TRANSFORMERS ALIGN IN SUBSPACES, NOT UNITS: ROTATION-AWARE MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

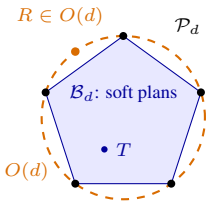
Model merging combines trained networks in weight space, and every alignment method implicitly picks a symmetry group for comparing them. We show that group is mis-specified for Transformers. Transport-based alignment searches the Birkhoff polytope $\mathcal{B}_d$; self-attention is also invariant under *continuous rotations* of its query–key and value–output subspaces, and $O(d) \cap \mathcal{B}_d$ is exactly the permutations, so no transport plan—however soft—expresses a non-trivial rotation. The LayerNorm residual stream, in turn, admits among isometries exactly the rotations fixing the mean axis, where a soft plan is *inexact*, not merely suboptimal. *Rotation-aware Optimal Transport* (RaOT) aligns in that quotient. Building on interface-wise rotation fusion (Zhang et al., 2025), it differs in four ways: activation-estimated alignments from one forward pass; head correspondence and within-head rotation solved *jointly* as one exact assignment; a mean-preserving Procrustes map on the residual stream; and identity competing with every estimate, which supplies a provenance switch and a test of whether alignment should act at all. This is a mechanism paper, not a leaderboard entry. No method—ours included—merges independently trained ImageNet ViTs to usable accuracy, so we read the barrier through *matched contrasts*: across nine public pairs and a four-width series inside one model family, what the isometries' estimator adds beyond the permutation's is larger at larger width. The polytope's own share shrinks only on the original three-pair ladder—across pairs it straddles zero, and a pre-registered same-recipe study reverses it—so we claim the isometries half alone, on recipe- or corpus-different parents, and not a width law. Alignment buys accuracy in frozen-feature transfer and, for parents differing in task *and* initialization, in trunk–head compatibility rather than in the backbone representation. On fine-tuned siblings of one initialization the gate rejects every transform: the role there is detection, not repair.

1 INTRODUCTION

Model merging consolidates several trained networks into one set of weights, without retraining and without the training data—a few unlabeled calibration images at most (Wortsman et al., 2022; Ilharco et al., 2023; Yang et al., 2026). Its success depends on a choice made before any averaging—the *quotient space*: weights should be compared only after factoring out the architecture's functional symmetries, so every alignment method is implicitly a commitment to one symmetry group. This paper is about the geometry of that commitment: the common permutation/transport quotient is *too small* for Transformers, and the isometric quotient that replaces it predicts when alignment succeeds and when it fails, and describes how its measured gain changes with width. For feed-forward blocks with a non-homogeneous activation such as GELU the symmetry group is generated by neuron permutations (ReLU networks add positive per-unit scalings), and a rich line of work aligns models by permutation matching (Ainsworth et al., 2023; Entezari et al., 2022) or by its convex relaxation, optimal transport (Singh & Jaggi, 2020; Nguyen et al., 2023; Imfeld et al., 2024).

A structural gap. Every OT-based merging method optimizes a transport plan in the *Birkhoff polytope* $\mathcal{B}_d$, yet Transformer self-attention is additionally invariant under *continuous orthogonal rotations* of its query–key and value–output subspaces (Zhang et al., 2025), and we prove (Proposition 1) that $O(d) \cap \mathcal{B}_d$ is exactly the permutations (Fig. 1a): no plan—however soft or rounded (Singh

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) $O(d) \cap \mathcal{B}_d = \mathcal{P}_d$: no plan expresses a rotation

residual stream: *one* $R_{\text{res}} \in O(d)_{\mathbf{1}}$ for all blocks, rotating every writer and reader — LN reads the stream only through $n(\cdot)$ (Proposition 2); the skips thread *one* stream, so one basis

attention (admits $GL(d_k) \times GL(d_v)$; a merge can use its isometries, Appendix E.5)
heads: $\pi \in \mathfrak{S}_H$, joint with the rotations — closed-form R leaves one exact assignment
QK: one *shared* $R_h \in O(d_k)$ — scores see $W_Q^\top W_K$; VO: $S_h \in O(d_v)$ — only $W_O W_V$

FFN: hidden-unit permutation — GELU is not positively homogeneous

LN: γ, β absorbed into the adjacent linears — function-preserving gauge

estimates: activation cross-covariances Σ, one forward pass, $\sim$64 unlabeled images · gate: apply a map only if $\|W_A - \hat{W}_B\|$ falls vs. I — shared init $\Rightarrow$ identity is correct

(b) what each parameter group receives, and why (grey)

Figure 1. **Choosing the quotient a merge can use. (a)** Permutation- and OT-based alignment optimizes over the Birkhoff polytope $\mathcal{B}_d$ (blue); its intersection with the rotations $O(d)$ (orange) is exactly the permutations $\mathcal{P}_d$ (Proposition 1), so a rotation separating two Transformers is expressible by *no* transport plan. **(b)** What RaOT does about it: each interface receives a map from the compact isometric class we align in for stable merging—for the heads the isometric subgroup of the strictly larger general-linear symmetry (Appendix E.5), for the LayerNorm stream the *mean-preserving* rotations (Proposition 2)—estimated in one calibration pass and gated against identity (Section 3.6).

& Jaggi, 2020; Imfeld et al., 2024)—can represent a rotation separating two parameterizations. The rotation-based fusion of Zhang et al. (2025) already acts on this observation, assigning rotations to attention and permutations to feed-forward layers. We take that interface-wise insight as established and sit in what it leaves open—*how* alignments are estimated, coupled and applied: **activation-driven estimation**, **joint head–rotation matching**, **the LayerNorm residual stream**, **the option not to act** (Section 3).

Our approach. We develop *Rotation-aware Optimal Transport* (RaOT), one training-free framework realizing all four. Its two technical cores are the residual stream and the gate (Fig. 1b): we characterize the linear maps a LayerNorm-normalized stream admits and align in their mean-fixing isometric subgroup $O(d)_{\mathbf{1}}$, with a closed-form constrained Procrustes solution, and let identity compete with every estimate on a weight-distance criterion. The gate yields *one* pipeline with a switch, not several methods: the maps are estimated once, and an operating point decides which to apply—all estimated maps (**RaOT-align**), only maps accepted by the weight-distance criterion (**RaOT-safe**), or maps selected after automatic regime detection (**RaOT-auto**). The weight gate detects the regime but ranks within one weakly, so it is the default when provenance is unknown, with the held-out criterion composed behind it. Head correspondence and within-head rotation are solved jointly—no alternation, no fractional plan to round: eliminating the rotations in closed form leaves one linear assignment on nuclear-norm costs, solved *exactly*, with the rotation a pair's query–key subspaces must share carried inside the cost (Section 3.4). Estimates come from activation cross-covariances on shared inputs; weight matching is a complementary signal, and swapping only the stream estimator to weights improves our ViT-S/B merges (Table 1). On the theory side the gate certifies that the aligned weight distance never exceeds direct averaging's, hence neither does the barrier bound under one smoothness constant (Theorem 1); we claim this *algorithmic* guarantee, not the *oracle* fact that exact minimization over $O(d)$ cannot lose to permutations (Proposition 4).

Contributions. Building on Zhang et al. (2025)'s interface-wise assignment of rotations to attention and permutations to feed-forward layers, we contribute three things: (1) **Symmetry mismatch**, three linked facts: transport plans cannot express attention rotations ($O(d) \cap \mathcal{B}_d = \mathcal{P}_d$, Proposition 1); the LayerNorm stream admits, among isometries, exactly the rotations fixing the mean axis, and we align in the sign-fixed subgroup of *mean-preserving* rotations (Proposition 2); and a soft plan there is *inexact*, not suboptimal, distorting every token by a gain $\rho(x) \geq 1$ (Proposition 3). (2) **Algorithm.** RaOT differs from rotation fusion in four places: alignments estimated from unlabeled activations, head correspondence and within-head rotation solved *jointly* in one exact pass, a mean-preserving Procrustes map on the LayerNorm stream, and identity competing with every estimate. (3) **Empirical trend.** Across nine public pairs of independently trained ImageNet ViTs and a four-width series within one model family, what the isometries' estimator adds beyond the permutation's is larger at the wider end of each series, while the permutation's own gain declines only on the original three-pair ladder; the ordering reappears in frozen-feature transfer and early fine-tuning, and its

sharpest application test is parents differing in task *and* initialization, where rotation-aware alignment most improves trunk–head compatibility (Table 2). **Independently trained Transformers carry alignment signal beyond what permutations capture, and it is rotational**; on the three-pair ladder a permutation's share of the rotational objective declines with width (η: $0.35 \to 0.24$), a trend we do not claim as a width law.

2 Related Work

Weight-space merging without alignment. When parents share an initialization, direct interpolation is often effective (Wortsman et al., 2022; Ilharco et al., 2023), weighted by parameter statistics in Fisher merging (Matena & Raffel, 2022) and RegMean (Jin et al., 2023), and much follow-up targets *task interference* (Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Zhang et al., 2026); these scale well but presuppose the shared coordinate system and address no architectural symmetry.

Alignment-based merging. A complementary line aligns neurons before averaging, exploiting the permutation symmetry of MLPs (Ainsworth et al., 2023) under linear mode connectivity (Entezari et al., 2022; Frankle et al., 2020). OTFusion generalizes hard permutations to soft transport plans and Wasserstein barycenters (Singh & Jaggi, 2020); CLAFusion adds cross-layer alignment (Nguyen et al., 2023); Imfeld et al. (2024) extend OT to Transformers but need post-merge fine-tuning; ZipIt! matches and "zips" correlated units across tasks (Stoica et al., 2024); and Gromov–Wasserstein variants align incomparable spaces (Fang et al., 2025). All parameterize alignment by (relaxations of) permutations, which Section 3 shows cannot express attention rotations.

Symmetries of Transformer parameterizations. Permutation symmetries' role in loss-landscape geometry is well documented (Entezari et al., 2022; Ainsworth et al., 2023; Ferbach et al., 2024). For Transformers, Tran et al. (2026) give a comprehensive account of functional equivalence in attention—including that rotary encodings materially shrink the group that sinusoidal and additive ones leave intact—and Zhang et al. (2025) showed that the equivalence class of attention layers includes continuous rotations of the query–key and value–output subspaces, and—crucially for us—already assigns rotations to attention and permutations to feed-forward layers, a paradigm we build on rather than claim. Their method matches from weights alone with a closed-form orthogonal solution, keeps the head correspondence positional rather than optimizing it, does not treat the LayerNorm-normalized residual stream, and always applies its transforms. RaOT extends it along exactly those four axes (Section 1), both formulations assuming equal widths; Table 5 makes the difference concrete, target by target, from their released implementation.

Optimal transport in vision and learning. Entropic (Cuturi, 2013) and unbalanced (Chizat et al., 2018) OT and joint Wasserstein–Procrustes alignment (Grave et al., 2019; Alvarez-Melis & Jaakkola, 2018) supply the tools; we repurpose them for the internal units of networks, in the symmetry class a merge can use (Appendix E.5).

3 Method

Algorithm 1 summarizes the full pipeline; all proofs are deferred to Appendix A.

3.1 Setup and Notation

We merge N trained models with architecturally identical layer graphs; for exposition $N = 2$, models A and B ($N > 2$: Appendix C.1). The *interface* $\ell = 1, \dots, L$ denotes the output space of layer ℓ, of dimension d_ℓ^A (resp. d_ℓ^B). Given N_{cal} unlabeled calibration images, we run one forward pass through each model and collect every token of every image as a row: activation matrices $H_A^{(\ell)} \in \mathbb{R}^{M \times d_\ell^A}$ and $H_B^{(\ell)} \in \mathbb{R}^{M \times d_\ell^B}$ with $M = N_{\text{cal}} T$ (T=197 tokens for the ImageNet ViTs), column-centered as $\bar{H} = H - \frac{1}{M} \mathbf{1}\mathbf{1}^\top H$. Both models process the *same* inputs, so rows are already in correspondence and our transport plans act on *features*, not samples (Grave et al., 2019). All alignment information is condensed in the *cross-covariance*

$$\Sigma^{(\ell)} \;=\; \bar{H}_A^{(\ell)\top} \bar{H}_B^{(\ell)} \;\in\; \mathbb{R}^{d_\ell^A \times d_\ell^B}. \tag{1}$$

Alignment-based merging seeks per-interface maps $G^{(\ell)}$ transporting model B's coordinates onto model A's; weights are pulled back through them and averaged,

$$\hat{W}_B^{(\ell)} = G^{(\ell)}\, W_B^{(\ell)} \left(G^{(\ell-1)}\right)^\top, \qquad W^\star = \lambda W_A + (1-\lambda)\hat{W}_B, \tag{2}$$

where Eq. (2) uses $G^{-1} = G^\top$, exact for square orthogonal maps and approximate for soft plans and the rectangular width-heterogeneous case (Appendix E.6).

3.2 Symmetry Groups of Transformer Interfaces

A reparameterization g is a *functional symmetry* if $f_{g\cdot\theta} = f_\theta$ for all inputs; the admissible g depends on the interface. At **post-activation FFN units** only permutations commute with an elementwise σ. At the **query–key subspace** the scores $s_{ij} = h_i^\top W_Q^\top W_K h_j$ depend only on $W_Q^\top W_K$, so $W_Q \mapsto AW_Q$, $W_K \mapsto A^{-\top} W_K$ is a symmetry for *any* $A \in GL(d_k)$; at the **value–output subspace** the path depends only on $W_O W_V$, so $W_V \mapsto BW_V$, $W_O \mapsto W_O B^{-1}$ is one for any $B \in GL(d_v)$; and **heads** may be permuted by $P_H \in \mathcal{P}_H$. The head interfaces' symmetries are therefore *not* compact. We align in the maximal compact subgroup, for reasons about merging rather than about symmetry (Appendix E.5),

$$\mathcal{G} = \left(\prod_{\ell=1}^{L} \mathcal{P}_{d_{\text{ffn}}}\right) \times \prod_{\ell=1}^{L} \left(\left(\prod_{h=1}^{H} O(d_k) \times O(d_v)\right) \rtimes \mathcal{P}_H\right) \times O(d)_{\mathbf{1}}, \tag{3}$$

where each block carries its own head factors relabeled by $\mathcal{P}_H$ (hence $\rtimes$), d_{ffn} is the FFN width, and the single $O(d)_{\mathbf{1}}$ (Section 3.5) acts on the whole stream.

Existing OT-based merging methods (Singh & Jaggi, 2020; Nguyen et al., 2023; Imfeld et al., 2024) constrain the alignment map at *every* interface to the transportation polytope, whose square, uniform-marginal case is the Birkhoff polytope $\mathcal{B}_d = \text{conv}(\mathcal{P}_d)$ of doubly stochastic matrices. The following proposition shows this constraint at odds with Eq. (3).

Proposition 1 (Representation gap)**.** *$O(d) \cap \mathcal{B}_d = \mathcal{P}_d$. Hence no non-trivial rotation is expressible by a transport plan, and any alignment method whose feasible set is contained in the transportation polytope cannot represent a rotation $R \notin \mathcal{P}_d$ of the query–key or value–output subspaces itself; for generic (non-degenerate) weights, whose stabilizer under Eq. (3) is trivial, this rules out exact realignment of two parameterizations differing by that rotation.*

Proposition 1 is about feasible sets, not any method's failure mode. It supplies a mechanism for two observations in the literature—soft OT alignment of Transformers benefits from fine-tuning (Imfeld et al., 2024), and rotation-based matching finds alignments permutation search misses (Zhang et al., 2025)—without claiming it is the only one at work.

3.3 Alignment as Symmetry-Constrained Projection

Both permutation-type and rotation-type alignment are instances of one template: maximize the inner product with the cross-covariance over the interface-appropriate feasible set $\mathcal{C}$ (the transport polytope for permutation interfaces, $O(d)$ for rotation interfaces),

$$G^\star = \arg\max_{G \in \mathcal{C}} \langle \Sigma, G \rangle_F. \tag{4}$$

(a) Permutation interfaces $\Rightarrow$ assignment LP: with $\mathcal{C} = \mathcal{T}(a,b) = \{T \geq 0 : T\mathbf{1} = a,\ T^\top \mathbf{1} = b\}$, Eq. (4) is *linear* in T, and every interface we align has equal-size unit sets and uniform marginals, so $\mathcal{T}(a,b) = \mathcal{B}_d = \text{conv}(\mathcal{P}_d)$ and the optimum is attained at a vertex—a permutation:

$$T^\star = \arg\max_{T \in \mathcal{B}_d} \langle \Sigma, T \rangle = P_{\pi^\star}, \qquad \pi^\star = \arg\max_{\pi \in \mathfrak{S}_d} \textstyle\sum_i \Sigma_{\pi(i)i}, \tag{5}$$

which the Hungarian algorithm solves *exactly* in $O(d^3)$. We do not use the entropic relaxation (Cuturi, 2013) that usually stands in for this LP—for exactness, not cost: a fractional plan mixes units, so it is not a reparameterization, and a merge built on one stops reproducing its parents at the endpoints ($|\Delta L| = 2.03$ against 0.000; Proposition 3 and Appendix E.1). Entropic and unbalanced OT appear only inside the extensions labelled approximate (Appendix E.6). **(b) Rotation interfaces $\Rightarrow$ orthogonal Procrustes**: with $\mathcal{C} = O(d)$ and SVD $\Sigma = U\Lambda V^\top$,

$$R^\star = UV^\top \in \arg\max_{R \in O(d)} \langle \Sigma, R \rangle, \tag{6}$$

since $\langle \Sigma, R \rangle = \langle \Lambda, U^\top R V \rangle \le \sum_i \lambda_i$, with equality iff $U^\top R V$ acts as the identity on the span of the singular vectors with $\lambda_i > 0$: $UV^\top$ is always a maximizer, unique iff Σ has full rank, and on the null directions the rotation is free (we take what the SVD returns).

The per-interface assignment—Eq. (5) for FFN units and head permutations, Eq. (6) for the per-head subspaces—recovers the interface-wise split of Zhang et al. (2025), which we adopt rather than claim. The rest of this section develops our four departures: a constrained residual-stream Procrustes, a joint head–rotation solver, the activation-driven Σ, and the gate.

