# WHEN DOES ORTHOGONAL TRAINING HELP MODEL MERGING? OPERATOR-RELATIVE DISENTANGLEMENT IN TASK ARITHMETIC

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Task arithmetic combines independently fine-tuned models through their parameter updates. Training methods and merge rules are usually assessed one component at a time, leaving open whether a training intervention benefits merge rules uniformly or changes their relative utility. We study orthogonality-regularized training under a stricter test: after allowing every merge rule to improve or deteriorate together, does training still change their gains over direct addition? We formalize this question as *operator-relative disentanglement* and test multiplicity-controlled differences between operator responses. Across image and language task suites, orthogonal training can change relative operator responses rather than producing only a common gain shift. Under our prespecified inference procedure, two vision backbones exhibit resolved preference reversals and a language suite exhibits resolved changes in relative gaps; a larger vision backbone is a null result under the primary metric. Reaggregating its frozen outputs by worst-quartile instead of average utility also changes the response pattern. Shared-radius searches, fixed-map equal-norm audits, validation-only selection, and identity-inclusive deployment separate compatibility from global scale and endpoint quality. A local second-order analysis shows why Euclidean orthogonality alone cannot guarantee compatibility after a merge rule transforms the updates. These results motivate evaluating training, merging, and functional metrics as a coupled system.

## 1 INTRODUCTION

Task arithmetic stores a task as the difference between a fine-tuned model and a shared pretrained model. These differences, called task vectors, can be added or negated to combine or remove behaviors without joint retraining (Ilharco et al., 2023). Work on task arithmetic usually follows one of two paths. Training-side methods try to produce updates that interfere less, for example with tangent-space training (Ortiz-Jimenez et al., 2023) or orthogonality regularization (Liu et al., 2026). Merge-side methods change the vectors after training: TIES drops weak coordinates and resolves sign conflicts (Yadav et al., 2023), LiNeS changes update strength across depth (Wang et al., 2025), TSV-Merge works in task-vector subspaces (Gargiulo et al., 2025), and aTLAS learns coefficients for parameter blocks (Zhang et al., 2024).

Training-side and merge-side studies naturally vary one component while holding the other fixed. This leaves a practical design question unanswered: does a training intervention shift the gains of all downstream merge rules together, or does it change their relative utility? A generic nonzero interaction is unsurprising and does not answer this question. The stricter test must allow an arbitrary common gain shift, then ask whether operator responses still differ.

We test these factors together. For a fixed training procedure, we ask a simple question: how much does a merge operator improve over direct task-vector addition? We call this quantity its *compatibility* with the training procedure. Every family is tuned only on validation data and shares the same radius grid, while unequal family-wise search budgets are disclosed. This control is important: a structured operator can appear useful simply because direct addition used the wrong global scale.

The main question is whether training merely shifts all operator gains together or selectively changes their relative utility. Multiplicity-controlled paired contrasts reject a common-shift account on both



