \sigma^2\big[r(1 - \cos\phi_{\min}) - 4\sqrt{r\epsilon}\big],$$

which is (19). □

### C.8 Proof of Theorem B.5

*Proof of Theorem B.5. Right inequality (deterministic).* For any partition $\pi$ with groups $\{G_1, \dots, G_K\}$, let $B_\pi^{(k)}$ minimize $\sum_{i \in G_k} \left\|(I - BB^\top)\Delta W_i\right\|_{\mathrm{F}}^2$ over orthonormal $B$, so $\mathrm{Err}(\pi) = \sum_k \sum_{i \in G_k} \left\|(I - B_\pi^{(k)} B_\pi^{(k)\top})\Delta W_i\right\|_{\mathrm{F}}^2$. Let $B^*$ be the analogous global minimizer for $\pi_{\mathrm{global}}$ (a single block of all $T$ tasks). Substituting $B = B^*$ as a feasible (sub-optimal) basis inside each group $G_k$,

$$\begin{aligned}\mathrm{Err}(\pi) \;=\; \sum_{k=1}^{K} \sum_{i \in G_k} \left\|(I - B_\pi^{(k)} B_\pi^{(k)\top})\Delta W_i\right\|_{\mathrm{F}}^2 \;\le\; \sum_{k=1}^{K} \sum_{i \in G_k} \left\|(I - B^* B^{*\top})\Delta W_i\right\|_{\mathrm{F}}^2 \\ = \sum_{i=1}^{T} \left\|(I - B^* B^{*\top})\Delta W_i\right\|_{\mathrm{F}}^2 \;=\; \mathrm{Err}(\pi_{\mathrm{global}}).\end{aligned}$$

This is exactly Theorem B.1; it is deterministic and holds in particular for $\pi = \pi_{\mathrm{rand}}$, hence $\mathbb{E}[\mathrm{Err}(\pi_{\mathrm{rand}})] \le \mathrm{Err}(\pi_{\mathrm{global}})$.

*Left inequality (in expectation, in the small-$\epsilon$ regime).* Let $\alpha := 1 - \cos\phi_{\min} \in (0, 2]$. From Theorem B.4 parts (1) and (2),

$$\mathbb{E}[\mathrm{Err}(\pi_{++})] \le 8(\ln K + 2)\,\sigma^2 T\epsilon, \qquad \mathbb{E}[\mathrm{Err}(\pi_{\mathrm{rand}})] \ge G\big(1 - \tfrac{G}{\binom{T}{m}}\big)\sigma^2[r\alpha - 4\sqrt{r\epsilon}],$$

the latter valid for $\epsilon \le r\alpha/4$. Sufficient for the left inequality is $\mathbb{E}[\mathrm{Err}(\pi_{\mathrm{rand}})] \ge \mathbb{E}[\mathrm{Err}(\pi_{++})]$, i.e.

$$G\big(1 - \tfrac{G}{\binom{T}{m}}\big)[r\alpha - 4\sqrt{r\epsilon}] \;\ge\; 8(\ln K + 2)\,T\epsilon.$$

Using $T = mG$ and $1 - G/\binom{T}{m} \ge 1/2$ for $G \ge 2$, $m \ge 2$, this is implied by $r\alpha - 4\sqrt{r\epsilon} \ge 16(\ln K + 2)\,m\,\epsilon$. The bracketed term is $\ge r\alpha/2$ whenever $\epsilon \le r\alpha^2/64$, and the right side is $\le r\alpha/2$ whenever $\epsilon \le r\alpha/(32(\ln K + 2)m)$. Combining these (using $\alpha \le 2$, $K \ge 2$, $m \ge 2$, $\ln K + 2 \le 4 \ln K$ for $K \ge 2$), both inequalities hold whenever $\epsilon \le c\,r\alpha^2/(m \ln K)$ for a sufficiently small absolute constant $c > 0$. Hence $\mathbb{E}[\mathrm{Err}(\pi_{++})] \le \mathbb{E}[\mathrm{Err}(\pi_{\mathrm{rand}})]$. □

### C.9 Non-uniform singular values (statement and proof)

**Proposition C.6** (Non-uniform singular values)**.** *Let $\kappa := \sigma_{\max}/\sigma_{\min}$ with $\sigma_{\max} := \max_{i,k} \sigma_{i,k}$, $\sigma_{\min} := \min_{i,k} \sigma_{i,k}$, and recall the optimal-basis merge error* (16)*. Then for any partition $\pi$,*

$$\sigma_{\min}^2\,\Phi(\pi) \;\le\; \mathrm{Err}(\pi) \;\le\; \sigma_{\max}^2\,\Phi(\pi), \tag{37}$$

*and consequently*

$$\mathbb{E}\big[\mathrm{Err}(\pi_{++})\big] \;\le\; 8(\ln K + 2)\,\kappa^2 \cdot \mathrm{Err}(\pi_w^*), \tag{38}$$