3.4 Coupled Rotation-Aware Transport

The multi-head interface carries a discrete symmetry (heads permute) *and* a continuous one (each head's subspace rotates), so we solve for both with an explicit block structure. Let $\pi \in \mathfrak{S}_H$ assign each head of B to a head of A, with permutation matrix P_π. A matched pair shares one $R_h \in O(d_k)$ on its query and key subspaces ($q \cdot k$ is invariant under a *common* rotation) and carries its own $S_h \in O(d_v)$ on the value–output path; with $R = \oplus_h R_h$, $S = \oplus_h S_h$ block-diagonal (R_h, S_h act on B's head h, which π then sends to A's head $\pi(h)$) and $\bar{Q}_B, \bar{K}_B, \bar{V}_B$ the head-stacked activations, the coupled problem is

$$\min_{\substack{\pi \in \mathfrak{S}_H \\ R_h \in O(d_k),\, S_h \in O(d_v)}} \left\| \bar{Q}_A - \bar{Q}_B R^\top (P_\pi \otimes I_{d_k})^\top \right\|_F^2 + \left\| \bar{K}_A - \bar{K}_B R^\top (P_\pi \otimes I_{d_k})^\top \right\|_F^2 + \left\| \bar{V}_A - \bar{V}_B S^\top (P_\pi \otimes I_{d_v})^\top \right\|_F^2. \tag{7}$$

One discrete and two continuous variables invite an alternating scheme; none is needed. P_π, R and S are orthogonal and preserve Frobenius norms, so Eq. (7) is equivalent to $\max_{\pi,R,S} \sum_h \left[\langle \Sigma^q_{\pi(h)h} + \Sigma^k_{\pi(h)h}, R_h \rangle + \langle \Sigma^v_{\pi(h)h}, S_h \rangle \right]$, where $\Sigma^q_{ab} = \bar{Q}^\top_{A,a} \bar{Q}_{B,b}$ and likewise Σ^k, Σ^v are the per-head cross-covariances; each inner maximum is Eq. (6), whose value is a nuclear norm, so eliminating R and S leaves a linear assignment in π alone:

$$\textbf{R-step:} \quad \max_{R_h \in O(d_k)} \langle \Sigma^q_{\pi(h)h} + \Sigma^k_{\pi(h)h}, R_h \rangle = \| \Sigma^q_{\pi(h)h} + \Sigma^k_{\pi(h)h} \|_*, \quad \max_{S_h \in O(d_v)} \langle \Sigma^v_{\pi(h)h}, S_h \rangle = \| \Sigma^v_{\pi(h)h} \|_*; \tag{8}$$

$$\textbf{T-step:} \quad \pi^\star = \arg\max_{\pi \in \mathfrak{S}_H} \sum_h C_{\pi(h)h}, \qquad C_{ab} = \| \Sigma^q_{ab} + \Sigma^k_{ab} \|_* + \| \Sigma^v_{ab} \|_*. \tag{9}$$

Hungarian solves Eq. (9) exactly in $O(H^3)$ and Eq. (8) then rotates each matched pair from its own block, in one pass—no alternation, no initialization, no local optimum, no fractional plan to round—and because Eq. (9) is scored at exactly the value a matched pair attains, $\pi^\star$ is the exact optimizer of Eq. (7) itself. The separable $\|\Sigma^q\|_* + \|\Sigma^k\|_* + \|\Sigma^v\|_*$ is an upper bound (tight iff Σ^q, Σ^k share a polar factor, $\Sigma^q = UP$, $\Sigma^k = UQ$ with $P, Q \succeq 0$) that changes 4 of 12 ViT-B assignments and moves the barrier by at most 0.07; we report it as an ablation of a relaxation we do not make (Appendix C.2). Restricting R_h, S_h to permutations recovers a hard-assignment *vertex* of the unregularized Birkhoff LP that OT fusion (Singh & Jaggi, 2020; Imfeld et al., 2024) relaxes—not entropically regularized OT, whose plans are interior to $\mathcal{B}_d$—and fixing $\pi = \mathrm{id}$ recovers rotation-only fusion's set (Zhang et al., 2025): RaOT's feasible set contains these restrictions, though the complete methods differ in estimation, regularization, normalization and residual handling. Figure 1b collects the whole assignment—every parameter group, the map it receives, and the reason.

Extensions (approximate, deferred). Relaxing the marginals with KL penalties (Chizat et al., 2018) flags *weakly matched* units by low transported mass, merged with a *utilization weight* $w_i \propto (u_i^A)^p$ rather than averaged—explicit hyperparameters, not automatic task-specificity detection—and a rectangular Stiefel Procrustes extends the step to different widths. Both are approximate rather than exact symmetries; their formulation, caveats and scope live in Appendix E.6.

3.5 Compatibility with Layer Normalization

LayerNorm (Ba et al., 2016) centers and rescales each token, so it might seem to obstruct rotation-aware alignment; the opposite holds—it does not forbid rotations but *prescribes* them, and excludes

soft plans on residual streams exactly. Normalization projects every token onto one sphere in the centered hyperplane; a mean-preserving rotation maps that sphere onto itself and commutes with the projection, so readers see an exactly rotated code, where a soft plan shrinks different tokens by different amounts.

With the centering projector $C = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$, write $\mathrm{LN}(x) = \mathrm{diag}(\gamma)\, n(x) + \beta$ with pure normalization $n(x) = \sqrt{d}\, Cx/\|Cx\|_2$. The elementwise affine parameters absorb exactly into the adjacent linear maps, $W\,\mathrm{LN}(x) + b = \tilde{W} n(x) + \tilde{b}$ with $\tilde{W} = W\mathrm{diag}(\gamma)$, $\tilde{b} = W\beta + b$; we henceforth assume this (function-preserving) reparameterization and emit the merged model in the canonical gauge $\gamma = \mathbf{1}$, $\beta = \mathbf{0}$. In pre-LN Transformers, the residual stream is read *only* through n, so the admissible alignment maps on the stream are those that intertwine with n.

Proposition 2 (LN-compatible maps, idealized ϵ=0 normalization)**.** *For an invertible linear G, there exists a linear G' with $n(Gx) = G'n(x)$ for all x with $Cx \neq 0$ if and only if $G = \alpha R + \mathbf{1}w^\top$ with $\alpha > 0$, $w \in \mathbb{R}^d$ such that $\alpha + w^\top R^\top \mathbf{1} \neq 0$, and $R \in O(d)_{\pm\mathbf{1}} := \{R \in O(d) : R\mathbf{1} = \pm\mathbf{1}\}$; in that case $G' = R$ on $\mathbf{1}^\perp$ (its action on $\mathrm{span}(\mathbf{1})$, which n does not see, we fix to R's).*

Proposition 2 identifies the stream's exact symmetry set as $\mathbb{R}_{>0} \times O(d)_{\pm\mathbf{1}}$ up to a mean gauge: a global scale α, invisible to every reader, a sign on the mean axis we fix positive as a gauge, and the *mean-preserving rotations* $O(d)_{\mathbf{1}} := \{R : R\mathbf{1} = \mathbf{1}\} \cong O(d-1)$ that RaOT aligns in, of which permutations ($P\mathbf{1} = \mathbf{1}$) are a measure-zero subset. The scale half is a property of the idealized normalizer: implementations divide by $\sqrt{\mathrm{var}(x) + \epsilon}$ and $n_\epsilon(\alpha x) \neq n_\epsilon(x)$ once $\epsilon > 0$. What survives is exactly the part RaOT uses: for $R \in O(d)_{\mathbf{1}}$, $\|CRx\| = \|Cx\|$, so $n_\epsilon(Rx) = Rn_\epsilon(x)$ for every $\epsilon \geq 0$. The mean-preserving *orthogonal* subgroup at α=1 is thus an exact symmetry of the implemented layer, and the only stream map any reported number applies; α is a gauge we fix, not a transform we exploit. Moreover, LayerNorm *sharpens* the representation gap of Proposition 1:

Proposition 3 (Sharpened gap under LN)**.** *A doubly stochastic T preserves the domain of n and satisfies $n(Tx) = Tn(x)$ on it iff T is a permutation. Otherwise, for x with $Cx \neq 0$ and $CTx \neq 0$, $n(Tx) = \rho(x)\, Tn(x)$ with a purely radial gain $\rho(x) = \|Cx\|/\|CTx\| \geq 1$ that is constant—and then absorbable by a linear reader—only when T acts on $\mathbf{1}^\perp$ as a scaled isometry, and token-dependent otherwise (its upper end, the $\epsilon > 0$ form, and the singular case are in Appendix A).*

Soft plans thus distort the stream by a per-token gain that downstream linear readers cannot undo unless it is constant (which, for the idealized normalizer, needs T to act on the centered subspace as a scaled isometry, as $aI + (1-a)J$ does; with $\epsilon > 0$ even that case varies with token variance), whereas mean-preserving rotations are *exact*. The scope is narrow: only *soft, non-permutation* plans reparameterizing a LayerNorm stream; methods that round to a permutation, rescale per channel, or never place a plan on the stream fall outside Proposition 3. It is one reason—not demonstrably the only one—why OT fusions that soft-fuse the stream benefit from fine-tuning (Imfeld et al., 2024). Theory then prescribes our interface assignment: **residual streams get constrained Procrustes, not a soft plan.** The constrained problem retains a closed form: parameterizing $R = \frac{1}{d}\mathbf{1}\mathbf{1}^\top + VQV^\top$ with V an orthonormal basis of $\mathbf{1}^\perp$ and $Q \in O(d-1)$,

$$R^\star = \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top + VU_cV_c^\top V^\top, \qquad V^\top \Sigma_c V = U_c \Lambda_c V_c^\top, \tag{10}$$

where $\Sigma_c = C\bar{H}_A^\top \bar{H}_B C$ is the feature-centered cross-covariance. In practice we feature-center activations and estimate a *single* R_{res} shared by all blocks from the pooled Σ_c, applied to every parameter that writes to or reads from the stream—which is what makes it an exact reparameterization of the whole ViT (writer/reader enumeration, caveats, and RMSNorm/post-LN variants in Appendix C).

3.6 Loss-Barrier Analysis

Let $\mathcal{B}(\lambda) = L(W^\star) - [\lambda L(W_A) + (1-\lambda)L(\hat{W}_B)]$ denote the barrier of the merged model.

Theorem 1 (Smoothness bound)**.** *If the loss L is β-smooth on the segment $[W_A, \hat{W}_B]$, then*

$$\mathcal{B}(\lambda) \;\leq\; \tfrac{\beta}{2}\,\lambda(1-\lambda)\,\big\|W_A - \hat{W}_B\big\|^2. \tag{11}$$

Theorem 1 reduces barrier control to control of the post-alignment weight distance $\|W_A - \hat{W}_B\|$.

Proposition 4 (Oracle set-inclusion). *Write* $d(G) = \sum_{W \in \mathcal{W}_G} \|W_A - G(W_B)\|_F^2$ *for the weight distance of the tensors a rotation-interface map* G *touches. Since hard permutations are orthogonal,* $\mathcal{P}_d \subset O(d)$*, so*

$$\min_{G \in O(d)} d(G) \;\le\; \min_{G \in \mathcal{P}_d} d(G). \tag{12}$$

This is an *oracle* statement: our estimators do not solve it—the assignment and Procrustes steps maximize agreement with the *activation* cross-covariance, a surrogate—so we claim only the guarantee the gate can enforce.

Weight-distance gating and its guarantee. An estimated map need not reduce d—with a shared initialization it usually increases it—so we let identity compete against every estimate. At interface ℓ let the candidate set be $\mathcal{G}^{(\ell)} \supseteq \{I\}$ (identity, the estimated rotation, and, where one is estimated, the permutation). Holding all other interface maps fixed, we select

$$G^{(\ell)} \leftarrow \arg\min_{G \in \mathcal{G}^{(\ell)}} d_\ell(G), \qquad d_\ell(G) = \sum_{W \in \mathcal{W}_G} \left\| W_A - \widehat{(W_B)}_G \right\|_F^2, \tag{13}$$

where $\widehat{(W_B)}_G$ is the *realized* pull-back (Eq. (2)). Sweeping interfaces in a fixed order (per-head rotations, then FFN, then the residual stream) makes Eq. (13) block-coordinate descent on the total $\|W_A - \hat{W}_B\|^2$: because identity is always a candidate ($I \in \mathcal{G}^{(\ell)}$), every accepted map weakly decreases the total. The head *correspondence* is applied ungated, so the sweep alone certifies nothing against the unaligned parent; a final comparison of the complete transform against identity (Algorithm 1) does, giving the guarantee

$$\|W_A - \hat{W}_B\| \le \|W_A - W_B\| \quad \text{(direct averaging)}, \tag{14}$$

Equation (14) is unconditional. The barrier-*bound* comparison it suggests needs one hypothesis more, since Theorem 1 instantiates a β per segment: when L is β-smooth with a *common* constant on a region containing both segments (their convex hull suffices), Eq. (14) makes the merged bound no worse than direct averaging's. The bound is still not the barrier: β is unknown and Eq. (11) is loose, so cutting the distance need not cut the *measured* barrier—in our own ladder RaOT-safe trails ungated RaOT-align at all three scales despite accepting only distance-cutting maps (Table 1), a surrogate gap whose cause these measurements do not identify. Every barrier we report is measured, never inferred; the guarantee is monotonicity along *this* path from identity, not a win over any other descent (Appendix C). The rule needs no labels and no extra forward passes; on shared-initialization parents it rejects everything (Table 8). The switch between the two operating points is *provenance, not a measurement*: known-independent parents get RaOT-align; anything else—checkpoints branched from one ancestor, near basins—gets RaOT-safe; the costs are asymmetric: ungated alignment of shared-initialization parents collapses the merge (Table 7), gating independently trained ones costs 0.38–0.68 barrier on the weight criterion and nothing on the held-out one (Table 9). A third operating point, *RaOT-auto*, reads the regime off the weight gate's acceptance statistics (Table 6) and then ranks candidates *within* the alignment regime by a held-out activation criterion; the stages are not symmetric—the second accepts *every* candidate on independent parents, and without the first it rotates one shared-initialization candidate in ten and lands *below* its base rule (65.9 against 66.6)—so the composition buys an automatic regime call, not a better within-regime rule, and RaOT-safe stays the default (Appendices E.3 and E.4).

Layerwise Lipschitz continuity makes Eq. (7)'s objective—the activation discrepancy—a control on the aligned models' functional deviation, so minimizing it is a data-driven surrogate for the barrier that the gate converts into a certified non-increase of the aligned weight distance (and, under the shared-β hypothesis, of Eq. (11)), without retraining (contrast Imfeld et al., 2024).

4 Experiments

Our experiments are six questions (Q1–Q6), mechanism to application. None is a leaderboard: each tests a prediction of the symmetry analysis, including that alignment should decline to act.

Protocol. Independently trained parents (primary): the AugReg and DeiT ImageNet-1k ViT-Ti/S/B releases (Dosovitskiy et al., 2021; Steiner et al., 2022; Touvron et al., 2021), merged on the full

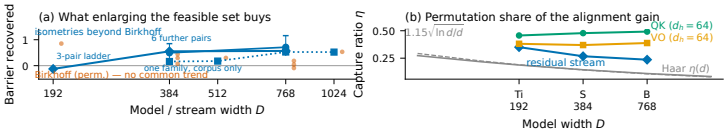


Figure 2. **What enlarging the feasible set buys, across three independently constructed series.** *(a)* blue: what the isometries' estimator adds beyond the permutation's at one calibration, on the three-pair AugReg/DeiT ladder, on six further public pairs (mean and range per width) and on a one-family four-width corpus series—rising with width in both series, the six pairs pointing the same way. Orange: the permutation estimator's own gain, drawn as points because they form no trend; we claim the blue half only. *(b)* capture ratio η against the Haar-random reference (Tables 14, 16, 18 and 19).

Table 1. **Independently trained ViTs, no fine-tuning.** AugReg and DeiT releases of ViT-Ti/S/B; parents reach 75.4/72.1, 78.8/79.8, 79.1/81.9. Barrier $\max_\lambda \mathcal{B}(\lambda)$ (primary) and midpoint top-1 (secondary) on the *full* 50k val split; transforms from 64 unlabeled *training* images, none from the eval split (Appendix E.19). *Controls* are Birkhoff *vertices* at a published method's interfaces, not reproductions; *released* is the authors' code (Appendix E.9). Bold is the best exact rung at 64 images; the last RaOT row swaps only the stream estimator to weights; RaOT-auto (= RaOT-align at 32 images) is in Table 9. †*Diagnostic* (every row above it is an exact reparameterization): entropic plans leave the transported parent at chance, so its barrier does not measure interpolation between the parents; scored against the *original* parents it reads 6.09/6.07/6.11 with midpoints at chance, and neither ranks against the rows above (Appendix E.9).

	ViT-Ti		ViT-S		ViT-B	
Method	Barrier	Acc. (%)	Barrier	Acc. (%)	Barrier	Acc. (%)
Direct averaging	7.34	0.11	6.44	0.13	6.29	0.14
Hard-assignment control, FFN units (OTFusion interfaces)	7.06	0.15	6.31	0.13	6.21	0.19
Hard-assignment control + matched heads (augmented)	**6.48**	0.14	6.14	0.18	6.10	0.12
Orth.-matching control, weights (rotation fusion's set)	6.99	0.11	6.42	0.11	6.27	0.14
Rotation fusion (Zhang et al., 2025), released code	6.87	0.11	6.34	0.10	6.28	0.10
RaOT-act (no residual rotation)	7.01	0.15	6.26	0.11	6.17	0.26
RaOT-safe (weight gate)	6.99	0.15	6.02	0.35	6.21	0.32
RaOT-align (ungated)	6.61	0.29	5.59	1.72	5.53	3.81
stream map from weights (Appendix B)	6.60	**0.47**	**5.25**	**3.02**	**4.64**	**7.99**
Transformer OT (Imfeld et al., 2024), released†	*not a reparameterization—endpoint, not barrier:*					
transported endpoint $\lvert\Delta L\rvert$, top-1 (%)	5.72	0.10	6.05	0.10	6.11	0.10

validation split against feasible-set controls and the released implementations of rotation fusion and Transformer OT (Appendices D and E.9). Shared initialization (secondary): eight fine-tuned CLIP (Radford et al., 2021) ViT-B/32 and ViT-L/14 models under one protocol. **Every ImageNet rung Table 22 lists is measured twice**, barrier and frozen-feature probe accuracy; the component-isolation rungs are measured by barrier alone.