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

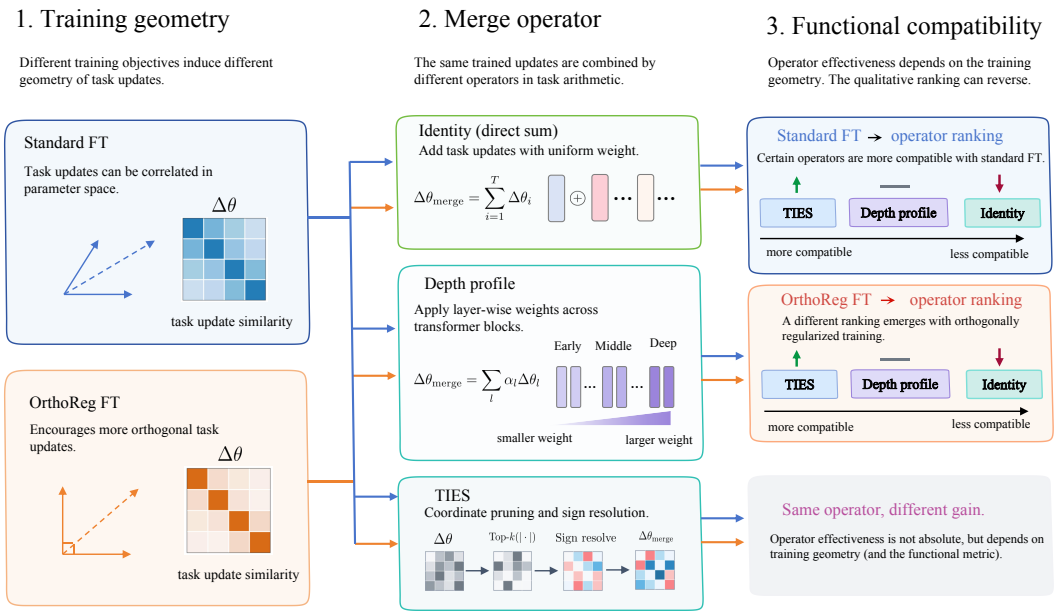


Figure 1: Operator-relative disentanglement. Training shapes task vectors, but direct addition, depth weighting, and coordinate sparsification use those vectors in different ways. Compatibility therefore depends on training, the merge operator, and the evaluation metric. The ranking glyphs are schematic; quantitative results appear in Table 1 and Table 2.

ViT-Base backbones and on the language suite. Two visual operator pairs also reverse order. ViT-Large does not reject the null under macro-average utility, which we count as a null result rather than evidence. On the same frozen ViT-Large outputs, however, a worst-quartile task aggregation produces a different response signature. A single training-only score therefore does not determine which merge readout is useful under a specified functional objective.

We contribute (i) a compatibility estimand, common-shift null, and pairwise contrast that make operator-relative disentanglement falsifiable; (ii) a paired factorial study with simultaneous inference across three vision backbones and a language suite; and (iii) fixed-output metric, scale, endpoint, and local curvature audits that delimit what the observed interactions do and do not establish. Existing operators serve as probes; we introduce no new merge rule.

## 2 Related Work

**Task vectors and training geometry.** Task arithmetic introduced addition and negation of fine-tuning deltas (Ilharco et al., 2023). Tangent Task Arithmetic relates composability to localized features in the neural tangent regime (Ortiz-Jimenez et al., 2023). OrthoReg identifies task-feature specialization as a sufficient condition for weight disentanglement and regularizes matrix-valued updates toward an internal orthogonal structure (Liu et al., 2026). Second-order compositionality links successful addition to curvature around the pretrained basin (Porrello et al., 2025), while curvature-aware regularization controls representation drift without retaining task data (Porrello et al., 2026). OrthoReg reports absolute improvements across several arithmetic methods; our within-geometry comparator asks the distinct question of whether each method adds more beyond its own tuned identity merge.

**Post-hoc merging operators.** Weight averaging and model soups demonstrate that models in a common basin can often be combined without additional training (Wortsman et al., 2022). TIES targets interference in a coordinate and sign basis (Yadav et al., 2023); DARE randomly drops and rescales model deltas (Yu et al., 2024); and TSV-Merge uses layerwise singular vectors and whitening to reduce task interference (Gargiulo et al., 2025). AdaMerging adapts taskwise and layerwise coefficients through test-time entropy minimization (Yang et al., 2024). LiNeS applies a

depth-increasing profile to parameter updates (Wang et al., 2025), while aTLAS learns parameter-block coefficients (Zhang et al., 2024). Sparse-mask and isotropic-subspace methods provide further coordinate and spectral readouts (Davari & Belilovsky, 2024; Marczak et al., 2025); recent approaches also align or preserve functionally important directions (Chaves et al., 2026; Li et al., 2026). Our fixed depth profile is not a new merging method. It is a controlled probe with a readout distinct from coordinate sparsification and singular-subspace correction.

**Alignment and evaluation regime.** Permutation alignment can move independently trained networks into a shared basin, while partial feature matching can merge models trained on disjoint tasks (Ainsworth et al., 2023; Stoica et al., 2024). Our models descend from the same checkpoint, so permutation alignment is held fixed rather than optimized. Large-scale studies show that base quality, model scale, and task pool can change the relative behavior of merge rules (Yadav et al., 2025; He et al., 2025; Hitit et al., 2025). We isolate a different factor: the interaction between source fine-tuning geometry and operator family under the same tasks, validation data, and radius search support.