*where $\pi_w^* := \arg\min_\pi \mathrm{Err}(\pi)$.*

*Proof.* Fix any orthonormal $B \in \mathbb{R}^{d_{\mathrm{out}} \times r}$ and any task $i$. Substituting the SVD $\Delta W_i = U_i \Sigma_i V_i^\top$ and using $V_i^\top V_i = I_r$,

$$\left\|(I - BB^\top)\Delta W_i\right\|_{\mathrm{F}}^2 = \mathrm{tr}\big(\Delta W_i^\top (I - BB^\top)\Delta W_i\big) = \mathrm{tr}\big(\Sigma_i^2\, A_i\big), \qquad A_i := U_i^\top (I - BB^\top) U_i \;\succeq\; 0,$$

with $\mathrm{tr}(A_i) = r - \left\|B^\top U_i\right\|_{\mathrm{F}}^2 = d_c^2(U_i, B)$ by (8). Since $\sigma_{\min}^2 I_r \preceq \Sigma_i^2 \preceq \sigma_{\max}^2 I_r$ in the Loewner order,

$$\sigma_{\min}^2\, d_c^2(U_i, B) \;\le\; \left\|(I - BB^\top)\Delta W_i\right\|_{\mathrm{F}}^2 \;\le\; \sigma_{\max}^2\, d_c^2(U_i, B). \tag{39}$$

Now fix a partition $\pi$. For the lower bound, let $B_w^{(k)} := \arg\min_B \sum_{i \in G_k} \left\|(I - BB^\top)\Delta W_i\right\|_{\mathrm{F}}^2$ achieve $\mathrm{Err}(\pi)$. By the lower side of (39),

$$\mathrm{Err}(\pi) \;=\; \sum_k \sum_{i \in G_k} \left\|(I - B_w^{(k)} B_w^{(k)\top})\Delta W_i\right\|_{\mathrm{F}}^2 \;\ge\; \sigma_{\min}^2 \sum_k \sum_{i \in G_k} d_c^2(U_i, B_w^{(k)}) \;\ge\; \sigma_{\min}^2\,\Phi(\pi),$$

where the last inequality uses that $\Phi(\pi)$ is the minimum of $\sum_k \sum_{i \in G_k} d_c^2(U_i, C^{(k)})$ over orthonormal $C^{(k)}$. For the upper bound, let $C_*^{(k)}$ be the chordal Fréchet mean of $G_k$ (achieving $\Phi(\pi)$). Since $C_*^{(k)}$ is a feasible (sub-optimal for the weighted problem) basis,

$$\mathrm{Err}(\pi) \;\le\; \sum_k \sum_{i \in G_k} \left\| (I - C_*^{(k)} C_*^{(k)\top}) \Delta W_i \right\|_{\mathrm{F}}^2 \;\le\; \sigma_{\max}^2 \sum_k \sum_{i \in G_k} d_c^2(U_i, C_*^{(k)}) \;=\; \sigma_{\max}^2 \, \Phi(\pi),$$

using the upper side of (39). This is (37).

For (38), let $\pi_w^* := \arg\min_\pi \mathrm{Err}(\pi)$ and let $\pi^* := \arg\min_\pi \Phi(\pi)$. Steps 1–3 of the proof of Theorem B.4 (1) are model-independent: they show $\mathbb{E}[\Phi(\pi_{++})] \le 8(\ln K + 2)\, \Phi(\pi^*)$ for any data $\{U_i\}$, using only the $k$-means++ guarantee [Arthur et al., 2007] and the Euclidean–Grassmannian comparison. Combining this with (37),

$$\begin{aligned} \mathbb{E}\big[\mathrm{Err}(\pi_{++})\big] \;&\le\; \sigma_{\max}^2 \, \mathbb{E}[\Phi(\pi_{++})] \;\le\; \sigma_{\max}^2 \cdot 8(\ln K + 2)\, \Phi(\pi^*) \\ &\le\; 8(\ln K + 2)\, \kappa^2 \, \sigma_{\min}^2 \, \Phi(\pi_w^*) \;\le\; 8(\ln K + 2)\, \kappa^2 \, \mathrm{Err}(\pi_w^*), \end{aligned}$$

using $\Phi(\pi^*) \le \Phi(\pi_w^*)$ (definition of $\pi^*$) and $\sigma_{\min}^2 \Phi(\pi_w^*) \le \mathrm{Err}(\pi_w^*)$ (the lower side of (37)). □

### C.10 Proof of Theorem B.6

*Proof.* Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables. Splitting each $h \in \mathcal{H}_a(\beta)$ as $h(x) = a^\top W_0 x + a^\top \Delta W\, x$, the $W_0$ part is a fixed function of $x$ (independent of $\Delta W$), so

$$\frac{1}{n} \sum_i \epsilon_i \, h(x_i) \;=\; \frac{1}{n} \sum_i \epsilon_i \, a^\top W_0 x_i \;+\; \frac{1}{n} \sum_i \epsilon_i \, a^\top \Delta W \, x_i, \tag{40}$$

and the first term factors out of the supremum over $\Delta W$. Taking expectation over $\epsilon$ and using $\mathbb{E}[\epsilon_i] = 0$ to kill the first term,

$$\hat{\mathfrak{R}}_n(\mathcal{H}_a(\beta)) \;=\; \mathbb{E}_\epsilon \left[ \sup_{\|\Delta W\|_{\mathrm{F}} \le \beta} \frac{1}{n} \sum_{i=1}^n \epsilon_i \, a^\top \Delta W \, x_i \right]. \tag{41}$$