Q1: Can transport plans express Transformer symmetries? On planted ground truth, no: a planted attention rotation is recovered to error 0.00 from 32 images where matching in $\mathcal{B}_d$ cannot (0.94), and a composed rotation-plus-permutation perturbation orders the barrier RaOT<perm<none (0.00<1.04<2.04; Appendix E.1).

Q2: Do independently trained Transformers differ by rotations, not permutations? Before any merging, the permutation capture ratio η (the share of the rotation optimum a permutation attains) decays where dimension grows and is flat at the fixed-dimension head interfaces (Fig. 2b), staying above—not approaching—the Haar-random reference $O(\sqrt{\ln d/d})$, whose finite-width value Appendix E.13 calibrates (Proposition 5). Subspace- against unit-level agreement concurs: net of the stronger correspondence-free null the rotational excess is 2.4–6.0× the permutational at every interface and scale (Appendix E.15 and Table 20). Byte-level *language* Transformers concur (rotational agreement 2.2–2.3× the permutational, no width decline; Appendix E.16), so the mismatch is architectural, not visual.

Q3: Does rotation-aware alignment improve merging geometry? Table 1 compares matched rungs of one alignment family on Q2's pairs: at 64 images the matched-heads permutation control has the lower ViT-Ti barrier and RaOT-align takes both columns at ViT-S/B; swapping only the

stream estimator to weights lowers the S/B barrier further ($5.25/4.64$) and doubles the midpoint ($3.02/7.99\%$), a hybrid, not either source, wins on these pairs. Enlarging the feasible set at a fixed estimator yields the predicted crossover (Fig. 2a): what the permutation estimator removes collapses while what the isometries' estimator adds beyond it grows—estimator contrasts, not symmetry components (numbers in Appendix E.12). Six further pairs and a one-family four-width corpus series separate the two trends: the isometries' term is positive on six of six and grows $0.16 \rightarrow 0.52$ with width, while the polytope term straddles zero across pairs and *grows* on same-recipe parents (Tables 14, 16 and 19). We claim the isometries half only, on recipe- or corpus-different parents (Appendices E.11 and E.12). Near-chance midpoints are the state of training-free merging, not one method's failure; REPAIR (Jordan et al., 2023) does not fix them, its statistical correction neutral or harmful in pre-LN ViTs (Appendix E.15).

Table 2. **Retained-head compatibility, with and without trunk merging.** Four ViT-S on disjoint CIFAR-100 tasks, DC-Merge backbone (α validation-selected); mean$\pm$SD over four draws (Appendix E.17). *Anchor+heads* leaves the anchor's backbone *unmerged* and transports only the other parents' heads into its coordinates; every row there shares that backbone and scores 55.3 ± 3.7 on the anchor's own task, so the block varies in the alignment alone. Anchor-only reuse, no head transport, scores 16.7 (3.9 on *Others*).

	Merged trunk		Anchor+heads (one backbone)	
Alignment	All	Others	All	Others
Direct averaging	16.2 ± 1.1	4.0 ± 0.5	16.7 ± 1.0	3.9 ± 0.5
Permutation	19.9 ± 0.9	8.7 ± 1.0	19.9 ± 1.1	8.1 ± 1.0
Rotation fusion (our port, Appendix E.9)	16.4 ± 0.9	4.1 ± 0.5	16.7 ± 1.0	3.9 ± 0.5
RaOT-align (ours)	$\mathbf{23.8 \pm 1.8}$	$\mathbf{13.8 \pm 1.9}$	$\mathbf{23.7 \pm 1.9}$	$\mathbf{13.2 \pm 1.7}$
RaOT-auto (ours)	23.1 ± 1.9	13.0 ± 1.9	23.2 ± 1.7	12.5 ± 1.7

Q4: Does the geometry correspond to better representations? The barrier ordering reappears in frozen-feature accuracy at ViT-S/B, RaOT's better point recovering $58\%/55\%$ of the floor-to-parent gap against every baseline's less, (ViT-Ti within 0.3 points; EuroSAT and GTSRB reproduce it, Table 22 and Appendix E.18). Better, not usable (51–52 at ViT-B against $77.9/81.5$), and a different question, since a probe retrains a head: what improves is the backbone as *feature extractor*.

Q5: When should alignment not be used? Where parents share an initialization identity is the right alignment and the gate says so, rejecting all $2{,}400$ and $6{,}336$ candidates: trailing interference resolution is predicted, not failure (Table 11). Forced open, nothing else changed, the merge falls $65.0 \rightarrow 57.4$ under averaging and $66.8 \rightarrow 4.5$ under task arithmetic (Table 7): alignment unable to decline to act is destructive. Two limits: the gate contributes avoided loss, not accuracy, and its surrogate is decisive *between* regimes ($\approx 5\times$ margins, no overlap) yet cannot rank *within* one (Table 6 and Appendix E.3).

Q6: Can alignment improve multi-task consolidation with retained task heads? The remaining corner—different tasks *and* initializations—is unoccupied: interference resolution presupposes coordinates such parents lack. We construct it (Table 2); every draw orders alike: unaligned and weight-estimated rows at or *below* anchor-only reuse, the permutation above, RaOT clear by 7.0. Alignment converts into *accuracy* here, not barrier, and survives seven label cuts (lead $7.0 \rightarrow 4.7/4.6 \rightarrow 3.6$, Appendix E.17). The cell is opened, not solved (23.8 vs. parents' 55.5), and the margin is head transport: merging the trunk adds nothing over leaving the anchor's backbone alone (23.8 against 23.7). In the *Anchor+heads* block, which fixes the backbone and varies only the alignment, *Others* runs $3.9 \rightarrow 8.1 \rightarrow \mathbf{13.2}$ (averaging and rotation fusion $\rightarrow$ permutation $\rightarrow$ RaOT): $3.4\times$ over no transport, $1.6\times$ over the polytope. Retraining heads on frozen features erases the ordering (Appendices E.17 and E.18); the residual map buys compatibility here, not a profitable merge.

5 Discussion and Conclusion

Scope (Appendix E.20): vanilla attention and additive positions; no usable accuracy on independent ImageNet ViTs, only frozen-feature transfer and trunk–head compatibility; the crossover's isometries half alone, on recipe- or corpus-different parents. Ship RaOT-safe unless parents are known independent; align in the isometries, not the full symmetry group.

Ethics Statement

This work studies a training-free procedure for combining networks that have already been trained: it collects no new data, involves no human subjects and releases no new model, and every parent and dataset is a public research release used as distributed (ImageNet-1k is known to contain images of people collected without individual consent, a property our use inherits rather than addresses). A merged model carries the biases and failure modes of every parent, possibly in combinations no parent exhibits alone; we measure loss barriers and accuracy only, audit no merged model on any subgroup, and nothing here should be read as evidence that a merge is safe to deploy. The paper's central negative result is itself the safety-relevant one: on independently trained ImageNet parents no method we test reaches usable accuracy, and a merge can collapse to near chance while the weight-space quantities a practitioner would inspect look unremarkable (Tables 6 and 7), which is why we recommend the conservative operating point whenever parent provenance is unknown.

Reproducibility Statement

Every number comes from a logged run of the accompanying code. Parents are public `timm` releases (Appendix E.10) or trained by recipes given in full (Appendices E.11, E.16 and E.17); the width study was pre-registered (the registration document, `runs/WIDTH_SCALING_PREREG.md`, ships with the code). The alignment ladder and gate are specified in Section 3 and Appendix C.2, calibration in Appendix E.19. Every run machine-checks endpoint exactness ($|\Delta L| \leq 1.4 \times 10^{-5}$ on ImageNet, $\leq 6.8 \times 10^{-5}$ elsewhere) and reproduces the gate-acceptance counts. All constants, seeds, software versions and checkpoint hashes are in the supplementary material.

Large Language Model Usage Disclosure

Large language models were used substantially, under the authors' direction and review: in implementing the estimators, ablation rungs and evaluation pipelines and in designing parts of the protocol (among them the component-isolation rungs and the cross-validated activation gate); in analyzing results; in parts of the derivations of the theoretical results, every proposition, theorem and proof of which was checked by the authors; and in drafting and revising the text and figures. Every reported number comes from logged runs of the released code, not from LLM output. The authors reviewed all LLM contributions and take full responsibility for the contents of this submission.

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Persi Diaconis and Mehrdad Shahshahani. On the eigenvalues of random matrices. *Journal of Applied Probability*, 31A:49–62, 1994.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Luyang Fang, Ehsan Latif, Haoran Lu, Yifan Zhou, Ping Ma, and Xiaoming Zhai. Efficient multi-task inferencing: Model merging with gromov-wasserstein feature alignment. In *International Conference on Artificial Intelligence in Education (AIED)*, 2025.

Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, and Aymeric Dieuleveut. Proving linear mode connectivity of neural networks via optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020.

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

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

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

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

Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. Transformer fusion with optimal transport. In *International Conference on Learning Representations (ICLR)*, 2024.

Tiefeng Jiang. Maxima of entries of Haar distributed matrices. *Probability Theory and Related Fields*, 131(1):121–144, 2005.

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

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations (ICLR)*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Matt Mahoney. Large text compression benchmark. http://mattmahoney.net/dc/text.html, 2011.

Michael S. Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Dang Nguyen, Trang Nguyen, Khai Nguyen, Dinh Phung, Hung Bui, and Nhat Ho. On cross-layer alignment for model fusion of heterogeneous neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research (TMLR)*, 2022.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Zipit! merging models from different tasks without training. In *International Conference on Learning Representations (ICLR)*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision (ECCV)*, 2022.

Viet-Hoang Tran, Vinh Khanh Bui, Van-Hoan Trinh, Tan Lai Ngoc, and Tan M. Nguyen. Functional equivalence in attention: A comprehensive study with applications to linear mode connectivity. In *International Conference on Machine Learning (ICML)*, 2026.

Ross Wightman. Pytorch image models. https://github.com/huggingface/pytorch-image-models, 2019.

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

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

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

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

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications and opportunities. *ACM Computing Surveys*, 58(8):216:1–216:41, 2026. doi: 10.1145/3787849.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024.

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *International Conference on Machine Learning (ICML)*, 2025.

Han-Chen Zhang, Zi-Hao Zhou, Mao-Lin Luo, Shimin Di, Min-Ling Zhang, and Tong Wei. DC-merge: Improving model merging with directional consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

How to read this appendix

This appendix is organized so that each claim in the main text has one place that carries it, together with the controls that would falsify it. It is not written to be read front to back, and nothing in it is needed to follow Sections 3 and 4: every part is pointed to from the sentence it supports.

If you read only what the claims rest on, five places suffice. Appendix A proves the two structural facts the paper turns on ($O(d) \cap \mathcal{B}_d = \mathcal{P}_d$, and the LayerNorm stream's mean-preserving rotations). Appendix C.2 establishes that the coupled head-and-rotation step is solved exactly, and names the one place it relaxes. Appendix E.10 specifies the independently trained parents behind the main ladder, and Appendix E.9 the baselines it is read against: rows marked *(port)* are our re-implementations sharing a published method's feasible set, not runs of its released code; the released implementations themselves are run there and in Table 1. Appendix E.11 is the same-recipe control and the pre-registered width study, which together scope the crossover. Appendix E.17 is the regime the paper argues is its sharpest application test.

Everything else is one of three things. Protocol and reproduction detail, answering *how* a number was produced (Appendices B, E.7, E.8 and E.19). Controls, each written against one objection we expect: that the ordering is a checkpoint-family artifact, that the gate is doing nothing, that a permutation would have sufficed, that the effect is an alignment of convenience rather than one a merge can use (Appendices E.1 to E.3, E.5 and E.12 to E.15). And boundary results that limit the claim rather than support it—the shared-initialization regime where alignment has nothing to find, the intermediate band where the regime test narrows, the language stream where the width trend is absent, the probe showing no representational gain, and the coverage list (Appendices E.4, E.16, E.18 and E.20).

This appendix is deliberately the short form. Each subsection keeps the tables the main text cites and the paragraph that reads them, and stops there; **its protocol, derivations, controls and discussion are in the supplementary material under the subsection's own title**, together with the further tables: the per-rung planted-rotation recovery, the permutation-against-rotation explained-variance breakdown, the separable-against-exact head-cost audit, the RegMean and scaling-coefficient curves, the per-interface language-model geometry and its per-cell ladder, the per-task downstream probe grid, the per-construction different-tasks table and the full merge sweep, the calibration source-and-size grid, and the per-pair intermediate-regime figure and table. All of them are in the supplementary material; none is a claim resting on itself.

A Proofs

Proof of Proposition 1. Let $T \in \mathcal{B}_d \cap O(d)$ and let p be any row of T. Double stochasticity gives $p \geq 0$ and $\sum_i p_i = 1$, hence $\|p\|_2^2 = \sum_i p_i^2 \leq \left(\sum_i p_i\right) \max_i p_i \leq 1$, with equality iff p is a standard basis vector. Orthogonality of T forces $\|p\|_2 = 1$ for every row, so every row is a standard basis vector and T is a permutation matrix. The converse inclusion is immediate. □

Proof of Proposition 2. (*If.*) Since $C\mathbf{1} = 0$ and R preserves $\mathrm{span}(\mathbf{1})$ and hence its orthocomplement, $CG = \alpha CR = \alpha RC$, so $n(Gx) = \sqrt{d}\,\alpha RCx/\|\alpha RCx\| = Rn(x)$. (*Only if.*) Since $n(x + c\mathbf{1}) = n(x)$ for every c, the assumed intertwining implies $n(Gx + cG\mathbf{1}) = G'n(x + c\mathbf{1}) = G'n(x)$ is independent of c. With G invertible this is possible at every x only if $CG\mathbf{1} = 0$: otherwise the direction of $C(Gx + cG\mathbf{1})$ varies with c (as $c \to \infty$ it tends to that of $CG\mathbf{1}$, and where the two stay parallel it flips sign). Hence $CG = CGC$ factors through $\mathbf{1}^\perp$. On the centered unit sphere the intertwining reads $G'v = CGv/\|CGv\|$; the left side is linear in v and so is CGv, so $\|CGv\|$ must be constant, $= \alpha > 0$, and $G'|_{\mathbf{1}^\perp} = CG|_{\mathbf{1}^\perp}/\alpha$ maps unit vectors to unit vectors—an isometry R of $\mathbf{1}^\perp$. Thus $CG|_{\mathbf{1}^\perp} = \alpha R$ while $(I - C)G = \mathbf{1}w^\top$ for some $w \in \mathbb{R}^d$, which is the stated form, the action on $\mathrm{span}(\mathbf{1})$ being the free part the statement fixes. □

Proof of Proposition 3. Double stochasticity gives $CT = TC$, so equivariance is equivalent to $\|TCx\| = \|Cx\|$ for all x, i.e., T is an isometry on $\mathbf{1}^\perp$; with $T\mathbf{1} = \mathbf{1}$ this makes T orthogonal, and Proposition 1 yields $T \in \mathcal{P}_d$. The identity follows since $CTx = TCx$ makes $n(Tx)$ and $Tn(x)$ parallel, differing only by the stated ratio; doubly stochastic maps are contractions on $\mathbf{1}^\perp$, so $\rho \geq 1$. The gain is bounded above by the inverse least singular value on the centered subspace,

$\rho(x) \in [1, \sigma_{\min}(T|_{\mathbf{1}^\perp})^{-1}]$; where $T|_{\mathbf{1}^\perp}$ is singular that upper end is $+\infty$ and ρ is undefined on its kernel, on which the plan destroys the token outright. With the implemented normalizer ($\epsilon > 0$) the same statement holds with $\rho_\epsilon(x) = \sqrt{\mathrm{var}(x) + \epsilon}/\sqrt{\mathrm{var}(Tx) + \epsilon}$, which is bounded by $\rho(x)$ and equals it as $\epsilon \to 0$, so the distortion is not an artifact of the idealization. For the idealized normalizer the gain is constant exactly when T acts on $\mathbf{1}^\perp$ as a scaled isometry—$T = aI + (1-a)J$ with $0 < a < 1$ is one such map, giving $\rho \equiv 1/a$—and a constant gain is absorbed by any linear reader; with $\epsilon > 0$ even that case gives $\rho_\epsilon(x) = \sqrt{(v+\epsilon)/(a^2 v + \epsilon)}$ with $v = \mathrm{var}(x)$, which varies with the token, so the exception is an idealization; what Proposition 3 names as the distortion is the token dependence, which such T lack and a generic doubly stochastic T has. □

Proof of Theorem 1. Apply the β-smoothness inequality $L(\lambda x + (1-\lambda)y) \le \lambda L(x) + (1-\lambda)L(y) + \frac{\beta}{2}\lambda(1-\lambda)\|x - y\|^2$ with $x = W_A$, $y = \hat{W}_B$. □

B THE RUNGS, IN ONE PLACE

Every row of every ladder in this paper is one of the sixteen rungs below, all of them the same code path with components switched on or off (`src/align.py`). They share the estimator, the calibration set and the interface list; what differs is the feasible set searched, which interfaces are touched, and whether a gate rules on the result. All sixteen are exact reparameterizations at both endpoints ($|\Delta L| \le 1.4 \times 10^{-5}$ on the ImageNet ladders, $\le 6.8 \times 10^{-5}$ on the CIFAR and language ones). `raot_ffn` serves two roles—OTFusion's port and the FFN-only isolation—and is listed once, so the table has sixteen entries for sixteen rungs.

Three are *operating points* a practitioner would ship—RaOT-align, RaOT-safe, RaOT-auto—and they differ only in the acceptance rule. The rest are controls, the component-isolation block using RaOT's estimators throughout.