**Functional notions of conflict.** Several methods recognize that Euclidean parameter conflict is an incomplete proxy for functional interference. Fisher merging uses a diagonal Fisher approximation to weight parameter directions (Matena & Raffel, 2022), while RegMean derives layerwise merge equations from second moments of input activations (Jin et al., 2023). TATR defines a trust region using task-loss gradients (Sun et al., 2025), Epistemic Merging uses an expected-Hessian metric in a task-vector subspace (Lu et al., 2026), and tangent-space analyses study the induced function change (Ortiz-Jimenez et al., 2023). These methods choose a particular geometry for merging. We make a different, narrower point: even with a fixed functional metric, the relevant geometry depends jointly on the source training rule and the merge operator. This dependence is the object measured by our factorial comparison. Adaptive projective gradient descent also uses functional retention to avoid discarding task-specific information (Wei et al., 2025), but does not isolate this training-by-operator contrast.

## 3 Operator-Relative Disentanglement

### 3.1 Setup

Let $G$ denote a training procedure and let $\mathcal{T}(G) = \{\tau_1^G, \dots, \tau_T^G\}$ be its task vectors. A merge operator $\mathcal{O}$ maps these vectors to a parameter displacement

$$d_{G,\mathcal{O}} = \mathcal{O}\big(\tau_1^G, \dots, \tau_T^G\big), \qquad \theta_{G,\mathcal{O}} = \theta_0 + d_{G,\mathcal{O}}. \tag{1}$$

The functional metric $\Phi$ specifies both the tasks and their aggregation. For the primary vision experiment, it is average normalized top-1 accuracy,

$$U(G, \mathcal{O}, \Phi) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathrm{Acc}_t(\theta_{G,\mathcal{O}})}{\mathrm{Acc}_t(\theta_t^G)}. \tag{2}$$

Normalization prevents tasks with high raw accuracy from dominating the average. Because single-task denominators depend on $G$, absolute utilities should be compared with care; our central quantity is a within-geometry gain. The language extension instantiates $\Phi$ with an unweighted mean of its task-specific metrics and is analyzed only within that suite.

For operator family $\mathcal{F}$, let $\mathcal{O}^\star_{G,\mathcal{F}}$ be its validation-selected member, including radius, and let $\mathcal{I}^\star_G$ be direct addition selected on the same grid. Define

$$C(G, \mathcal{F}, \Phi) = U(G, \mathcal{O}^\star_{G,\mathcal{F}}, \Phi) - U(G, \mathcal{I}^\star_G, \Phi). \tag{3}$$

This operational criterion indexes the measured training–operator–metric tuple; a positive value is not a certificate of pointwise weight disentanglement. A singleton family is a fixed map; Appendix E also fixes configuration and absolute radius across geometries. The training–operator interaction is the paired difference

$$\Delta_{\mathrm{int}}(\mathcal{F}, \Phi) = C(G_{\mathrm{Ortho}}, \mathcal{F}, \Phi) - C(G_{\mathrm{Std}}, \mathcal{F}, \Phi). \tag{4}$$

A positive value means that the family gains more from OrthoReg geometry. It says nothing about absolute utility or ordering under another metric. Because the selected family member may depend

on $G$, this operational interaction contains both a response to a fixed merge map and any change in the validation-selected configuration. We expose the selected configurations and separately audit fixed maps at common absolute norms in Appendix E.

An individual nonzero interaction is not our test: changing source training can unsurprisingly change any downstream method. Instead, for two operator families, define the operator-response contrast

$$\mathcal{H}_{ij}(\Phi) = \Delta_{\text{int}}(\mathcal{F}_i, \Phi) - \Delta_{\text{int}}(\mathcal{F}_j, \Phi). \tag{5}$$

The null permits an arbitrary training-dependent common shift,

$$H_0 : \quad \Delta_{\text{int}}(\mathcal{F}_k, \Phi) = \kappa(\Phi) \quad \text{for every } k, \tag{6}$$

and therefore implies $\mathcal{H}_{ij} = 0$ for every pair. This null is more permissive than zero interaction: source quality, calibration, or saturation may move all gains together. Rejecting it therefore requires evidence that training changes one operator relative to another. We test all six pairs among four families with simultaneous bootstrap intervals. A resolved nonzero contrast rejects the common-shift null; a preference reversal is a stronger special case. An unresolved contrast is not evidence.

Metric dependence is tested on the same frozen outputs, not by comparing different domains. For two metrics, we examine

$$\mathcal{Q}_{ij}(\Phi, \Phi') = \mathcal{H}_{ij}(\Phi) - \mathcal{H}_{ij}(\Phi'). \tag{7}$$

Thus the claim requires a change in the relative operator response when only the functional aggregation changes. We suppress the backbone index in the notation, but retain it experimentally because functional curvature and parameterization change with architecture.