**Step 1: rewrite as a Frobenius inner product.** For any $\Delta W$, $a^\top \Delta W\, x_i = \mathrm{tr}(a^\top \Delta W x_i) = \langle \Delta W,\, a x_i^\top \rangle_F$, where $\langle A, B \rangle_F = \mathrm{tr}(A^\top B)$. Linearity gives

$$\frac{1}{n} \sum_{i=1}^n \epsilon_i \, a^\top \Delta W \, x_i \;=\; \Big\langle \Delta W,\; \frac{1}{n} \sum_{i=1}^n \epsilon_i \, a x_i^\top \Big\rangle_F. \tag{42}$$

**Step 2: supremum via Cauchy–Schwarz.** Let $G_n := \frac{1}{n} \sum_{i=1}^n \epsilon_i \, a x_i^\top \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$. By the Cauchy–Schwarz inequality on the Frobenius inner product,

$$\sup_{\|\Delta W\|_{\mathrm{F}} \le \beta} \langle \Delta W, G_n \rangle_F \;=\; \beta \, \|G_n\|_{\mathrm{F}}. \tag{43}$$

Moreover, $G_n = a\big(\frac{1}{n} \sum_i \epsilon_i x_i^\top\big)$ is a rank-one matrix of the form $a v^\top$ with $v = \frac{1}{n} \sum_i \epsilon_i x_i$, so $\|G_n\|_{\mathrm{F}} = \|a v^\top\|_{\mathrm{F}} = \|a\|_2 \, \|v\|_2 = \|v\|_2$ (using $\|a\|_2 = 1$).

**Step 3: bound $\mathbb{E}\|v\|_2$ via Jensen and independence.** By Jensen's inequality,

$$\mathbb{E}_\epsilon \|v\|_2 \;\le\; \sqrt{\mathbb{E}_\epsilon \|v\|_2^2} \;=\; \sqrt{\mathbb{E}_\epsilon \frac{1}{n^2} \sum_{i,j} \epsilon_i \epsilon_j \, x_i^\top x_j} \;=\; \sqrt{\frac{1}{n^2} \sum_{i=1}^n \|x_i\|_2^2} \;\le\; \frac{R}{\sqrt{n}}, \tag{44}$$

where the cross-terms vanish because $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}$.

**Step 4: combine.** Taking expectation over $\epsilon$,

$$\hat{\mathfrak{R}}_n(\mathcal{H}_a(\beta)) \;=\; \beta \, \mathbb{E}_\epsilon \, \|G_n\|_{\mathrm{F}} \;\le\; \beta \, \frac{R}{\sqrt{n}}, \tag{45}$$

which is (21). □

Table 6: Sensitivity to the number of groups $K$ on Flan-T5-Base and Flan-T5-Large / GLUE; each cell reports Base / Large. Partitions come from the same clustering GRAM uses; every group is re-merged with the adaptive-rank merge ($r$=16) and freshly evaluated with the Table 1 pipeline, which reproduces the published $K$=2 row exactly on Large (on Base, re-merged $K$=1 scores 81.9 vs. 81.7 for the original checkpoint—SVD tie-breaking on near-degenerate spectra). The Weight Averaging row and the Large $K \in \{1, 8\}$ anchors are the corresponding Table 1 rows.

| $K$ | CoLA | MNLI | MRPC | QNLI | QQP | RTE | SST-2 | STS-B | Avg. | Extra params |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 (weight averaging) | 69.7 / 74.6 | 59.7 / 84.3 | 78.9 / 84.1 | 90.1 / 92.8 | 83.8 / 86.2 | 80.5 / 85.9 | 91.2 / 94.8 | 72.0 / 88.0 | 78.2 / 86.3 | — |
| 1 (global merge) | 69.1 / 79.2 | 74.2 / 86.0 | 80.4 / 87.3 | 89.7 / 93.6 | 83.5 / 85.6 | 80.9 / 87.0 | 92.7 / 95.3 | 84.6 / 88.7 | 81.9 / 87.8 | — |
| **2 (GRAM)** | 69.5 / 79.3 | 82.8 / 89.0 | 84.3 / 88.7 | 90.2 / 93.8 | 84.0 / 85.7 | 79.8 / 88.1 | 93.2 / 95.4 | 86.7 / 89.9 | **83.8 / 88.7** | +1.8M / +4.7M |
| 3 | 69.5 / 79.3 | 82.7 / 88.5 | 84.3 / 88.7 | 90.2 / 93.8 | 84.0 / 85.7 | 79.8 / 88.1 | 93.2 / 95.4 | 87.4 / 90.9 | 83.9 / 88.8 | +3.5M / +9.4M |
| 4 | 69.1 / 79.9 | 82.7 / 88.5 | 84.8 / 89.5 | 90.2 / 93.8 | 83.7 / 85.8 | 78.0 / 88.4 | 93.2 / 95.2 | 87.4 / 90.9 | 83.6 / 89.0 | +5.3M / +14.2M |
| 5 | 69.1 / 80.1 | 82.7 / 88.5 | 84.8 / 89.5 | 90.2 / 93.8 | 83.7 / 85.8 | 78.0 / 88.4 | 92.9 / 95.2 | 87.4 / 90.9 | 83.6 / 89.0 | +7.1M / +18.9M |
| 6 | 69.1 / 80.1 | 82.7 / 88.5 | 83.3 / 89.5 | 90.9 / 93.9 | 83.5 / 85.8 | 79.8 / 90.6 | 92.9 / 95.2 | 87.4 / 90.9 | 83.7 / 89.3 | +8.8M / +23.6M |
| 7 | 69.1 / 80.1 | 82.7 / 88.5 | 83.1 / 88.7 | 90.9 / 94.4 | 84.1 / 87.1 | 79.1 / 88.8 | 92.9 / 95.2 | 87.4 / 90.9 | 83.7 / 89.2 | +10.6M / +28.3M |
| 8 (no merging) | 69.1 / 80.1 | 82.7 / 88.5 | 85.5 / 89.2 | 90.9 / 94.4 | 84.1 / 87.1 | 84.5 / 91.3 | 92.9 / 95.2 | 87.4 / 90.9 | 84.6 / 89.6 | +12.4M / +33.0M |