Table 3. **The sixteen rungs.** *Est.* is the estimation source (act. = activation cross-covariance on the shared calibration images, wt. = weight cross-products). *Set* is the feasible set at the rotation-legal interfaces. *Res.* is whether the LayerNorm residual stream is transformed. The three shipped operating points are in **bold**; `perm` is Table 1's *Hard-assignment control + matched heads* row—Transformer OT's interfaces under hard assignment, with the head correspondence it omits, estimated as RaOT estimates it—and differs from RaOT-align in *Set* alone. [†]Head and FFN maps from weights; the stream rotation, like every original rung's, from activations. [‡]Stream rotation estimated from weights (estimator ablation below). [§]The held-out activation gate alone; the shipped RaOT-auto adds the stage-one regime call of Algorithm 1, which on the independently trained pairs declares *independent* and hands over to this rung, so the two coincide there and differ on shared-initialization parents (Appendix E.3).

Rung (code name)	Paper name	Est.	Set	Res.	Gate
`naive`	Direct averaging	—	—	no	—
`perm`	Transformer OT + head match.	act.	$\mathcal{P}_d$	perm	—
`transformer_ot`	Transformer OT (Imfeld et al., 2024) (port)	act.	$\mathcal{P}_d$	no	—
`raot_ffn`	OTFusion (Singh & Jaggi, 2020) (port)	act.	—	—	—
`rf_weight`	Rotation fusion (port)	wt.	$O(d)$	no	—
`rf_weight_res`	+ residual rotation	wt.[†]	$O(d)$	rot.	—
`raot_act`	RaOT-act	act.	$O(d)$	no	—
`raot_full`	**RaOT-align**	act.	$O(d)$	rot.	none
`raot_gated`	**RaOT-safe**	act.	$O(d)$	rot.	weight, Eq. (13)
`raot_gated_act`	**RaOT-auto**, stage two[§]	act.	$O(d)$	rot.	held-out act.
`raot_gated_both`	*control:* both	act.	$O(d)$	rot.	both
Component isolation (Appendix E.14); FFN only = `raot_ffn` above					
`raot_nohead`	no head matching	act.	$O(d)$	rot.	—
`raot_wres`	stream map from weights	act.[‡]	$O(d)$	rot.	—
`rf_weight_wres`	+ stream, all from weights	wt.[‡]	$O(d)$	rot.	—
`res_w`	stream map only, from weights	wt.[‡]	$O(d)$	rot.	—
`raot_heads`	heads + QK/VO only	act.	$O(d)$	no	—
`raot_res`	residual only	act.	—	rot.	—
Feasible-set ladder (Appendix E.5), value–output interface only					
`raot_vodiag`	+ scaling	act.	$\mathrm{diag} \cdot O$	rot.	—
`raot_vogl`	+ general linear	act.	GL	rot.	—

Two readings of the table are worth stating, because both answer questions the ladders are meant to settle but state across several tables.

Permutation against rotation, everything else fixed. `perm` and `raot_full` differ in the *Set* column and in nothing else—same activation estimator, same 64 images, same interfaces including the residual stream, same absence of a gate. Their difference is therefore the feasible set, not the method, and it is what Table 1's *Hard-assignment control + matched heads* and *RaOT-align* rows compare. `rf_weight` is the separate question of estimation *source*: it searches $O(d)$ like RaOT but from weights, and `rf_weight_res` adds the residual interface to it ($6.88/6.16/6.32$ barrier and $0.14/0.24/0.18\%$ at Ti/S/B, against `rf_weight`'s $6.99/6.42/6.27$). That residual map is estimated from activations, as in every rung of the original ladder that carries one, so `rf_weight_res` is weight-estimated head and FFN maps under the activation-estimated stream rotation, and the `rf_weight_res`-against-`raot_full` gap isolates the estimation source of the head and FFN maps together with the head correspondence, which `rf_weight_res` fixes by position where `raot_full` searches it; the two rungs share the stream estimator.

The stream estimator, isolated. Because every original rung estimates the stream rotation from activations, the ladder never compared estimators *on the stream*. We add that comparison: a weight-estimated stream rotation solves the same mean-preserving Procrustes problem with the objective built from the weights that read from and write to the stream—readers contribute $W_A^\top W_B$, writers $W_A W_B^\top$, token embeddings and writer biases their outer products, over exactly the parameters Appendix C enumerates—so no activation is read for that map. Three rungs use it: `res_w` (stream only), `raot_wres` (RaOT-align with *only* the stream estimator swapped) and `rf_weight_wres` (rotation fusion's weight-estimated heads and FFN under the weight-estimated stream, i.e. every map

from weights). All three are exact reparameterizations ($|\Delta L| \leq 10^{-3}$ at three decimals on every pair). Table 4 is the result, on Table 1's pairs, calibration and full evaluation split.

Table 4. **Swapping only the stream estimator.** Maximum loss barrier and midpoint top-1 (%) on the three-pair ImageNet ladder at 64 training-split calibration images; *act./wt.* is the estimator of the stream rotation, everything else as named. The hybrid—activation-estimated heads and FFN under a weight-estimated stream—is the best exact rung at ViT-S and ViT-B by a margin larger than any rung difference in Table 1; estimating every map from weights is not, and the weight-estimated stream alone is worse than direct averaging at Ti and only marginally below it at S/B, its midpoint at chance. The estimator comparison is therefore interface-dependent, and neither source is uniformly better.

		Barrier ↓			Midpoint acc. (%) ↑		
Rung	Stream est.	ViT-Ti	ViT-S	ViT-B	ViT-Ti	ViT-S	ViT-B
Direct averaging	—	7.335	6.436	6.287	0.11	0.13	0.14
RaOT-align (`raot_full`)	act.	6.609	5.589	5.534	0.29	1.72	3.81
stream from weights (`raot_wres`)	wt.	**6.601**	**5.248**	**4.640**	**0.47**	**3.02**	**7.99**
All maps from weights (`rf_weight_wres`)	wt.	7.317	6.399	6.167	0.15	0.12	0.13
Stream only, weights (`res_w`)	wt.	7.627	6.318	6.194	0.13	0.14	0.18
Stream only, activations (`raot_res`)	act.	7.560	6.117	6.489	0.10	0.26	0.27

Two readings survive and one does not. The stream's alignment signal is recoverable from weights, and on these pairs better than from activations once the maps beneath it are activation-estimated (0.34 and 0.89 barrier at ViT-S/B, midpoint doubled); the all-weights pipeline is near direct averaging, so the activation estimate still carries the head and FFN maps; and Table 1's original ordering is unchanged, since every original rung shares the activation stream estimator. What does not survive is any blanket statement about estimation source: we withdrew "activations, not weights" for the stream before this table existed, and it now reads the other way there. The weight estimate is good only *given* aligned readers and writers: alone (`res_w`) or over weight-estimated heads and FFN (`rf_weight_wres`) it is no better than direct averaging, because its objective sums $W_A^\top W_B$ over reader rows and $W_A W_B^\top$ over writer columns that only agree once the head and FFN maps have put them in correspondence. The same holds on the retained-head control: with activation-aligned heads and FFN beneath it the weight-estimated stream lifts *Others* to 14.6 ± 0.7 against the activation stream's 13.2 ± 1.7 (All 24.8 ± 0.7 against 23.7 ± 1.9), while alone it scores 4.6 ± 0.5, within noise of no transport (Appendix E.17). Why the weight estimate is the better one once that condition holds—every reader and writer of the stream enters at once, where 64 images sample the stream through 197×64 activation rows—we leave as a hypothesis this table does not test.

The three operating points differ in the acceptance rule and in the estimation budget that rule forces. RaOT-align applies every map, RaOT-safe those Eq. (13) accepts, RaOT-auto those a held-out activation criterion accepts once Eq. (13)'s aggregate statistics have established the regime. The first two estimate on all 64 calibration rows; the held-out criterion needs a split, so RaOT-auto refits on the first 32 and gates on the other 32—a budget difference Table 9 isolates, and the reason its row is not a like-for-like comparison of acceptance rules.

One map escapes Eq. (13) by construction: the head *correspondence* must be fixed before any per-head rotation can be scored against identity, so it is applied ungated, and the per-interface descent alone therefore does not establish Eq. (14). RaOT-safe closes this by letting identity compete with the *composite*: after the sweep it recomputes $\|W_A - \hat{W}_B\|^2$ over every parameter and returns the unaligned model if the composite did not cut it. The check is what makes Eq. (14) a property of the shipped rule rather than of a nearby one; it has not fired on any pair we report.

C Residual-Rotation Implementation, and $N > 2$

In practice we fix the gauge by feature-centering activations, and estimate a *single* R_{res} shared by all blocks (the stream is threaded through every skip connection) from the pooled $\Sigma_c = \sum_\ell \omega_\ell C \bar{H}_A^{(\ell)\top} \bar{H}_B^{(\ell)} C$ with $\omega_\ell = 1$ at every stream read point (an unweighted sum, no per-layer variance normalization; a common positive scale would leave the Procrustes optimizer unchanged).

Applying R_{res} consistently is what makes it an exact reparameterization of the whole ViT: *every* parameter that writes to or reads from the stream must be transformed by the same basis. Writers—patch embedding, positional embeddings, the CLS token, and each block's W_O and W_2—are multiplied by R_{res} on their output side. Readers—each block's $\tilde{W}_Q, \tilde{W}_K, \tilde{W}_V, \tilde{W}_1$ after γ-absorption, the final LayerNorm's absorbed head, and any external readout such as a CLIP visual projection—are multiplied by $R_{\text{res}}^\top$ on their input side. Two implementation caveats matter. First, the γ, β absorption of Proposition 2 is valid because each LayerNorm output feeds a single linear map along each path. When a normalized signal fans out to several branches, the same γ is absorbed into each reader consistently; this remains exact, but it must be applied to *all* of them. The β half carries its own assumption: $\tilde{b} = W\beta + b$ needs somewhere to live. This absorption is exact when every adjacent reader has a bias term, as in all architectures evaluated here (and the per-run endpoint check would catch a violation). For bias-free readers, it requires either $\beta = 0$ or an architecture-preserving mechanism for representing the induced constant offset. Second, some architectures normalize *before* the first block that writes directly into the stream (e.g. CLIP's `pre_layernorm`). That places a γ on the write side which does *not* commute with R_{res}, so the exact form requires a generalized writer $R_{\text{res}}\,\mathrm{diag}(\gamma)\,R_{\text{res}}^\top$. In the shared-initialization regime we simply set $R_{\text{res}} = I$ (Appendix E.2). For RMSNorm, which omits centering, the constraint disappears and the full $O(d)$ of Eq. (6) applies; for post-LN architectures Propositions 2 and 3 hold verbatim without the scale freedom.

C.1 Merging $N > 2$ Models

For N parents one anchor basis is fixed (the pre-trained release where one exists, otherwise a distinguished parent) and every other parent is aligned to it independently in a single pass from the shared calibration capture; the gate rules on each parent–anchor pair as at $N{=}2$, each endpoint is verified exact, and the aligned parents are averaged with uniform weights or handed to an interference-resolving backend. Re-aligning to the running merge would yield a symmetry-aware Fréchet barycenter; we use it in no reported number. RaOT needs one forward pass per model and no gradients; the rest is $O(d^3)$ Hungarian and $O(d_k^3)$ Procrustes per interface. Equation (14) is a monotonicity statement about this descent path from identity and neither certifies a win over a permutation-only descent nor reaches the global $\min_{O(d)} d$; the per-candidate comparison is clean only where interfaces decouple.

C.2 The Coupled Solver: Exactness, and the One Place It Relaxes

Section 3.4 eliminates R in closed form and hands what remains to the Hungarian algorithm, which is exact only because π is a permutation: under a fractional plan the R-step would Procrustes a plan-weighted mixture whose maximizer aligns no actual pair of heads. The head sets have equal size and uniform marginals, so the relaxation buys nothing the vertex solution does not already give, and no alternation is needed—which is the concrete sense in which the solver differs from the separate assignment and rotation steps we adopt from Zhang et al. (2025). A matched pair's query–key subspaces share one R_h, since $q{\cdot}k$ is invariant under a common rotation, while VO carries its own S_h, so the value a pair (a, h) attains is

$$V_{ah} = \left\|\Sigma^q_{ah} + \Sigma^k_{ah}\right\|_* + \left\|\Sigma^v_{ah}\right\|_*, \tag{15}$$

and evaluating Eq. (9) at V_{ah} makes $\pi^\star$ the exact optimizer of Eq. (7). The separable $\|\Sigma^q\|_* + \|\Sigma^k\|_* + \|\Sigma^v\|_*$ is an upper bound (equality iff Σ^q, Σ^k share a polar factor, $\Sigma^q = UP$, $\Sigma^k = UQ$ with one orthogonal U and $P, Q \succeq 0$) whose slack is confined to the QK half; it exceeds V_{ah} by 7–9%, changes 4 of 12 ViT-B assignments, forfeits under 0.7% of the exact objective on average, and moves the end-to-end barrier by at most 0.07 in either direction, so the exact cost removes a caveat rather than a number. Cost is not a design constraint: one alignment takes 4.6, 20.0 and 123.3 seconds at ViT-Ti/S/B. Exactness here is for the activation surrogate, not the barrier; nothing in the method certifies a smaller barrier, and the gate's guarantee is only Eq. (14).

D Transformation-Target Comparison to Rotation Fusion

Table 5 details, by parameter group, what the released implementation of Zhang et al. (2025) transforms versus RaOT. The reading is exact: their per-head rotations are estimated from weight cross-products $W^{(1)}_{Q,h} W^{(2)\top}_{Q,h} + W^{(1)}_{K,h} W^{(2)\top}_{K,h}$ (analogously for VO), applied inside the d_k/d_v head

dimensions with heads paired by *position*; the FFN is permuted (Hungarian on weight cross-products); and the embeddings, both LayerNorms, and the residual-stream basis are copied unchanged. This confirms that the interface-wise rotation/permutation split is shared prior art, while RaOT's residual-stream rotation, head matching, activation-driven estimation, and gate are the additions.

Table 5. Transformation targets: rotation fusion (Zhang et al., 2025) vs. RaOT, by parameter group, read from rotation fusion's released code. "rot." = orthogonal rotation, "perm." = permutation. Our rotation-fusion port (the *orth.-matching control*) applies the orthogonal matching and the weight-estimated FFN permutation only, without the released pipeline's bias terms and scaling stage; the released code itself, scaling stage included, is run in Table 1 (Appendix E.9).

Interface / parameter	Rotation fusion	RaOT (ours)
QK head subspace (d_k)	rot., weight	rot., activation
VO head subspace (d_v)	rot., weight	rot., activation
Head correspondence	fixed (positional)	OT, joint w/ rot.
FFN hidden units	perm., weight	perm., activation
Residual stream (d)	**none** (copied)	**mean-pres. rot.**
Embeddings / CLS / pos.	copied	rotated (R_{res})
LayerNorm	copied	symmetry derived
Estimation source	weights	activations
Applied when	always	gated (vs. identity)
Different widths	equal only	Stiefel (ext.)

E Empirical Verification

We verify the theoretical claims of this paper in controlled settings where every quantity can be measured exactly, then report preliminary larger-scale runs. All numbers are actual measurements from the released implementation, as are the main-paper Tables 1–4, which report the full-scale evaluation; Appendix E.20 states what that evaluation does and does not cover.

E.1 Controlled Rotation Recovery

A rotation planted in a trained CIFAR-100 ViT-Ti is recovered to alignment error 0.00 from 32 images by RaOT, where permutation matching cannot (0.94) and soft OT breaks the endpoint ($|\Delta L| = 2.03$, function error 0.53); on two independently trained ViTs a best-fit rotation removes 1.7–$2.5\times$ more activation mismatch than a best-fit permutation; and under a composed rotation-plus-permutation perturbation the barrier among exact rungs orders none $2.04 >$ permutation $1.04 >$ RaOT $\mathbf{0.00}$, RaOT recovering the composite exactly and merging losslessly. Means over three planting seeds.

E.2 Shared-Initialization Regime and the Barrier-Surrogate Gate

The complementary regime and the gating rule of Eq. (13), on the eight-task CLIP benchmark. The tables below carry the three facts the main text uses.

Table 6. Gate acceptance and the margin behind it. $r_g = d_{\text{aligned}}/d_{\text{identity}}$ is the per-candidate ratio Eq. (13) thresholds; a candidate is accepted iff $r_g < 1$. For the shared-initialization rows we report the range of *per-task, per-interface medians* and the smallest r_g observed anywhere, excluding interfaces where the Hungarian map is exactly the identity ($r_g = 1.0000$ by construction, a no-op: one such interface at B/32, four at L/14). For independently trained rows we report the full range over every candidate. Calibration is 16 images/task (CLIP; the alignments' budget and the one these counts are measured at, where the shipped merge's gate reads 64, Table 11) and 64 images (ImageNet), matching the runs each row describes, and every row is measured under the shipped configuration (exact tied-QK cost, training-split calibration). The trio rows subsume the pair rows: their AugReg/DeiT sub-blocks reproduce the ViT-S and ViT-B pair rows candidate for candidate ($61/156$ and $146/300$).

Regime	Parents	Cand.	Acc.	r_g
Shared init.	CLIP B/32, 8 tasks	2,400	0	med. 65–368; min 14.9
	CLIP L/14, 8 tasks	6,336	0	med. 2.7–844; min 5.4
Independent	ViT-Ti, AugReg/DeiT	84	54.8%	[0.971, 1.021]
	ViT-S, AugReg/DeiT	156	39.1%	[0.981, 1.035]
	ViT-B, AugReg/DeiT	300	48.7%	[0.985, 1.019]
	ViT-S trio (N=3)	312	45.2%	[0.981, 1.035]
	ViT-B trio (N=3)	600	46.5%	[0.985, 1.019]

Table 6 records the per-candidate ratio the gate thresholds, $r_g = d_{\text{aligned}}/d_{\text{identity}}$, at both regimes and all scales. *Whether to align at all* it decides unambiguously—all 996 independent-parent candidates in $[0.971,\ 1.035]$, no shared-initialization candidate below 5.4, medians 223 and 117, so any threshold in $[1.04,\ 5.38]$ gives the same verdict on all 9,732. *Which maps to keep once alignment applies* it decides at the noise floor (r_g within 4% of unity, acceptance pinned at 39–55%), which is why RaOT-safe trails RaOT-align and the ladder carries both.