### 3.2 Empirical predictions and failure conditions

The relational view makes three falsifiable predictions. First, training may produce a common gain shift, but at least one multiplicity-controlled $\mathcal{H}_{ij}$ must differ from zero. If all simultaneous intervals contain zero, the experiment does not support operator relativity. Second, a preference reversal between two operators is direct evidence that their response cannot be represented by a common shift, although a resolved change in their gap is sufficient. Third, metric relativity requires a resolved $\mathcal{Q}_{ij}$ on identical outputs and fixed configurations. Cross-domain differences and metric contrasts whose intervals contain zero do not count.

Scale remains a separate threat. If a contrast arises only because families use unequal scale ranges, it should disappear when they traverse the same radius grid or when configurations are evaluated at common absolute norms. These controls refine the estimand; they do not turn an unresolved contrast into positive evidence.

These predictions also delimit what the paper does not claim. The sign of one operator's interaction is descriptive, not by itself evidence for the central hypothesis. Nor does a resolved pairwise contrast imply that either operator is universally useful. We therefore report endpoint utility, within-geometry gain, individual interactions, and multiplicity-controlled operator contrasts separately.

### 3.3 Why Euclidean orthogonality is insufficient

Consider a task loss around $\theta_0$,

$$L_t(\theta_0 + d) \approx L_t(\theta_0) + g_t^\top d + \tfrac{1}{2} d^\top H_t d. \tag{8}$$

For a shared linear operator $A$, the transformed sum is $d = A \sum_i \tau_i$. Its pairwise second-order interaction on task $t$ is

$$\mathcal{I}_{t,A}(i,j) = \tau_i^\top A^\top H_t A \tau_j. \tag{9}$$

Thus $\tau_i^\top \tau_j \approx 0$ does not imply $\mathcal{I}_{t,A}(i,j) \approx 0$. For jointly nonlinear operators such as TIES, the corresponding local form contains operator Jacobians $J_{\mathcal{O},i}^\top H_t J_{\mathcal{O},j}$. The operative metric is induced jointly by the functional curvature and the merge rule.

The same expansion connects this mechanism to our empirical contrast. Let $s_G = \sum_i \tau_i^G$ and compare a shared linear operator $A$ with identity addition. Their local task-loss difference is

$$\begin{aligned} D_t(G, A) &\equiv L_t(\theta_0 + A s_G) - L_t(\theta_0 + s_G) \\ &\approx g_t^\top (A - I) s_G + \tfrac{1}{2} s_G^\top (A^\top H_t A - H_t) s_G. \end{aligned} \tag{10}$$

The training–operator interaction compares this quantity across source geometries. It can therefore change through first-order alignment, induced curvature, or both. Appendix A.1 gives the derivation and the corresponding nonlinear form. Lemma 1 also bounds the approximation error: at matched radius, the error in this contrast is cubic in the displacement norm when the Hessian is locally Lipschitz.

**Proposition 1** (Orthogonality does not determine operator-induced interaction)**.** *There exist Euclidean-orthogonal task vectors, a positive-definite functional metric, and two diagonal operators such that one operator eliminates pairwise functional interaction while the other introduces nonzero interaction. The construction is given in Appendix A.*

**Proposition 2** (Universal preservation is isotropic)**.** *Let $H \succ 0$, let $A$ be linear, and set $M = A^\top H A$. Then $x^\top M y = 0$ for every pair $x \perp y$ if and only if $M = cI$ for some $c \geq 0$. Therefore an operator preserves all Euclidean orthogonality in the functional metric only when its induced metric is isotropic. The proof is in Appendix A.*

**Proposition 3** (The interaction sign can be architecture-dependent)**.** *For fixed Euclidean-orthogonal vectors and a fixed operator, there exist two positive-definite functional metrics for which the pairwise interaction in Equation 9 has opposite signs. A two-dimensional construction is given in Appendix A.*

The propositions are local statements, not models of TIES or LiNeS in full. They establish the logical point required by our empirical analysis: an operator-independent Euclidean score cannot certify low functional interference for every operator, nor can an interaction sign established on one architecture be assumed on another. Proposition 2 further shows that universal preservation requires the exceptional condition $A^\top H A = cI$. Figure 1 summarizes this mechanism at the level of the full experimental pipeline.