Table 7: GRAM vs. random grouping quality. Intra-sim and Inter-sim denote the average subspace similarity ($1 - d_c^2/r$) within and between groups, respectively. Ratio = Intra-sim / Inter-sim. Higher ratio indicates better separation.

| Setting | Intra-sim | Inter-sim | Ratio | Avg Acc |
|---|---|---|---|---|
| No grouping ($K$=1, global merge) | 0.097 | — | — | 81.7 |
| GRAM $K$=2 | 0.143 | 0.035 | 4.06 | 83.8 |
| Random $K$=2 | 0.089 | 0.102 | 0.87 | 82.4 |
| GRAM $K$=3 | 0.150 | 0.036 | 4.21 | 83.9 |
| Random $K$=3 | 0.119 | 0.089 | 1.34 | 83.1 |

**Remark on rank.** The bound makes no use of $\mathrm{rank}(\Delta W) \leq r$, so it applies uniformly to any Frobenius-bounded linear adapter, low-rank or not. A rank constraint would buy a sharper bound only if it were combined with a nuclear-norm constraint $\|\Delta W\|_* \leq \sqrt{r}\beta$ via standard low-rank Rademacher arguments; we do not pursue this since for our purposes the rank-free bound already suffices to show that grouping does not enlarge the hypothesis class.

# D Additional Experimental Details

## D.1 Grouping Analysis

Table 7 quantifies the grouping quality produced by GRAM versus random assignment. GRAM with $K$=2 achieves an intra-group to inter-group similarity ratio of 4.06, compared to 0.87 for random grouping—a factor-of-four difference. This separation increases slightly to 4.21 at $K$=3. Across the four grouped settings, higher ratio correlates with better downstream accuracy: geometric coherence within groups translates directly to merge quality.

Random grouping at $K$=2 produces groups whose intra-group similarity (0.089) is actually *lower* than the global average (0.097), because some random partitions place highly dissimilar tasks together. GRAM avoids this failure mode by construction.

## D.2 Deployment without Oracle Task Identity

GRAM targets task-aware deployments—e.g., a multi-institution medical system where each request carries its provenance, so task identity is deployment *metadata* rather than an inference problem. Nevertheless, GRAM extends to the task-agnostic setting at no accuracy cost via a lightweight router. We instantiate the Mahalanobis task mapper of Lee et al. [2018], as deployed for task recognition in shared-knowledge lifelong learning (SKILL) [Ge et al., 2023]: class-conditional Gaussians in the *frozen base encoder's* feature space (mean-pooled hidden states; no extra model), class means estimated from only $m$=5 training samples per class ($\approx 85$ raw examples total across all 8 tasks), and one tied covariance (Ledoit–Wolf). A test input is routed to the group owning its nearest class centroid. The router involves no gradient training and no test-time task labels; merging itself remains

fully data-free, and the router's $\approx 85$-example requirement is strictly weaker than the 64-samples-*per-task* calibration consumed by the data-using baselines (RegMean, Fisher-Merge) in section 5.4. On Flan-T5-Base / GLUE at $K{=}2$, the router routes $100.0\%$ of validation examples of every task to the correct group, so task-agnostic GRAM *equals* oracle GRAM—$83.8$ on every task, to the decimal.

Table 8: Group-routing accuracy (%) of the Mahalanobis task mapper under the partition chain for every $K$, Flan-T5-Base / GLUE. $K{=}8$ requires full task identification.

| $K$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 (full task ID) |
|---|---|---|---|---|---|---|---|
| Group-routing acc. | 100.00 | 100.00 | 100.00 | 100.00 | 99.99 | 99.98 | 99.98 |

Small $K$ is what makes routing easy. The router encodes the same formatted request the instruction-tuned model itself consumes, and it only needs to resolve *groups*, not tasks: at $K{=}2$ it faces a binary decision, and any confusion between tasks in the same coalition is harmless by construction. Table 8 reports the mapper's group-routing accuracy under the partition chain for every $K$. The mapper's *only* errors, at any $K$, are QQP$\leftrightarrow$MRPC confusions (72 of $40{,}430$ QQP examples, $0.18\%$)—two semantically adjacent paraphrase-detection tasks. These errors are invisible for $K \leq 5$ because GRAM's geometry-aware clustering places QQP and MRPC in the *same* coalition, and only surface once $K \geq 6$ separates them. This reveals a structural synergy: a router's confusions occur precisely between similar tasks, and similar tasks are precisely what GRAM groups together—the coalition structure absorbs router errors by design. A per-task deployment ($K{=}8$) forgoes this protection and demands full task identification, exactly the harder regime studied in SKILL ($87.1\%$ over 102 tasks) [Ge et al., 2023].