Table 7. **What the gate is worth: three settings of it.** Eight-task CLIP, full test sets, per-task top-1 (%). Gated rows are the published ones. *Gate forced open* applies every candidate the gate rejects (2,400 at B/32, 6,336 at L/14); *activation criterion only* replaces Eq. (13) by the held-out activation criterion of this subsection and applies the 238 candidates *it* accepts (B/32; the L/14 counterpart is not run). Every other component is identical in all rows. Removing the gate degrades averaging (-7.6 at B/32, -1.5 at L/14) and collapses task-arithmetic scaling (-62.3 and -67.2), at B/32 with Cars and SUN397 at chance for 196 and 397 classes; the two-tier outcome is scale-independent and the collapse is the deeper of the two at L/14. Weakening the gate to the activation criterion does neither—and still costs the method its margin, putting task arithmetic *below* the 66.6 base rule it merges with.

Method	MNIST	SVHN	GTSRB	EuroSAT	RESISC45	DTD	Cars	SUN397	Avg.
ViT-B/32									
Zero-shot	49.3	26.3	30.8	43.6	55.2	42.3	58.6	61.8	46.0
Weight averaging	84.6	63.7	51.7	76.3	67.7	47.4	61.8	64.5	64.7
Task Arithmetic	95.5	77.2	66.6	75.2	62.6	44.9	54.4	56.6	66.6
RaOT, averaging (gated)	85.2	64.0	52.0	76.5	68.0	47.8	61.9	64.6	**65.0**
activation criterion only	84.8	64.4	51.8	75.8	67.9	47.9	62.0	64.5	64.9
gate forced open	79.7	54.6	43.1	69.6	56.7	44.0	56.9	54.7	57.4
RaOT, task arith. (gated)	95.8	77.6	67.3	75.1	62.6	45.3	54.2	56.5	**66.8**
activation criterion only	95.5	76.4	65.8	72.2	62.0	45.1	54.0	56.6	65.9
gate forced open	9.8	7.7	3.5	9.2	2.7	2.2	0.5	0.3	4.5
ViT-L/14									
Zero-shot	70.6	61.4	50.4	58.1	66.8	52.7	76.6	66.7	62.9
Weight averaging	95.6	82.8	73.8	88.0	78.7	59.9	80.3	71.3	78.8
Task Arithmetic	97.9	87.6	83.2	84.4	78.4	58.7	78.1	71.6	80.0
RaOT, averaging (gated)	95.6	83.0	73.8	88.1	79.0	59.9	80.3	71.3	**78.9**
gate forced open	95.9	80.4	72.0	82.5	78.4	59.3	79.7	70.6	77.3
RaOT, task arith. (gated)	97.8	87.9	83.6	84.6	79.0	59.4	78.2	71.5	**80.3**
gate forced open	8.5	11.2	6.9	25.0	18.0	22.7	2.0	10.0	13.0

Table 7 forces the gate open on the eight-task protocol with everything else fixed: averaging degrades gracefully ($65.0 \to 57.4$), task-arithmetic scaling collapses ($66.8 \to 4.5$; $80.3 \to 13.0$ at L/14),

and the ungated OT baselines land at the same chance level (5.6, 5.5). The mechanism is the r_g statistics—a forced candidate displaces weights by a median factor of 223, and task arithmetic sums eight such displacements at an effective scale $N\alpha = 2.4$ with nothing to contract them.

Table 8. **What each component contributes, shared-initialization CLIP-8** (average over eight *full* test sets, %). Each row is a base merging rule; columns add RaOT's components. *Gate only* sets the weighting exponent to p=0; at both scales it reproduces the base rule bit-exactly (max per-task difference 0), so the gate's accuracy contribution in this regime is exactly zero. *Gate + weighting* is p=1; the published Table 11 rows use p=4 and reach 67.06 and 80.87. *Gate forced open* is the counterfactual of Table 7, run at both scales. The gate's value is the last column, not the third.

Base merging rule	Base	Gate only (p=0)	Gate + weighting	Gate forced open
ViT-B/32				
Weight averaging	64.71	64.71 (+0.00)	64.98 (+0.27)	57.41 (−7.30)
Task arithmetic (α=0.3)	66.63	66.63 (+0.00)	66.79 (+0.16)	4.49 (−62.14)
ViT-L/14				
Weight averaging	78.80	78.80 (+0.00)	78.87 (+0.07)	77.35 (−1.45)
Task arithmetic (α=0.3)	80.00	80.00 (+0.00)	80.26 (+0.26)	13.05 (−66.95)

Table 8 separates the gate from the utilization weighting: with p=0 RaOT reproduces the base merging rule *bit-exactly* (64.71, 66.63; max per-task difference 0), so the +2.35 over weight averaging at B/32 is +1.92 merging rule, +0.43 weighting and +0.00 gate (+1.20/+0.87/+0.00 at L/14). What the gate is worth is the last column—−7.30 and −62.14 when forced open—a safety property invisible in exactly the table where the method looks good.

E.3 A Cross-Validated Activation Gate

The activation residual the transform minimizes is vacuous as an in-sample gate (it accepts everything in both regimes), so we evaluate it out of sample: fit on the first 32 of the 64 calibration images, keep a candidate only if it cuts the *centered* residual on the held-out 32, identity the per-candidate fallback. Synthetic controls fix its arithmetic (nothing accepted on identical networks; a planted residual rotation accepted in every run).

Table 9. **What the gate costs, by surrogate.** Maximum loss barrier and midpoint accuracy on the independently trained pairs, for the ungated operating point and the three gated ones. *Weight* is Eq. (13) as published (RaOT-safe); *Act.* is the cross-validated activation criterion of this subsection; *Both* requires a head or FFN candidate to pass the two; on the residual map it applies the activation criterion alone, the weight verdict being recorded but not enforced. Candidate counts here include the residual-stream map (85/157/301); Table 6 counts the head and FFN candidates only (84/156/300), the residual map being accepted by the weight gate at all three scales. The activation criterion attains the lowest barrier at all three scales—and the control row says why: because it estimates on half the calibration images, RaOT-align at that same halved budget is *the same merge*, to the last bit at ViT-Ti and ViT-B and to 6×10^{-7} at ViT-S. What the *Act.* row measures here is the estimation budget, not the gate: at a matched budget—fit on the same 64 images as RaOT-align, gate on 64 further disjoint training images—it again accepts every candidate (85/85, 157/157, 301/301) and its merge is RaOT-align's to the digit. The conjunction tracks the weight criterion, which is therefore the binding one. The hierarchical RaOT-auto of this subsection coincides with the *Act.* row here and with the base merging rule on shared initialization.

	Barrier ↓			Midpoint acc. (%) ↑		
Gate	ViT-Ti	ViT-S	ViT-B	ViT-Ti	ViT-S	ViT-B
None (RaOT-align)	6.609	5.589	5.534	**0.29**	**1.72**	**3.81**
Weight (RaOT-safe)	6.988	6.021	6.205	0.15	0.35	0.32
Activation, held-out	**6.394**	**5.569**	**5.316**	0.24	1.54	3.76
control: no gate, half budget	6.394	5.569	5.316	0.24	1.54	3.76
matched budget: 64 fit, 64 held-out	6.609	5.589	5.534	0.29	1.72	3.81
Both	6.501	6.052	6.108	0.14	0.16	0.50
Direct averaging	7.335	6.436	6.287	0.11	0.13	0.14

Three facts. *It never rejects where alignment applies*: $85/85$, $157/157$ and $301/301$ candidates accepted at ViT-Ti/S/B—and the same counts when the gate is given 64 further disjoint images to judge on while the maps are fit on all 64 (the matched-budget row of Table 9, whose barrier and midpoint equal RaOT-align's exactly), so the lower barrier of the shipped $32/32$ split is the smaller fit, not selection— the median cutting the held-out residual to 0.53 of identity's, so the *Act.* row of Table 9 is RaOT-align at half the calibration budget—identical to the control row to every bit at ViT-Ti and ViT-B—and its edge over the ungated rung is a budget effect, not gating. *It cannot detect the regime*: on the CLIP towers it accepts $238/2400$ (9.9%) where the weight gate accepts 0, and applying those candidates lands the merge *below* its base rule (65.94 against 66.63). *RaOT-auto* therefore composes the two: stage one declares **shared** if the weight gate's acceptance fraction $\hat{f} < \tau$, with $\tau = 0.10$ fixed everywhere. Three thresholds thus act on three statistics and should not be conflated: the per-candidate rule accepts iff $r_g < 1$; the regime rule tests $\hat{f} < \tau$; and the equivalent regime rule on the median r_g has the window $[1.04,\ 5.38]$ at the two endpoints (any $\tau \in (0\%, 21.2\%]$ classifies them identically), narrowing to $[1.04,\ 1.061)$ in the intermediate band of Appendix E.4, and stage two ranks candidates inside the alignment regime by the held-out criterion. On shared initialization RaOT-auto reproduces Table 11 exactly; on independent pairs it *is* the *Act.* row. It contributes an automatic regime call, not a better within-regime rule, and one measured failure at the far end of the intermediate band keeps RaOT-safe the recommendation when provenance is unknown.

E.4 The Intermediate Regime: A Continuous Divergence Axis

Twenty-seven branch pairs from one non-converged CIFAR-100 ViT-S ancestor, spanning $\delta \in [0.18, 0.81]$ along seed, class-split and data-split arms, have no barrier at any rung ($\leq 2 \times 10^{-5}$), so the whole band is the shared regime and the right action is to align nothing. Stage one calls 26 of 27 pairs *shared*, but the threshold window that also classifies this band narrows from $[1.04,\ 5.38]$ to $[1.04,\ 1.061)$—threshold-independence is an endpoint property—and the one failure (seed arm at 100 epochs, $r_g = 1.002$, acceptance 49.4%) costs -1.0 point. This band is also where we found that our implementation applied the residual-stream map without letting Eq. (13) rule on it, so Eq. (14) did not hold of our own merge at any of the 27 pairs; gating that map as the method always stated leaves every published number unchanged ($\max|\Delta| = 0$ on the ImageNet pairs) and returns RaOT-safe to within 0.2 points of direct averaging at every pair here, exactly at 20 of 27.

E.5 Which Subgroup Should Merging Align In?

Equation (3) is the group RaOT aligns in, not the group the architecture admits: the head interfaces admit $GL(d_k) \times GL(d_v)$, which is non-compact and strictly larger than the isometries—and even that is not the whole gauge, since with a key bias shifting b_K adds a per-row constant to the scores that the softmax discards; Tran et al. (2026) give the complete account. Three things pick out the isometries for *merging*: Theorem 1 bounds the barrier through a weight distance a general-linear map can shrink by rescaling rather than aligning (the bound stays valid; its value stops being comparable across maps), Eq. (13)'s verdict, read on that same distance, is then a comparison of orientations rather than of scales—a motivation, not a theorem, which Table 10 tests—and $O(d)$ is compact where $GL(d)$ is not even closed.

Table 10. **Enlarging the feasible set improves the fit and raises the barrier.** ImageNet-1k maximum loss barrier and midpoint accuracy for three nested transform families on the value–output interface ($\mathrm{diag} \cdot O(d_v)$ is a family, not a group). Each feasible set contains the one above; the scaling rung refines the orthogonal solution in place, so its activation fit cannot be worse, while the GL row minimizes a 10^{-3}-ridged objective for conditioning and its fit is compared only descriptively; the barrier moves the other way, monotonically, at every scale, and the general-linear optimum—the least-squares map over the multiplicative $GL(d_v)$ subgroup, ridged toward the identity's scale at 10^{-3} for conditioning—is worse than direct averaging at all three scales with its midpoint at chance. Endpoint exactness holds throughout ($|\Delta L| \leq 1.4 \times 10^{-5}$). Cost is not the discriminator (4.9/4.6/5.0, 19.7/19.5/20.2 and 117/116/118 s). The RaOT-align row reproduces the main independent-pair row exactly.

		Barrier ↓			Midpoint acc. (%) ↑		
Rung	Family	ViT-Ti	ViT-S	ViT-B	ViT-Ti	ViT-S	ViT-B
Direct averaging	—	7.335	6.436	6.287	0.11	0.13	0.14
RaOT-align	$O(d_v)$	**6.609**	**5.589**	**5.534**	**0.29**	**1.72**	**3.81**
+ scaling	$\mathrm{diag} \cdot O(d_v)$	6.826	5.728	5.824	0.24	1.01	2.73
+ general linear	$GL(d_v)$	8.031	6.551	7.428	0.10	0.10	0.09

Table 10 measures it on the value–output interface with three nested transform families: enlarging the feasible set improves the activation fit by construction and *raises* the barrier, monotonically, at every scale, the general-linear optimum landing below direct averaging with its midpoint at chance.

E.6 Notes on the Unbalanced-Mass and Width Extensions

Three extensions Algorithm 1 deliberately omits, being approximate rather than exact symmetries; only the first enters any reported number, in the shared-initialization experiment alone.

Unbalanced T-step. Fine-tuned parents contain task-specific units with no counterpart, which balanced OT forcibly matches, so the T-step's marginals relax with KL penalties (Chizat et al., 2018):

$$\min_{T \geq 0} \langle C, T \rangle - \epsilon \, \mathrm{H}(T) + \tau \, \mathrm{KL}(T\mathbf{1} \,\|\, a) + \tau \, \mathrm{KL}(T^\top \mathbf{1} \,\|\, b). \tag{16}$$

Equation (16) is a *diagnostic* here, not part of the merge: it scores how well each unit is matched, and low mass is a score for weak matching, not proof of task-specificity. The rule that produces the reported numbers is simpler and ungated by mass—*every* FFN slot is merged with utilization weight $w_i \propto (u_i^A)^p$, u_i^A the mean absolute activation of unit i on the calibration images, normalized across models. No transport plan, transported mass or $m_{\min}$ threshold enters the shipped merger.

Width-heterogeneous transfer. When $d^A \neq d^B$ a rectangular Stiefel map preserves the smaller model's function only on the subspace it spans—an approximate representation transfer, not a reparameterization, to which Theorem 1 does not apply verbatim. We report no table for it, because one would imply an evaluation we have not run.

Table 11. Shared-initialization eight-task CLIP merging: average accuracy over the eight *full* test sets—MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), GTSRB (Stallkamp et al., 2012), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), DTD (Cimpoi et al., 2014), Cars (Krause et al., 2013), SUN397 (Xiao et al., 2010)—one protocol at both scales, 16 unlabeled calibration images per task for the merging rule; RaOT's gate is estimated on 64; re-running the B/32 rows with the gate at 16 moves every entry by at most 0.06 points and no gate decision, and L/14 was not re-run (Appendix E.8). "Sym." is the feasible alignment set per interface. **RaOT does not lead here, and is not expected to**: its row is the *fallback path* (Table 12 completes it). §Ungated, better of two merges; ¶not a symmetry, B/32 only, second entry a *control, not a method* (Appendix E.9); ‡coefficient selected on a validation split, never the test sets (Appendix E.7).

Method	Sym.	B/32	L/14
Zero-shot	—	46.0	62.9
Individual (upper bound)	—	89.9	94.1
Weight averaging	—	64.7	78.8
Task Arithmetic	—	66.6	80.0
DC-Merge‡	—	**84.1**	**92.0**
OTFusion (port)§	$\mathcal{B}$	58.4	77.6
Transformer OT (port)§	$\mathcal{B}$	58.4	77.6
ZipIt!¶	merge M	6.4	—
correct correspondence¶	identity	64.7	—
Rotation fusion (port)	$O(d)$	66.6	80.0
RaOT-safe (ours)	per-iface	67.1	80.9

Table 12. **Full shared-initialization comparison** behind Table 11: eight-task CLIP merging, average accuracy over the eight full test sets, 16 unlabeled calibration images per task except RegMean† and RaOT's gate, which uses 64 (at most 0.06 points on the B/32 rows re-run at 16, Appendix E.8; L/14 not re-run). Table 11 shows the rows the main text discusses; interference-resolution baselines are completed here. †At 64 images/task, four times the shared budget, because at 16 it falls below weight averaging; its curve has not saturated at 256 (Appendix E.8). §Ungated, composed with the better of the two merges (Appendix E.9). ¶Not a symmetry, B/32 only; the second entry is a control, not a method (Appendix E.9). ‡Coefficients (TSV-M's and DC-Merge's α, RaOT's p) selected on a *validation* split of 512 labelled training images per task at B/32 and transferred to L/14; the test sets are never used for selection (Appendix E.7).

Method	Sym.	B/32	L/14
Zero-shot	—	46.0	62.9
Individual (upper bound)	—	89.9	94.1
Weight averaging	—	64.7	78.8
Task Arithmetic (Ilharco et al., 2023)	—	66.6	80.0
TIES (Yadav et al., 2023)	—	71.4	83.6
DARE (Yu et al., 2024)	—	66.5	80.0
RegMean (Jin et al., 2023)†	—	68.8	69.5
TSV-M (Gargiulo et al., 2025)‡	—	82.3	90.2
DC-Merge (Zhang et al., 2026)‡	—	**84.1**	**92.0**
OTFusion (Singh & Jaggi, 2020) (port)§	$\mathcal{B}$	58.4	77.6
Transformer OT (Imfeld et al., 2024) (port)§	$\mathcal{B}$	58.4	77.6
ZipIt! (Stoica et al., 2024)¶	merge M	6.4	—
with the correct correspondence¶	identity	64.7	—
Rotation fusion (Zhang et al., 2025) (port)	$O(d)$	66.6	80.0
RaOT-safe (ours)	per-iface	67.1	80.9

E.7 Scaling-Coefficient Sweeps for TSV-M and DC-Merge

TSV-M and DC-Merge apply their task vector with a coefficient α their references select by sweeping, and RaOT has a utilization temperature p. Tables 11 and 12 report every swept row at the coefficient that maximizes accuracy on a *validation* split—512 labelled training images per task, drawn from the training split under a different seed from the 16 calibration images and from the test sets—selected at

B/32 and transferred unchanged to L/14: TSV-M α=1.0, DC-Merge α=2.0, RaOT p=4. The test-set sweep behind them (an oracle, kept in the supplement) selects the same three values, so no reported number changes. Table 2 fixes DC-Merge's α the same way, on 2,000 labelled training images per task (Appendix E.17).