**Nonlinear merge rules.** The argument extends locally to differentiable merge rules by replacing $A$ with task-specific operator Jacobians and adding a term for the operator's mixed second derivative. Piecewise-affine rules such as TIES have this form away from mask and sign boundaries. Consequently, two rules may preserve, rotate, or suppress different parts of the same training geometry even when no single global matrix represents either rule. Appendix A.1 gives the chain-rule expression and its assumptions. The exact Hessian–vector diagnostic in Appendix B tests selected output directions; it does not estimate the operator's input-side Jacobian.

### 3.4 Separating scale from shape

Let $v_{G,\mathcal{O}}$ be the raw displacement returned by an operator and $v_{G,\mathcal{I}}$ the identity sum. For relative radius $r$, we evaluate

$$d_{G,\mathcal{O}}(r) = r\,\alpha_{\mathrm{ref}}\frac{\|v_{G,\mathcal{I}}\|_2}{\|v_{G,\mathcal{O}}\|_2}v_{G,\mathcal{O}}. \tag{11}$$

The operator and identity therefore traverse the same grid of parameter-space radii. Here $\alpha_{\mathrm{ref}}$ is the validation-selected identity coefficient for the same training geometry and paired run. Equation 11 does not make their directions equivalent; it controls global norm without assuming that equal-norm directions have equal functional effect. To make this distinction explicit, define $u_{G,\mathcal{O}} = v_{G,\mathcal{O}}/\|v_{G,\mathcal{O}}\|_2$ and let $\rho = r\alpha_{\mathrm{ref}}\|v_{G,\mathcal{I}}\|_2$. Substitution into Equation 8 gives the equal-radius local contrast

$$\begin{aligned} &L_t(\theta_0 + \rho u_{G,\mathcal{O}}) - L_t(\theta_0 + \rho u_{G,\mathcal{I}}) \\ &\quad\approx \rho g_t^\top(u_{G,\mathcal{O}} - u_{G,\mathcal{I}}) + \frac{\rho^2}{2}\left(u_{G,\mathcal{O}}^\top H_t u_{G,\mathcal{O}} - u_{G,\mathcal{I}}^\top H_t u_{G,\mathcal{I}}\right). \end{aligned} \tag{12}$$

Thus matching radius removes a difference in update norm, not the direction-dependent first-order sensitivity or curvature. At a stationary point, only the operator-induced directional curvature remains.

**Proposition 4** (Equal radius does not equalize functional effect)**.** *Let $g_t = 0$ and let the local loss in Equation 8 be exact. For any $\rho > 0$ and unit directions $u, w$,*

$$L_t(\theta_0 + \rho u) - L_t(\theta_0 + \rho w) = \tfrac{1}{2}\rho^2(u^\top H_t u - w^\top H_t w).$$

*This contrast vanishes for every unit pair $u, w$ if and only if $H_t = cI$ for some scalar $c$. The proof is in Appendix A.*

The proposition gives the precise scope of our scale control. At each candidate $r$, the operator and identity have equal displacement norm, while their directions and functional effects can differ. In the primary protocol, each family independently selects $r$ on validation data from the same grid; the reported endpoints therefore have common radius support but need not have the same selected radius. We evaluate each selected member once on test data.

## 4 Experimental Design

**Models and tasks.** We use OpenAI-pretrained CLIP ViT-B/32, ViT-B/16, and ViT-L/14 encoders (Radford et al., 2021). Following the task-arithmetic benchmark, we fine-tune separate encoders on Cars, DTD, EuroSAT, GTSRB, MNIST, RESISC45, SUN397, and SVHN. Each merged encoder is evaluated with the corresponding fixed task-specific classification head. We compare standard nonlinear fine-tuning with OrthoReg fine-tuning at $\lambda = 10$. Both use AdamW, learning rate $10^{-5}$, weight decay $0.1$, a 500-step warmup, the dataset-specific epoch schedule inherited from the OrthoReg implementation, and three independent random initializations. A separate ViT-B/32 audit repeats the complete factorial at $\lambda \in \{0, 1, 3, 10, 30\}$ to distinguish a training-geometry response from an isolated choice of regularization strength.