### D.3 Per-Task Intrinsic Rank Heterogeneity

Figure 4 reports the number of singular values needed to capture $\geq 99\%$ of the per-layer Frobenius energy, computed over the 8 GLUE tasks and 72 LoRA layers of Flan-T5-Base. Uniform per-task allocation $r/|G_k|$ inside a coalition either wastes slots on compact-spectrum tasks or truncates long-tail ones; the adaptive allocator of section 4.3 reads off this heterogeneity directly from the singular values and assigns rank where it actually reduces error.

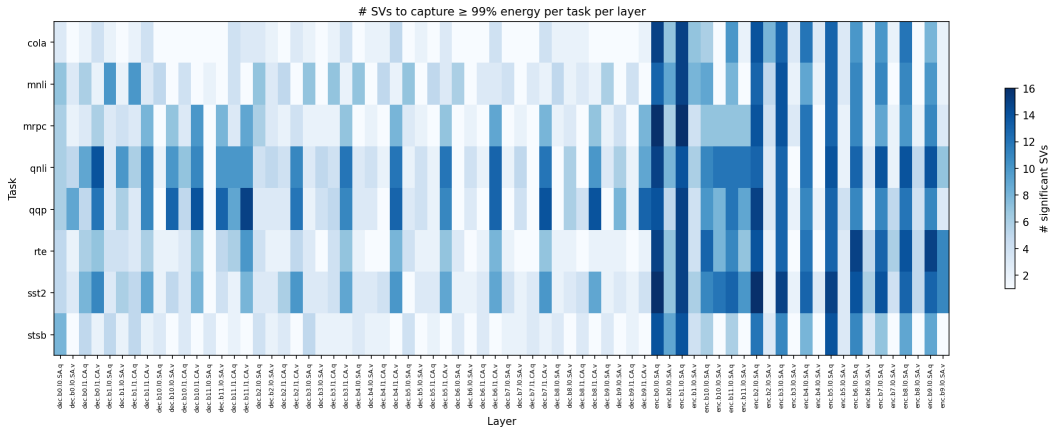



Figure 4: Number of singular values required to capture $\geq 99\%$ of Frobenius energy, per task (rows) and per LoRA layer (columns), on Flan-T5-Base. Decoder layers (left half) are typically low-rank (1–4 significant directions); encoder layers (right half) are higher-rank (10–16). The cross-task and cross-layer heterogeneity motivates the adaptive rank allocator in section 4.3.

### D.4 Pairwise Projection Error on ViT-B/32

Figure 5 visualizes the pairwise projection error $\sum_s \sigma_{i,s}^2 \sin^2 \theta_s$ between task subspaces on ViT-B/32 under $K{=}1$ (TSV-Merge) vs. $K{=}2$ (GRAM); the block structure that emerges at $K{=}2$ is described in the caption.

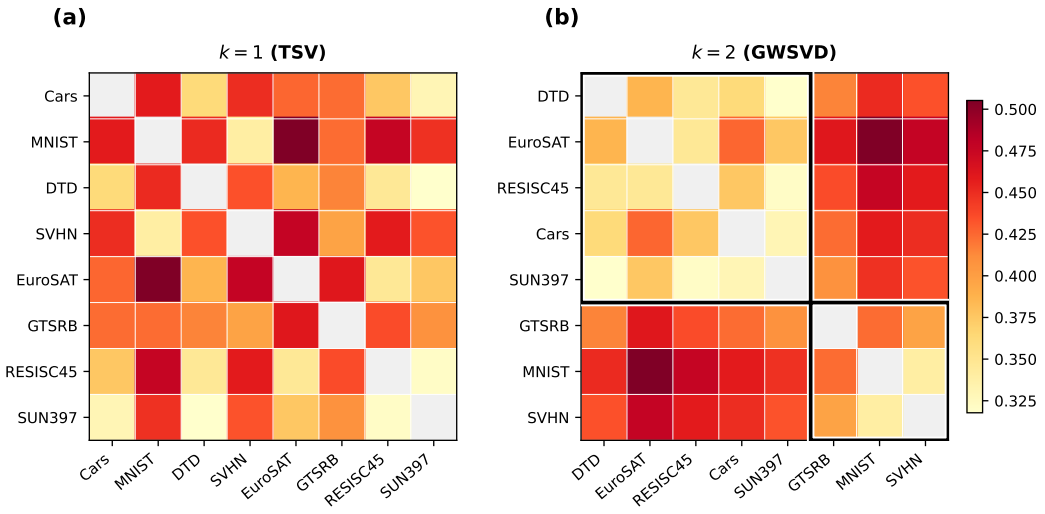



Figure 5: Pairwise projection error $\sum_s \sigma_{i,s}^2 \sin^2 \theta_s$ between task subspaces on ViT-B/32, under $K{=}1$ (TSV-Merge, left) and $K{=}2$ (GRAM, right). Reordering at $K{=}2$ by GRAM's partition exposes a block-diagonal structure: within-group pairs incur lower error than across-group pairs, consistent with Theorem B.3.