E.8 RegMean's Calibration-Size Dependence

Implementation, for the record: per linear layer $W_M = (\sum_k G_k^{\mathrm{r}})^{-1} \sum_k W_k G_k^{\mathrm{r}}$ with each task's Gram G_k built from every token of its calibration images and shrunk toward its diagonal, $G_k^{\mathrm{r}} = 0.9\, G_k + 0.1\, \mathrm{diag}(G_k)$ (the reference's default); no ridge term is added, the inverse is a pseudo-inverse, and neither the shrinkage nor a ridge was swept. The rise with calibration size therefore reflects this fixed implementation and does not separate sampling error from conditioning.

The same budget question applies to RaOT's gate, which reads 64 images per task where the merge rows use 16. Re-running every B/32 row of Table 12 with the gate at 16 moves the reported accuracy by at most 0.06 points (the largest change, 66.96 → 67.01, is on the p=2 row; the published p=4 row moves 67.06 → 67.10) and changes no gate decision (0/2,400 accepted either way); the L/14 rows were not re-run, so the bound is a B/32 statement.

RegMean (Jin et al., 2023) is the one baseline estimated from calibration statistics, so its score depends on the budget where the others do not. At the shared 16-image budget it falls *below* plain weight averaging at both scales (64.06 against 64.71 at B/32, by 13.4 points at L/14), which is why Table 12 breaks its own protocol for this row and grants it 64—an asymmetry that runs against us. Its curve has not saturated at 256, so the entry is a lower bound, but at L/14 even 71.97 is 6.8 points below averaging, and closing that would need on the order of 10^4 images per task.

E.9 The Matching Baselines: Ports, Transport-Map Statistics, and the ZipIt! Control

OTFusion (Singh & Jaggi, 2020) and Transformer OT (Imfeld et al., 2024) are ported as *hard-assignment controls* sharing each method's feasible set, not as reproductions—both originals use soft (entropic) plans, and Transformer OT's residual and LayerNorm handling is not reproduced: equal unit sets and uniform marginals make the transport problem an assignment LP, solved exactly by Hungarian on the squared activation-distance cost. OTFusion permutes FFN hidden units; Transformer OT additionally permutes each head's QK and VO subspaces at the same head index, leaving the residual basis fixed. Every alignment is an exact reparameterization ($\max|\Delta| = 3.3 \times 10^{-6}$ on shared-initialization towers, endpoint $|\Delta L| \le 2.8 \times 10^{-8}$ on the independent pairs). On the shared-initialization towers the maps are near-identity—99.97% of QK and 99.96% of VO coordinates fixed, 93.1% of FFN units—and the displaced 6.9% costs −6.3 points under averaging and is catastrophic composed with task arithmetic (5.6). On the independently trained pairs our OTFusion port removes 0.27/0.13/0.07 barrier and our Transformer OT port 0.18/0.04/0.08 (7.15/6.40/6.21, 0.11/0.08/0.15%; Table 1 keeps the former): adding per-head permutations between heads at the same index *hurts* on parents whose head order is arbitrary, and supplying the correspondence the reference omits (Table 1's *Hard-assignment control + matched heads* row) moves the same per-head feasible set (that row also permutes the residual stream where the port leaves it fixed, Table 3, so head correspondence is bounded here rather than isolated) to 0.85 at ViT-Ti. We report both so as neither to understate a published method on a regime its authors did not target nor to credit it with a step it does not perform.

The released implementations. Two of the three references have public code, and Table 1 runs both as released. *Rotation fusion*: the authors' matching functions—per-head weight Procrustes with the bias terms, weight-Hungarian FFN permutation, embeddings, LayerNorms and residual basis copied, heads paired by position—are called unchanged on the parents' own weights, with and without their scaling stage; only the head geometry is passed explicitly, the released driver hard-coding the 12 × 32 layout of its CIFAR-10 model. Every map is an exact reparameterization (endpoint $|\Delta L| = 0$). It removes 0.47/0.10/0.01 barrier at Ti/S/B against our port's 0.35/0.02/0.02—the bias terms in the cross-product help at small width—the scaling stage moves it by at most 0.003, evaluating it in the released pipeline's own LayerNorm gauge instead of ours changes it by ≤0.07 (6.83/6.27/6.25), and it stays above RaOT-align by 0.26/0.75/0.75. *Transformer OT*: the authors'

pipeline is run under their released experiment file (entropic plans, Sinkhorn regularization 0.08, residual policy *mean*, LayerNorm and embedding transport) on the same 64 calibration images, each parent converted losslessly to the Hugging Face class it expects (logit difference 0). Its plans are far from permutations: the transported parent sits at chance at all three widths (loss $6.93 \approx \ln 1000$, top-1 0.1%, agreeing with B on $<0.1\%$ of images), and so does the output when the pipeline is handed *two copies of the same parent* ($\max|\Delta W| = 6.1$)—the plan, not the parents' disagreement, moves the weights—consistent with Proposition 3, though the pipeline also transports LayerNorm and embeddings, so this control does not isolate the stream plan. Its curve therefore runs from A to a uniform predictor, so Section 3.6's $\mathcal{B}(\lambda)$ does not apply to it: taken literally, with $L(\hat{W}_B)$ the *transported* endpoint's own loss, the chord would end at $\ln 1000$ and the barrier would read ≈ 0 by construction. We therefore state the deviation rather than hide it. The number quoted for it in Table 1's caption ($6.09/6.07/6.11$) is $\max_{\lambda\in\{0,0.1,\ldots,1\}} L(W^\star_\lambda) - [\lambda L(W_A) + (1-\lambda)L(W_B)]$ with the *original* $L(W_B)$ and with the endpoints included in the maximum; it is attained in the interior ($\lambda = 0.6/0.3/0.1$ at Ti/S/B) and so is not the endpoint gap $|\Delta L|$ ($5.72/6.05/6.11$, the row Table 1 does list), though at ViT-B the two agree to two digits. It measures how far the pipeline moved the parent, not how the two parents interpolate, and it must not be ranked against the exact rows—which is why Table 1 sets the row apart. Its midpoint top-1 is at chance at every width ($0.28/0.43/0.37\%$ at Ti/S/B): the curve has no informative interior. The exact-plan variant of the same pipeline (`emd`) breaks the endpoint too ($|\Delta L| = 6.00/6.29/6.29$, midpoint top-1 $\leq 0.13\%$), through the residual policy and LayerNorm transport rather than softness.

A released-code defect we had to patch. Upstream's `encoder_fusion` and `act_attention_fusion` never forward their `alpha` to the `ln_fusion` and `fc_fusion` calls inside them, so those keep their default $\alpha = 0.5$ and `hf_vit_fusion(alpha=0)` returns a *hybrid*—embeddings, final LayerNorm and head transported, but every encoder block a $50/50$ blend of the two parents—rather than the transported endpoint our curve needs. We forward `alpha` at all 17 call sites (patch shipped with the code) and report the patched run throughout; the numbers above are the patched ones. The correction leaves the endpoint untouched and the interior unrecognizable. $|\Delta L|$ is unchanged to three digits ($5.72/6.05/6.11$ either way), because both versions sit at the uniform predictor's loss, $\ln 1000$; the caption number moves only $5.75/6.05/6.11 \rightarrow 6.09/6.07/6.11$. What the blend was hiding is the interior: unpatched midpoints read $1.5/16.6/37.4\%$ and *rose* with width, which invited reading them as a residue of A surviving the merge, whereas the true transported endpoint gives $0.28/0.43/0.37\%$ with no width trend at all. The unpatched maximum also sat at $\lambda = 0.2/0/0$—at the endpoint for two of three widths—where the patched one sits at $\lambda = 0.6/0.3/0.1$. None of this touches the conclusion: the two-copies-of-one-parent control is unaffected by α (a $50/50$ blend of identical weights is the identity), so it alone still shows the plan, not the parents' disagreement, moves the weights. OTFusion has no Transformer implementation; Transformer OT is its extension by the same group. ZipIt! (Stoica et al., 2024) is ported at the FFN units and scores 6.4 at B/32; run with the correspondence that is correct by construction on shared-initialization towers it returns 64.71, weight averaging to the last digit, so the 58-point deficit belongs entirely to its estimated matching.

E.10 The Independently Trained Pairs

Every ImageNet (Deng et al., 2009) result uses *one* pair per width from public `timm` (Wightman, 2019) releases: DeiT `deit_{tiny,small,base}_patch16_224.fb_in1k` against AugReg `vit_tiny_patch16_224.augreg_in21k_ft_in1k`, `vit_small_patch16_224.augreg_in1k` and `vit_base_patch16_224.augreg_in1k`—the same architecture trained by different groups, codebases, initializations and schedules. The ViT-Ti pair is not matched the way ViT-S and ViT-B are: no ImageNet-1k-only AugReg ViT-Ti exists, so that parent is pretrained on ImageNet-21k and differs from its DeiT counterpart in *corpus* as well as recipe. That is where the inversions we report occur (joint head matching hurts, the residual rotation in isolation is harmful, the probe barely separates the rungs). The corpus difference is a documented association, not an isolated cause: the same 21k/1k disparity at d=384 and d=768 (Table 14) gives the *largest* rotation-over-permutation contrasts in that table ($+0.85$, $+1.16$), so it cannot by itself explain rotation losing to permutation at ViT-Ti. Calibration variability accounts for most of it (Table 13): over five independent draws of training images, the sign of the permutation-over-RaOT-align contrast at 64 images depends on the draw ($3/5$ favour RaOT-align;

the published draw is one of the two that do not, and its 6.61 sits at the top of the draws' 6.26–6.65), at 32 images the contrast is centred on zero with a spread of 0.38, and at 128 images RaOT-align wins on $5/5$ draws with the spread down to 0.09. Head correspondence (disabling it improves ViT-Ti at 32 and 64 images, Appendix E.14) is part of the same story—its advantage is gone at 128—and the width trend is read on ViT-S→ViT-B.

Table 13. **The ViT-Ti pair under five independent calibration draws.** Maximum loss barrier (mean±SD over draws; full 50k evaluation) at three calibration budgets, each draw one image from each of 128 distinct ImageNet-1k training classes with its own seed (manifest in the code), the published draw excluded. Right: the paired permutation → RaOT-align contrast, with the number of draws on which RaOT-align has the lower barrier. The inversion Table 1 reports at ViT-Ti is one draw's outcome at 64 images; it is gone at 128, and the stream-from-weights hybrid has the lowest mean barrier at every budget (the permutation rung the smallest spread), with a midpoint 3–4× the others'.

Images	Permutation	RaOT-align	no head matching	stream (wt.)	perm.→RaOT-align
32	6.65 ± 0.17	6.69 ± 0.41	6.43 ± 0.24	$\mathbf{6.41 \pm 0.26}$	-0.03 ± 0.38 (3/5)
64	6.63 ± 0.17	6.47 ± 0.19	6.40 ± 0.35	$\mathbf{6.37 \pm 0.19}$	$+0.15 \pm 0.34$ (3/5)
128	6.65 ± 0.08	6.42 ± 0.13	6.48 ± 0.17	$\mathbf{6.32 \pm 0.16}$	$+0.23 \pm 0.09$ (5/5)
published draw, 64	6.48	6.61	6.38	6.60	−0.13

Six further pairs at the two matched widths. One pair per width is this ladder's weakest point, so we add six more public `timm` pairs at d=384 and d=768: AugReg / DeiT-III, DeiT / DeiT-III and AugReg-21k / DeiT at each width. All six are run on the full 50k evaluation set with *this* table's calibration (64 training-split images, Appendix E.19), so they read against Table 1 without a protocol change.

Table 14. **Six additional independently trained pairs, run at the main ladder's calibration.** Maximum loss barrier and midpoint top-1 on the full ImageNet-1k validation set; Δ_{perm} and Δ_{rot} are the matched feasible-set contrasts of Table 18 (direct averaging → permutation, and permutation → RaOT-align). **RaOT-align takes the lowest barrier and the highest midpoint on all six**, and Δ_{rot} is positive on all six; Δ_{perm} changes sign across pairs—on two of the six a permutation is *worse* than doing nothing. Chance is 0.1%. Best barrier per row in **bold**.

		Parents (%)		Barrier ↓ / midpoint acc. (%)				
d	Pair	A	B	Direct	Perm.	RaOT-align	Δ_{perm}	Δ_{rot}
384	AugReg / DeiT-III	78.8	81.4	6.50 / 0.10	6.82 / 0.18	**6.44** / 0.81	−0.32	0.37
384	DeiT / DeiT-III	79.8	81.4	6.45 / 0.10	6.38 / 0.08	**6.12** / 0.72	0.07	0.26
384	AugReg-21k / DeiT	81.4	79.8	7.15 / 0.10	6.69 / 0.13	**5.84** / 0.98	0.46	0.85
768	AugReg / DeiT-III	79.1	83.8	6.42 / 0.11	6.19 / 0.25	**5.75** / 4.77	0.23	0.44
768	DeiT / DeiT-III	81.9	83.8	6.50 / 0.11	6.32 / 0.15	**5.79** / 4.61	0.18	0.54
768	AugReg-21k / DeiT	84.5	81.9	6.61 / 0.12	7.30 / 0.22	**6.14** / 3.87	−0.69	1.16

Three things separate here. The *ordering* replicates without exception: RaOT-align takes both columns on six of six pairs, and Δ_{rot} is positive on six of six (0.26 to 1.16). The *permutation* term does not survive the enlargement at all— $+0.07 \pm 0.39$ at d=384 and -0.09 ± 0.52 at d=768 over three pairs each, straddling zero at both widths, with a permutation actively harmful on two pairs. Table 18's single-pair values at these widths (0.29 and 0.18) sit inside that spread, so they are not outliers, but neither are they evidence of a trend: what the extra pairs show is that Δ_{perm} is pair noise at these widths, while Δ_{rot} is not. The *width* contrast points the way Table 18 reports—Δ_{rot} rising $0.50 \pm 0.31 \rightarrow 0.71 \pm 0.39$—but with three pairs per width it is not resolvable, and we do not claim it from this table. Finally, the midpoints reproduce Table 1's pattern rather than extending it: 0.72–0.98% at d=384 and 3.87–4.77% at d=768, against 0.1% chance and 0.08–0.25% for every other rung, where that table's RaOT rows read 1.72 and 3.81. The merges that leave chance at all are the rotation-aware ones, and they now do so on nine pairs, not three. Usable accuracy this is not (Section 5); a barrier ordering read on a curve that is flat at chance, it also is not.

E.11 A Same-Recipe Control, and the Pre-Registered Width Study

Do the orderings of Table 1 survive when parents differ in *nothing but the seed*? Four ViT-Ti and four ViT-S are trained from scratch on CIFAR-100 under one recipe (AdamW 10^{-3}, weight decay 0.05, cosine, 100 epochs, label smoothing 0.1, bf16), and Table 15 runs the ladder on four pairs per width.

Table 15. **Same architecture, same recipe, same task—only the seed differs.** Maximum loss barrier and midpoint top-1 on CIFAR-100 at two widths, mean±SD over four seed pairs each ($(0,1)$, $(0,2)$, $(1,2)$, $(2,3)$; parents reach 43.1–46.2% at ViT-Ti and 43.0–45.5% at ViT-S), with the ladder and estimators of Table 1, calibrated on 256 images from the evaluation split rather than that table's 64 training-split ones (Appendix E.19). **The margin grows with width**—RaOT-align clears direct averaging by 5.1 accuracy points at ViT-Ti and 10.7 at ViT-S—with the recipe difference, and Appendix E.10's corpus difference, removed. Under the shipped exact cost the barrier margin grows too (0.19 to 0.32, where the separable cost gave 0.18 at both), but the rotated rungs carry a spread an order of magnitude larger than the unrotated ones', so we still read the trend here on accuracy. Best per column in **bold**.

	ViT-Ti (d=192)		ViT-S (d=384)	
Method	Barrier ↓	Acc. (%)	Barrier ↓	Acc. (%)
Direct averaging	1.922 ± 0.021	1.2 ± 0.2	1.860 ± 0.030	1.4 ± 0.2
Permutation	1.889 ± 0.054	2.2 ± 0.7	1.762 ± 0.036	4.2 ± 0.2
Rotation fusion (port)	1.920 ± 0.024	1.2 ± 0.3	1.871 ± 0.025	1.3 ± 0.2
+ residual rotation	1.986 ± 0.030	1.2 ± 0.1	1.907 ± 0.050	2.3 ± 0.1
RaOT-act	1.871 ± 0.045	1.8 ± 0.8	1.786 ± 0.017	4.3 ± 0.6
RaOT-safe (weight gate)	1.832 ± 0.109	3.4 ± 1.9	1.822 ± 0.335	5.9 ± 4.3
RaOT-auto (held-out gate)	1.754 ± 0.100	5.3 ± 2.3	$\mathbf{1.511 \pm 0.189}$	$\mathbf{12.5 \pm 2.5}$
RaOT-align	$\mathbf{1.736 \pm 0.118}$	$\mathbf{6.3 \pm 2.9}$	1.545 ± 0.179	12.1 ± 2.7

The rotation-aware rungs lead on all eight pairs in midpoint accuracy, weight-space rotation fusion again fails to separate from direct averaging, and the margin over averaging grows with width (5.1 → 10.7 points, 6.8 → 12.6 on the two model-disjoint pairs). The crossover does *not* replicate: the permutation estimator's term *grows* (0.033 → 0.098) where the AugReg/DeiT pairs give 0.85 → 0.29 → 0.18, so what widens here is the whole decomposition.