**Operator probes.** Identity task arithmetic sums all eight vectors. The *fixed depth* probe multiplies the first, middle, and final thirds of each task vector's visual-transformer blocks by $(0.30, 0.75, 1.25)$, preserves each vector's $\ell_2$ norm, and then sums. It is a fixed diagnostic probe from the monotone family $(0.30^s, 0.75^s, 1.25^s)$, not a new merge method. Its shape was fixed before the three-run evaluation. Appendix M discloses its exploratory origin, and Appendix H reports a multi-shape robustness audit. LiNeS applies its official multi-task linear depth profile to the summed vector; $\beta$ is selected from $\{0.1, 0.2, \ldots, 1.0\}$ (Wang et al., 2025). TIES is evaluated with summed reduction and densities $\{0.2, 0.4, 0.6\}$ (Yadav et al., 2023). TSV-Merge retains the leading $1/T$ singular fraction of every task update per matrix layer, jointly orthogonalizes the left and right singular bases, and averages non-matrix parameters (Gargiulo et al., 2025). The families therefore read task vectors through global addition, depth, coordinate/sign sparsity, and layerwise low-rank subspaces.

**Selection and reporting.** For every $(G, \mathcal{O})$ pair, relative radius is selected from $\{0.5, 0.625, \ldots, 1.5\}$. All radius and family hyperparameters are selected using only the eight validation splits. We then evaluate the selected member of each family once on the eight test splits. The main vision table reports the mean across three independently trained, fully paired runs; sample standard deviations and run-level values are reported in the appendix. The primary statistic is the test gain in average normalized top-1 accuracy over the independently selected identity comparator, reported in percentage points (pp), with interactions computed within run before aggregation. The family-level compatibility analysis applies the selected operator even when validation prefers identity; Appendix L also reports the practical identity-inclusive fallback. The method-faithful family search spaces are unequal: fixed depth and TSV-Merge use 9 validation evaluations, TIES uses 27, and LiNeS uses 90. We estimate each stated family's validation-selected envelope, not a budget-neutral search comparison; complete curves appear in Appendix I.

**Inference for operator heterogeneity.** For each backbone, we form all six contrasts in Equation 5. A paired hierarchical bootstrap resamples the three runs and eight tasks with shared indices across geometries and operators. We report 20,000-draw percentile intervals and simultaneous 95% intervals obtained from the maximum absolute bootstrap error across all six contrasts. The common-shift null is rejected only by a simultaneous interval that excludes zero. Rank reversal is reported separately and is not required for rejection. The language analysis uses the same run–task resampling. With three training runs, these intervals are prespecified finite-sample summaries over the observed run–task hierarchy, not asymptotic guarantees over unseen architectures or task distributions.

For the fixed-output metric audit, configurations remain those selected by the primary validation metric. We recompute utility as raw macro accuracy, specialist-normalized macro accuracy, and worst-quartile normalized utility (the mean of the two lowest task utilities per run). We report every metric pair and operator pair, with simultaneous correction over all 18 comparisons within each backbone.

**Language extension.** To test whether the effect is restricted to visual encoders, we use the FusionBench implementation (Tang et al., 2025) to merge eight instruction-tuned Flan-T5-base checkpoints (Chung et al., 2022) for the GLUE tasks (Wang et al., 2019). Standard and OrthoReg checkpoints are trained with three independent random initializations. We use the same identity, fixed-depth, LiNeS, TIES, and TSV operator families. Settings are selected once on development data from the first run, then frozen for all three held-out evaluations. The language metric is the unweighted mean of the FusionBench task metrics: exact accuracy for seven tasks and Spearman correlation for STS-B. The detailed protocol, per-run outcomes, and a common-radius TIES control appear in Appendix K.

**Controls.** The paired $2 \times 5$ factorial shares validation data and radius grids across families, discloses unequal configuration counts, reselects identity within each geometry, and touches test once per selected configuration. Run-, task-, selection-, and deployment-level audits appear in Appendices J–L.