Table 9: Eight-task vision benchmark results (test accuracy) using CLIP ViT-B/16 with full fine-tuning. **Bold**: best merged result per column; underline: second best. Individual adapters use one task-specific model per task (oracle upper bound).

| Method | MNIST | SVHN | GTSRB | DTD | Cars | RESISC45 | EuroSAT | SUN397 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Individual (oracle) | 99.8 | 97.8 | 99.0 | 98.3 | 87.5 | 97.0 | 99.8 | 78.8 | 94.8 |
| Weight Averaging | 94.4 | 72.6 | 60.6 | 54.0 | 69.9 | 75.2 | 81.3 | 68.6 | 72.1 |
| Task Arithmetic | 98.3 | 84.5 | 76.1 | 60.4 | 68.1 | 73.7 | 86.4 | 64.4 | 76.5 |
| TIES-Merging | 98.3 | 84.8 | 73.9 | 64.8 | 71.7 | 77.8 | 89.4 | 68.0 | 78.6 |
| AdaMerging++ | 98.4 | 93.1 | 92.3 | 64.0 | 79.2 | 81.1 | 95.6 | 69.2 | 84.1 |
| TSV-Merge | 99.3 | 92.9 | 94.5 | 85.2 | 81.0 | 88.5 | 97.4 | 70.6 | 88.7 |
| Iso-C | 99.1 | 90.7 | 94.9 | 89.9 | 82.6 | 91.6 | 97.4 | 73.5 | 90.0 |
| Iso-CTS | 99.2 | 90.4 | 96.2 | **92.3** | **83.2** | **92.8** | 97.7 | **74.1** | 90.7 |
| LiNeS-Merge | 98.6 | 85.6 | 79.7 | 64.7 | 72.5 | 77.5 | 87.6 | 65.4 | 79.0 |
| LOT-Merging | 99.0 | 94.8 | 70.8 | 68.7 | 75.9 | 82.6 | 84.4 | 70.7 | 80.9 |
| **GRAM** (ours) | **99.5** | **96.0** | **96.8** | 92.2 | 82.7 | 91.4 | **98.8** | 72.0 | **91.2** |

### D.5 CLIP ViT-B/16 Full Fine-tuning Results

For completeness across the standard CLIP backbone family, table 9 reports per-task accuracy on the 8-task vision benchmark using a fully fine-tuned CLIP ViT-B/16 checkpoint. GRAM reaches $91.2\%$ average—the highest among all compared methods—outperforming Iso-CTS ($90.7\%$), Iso-C ($90.0\%$), TSV-Merge ($88.7\%$), and the remaining single-model baselines. The win pattern mirrors the B/32 and L/14 results in table 2: the largest absolute gains over global merging concentrate on the high-chordal-distance tasks (GTSRB, EuroSAT, SVHN), exactly the cells that the alignment-penalty analysis of Theorem B.3 predicts will benefit most from grouping. The consistency of the gain pattern across three ViT scales (B/32, B/16, L/14) is consistent with the partition-quality mechanism being scale-invariant rather than tied to a specific backbone.

### D.6 Implementation Details

**Flan-T5 GLUE adapters.** On both Flan-T5-Base and Flan-T5-Large we use LoRA with rank $r = 16$ and $\alpha = 32$ (scaling factor 2.0), applied to the query and value projections in self-attention and (in the decoder) cross-attention. Each LoRA delta has shape $768 \times 768$ on Base and $1024 \times 1024$ on Large; the number of adapted layers is $L = 72$ on Base (12 encoder + 12 decoder blocks) and

$L = 144$ on Large (24 encoder + 24 decoder blocks). Per-task adapters are trained with AdamW (learning rate $5 \times 10^{-4}$, batch size 16) for 10 epochs; we report on the GLUE validation splits, with Spearman $\rho\times100$ on STS-B and accuracy on the remaining tasks.

**CLIP ViT vision adapters.** We use the per-task fully fine-tuned CLIP ViT-B/32 and ViT-L/14 checkpoints released by TSV-Merge [Gargiulo et al., 2025], evaluated on the standard 8-task vision benchmark (MNIST, SVHN, GTSRB, DTD, Cars, RESISC45, EuroSAT, SUN397). The ViT-B/16 results in section D.5 are produced under the same protocol on a fully fine-tuned ViT-B/16 backbone. Since the per-task checkpoints are fully fine-tuned, the per-layer task delta $\Delta W_i^{(\ell)} = W_i^{(\ell)} - W_0^{(\ell)}$ is full-rank; we apply SVD truncation at merge time with rank budget $r$=16 per layer (matching the GLUE setting), which is the rank actually flowing into the chordal distance and the per-coalition merge. As stated in section 3, the matrix formulation applies to linear layers: SVD-based merging is applied to all 2-D weight matrices (MLP blocks and attention projections), while the remaining non-matrix parameters (LayerNorm gains, biases) are merged by simple averaging. All vision results report test accuracy. No additional task-specific training is performed; GRAM and all baselines operate on these checkpoints directly.