The non-replication is not an artifact of the calibration source. This control and the width study below calibrate on 256 evaluation-split images where Table 1 uses 64 training-split ones (Appendix E.19), so the comparison that scopes the crossover crossed a protocol boundary: a term that fails to replicate here might be failing because of the budget and the split, not because the parents share a recipe. We therefore re-ran the whole study—all six pairs of all eight configurations, every rung—on the main ladder's own calibration tensor.

Table 16. **The same-recipe study under both calibration sources.** Maximum loss barrier on CIFAR-100, mean over the six seed pairs per configuration, with the terms of Table 17 (Δ_{perm} direct averaging → permutation, Δ_{res} RaOT-act → RaOT-align). Left: the published protocol. Right: Table 1's 64 training-split images. Direct averaging uses no calibration and is identical by construction. **No term changes sign, and neither trend changes direction**: Δ_{perm} stays positive at every width under both, where the crossover predicts it to fall, and Δ_{res} still grows with width. The smaller budget costs the rotated rung at the widest configurations (0.94 → 1.10, 1.23 → 1.34), which is the expected estimation cost and not a reordering.

			256 eval-split images (published)				64 train-split images (Table 1)			
d	h	L	Direct	Perm.	Δ_{perm}	Δ_{res}	Direct	Perm.	Δ_{perm}	Δ_{res}
128	2	12	2.230	2.101	0.129	0.323	2.230	2.102	0.128	0.293
192	3	12	1.907	1.858	0.048	0.132	1.907	1.845	0.062	0.138
192	3	24	1.762	1.690	0.072	0.144	1.762	1.720	0.042	0.118
384	3	12	1.751	1.632	0.119	0.156	1.751	1.660	0.091	0.148
384	6	6	2.082	1.954	0.128	0.241	2.082	1.952	0.131	0.271
384	6	12	1.848	1.752	0.097	0.309	1.848	1.765	0.084	0.321
768	3	12	1.833	1.714	0.119	0.496	1.833	1.727	0.106	0.407
768	12	12	1.848	1.627	0.221	0.672	1.848	1.677	0.170	0.549

Δ_{perm} is positive at all eight configurations under both calibrations and still does not fall with width ($0.13/0.05/0.10/0.14$ by width under the matched protocol, against $0.13/0.06/0.11/0.17$ under the published one), and Δ_{res} still grows ($0.29/0.13/0.25/0.48$ against $0.32/0.14/0.24/0.58$). The scoping statement above therefore stands on a within-protocol comparison: what fails to replicate on same-recipe parents does not fail because the two studies were calibrated differently; the other differences between them remain.

The pre-registered width study. To settle the width question the recipe, inclusion bar and analysis were fixed in `runs/WIDTH_SCALING_PREREG.md` before any barrier was computed: Series A holds d_h=64 over $d \in \{128, 192, 384, 768\}$, Series B holds h=3, two shape controls sit outside both, all six pairs per configuration, learning rate chosen per width on seed-0 parent accuracy alone. One configuration (d=384, h=3) fails the pre-registered 5-point seed-spread bar and is reported but excluded—its weak seed carried half of the only large slope in the study. Two departures from that document should be read with the fit: it registers the *capture ratio* as a third outcome, which we do not fit, and reports the accuracy margin, which it does not register; and its second, covariate-adjusted fit (parent accuracy and parameter count, whose sign the effect was to survive) was not run, so the trend below rests on the unadjusted fit alone.

Table 17. **The pre-registered width study, all six pairs per configuration.** Maximum loss barrier on CIFAR-100, mean over the six seed pairs (SD shown for RaOT-align, the noisiest rung); Δ_{perm} is the permutation-recoverable term (direct averaging → permutation) and Δ_{res} the residual-rotation term (RaOT-act → RaOT-align), the registered outcome, which is *not* Table 18's Δ_{rot} (permutation → RaOT-align); the last column is RaOT-align's midpoint accuracy margin over direct averaging. "Spread" is the range of the four parent accuracies, and † marks the one configuration failing the pre-registered 5-point bar, excluded from both fits and shown for completeness. The last two rows are matched-parameter shape controls belonging to neither series.

				Spread	Barrier ↓					Acc. margin
d	h	L	Series	(pt)	Direct	Perm.	RaOT-align	Δ_{perm}	Δ_{res}	(pt)
128	2	12	A	2.10	2.230	2.101	1.839 ± 0.067	0.129	0.323	7.4
192	3	12	A,B	3.00	1.907	1.858	1.734 ± 0.094	0.048	0.132	5.0
384	6	12	A	2.55	1.848	1.752	1.458 ± 0.204	0.097	0.309	12.3
768	12	12	A	1.15	1.848	1.627	0.936 ± 0.094	0.221	0.672	21.7
384	3	12	B	**6.10**†	1.751	1.632	1.524 ± 0.086	0.119	0.156	6.8
768	3	12	B	2.20	1.833	1.714	1.226 ± 0.097	0.119	0.496	15.0
192	3	24	–	1.60	1.762	1.690	1.608 ± 0.073	0.072	0.144	5.8
384	6	6	–	2.40	2.082	1.954	1.755 ± 0.145	0.128	0.241	8.7

The registered fit is a mixed-effects model with crossed parent random effects at **two** between-width degrees of freedom, not the twenty-two a pair-level OLS would claim. The exploratory, unregistered accuracy margin over direct averaging grows with width, $7.4/5.0/12.3/21.7$ points at $d = 128/192/384/768$, about 6 points per doubling of d ($p = 0.03$; one nominally significant slope among four tested, not surviving a Bonferroni 0.0125), while the registered barrier outcomes are inconclusive under the two-degree-of-freedom reference: Δ_{perm} $\beta = +0.042$ per doubling ($p = 0.13$) and Δ_{res} $\beta = +0.154$ ($p = 0.10$); the registered capture-ratio fit and the covariate-adjusted fit were not run. The permutation term's point estimate is *positive* in both series where the crossover predicts negative, and the relative form is flat ($33.1/28.1/24.7/24.2\%$) and was ruled out in advance. So the crossover is scoped to parents differing in recipe or corpus, and not claimed of width alone.

E.12 The Permutation-to-Rotation Crossover

Two rungs of Table 1 differ in feasible set alone: the permutation rung and RaOT-align share the estimator, the interfaces and the calibration, and enlarge the set from the polytope's vertices, the permutations, to the isometries. Their differences against direct averaging and against each other are therefore an order-free pair of matched contrasts (Table 18, top): what this permutation estimator removes and what this rotation estimator adds beyond it at the same calibration. They are estimator-specific quantities, not the barrier reduction achievable over a symmetry family—the estimators fit activations, not the barrier, and at ViT-Ti the larger family scores worse (6.61 against 6.48).

Table 18. **The permutation-to-rotation crossover with width.** Top: two matched contrasts by *feasible set*, read at one estimator and one interface set—what the permutation estimator removes (direct averaging → permutation rung) and what the isometries' estimator adds beyond it (permutation → RaOT-align); the two rungs differ in feasible set alone, so the pair is order-free, and it is estimator-specific rather than a symmetry-removable component. Middle: increments along the shipped rung order, which *are* order-dependent because the interfaces interact (attention-only alignment is worse than nothing at ViT-B, Appendix E.14); the residual-rotation rows are read at a fixed estimation budget, the shipped RaOT-auto fitting on half the calibration images. Bottom: permutation capture ratio η at each rotation-legal interface with the Haar-random reference and its $1.15\sqrt{\ln d/d}$ estimate (Proposition 5); per-head dimension is fixed (d_h=64), the stream widens (D=192/384/768).

Barrier removed, by estimator	Ti	S	B
By feasible set (order-free)			
Transport polytope (naive→perm)	0.85	0.29	0.18
Isometries beyond it (perm→RaOT-align)	**−0.13**	**0.55**	**0.57**
Along the shipped rung order (order-dependent)			
FFN permutation (naive→OTFusion)	0.27	0.13	0.07
Head matching + rotation, given FFN (OTFusion→RaOT-act)	0.05	0.05	0.04
Residual rotation, given the rest, 64 images	0.40	0.67	0.64
the same at 32 images	0.39	0.64	0.91
estimation budget alone	0.22	0.02	0.22
shipped point (RaOT-act→RaOT-auto)	0.62	0.69	0.85
added to weight-estimated heads and FFN instead	0.12	0.27	−0.04
Capture ratio η	Ti	S	B
QK (d_h=64, fixed)	0.46	0.48	0.49
VO (d_h=64, fixed)	0.38	0.37	0.39
Residual stream (D grows)	0.35	0.27	0.24
Haar-random reference $\eta_{\text{Haar}}(D)$	0.19	0.14	0.11
$1.15\sqrt{\ln D/D}$	0.19	0.14	0.11

What the permutation estimator removes collapses as the model widens ($0.85 \to 0.29 \to 0.18$), while what the isometries' estimator adds beyond it grows ($-0.13 \to 0.55 \to 0.57$)—negative at ViT-Ti, where rotation loses to permutation on the published calibration draw and not on most others (Table 13). Increments along the shipped rung order are *not* a decomposition by symmetry class and are reported only as increments: the FFN permutation alone removes $0.27/0.13/0.07$, head matching with rotation adds $0.05/0.05/0.04$ given it, and the residual rotation given the rest—read at a *fixed* 64-image budget, because the shipped RaOT-auto fits on half the rows and RaOT-act → RaOT-auto therefore also carries a $64 \to 32$ budget change—is $0.40 \to 0.67 \to 0.64$, rising to ViT-S then level; at 32 images it keeps rising and the budget term alone is $0.22/0.02/0.22$. We claim the rise to ViT-S, which survives at either budget, and not its continuation. The same activation-estimated residual rotation added on top of weight-estimated head and FFN maps turns harmful at ViT-B (-0.04)—an interaction with the maps beneath it, not an estimator comparison; the estimator ablation of Appendix B, which swaps only the stream estimator, points the other way at ViT-S/B. The capture ratio η of Appendix E.13 is scale-invariant at the fixed-dimension head interfaces and decays on the widening stream ($0.350 \to 0.268 \to 0.236$), above the Haar-random reference and not at its rate. With the decay absent on same-recipe parents (Appendix E.11) and on the language stream (Appendix E.16), what this ladder measures is a crossover on three public pairs differing in recipe or corpus. Two further sets of runs test how much of it is the pairs: six more pairs at the two matched widths (Table 14), and the second series below.

A second width series: one family, one recipe, corpus difference alone. The ladder above confounds width with whatever else separates a DeiT from an AugReg checkpoint at each scale. A cleaner series exists in public releases: DeiT-III `fb_in1k` against DeiT-III `fb_in22k_ft_in1k` at {`small`, `medium`, `base`, `large`}, one codebase, one recipe, one architecture family, differing in *pretraining corpus* and nothing else, at $d \in \{384, 512, 768, 1024\}$ with $h \in \{6, 8, 12, 16\}$—so per-head dimension is fixed at d_h=64 across the series, as in Series A of Appendix E.11. Four widths, four pairs, the main ladder's calibration, the full 50k evaluation set.

Table 19. **A four-width series within one model family, differing in pretraining corpus alone.** DeiT-III `in1k` against DeiT-III `in22k_ft_in1k`; maximum loss barrier and midpoint top-1 on the full ImageNet-1k validation set, at Table 1's calibration. d_h=64 is held fixed; the stream, head count and FFN width grow together, as in any width series. **Of the crossover's two halves this series reproduces one**: the isometries' term Δ_{rot} roughly triples across the series, while the transport polytope's term Δ_{perm} has no trend and is smallest in the middle. At d=1024 RaOT-align is the only rung whose midpoint leaves chance (0.1%). Best barrier per row in **bold**.

			Parents (%)		Barrier ↓ / midpoint acc. (%)			
Size	d	h	`in1k`	`22k`	Direct	Perm.	RaOT-align	Δ_{perm} / Δ_{rot}
S	384	6	81.4	82.8	6.95 / 0.11	6.51 / 0.11	**6.35** / 0.10	0.44 / 0.16
M	512	8	83.0	84.3	6.79 / 0.08	6.50 / 0.10	**6.32** / 0.10	0.30 / 0.18
B	768	12	83.8	85.6	6.85 / 0.10	6.81 / 0.12	**6.29** / 0.27	0.04 / 0.53
L	1024	16	84.8	86.9	6.97 / 0.10	6.44 / 0.13	**5.91 / 3.54**	0.53 / 0.52

This series splits the crossover into the half that replicates and the half that does not. Δ_{rot}—what enlarging the feasible set from the permutations to the isometries buys—rises with the stream to a plateau, 0.16/0.18/0.53/0.52, roughly tripling between d=512 and d=768 and then level; RaOT-align takes the lowest barrier at all four widths. Δ_{perm} does *not* collapse: 0.44/0.30/0.04/0.53 is non-monotone and ends where it started, so the AugReg/DeiT ladder's $0.85 \to 0.29 \to 0.18$ is not a property of width in general. Together with Table 14, where Δ_{perm} straddles zero across pairs at a fixed width, we now read the crossover as **one claim, not two**: enlarging to the isometries buys more as the stream widens, on parents differing in recipe or corpus. That the transport polytope's share collapses in the same direction holds on the three-pair ladder and is not reproduced elsewhere, so we report it as that ladder's behaviour and not as a scaling law. The one place the series speaks to usability is its widest pair, where the isometries alone take the midpoint from chance to 3.54%—twenty-seven to thirty-five times the other two rungs, and still far from either parent's 85% (Section 5).

E.13 The Permutation Capture Ratio: Protocol and the Haar Baseline

At each rotation-legal interface the alignment gain $\langle M, T \rangle$ is maximized by $\|M\|_*$ over $O(d)$ and by the assignment optimum over permutations, against $\operatorname{tr} M$ for doing nothing, so $\eta = (\text{perm} - \text{id})/(\text{rot} - \text{id})$ is the fraction of the rotational objective a permutation can express. Two protocol facts bound what these diagnostics say about the merger's map: on the stream, M is the same pooled, feature-centered cross-covariance the merger uses, so η is a global statement, but its rotational optimum is taken over all of $O(d)$ rather than the mean-preserving subgroup the merger applies (the constraint can only lower the rotational value); and the per-unit correlations of Table 21 are fitted *per read point*, over all of $O(d)$, on centered columns with unit normalization applied after rotation, so they describe representational similarity at each depth and not a single map the merger could apply.

Proposition 5 (Permutations capture a vanishing share of a random rotation)**.** *Let Q be Haar-distributed on $O(d)$ and take $M = Q$. Then, with probability $1 - o(1)$, $\eta(Q) \le (2 + o(1))\sqrt{\ln d/d} \xrightarrow{d\to\infty} 0$.*

Proof. The rotation optimum is $\|Q\|_* = d$, while $\operatorname{tr} Q$ is asymptotically standard normal (Diaconis & Shahshahani, 1994), hence $O_P(1)$. By Jiang (2005), $\sqrt{d/\ln d}\,\max_{ij}|Q_{ij}| \to 2$ in probability, so *any* assignment attains at most $d\max_{ij}|Q_{ij}| = (2 + o(1))\, d\sqrt{\ln d/d}$, and $\eta(Q) = (\text{perm} - \operatorname{tr} Q)/(d - \operatorname{tr} Q) \le (2 + o(1))\sqrt{\ln d/d}$. □

The rate's constant is calibrated numerically to $\approx 1.15\sqrt{\ln d/d}$ (measured Haar 0.188/0.144/0.110 against 0.190/0.143/0.107 at $d = 192/384/768$), comfortably inside the proposition's $(2 + o(1))$; the sampler draws Q by Gaussian QR *with the column signs corrected by* $\operatorname{sign}\operatorname{diag} R$, without which the draw is not Haar and η reads $\approx$0.03 high, since $\operatorname{tr} Q$ and the assignment optimum both see the signs. Real pairs sit above the Haar reference and decline more slowly than it, their ratio to it growing $1.87 \to 1.87 \to 2.18$; it is a null baseline, not a limit they approach.

Table 20. **Rotational alignment against correspondence-free nulls.** Normalized alignment a on the pairs of Table 1, on 64 images from the evaluation split—the geometry tables read `cal_x_u8.pt` where the merging ladder reads the training-split `cal_train_x_u8.pt` (Appendix E.19). Each cell is the rotational optimum over the permutational one. *Shuffled* destroys the image correspondence while keeping both marginals; *rand. init* keeps the correspondence—the untrained network sees the same images in the same order—and removes the shared training instead, so the two nulls bound different explanations. *Excess* is real a_{rot} minus the larger null. *Held out* fits R on the first 32 images and scores it on the last 32, against the rotational optimum on those same images; the matched permutation, fitted and scored the same way, is not measured, so this column tests out-of-sample transfer of the rotation, not the rotation-over-permutation excess out of sample. Identity scores $|a| < 0.005$ everywhere and is omitted.

	a_{rot} / a_{perm}				
Interface	Real	Shuffled	Rand. init	Excess	Held out / opt.
ViT-Ti					
Per-head QK (shared R_h)	0.49 / 0.23	0.23 / 0.13	0.20 / 0.12	0.27	0.46 / 0.51
Per-head VO	0.49 / 0.19	0.13 / 0.05	0.22 / 0.13	0.27	0.45 / 0.51
Residual stream	0.14 / 0.05	0.07 / 0.02	0.05 / 0.02	0.08	0.12 / 0.15
ViT-S					
Per-head QK (shared R_h)	0.48 / 0.24	0.22 / 0.13	0.22 / 0.14	0.26	0.45 / 0.49
Per-head VO	0.49 / 0.19	0.13 / 0.06	0.24 / 0.14	0.26	0.44 / 0.51
Residual stream	0.43 / 0.12	0.15 / 0.04	0.13 / 0.06	0.28	0.36 / 0.46
ViT-B					
Per-head QK (shared R_h)	0.43 / 0.21	0.20 / 0.12	0.21 / 0.12	0.22	0.38 / 0.44
Per-head VO	0.45 / 0.17	0.14 / 0.06	0.23 / 0.13	0.22	0.38 / 0.47
Residual stream	0.47 / 0.11	0.22 / 0.06	0.17 / 0.07	0.25	0.31 / 0.49

Table 20 asks whether the measured alignment is real: against a shuffled-correspondence null, a random-init null and a held-out fit, the real alignment sits at 0.43–0.49 against 0.13–0.24 under the nulls, and net of the larger null the rotational excess is 2.4–6.0× the permutational at every interface and scale—the statistic that establishes Table 21's ordering once $O(d)$'s extra freedom is priced in.