# 5 RESULTS

## 5.1 VISION: OPERATOR RESPONSES DIFFER ACROSS BACKBONES

Table 1: Mean vision compatibility over three paired runs under a shared radius grid (pp). Identity is normalized top-1 utility; all other entries are gains over the independently validation-selected identity. $\Delta$ is the paired OrthoReg-minus-Standard gain. Identity values are not comparable across training geometries because their normalizers differ. Mean $\pm$ sample standard deviation is reported in Appendix 4.

| ViT-B/32 | | | | ViT-B/16 | | | | ViT-L/14 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Operator | Std. | Ortho. | $\Delta$ | Operator | Std. | Ortho. | $\Delta$ | Operator | Std. | Ortho. | $\Delta$ |
| Identity | 77.61 | 96.06 | – | Identity | 81.21 | 96.15 | – | Identity | 89.40 | 109.24 | – |
| Fixed depth | 1.95 | 3.39 | +1.44 | Fixed depth | 0.21 | 2.17 | +1.96 | Fixed depth | 2.15 | 1.15 | −1.00 |
| LiNeS | 4.48 | 3.23 | −1.25 | LiNeS | 2.45 | 2.22 | −0.24 | LiNeS | 2.30 | 1.02 | −1.28 |
| TIES | 1.80 | 0.25 | −1.54 | TIES | 1.09 | −0.33 | −1.42 | TIES | 0.80 | −0.65 | −1.45 |
| TSV-Merge | 8.24 | 8.88 | +0.65 | TSV-Merge | 6.36 | 7.70 | +1.34 | TSV-Merge | 4.43 | 2.87 | −1.56 |

The individual interactions in Table 1 are descriptive; their being nonzero is not the central test. Table 2 instead compares operators directly after allowing an arbitrary common training shift. On ViT-B/32, the fixed-depth response exceeds the LiNeS and TIES responses by 2.69 and 2.98 pp. On ViT-B/16, fixed depth exceeds TIES by 3.38 pp, while TIES falls 2.76 pp below TSV-Merge. All four simultaneous intervals exclude zero. The fixed-depth–LiNeS ordering reverses on B/32, and the fixed-depth–TIES ordering reverses on B/16.

The ViT-B/32 strength sweep supports the same qualitative pattern without reducing training geometry to two recipes. As $\lambda$ increases from 0 to 30, a scale-free within-update isotropy error decreases from 0.956 to 0.721 and raw identity accuracy rises from 70.14 to 76.23. Nevertheless, the raw fixed-depth interaction is positive at every nonzero strength, the LiNeS and TIES interactions are negative at every nonzero strength, and TSV-Merge changes sign. The interaction pattern is therefore neither a $\lambda = 10$ accident nor a monotone function of one orthogonality statistic (Appendix D).

ViT-L/14 is a null result for the primary hypothesis: none of its six simultaneous operator contrasts excludes zero under normalized macro utility. Although every individual operator adds less beyond the OrthoReg identity merge, that common downward movement does not establish operator relativity. The normalized identity entries also cannot establish which training procedure has the stronger endpoint because their denominators differ; the common-scale audit in Table 5 uses raw top-1 accuracy.

**Absolute utility, compatibility, and deployment answer different questions.** On a common raw-accuracy scale, OrthoReg produces a stronger identity merge and a stronger TSV-Merge endpoint on all three backbones (Table 5). Hierarchical bootstrap intervals for every paired endpoint difference exclude zero. With a validation-gated identity fallback, Standard/OrthoReg TIES gains are $1.80/0.18$, $1.09/0$, and $0.80/0$ pp on B/32, B/16, and L/14. The safeguard prevents harmful deployment while preserving the geometry-dependent loss of marginal utility; full outcomes appear in Appendix L.

Table 2: Multiplicity-controlled evidence against the common-shift null. We list every operator pair whose simultaneous 95% interval excludes zero, plus the ViT-L/14 null result. Appendix G reports all pairs. A reversal means that the gain ordering changes across training geometries.

| Backbone | Operator contrast | $\mathcal{H}_{ij}$ (pp) | Simultaneous 95% CI | |
|---|---|---|---|---|
| ViT-B/32 | Fixed depth − LiNeS | $+2.69$ | $[+0.14, +5.23]$ | reversal |
| | Fixed depth − TIES | $+2.98$ | $[+0.44, +5.53]$ | |
| ViT-B/16 | Fixed depth − TIES | $+3.38$ | $[+0.85, +5.91]$ | reversal |
| | TIES − TSV-Merge | $-2.76$ | $[-5.28, -0.23]$ | |
| ViT-L/14 | No resolved pair | – | – | |
| Language | Fixed depth − TIES | $+6.61$ | $[+0.86, +12.36]$ | |
| | LiNeS − TIES | $+6.29$ | $[+0.53, +12.04]$ | |