**Qwen2 medical adapters.** We use four publicly released LoRA adapters on Qwen2-7B-Instruct, one per medical sub-domain: MedQA, CMExam, HealthcareMagic, and cMedQA2. Each adapter uses LoRA with rank $r = 8$ and $\alpha = 16$, applied to the attention $(q, k, v, o)$ and MLP (gate, up, down) projections of every decoder block. GRAM operates directly on these adapters without further training.

**Calibration set for data-using baselines.** Fisher-Merge [Matena and Raffel, 2022] and RegMean [Jin et al., 2022] require a small calibration set per task at merge time. We use a uniform 64-sample-per-task calibration set drawn from each task's training split: Fisher-Merge uses these samples to estimate per-parameter Fisher information, and RegMean uses them to estimate input second-moment matrices for the closed-form layer-wise solver. The data-free baselines (Weight Averaging, Task Arithmetic, TIES-Merging, Iso-C, Iso-CTS, TSV-Merge) and GRAM see no calibration data.

**Clustering.** For each model family we compute the $T \times T$ pairwise chordal distance matrix (averaged across all $L$ adapted layers) and run $k$-means++ with 20 random initializations on the projection-matrix embeddings. We set $K = 2$ as the number of groups throughout.

**Computational cost.** For the Flan-T5-Base/GLUE setup ($T$=8, $L$=72), SVD decomposition of all $T \times L$ task deltas takes approximately 20 seconds on CPU; distance-matrix construction and $k$-means++ clustering complete in under 1 second; per-group SVD merging adds negligible overhead. Total wall-clock time for the GRAM pipeline is under 30 seconds, compared to hours of adapter training per task. The other model families exhibit the same qualitative cost structure: SVD dominates and is bounded by the per-layer cost $O(d_{\text{out}} d_{\text{in}} r)$ summed over $TL$ deltas.

### D.7 Medical Benchmark Details

**Task suite.** The medical benchmark covers two task families crossed with two languages:

- *Multiple-choice QA (MCQA).* MedQA (English) and CMExam (Chinese): the model selects one option from a candidate set; reported metric is exact-match accuracy.
- *Doctor–patient dialogue (open-ended generation).* HealthcareMagic (English) and cMedQA2 (Chinese): the model generates a free-form answer to a patient query; reported metrics are BLEU-4 and ROUGE-1, ROUGE-2, and ROUGE-L (F1) against the reference response.

The 4 task families with 1 metric each (MCQA accuracy) plus 4 metrics each (BLEU-4, ROUGE-1/2/L) on the two open-ended tasks yield the 10 heterogeneous metric columns reported in table 3.

**Summary metrics.** *Avg Rank* is the mean rank across the 10 columns, where rank 1 is the best per-column merging method (the per-task Individual oracle is excluded from ranking). *Avg Impr* is the per-column relative improvement over the un-adapted Pretrained baseline, averaged across the 10 columns: for each metric column $j$ and method $M$, we compute $(\text{score}_{M,j} - \text{score}_{\text{Pretrained},j})/\text{score}_{\text{Pretrained},j}$, then average over $j$. This normalizes for the wide range of metric magnitudes (MCQA accuracy in tens of points; ROUGE-2 in single digits).