E.14 Which Interface Carries the Gain: Component Isolation

Switching interfaces off one at a time, with protocol, split, calibration and barrier grid fixed, shows a superadditivity at the largest scale: at ViT-B no single interface recovers more than 10% of the full gain, two of the three are worse than doing nothing, and only the three together move the barrier, joint head matching then supplying most of the remainder (6.02 → 5.53); at ViT-Ti and ViT-S single interfaces recover 32–38%, and which one carries most is not stable across scales. Removing joint head matching *improves* ViT-Ti but costs 0.10 at ViT-S and 0.49 at ViT-B.

E.15 Unit- and Subspace-Level Agreement

The rotations scored here are fitted per read point over all of $O(d)$ on centered columns (Appendix E.13); they are a representational diagnostic, not the merger's single mean-preserving map, and the held-out column of Table 20 transfers the rotation only, a matched held-out permutation being unmeasured.

The crossover has a representational counterpart that needs no merging. On the same pairs, the mean per-unit correlation attainable under the best permutation of stream units is 0.19/0.20/0.20 (Ti/S/B) and flat in width, while under the best rotation it reaches 0.60/0.68/0.69—rotation beating permutation by 2.4–4.4× at every read point, the gap growing where the barrier crossover says it should.

Table 21. **Independent ViTs align in subspaces, not on axes.** Read points grouped into quarters of the network (depth →). CKA is linear CKA; CKA^{-1} is the same after projecting out each parent's leading principal direction; a_{perm} and a_{rot} are the mean per-unit correlations after the best permutation and the best rotation. σ_1 is the variance share of the AugReg parent's leading stream direction. The raw-CKA collapse in the last quarter tracks σ_1, not representational divergence: CKA^{-1} declines only gradually. $a_{\text{rot}} \gg a_{\text{perm}}$ everywhere; the ratio alone does not price in the larger group's extra freedom, for which see the excess statistic of Table 20.

	Read points		CKA	CKA^{-1}	a_{perm}	a_{rot}
Ti	0–6	(σ_1=0.07)	0.82	0.79	0.27	0.80
	7–12	(0.07)	0.78	0.74	0.20	0.66
	13–18	(0.52)	0.28	0.58	0.17	0.52
	19–24	(0.95)	0.01	0.41	0.12	0.38
S	0–6	(0.07)	0.89	0.83	0.29	0.86
	7–12	(0.17)	0.52	0.78	0.18	0.69
	13–18	(0.64)	0.16	0.56	0.15	0.59
	19–24	(0.54)	0.18	0.40	0.14	0.53
B	0–6	(0.09)	0.72	0.73	0.29	0.80
	7–12	(0.49)	0.16	0.68	0.17	0.65
	13–18	(0.70)	0.09	0.56	0.15	0.63
	19–24	(0.53)	0.15	0.44	0.19	0.66
Mean over all read points:		Ti	0.49	0.64	0.19	0.60
		S	0.46	0.65	0.20	0.68
		B	0.30	0.61	0.20	0.69

Rotation-invariant linear CKA (Kornblith et al., 2019) confirms agreement at the level of subspaces (0.72–0.89 early, 0.40–0.44 late); its collapse to ≈0 in the last quarter is an artifact of a single massive-norm direction (Sun et al., 2024) that projecting out (CKA^{-1}) removes. Independently trained ViTs thus place related information in related subspaces, not on the same axes. Of the barrier's three components, the *statistical* one that REPAIR (Jordan et al., 2023) repairs in BatchNorm CNNs is not one the tested corrections reduce here—a faithful LayerNorm port is neutral at best and harmful where there is signal (44.3% → 33.4%)—and the remainder after RaOT is removable by no map we can estimate, the general-linear optimum *raising* the barrier (Table 10).

E.16 Outside Vision: Byte-Level Language Transformers

Causal byte-level Transformers (enwik8 (Mahoney, 2011), depth 6, $d \in \{192, 384\}$, four seeds per width, paired two ways) run the ladder verbatim and exactly ($|\Delta L| = 0.00$ at both endpoints while moving 48–58% of the parameters), which is the machine-checked form of "causal masking changes no weight-space symmetry". The pre-merging geometry matches vision at matched width—rotational agreement 2.2–2.3× the permutational—and the share of the barrier a permutation removes is smaller than the best rotation-aware share in every cell, by a factor of 1.5 to 4.8. Two things do not transfer: the query–key capture ratio is far more permutation-like (0.73–0.78 against 0.46–0.49), and the residual-stream η shows no width trend over the two widths, so the ViT decay is reported as a ViT result. The gate's r_g sits in the same four-percent band around unity as on ImageNet, with the same cost.

E.17 Different Tasks and Different Initializations: Protocol and Full Sweep

CIFAR-100 (Krizhevsky, 2009) is cut into four disjoint 25-class tasks and one ViT-S is trained per task from its own seed, so parents differ in initialization, data order *and* task; four draws use different class groupings and seed blocks (100–103, 300–303, 500–503, 700–703), 2,000 test and 64 calibration images per task. Parent 0 is the anchor; each other parent is aligned to it (its head transported too, endpoint $\max|\Delta\text{Acc}| \leq 5 \times 10^{-4}$), only the backbone is merged by DC-Merge at α=0.125—selected on a validation split of 2,000 labelled *training* images per task, never the test split, where it is the optimum for every alignment (on the test sets it is the per-draw optimum in 15 of 16 combinations)—and each task is read out with its own aligned head. The transported-heads control prices the merge itself: the *unmerged* anchor backbone read out with the same transported heads matches the merged trunk to within noise at every alignment (23.7 ± 1.9 against 23.8 ± 1.8 for RaOT-

align), so the gain is coordinate-compatible head transport and the backbone merge adds ≈ 0. That control is also the cleanest isolation of the alignment we have—specifically of the residual-stream map, the one component that reaches an external head, since head matching and FFN permutations act inside the fixed backbone and rotation fusion, which rotates no stream, is structurally equal to no transport here—because it holds the backbone fixed: all five rows share one set of trunk weights and score 55.3 ± 3.7 on the anchor's own task, and the only thing that varies is the map carrying each other parent's head into the anchor's coordinates. Run with the residual map alone (`raot_res`), that block reproduces RaOT-align's to the last digit on every draw and task (23.7 ± 1.9, 13.2 ± 1.7 on *Others*), so the compatibility gain is the residual map's entirely; `rf_weight_res` reproduces it likewise, its stream map being the same activation-estimated one (Appendix B). The weight-estimated stream of Table 4, applied over activation-aligned heads and FFN (`raot_wres`), scores 24.8 ± 0.7 (All) and 14.6 ± 0.7 (*Others*) here, above RaOT-align's 23.7 ± 1.9 and 13.2 ± 1.7 on three of the four draws and within 0.1 on the fourth; applied alone (`res_w`) or over weight-estimated heads and FFN it scores 17.3 ± 0.9 and 17.0 ± 0.7, within noise of no transport (16.7 ± 1.0). Across that column *Others* runs 3.9 ± 0.5 for direct averaging and for rotation fusion—whose weight-estimated maps leave the heads no better placed than no alignment at all—8.1 ± 1.0 for the permutation, and 13.2 ± 1.7 for RaOT-align (12.5 ± 1.7 for RaOT-auto), a $3.4\times$ spread attributable to the feasible set and the estimator alone. The merge is what adds nothing here; the alignment is not. Three less convenient constructions (whole superclasses; four different datasets) preserve the ordering with a monotonically shrinking margin ($7.0 \to 4.7/4.6 \to 3.6$), and across the full sweep RaOT is the best alignment in 20 of 21 usable cells.

E.18 Linear-Probe Evaluation of Merged Models

The backbone is frozen and one linear head is trained per row under an identical budget (AdamW, 100 epochs, 10^{-3}, batch 1024), features cached under the test transform; extraction is exact (a parent probed on its own features recovers its accuracy to four decimals). Full fine-tuning was run first and reported as a negative result: 20 epochs drive every rung to parent accuracy and a random initialization to 44.1, so it does not discriminate, whereas the frozen probe sends that initialization to 15.8 and separates the rungs.

Table 22. **Merged models as frozen feature extractors.** Each Table 1 midpoint frozen, one linear head trained on CIFAR-100 at 224 px (Appendix E.18); *Scratch* is the floor.

Alignment	Ti	S	B
Scratch (floor)	11.9	15.7	18.9
Parents A/B	77.7/72.6	77.2/77.9	77.9/81.5
Direct averaging	21.5	30.6	28.5
Rotation fusion (port)	22.3	32.6	33.1
Rotation fusion, released code	21.1	33.0	36.0
Permutation	26.9	38.8	41.0
RaOT-align (ours)	**27.1**	**51.5**	51.1
RaOT-auto (ours)	26.9	50.2	**52.2**

What the probe establishes is that the merged features retain linearly accessible information under a retrained readout on a different task; because that task and head differ from the standalone ImageNet evaluation, it does not show that the ImageNet failure is only a classifier mismatch—cumulative feature drift across depth, incompatible intermediate statistics and readout mismatch all remain candidate explanations, and the REPAIR control (Appendix E.15) rules out only the statistical corrections it tests. RaOT's better point recovers 58% of the floor-to-parent gap at ViT-S and 55% at ViT-B, every baseline less; the barrier separates the methods along the whole path (52.6% of parent accuracy retained at $\lambda{=}0.3$ against averaging's 18.6%). EuroSAT and GTSRB reproduce the ordering at ViT-S/B, and at ViT-Ti CIFAR-100 and GTSRB are near-ties while EuroSAT is not (82.2 against permutation's 76.6). On Table 2's backbones the probe finds nothing—all rungs within 0.4 points once the heads are retrained, the unmerged anchor leading—which is why that cell claims trunk–head compatibility rather than better backbone features.

E.19 Where the Calibration Images Come From

The 64 calibration images behind the ImageNet ladder are ImageNet-1k *training* images, one from each of 64 distinct classes (sampled deterministically; the class list and seed ship with the code), so no transform any reported ImageNet merge applies has seen an image the barrier is evaluated on (the CIFAR same-recipe study is the disclosed exception, Table 15, and Table 16 re-runs it on training-split images); earlier versions used the first 64 *validation* images (0.13% of the split), and we changed it rather than defend it. The switch is a one-factor swap that costs the ViT-B headline 0.36 barrier ($5.142 \rightarrow 5.504$ under the separable cost of that sweep; the shipped exact cost reads 5.534) and leaves every ordering intact, and neither the source nor the size carries the result. The pre-merging geometry is still measured from validation images, which is a different exposure: those are properties of the parent pair reported directly, not transforms scored on the images they came from.

E.20 Coverage and Limitations of the Evaluation

Every number in this paper is a measurement, reported in Section 4 and Appendix E. What the evaluation does *not* reach we list here rather than bury.

Architectures the theory does not cover. Propositions 1 and 2 and Eq. (3) describe *vanilla* multi-head attention with *additive* positional embeddings, and every model measured here—the ViTs and the byte-level Transformers—has that form. Three common departures change the group and we claim nothing about them. *Rotary embeddings* (Su et al., 2024) apply a position-dependent rotation inside the query–key product, so an admissible map must additionally commute with it; the group is strictly smaller than Eq. (3)'s, and Tran et al. (2026) characterize what remains, finding that sinusoidal and additive encodings preserve the vanilla structure where rotary ones materially shrink it. *Grouped- and multi-query attention* share one key–value projection across several query heads, which ties those heads together: they may no longer be permuted independently, and a per-head rotation must be common to the group that shares the projection. *Gated feed-forward blocks* (SwiGLU (Shazeer, 2020) and relatives) split the hidden layer into two multiplied branches, so a hidden unit is not the permutable object the FFN rung assumes; a permutation must act identically on both branches. None of these is measured here, and the title's scope should be read accordingly.

What the evaluation does not cover.

- Training-based references such as AdaMerging (Yang et al., 2024) are not ported, so the eight-task comparison is *training-free only*. TIES and DARE are measured on the shared-initialization CLIP benchmark (Table 12); our permutation controls are activation-estimated Birkhoff vertices and do not reproduce Git Re-Basin's iterative weight matching (Ainsworth et al., 2023); no out-of-distribution suite (ImageNet-A/R/Sketch) and no modern foundation language model is evaluated.
- ZipIt! is ported and measured at B/32 only: its greedy agglomeration scales with the square of the concatenated FFN width, and the B/32 control already localizes the row's collapse in the matching rather than the merge (Appendix E.9).
- RegMean is the one row reported off the shared 16-image budget—at 64, because it does not function at 16—with a curve still rising there, so its entry is a lower bound (Appendix E.8).
- $N{=}3$ covers ViT-S and ViT-B, DeiT-III (Touvron et al., 2022) supplying the third release; RaOT attains the lowest centroid barrier at both (6.142 and 6.063 against direct averaging's 6.251 and 6.319, and the permutation's 6.664 and 6.219—which at ViT-S is worse than not aligning at all), the margin over direct averaging again growing with scale ($0.11 \rightarrow 0.26$), although which rung attains it does not: RaOT-auto at ViT-S, RaOT-align at ViT-B. Rotation fusion is measured at $N{=}3$ under the previous calibration source only, so no matched comparison against it is available at this parent count. $N{=}4$ wants a fourth ImageNet-1k release with a matching state-dict signature, which does not exist.
- Width-heterogeneous merging is not an unrun experiment but a regime we deliberately do not claim: Theorem 1 does not apply once widths differ (Appendix E.6).
- The transported-heads single-parent reference for Table 2 is measured on the four label-split draws and matches the merged trunk to within noise (Appendix E.17), which is why Section 4

claims trunk–head compatibility and prices the merge itself at ≈ 0 there; the E/F/X constructions are not re-run under it.

- The causal validation of private-unit detection is a validation we owe for a mechanism we do report (Appendix E.6).

Algorithm 1: RaOT-Merge (training-free). RaOT supplies the alignment maps and the gate; the merge backend Merge_ϕ of the last step is a parameter of the pipeline, not part of it: linear interpolation is used in Table 1, and Table 2 composes RaOT alignment with DC-Merge. The extensions are deliberately absent from these lines, being approximate rather than exact symmetries—*private-unit weighting*, *unbalanced mass* and *curvature weighting* are specified in Appendix E.6; only private-unit weighting enters any reported number, in the shared-initialization experiment alone, where Table 8 isolates its $+0.43$ and $+0.87$ contribution. No independently trained result uses any of them.

Input: models $\{\theta_k\}_{k=1}^N$ (anchor θ_1), unlabeled calibration images $\mathcal{D}$, merge backend Merge_ϕ with its coefficients
Output: merged parameters $\theta^\star$

Fix gauges: absorb every LayerNorm affine into the linear it feeds, canonical positional gauge (Appendix C);
Forward $\mathcal{D}$ through all models once; collect every token of every image as a row of $\bar{H}_k^{(\ell)}$ at every interface and stream read point;
if *operating point is* RaOT-auto **then** *stage one:* estimate every candidate, let $\hat{f}$ be the fraction Eq. (13) accepts over all of them; declare **shared** if $\hat{f} < \tau$ and **independent** otherwise (τ=0.10 throughout; Appendix E.3);
if RaOT-auto *and* ***shared*** **then return** $\mathrm{Merge}_\phi(\{\theta_k\})$ unaligned, with RaOT's utilization weighting (Appendix B) — identity is the whole alignment map in this regime;
foreach *non-anchor parent* k **do**
  foreach *block* ℓ **do**
    heads (coupled): Hungarian on nuclear-norm costs Eq. (9) $\to \pi^{(\ell)}$, then one Procrustes per matched pair Eq. (8) $\to R_h^{(\ell)}, S_h^{(\ell)}$ — exact, single pass;
    FFN units: Hungarian on Eq. (5) $\to P^{(\ell)}$;
  stream (global): pool the centered cross-covariance over every read point, $\Sigma_c = \sum_\ell C\bar{H}_1^{(\ell)\top}\bar{H}_k^{(\ell)}C$ (ω_ℓ=1), and solve the constrained Procrustes Eq. (10) $\to$ one R_{res} (from the weights that read from and write to the stream instead, for the estimator ablation of Appendix B);
  gate, by operating point, in the order heads $\to$ FFN $\to$ stream, each candidate on the realized pull-back with the maps already accepted in place:;
    RaOT-align: accept every estimate;
    RaOT-safe: keep a map iff it lowers d_ℓ against identity (Eq. (13)); the head *correspondence* $\pi^{(\ell)}$ is applied ungated;
    RaOT-auto (**independent**): refit each candidate on the fitting images and keep it iff it cuts the centered activation residual on held-out images, else identity (Appendix E.3);
  apply: permute heads by $\pi^{(\ell)}$; rotate each head's QK and VO subspaces by the accepted $R_h^{(\ell)}, S_h^{(\ell)}$ and permute FFN units by the accepted $P^{(\ell)}$ inside each block; apply the accepted R_{res} to every stream writer and reader—patch embedding, tokens, attention and FFN outputs with their biases, qkv/fc_1 inputs, and the external head (Appendix C)—giving $\hat{\theta}_k$;
  if RaOT-safe *and* $\|W_1 - \hat{W}_k\| > \|W_1 - W_k\|$ **then** $\hat{\theta}_k \leftarrow \theta_k$: hand the *unaligned* parent to the backend (Eq. (14); the head correspondence being ungated, this composite check—not the sweep—certifies the guarantee; it has never fired on a pair we report, Appendix E.4);
  Verify the endpoint: $|L(\hat{\theta}_k) - L(\theta_k)|$ at machine precision (Appendix E.19);
Merge $\theta^\star = \mathrm{Merge}_\phi\left(\theta_1, \{\hat{\theta}_k\}_{k=2}^N\right)$: linear interpolation in Table 1, the retained-head backends of Table 2; N>2 aligns each parent to the fixed anchor (Appendix C.1);
return $\theta^\star$ (no fine-tuning);