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and section 1 state three contributions: (i) the GRAM algorithm (Grassmannian $k$-means++ partitioning with adaptive rank allocation, formalized in section 4); (ii) an $O(\log K)$ partition-error bound and a strict-in-expectation improvement over global merging (section B, Theorem B.4, Theorem B.5); and (iii) consistent gains over data-free baselines on GLUE, 8-task CLIP vision, and the medical Qwen2-7B benchmark (section 5). All three are supported by content in the paper; the abstract scopes the medical claim to the data-free comparison family and to the four LoRA experts evaluated.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: section 6 contains an explicit "Limitations" paragraph covering three points: (i) single-seed reporting (caveat on small Flan-T5-Large margins); (ii) one backbone family per modality and the medical evaluation being a public-benchmark case study rather than a real multi-site federated deployment; (iii) the theory analyzes the optimal-basis partition objective rather than GRAM's exact column-stacked merge with adaptive allocation, so it should be read as geometric motivation rather than an end-to-end optimality guarantee.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems, lemmas, propositions, and corollaries are numbered, cross-referenced, and accompanied by full proofs. section B contains the statements (Theorem B.1, Theorem B.2, Theorem B.3, Theorem B.4, Theorem B.5, Theorem B.6, Theorem B.7); the deterministic latent-group model and all hypotheses (rank budget $\sum k_i \leq r$, prototype principal-angle separation $\phi_{\min}$, within-group spread $\epsilon$, regime $\epsilon \leq r(1-\cos\phi_{\min})/4$, etc.) are stated alongside each result. Complete proofs appear in section C, with the supporting lemmas (Theorem C.1, Theorem C.2, Theorem C.3, Theorem C.5) and a non-uniform singular-value extension (Theorem C.6).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: section D.6 specifies LoRA configurations for every model family used (Flan-T5-Base/Large: $r$=16, $\alpha$=32, q/v projections, with $L$=72/144 adapted layers; CLIP ViT checkpoints from the TSV-Merge release; Qwen2-7B medical adapters: $r$=8, $\alpha$=16, all attention + MLP projections), the AdamW optimizer / batch size / epoch settings for the Flan-T5 adapters, the GLUE validation-split protocol, the $k$-means++ clustering setup ($n_{\text{init}}$=20, $K$=2), and the calibration-set size for Fisher-Merge / RegMean. section D.7 defines the medical metric set (BLEU-4, ROUGE-1/2/L) and the Avg Rank / Avg Impr summaries. section B.2 reports the computational cost. The GRAM algorithm itself is fully specified in section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is not released with the submission. All datasets used (GLUE, the standard 8-task CLIP vision benchmark, MedQA, CMExam, HealthcareMagic, cMedQA2) are publicly available, and the four Qwen2-7B medical LoRA adapters are publicly released checkpoints; the CLIP ViT checkpoints come from the TSV-Merge release [Gargiulo et al., 2025]. We will release a reference implementation of the GRAM pipeline upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: section 5 states the high-level evaluation protocol (validation accuracy on GLUE, Spearman $\rho \times 100$ for STS-B, test accuracy on vision, the BLEU/ROUGE metric set on medical), and section D.6 together with section D.7 provide the full hyperparameters: LoRA rank/$\alpha$/target modules per model family, AdamW learning rate, batch size, epochs for the Flan-T5 adapter training, $k$-means++ initialization count, $K$ throughout, and the calibration-set size for the data-using baselines.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: All experiments are reported as single-seed runs and the paper does not include error bars or confidence intervals. The Limitations paragraph in section 6 explicitly flags this and cautions that small Flan-T5-Large margins should be read with care. The gap pattern between GRAM and the baselines is consistent across three model families (Flan-T5, CLIP ViT, Qwen2-7B) and across ten metric columns on the medical benchmark, which we view as the dominant evidence for the conclusions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: section B.2 reports the asymptotic cost ($O(TLd_{\text{out}}d_{\text{in}}r)$ for SVD, dominant) and section D.6 (Computational cost paragraph) gives concrete wall-clock numbers for the Flan-T5-Base/GLUE pipeline: SVD of all $T \times L$ task deltas in approximately 20 seconds on CPU, distance-matrix construction and $k$-means++ clustering in under 1 second, and

total GRAM pipeline wall-clock under 30 seconds. The other model families exhibit the same qualitative cost structure since SVD dominates and is bounded by the per-layer cost summed over $TL$ deltas. GRAM does not require GPU at merge time; the underlying per-task fine-tuned checkpoints come from prior released artifacts. fig. 3 additionally profiles wall-clock time and peak GPU memory of the grouping pipeline on a single RTX 3080 as $T$ grows from 8 to 1024.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses publicly available benchmarks (GLUE, the standard CLIP 8-task vision benchmark, MedQA / CMExam / HealthcareMagic / cMedQA2) and publicly released model checkpoints (Flan-T5, CLIP ViT, Qwen2-7B-Instruct, the four medical LoRA adapters). No human-subjects data was collected; no personally identifying information was processed; no scraped data was generated. Authors confirm they have reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper proposes a model-merging algorithm whose societal impact is mediated by the underlying models being merged rather than by the merging procedure itself. The medical case study uses publicly available LoRA experts on a public benchmark and is a methodological case study of merging quality, not a deployment recommendation; the Limitations paragraph in section 6 notes that real federated medical deployment remains future work. We do not see direct positive or negative societal impacts specific to the GRAM algorithm beyond those of any task-aware multi-task merging method.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new high-risk models or scraped datasets. GRAM is an algorithm that merges existing publicly released LoRA / fine-tuned checkpoints; no new models or data are produced for distribution as part of this submission.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Every dataset and model checkpoint used is publicly available and credited to its original source by citation: GLUE [Wang et al., 2018], Flan-T5 [Chung et al., 2024], LoRA [Hu et al., 2022], the CLIP ViT checkpoints from the TSV-Merge release [Gargiulo et al., 2025], the FusionBench library [Tang et al., 2025], and the medical benchmarks (MedQA, CMExam, HealthcareMagic, cMedQA2) along with the four publicly released Qwen2-7B medical LoRA adapters. We use each asset under its original release terms (typically Apache-2.0 or CC-BY for the datasets / model weights involved).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission introduces an algorithm and analysis, not a new dataset, model, or pretrained checkpoint. We do not release new assets at submission time; a reference implementation will accompany the paper upon acceptance (see Q5).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All evaluation uses pre-existing public benchmarks and pre-existing model checkpoints; no human annotators or participants were recruited.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects. No IRB review is therefore required.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are an *object* of the experiments (Qwen2-7B-Instruct is one of the model families on which GRAM is evaluated; Flan-T5 is another), not a component of the GRAM methodology itself. The GRAM algorithm is a deterministic pipeline of SVD, $k$-means++ on chordal distances, and a greedy rank allocator (section 4); no LLM is used inside the algorithm or in any non-standard way for writing or analysis.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.