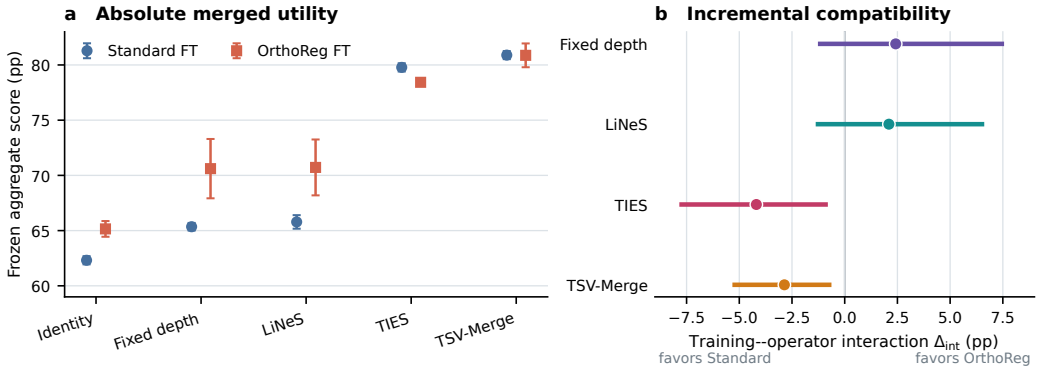


Figure 2: Language results separate final utility from incremental compatibility. Left: frozen aggregate score after merging; markers and bars show the mean and sample standard deviation over independent runs. Right: paired training–operator interactions with hierarchical bootstrap 95% intervals over runs and tasks. TIES and TSV-Merge remain the strongest final models, but both add less beyond the OrthoReg identity merge.

### 5.2 Language: coordinate and subspace corrections lose marginal value

Figure 2 extends the analysis beyond CLIP encoders; the complete numerical matrix is reported in Appendix K. The central evidence is not that TIES has a negative individual interaction. Rather, the fixed-depth–TIES and LiNeS–TIES operator contrasts are $+6.61$ and $+6.29$ pp, with simultaneous 95% intervals $[+0.86, +12.36]$ and $[+0.53, +12.04]$. The common-shift null is therefore rejected for this suite. No language operator pair reverses order, so the result supports a selective change in relative gaps, not a rank reversal. The individual fixed-depth and LiNeS intervals cross zero and are not counted as evidence on their own. Every language family beats identity during development selection, so the identity-inclusive deployment view is unchanged.

Changing only the OrthoReg TIES radius to the Standard value of $0.2$ leaves a $-5.05$ pp interaction, with a 95% interval of $[-8.73, -1.66]$ (Appendix K). This control removes the selected-radius difference, while leaving each geometry's selected sparsity setting unchanged. Mean task-vector cosine similarity also decreases from $0.135$ under Standard FT to $0.0195$ under OrthoReg. These observations rule out the radius difference as the sole explanation; they are not a causal decomposition of the operator interaction.

### 5.3 Resolved evidence, null results, and metric dependence

The evidence has a deliberately asymmetric interpretation. ViT-B/32 and ViT-B/16 reject the common-shift null, with one resolved preference reversal on each backbone. The language suite rejects the same null through two resolved relative-gap changes but has no reversal. ViT-L/14 under normalized macro utility does not reject the null. Individual interactions or trends whose simultaneous contrasts cross zero are descriptive null results, not a third form of positive evidence.

The metric claim is evaluated within ViT-L/14, where the macro result is null. We hold every trained checkpoint and selected operator configuration fixed and replace normalized macro utility by worst-quartile normalized utility. The latter averages the two lowest task utilities in each run. Three operator-pair responses change after simultaneous correction over all 18 metric-by-pair comparisons. For example, the fixed-depth–TIES contrast changes from $+0.45$ pp under macro utility to $-4.45$ pp under worst-quartile utility; the simultaneous 95% interval for the macro-minus-worst-quartile difference is $[+1.20, +8.60]$ pp. Thus metric relativity is not inferred from comparing vision with language: it is observed when only the aggregation of identical held-out outputs changes. The three resolved contrasts and complete interaction signatures are reported in Appendix G (Tables 11 and 12).

Across both domains, the within-geometry identity comparator, common radius support, and paired inference isolate selective responses beyond an arbitrary common shift. They do not identify a unique causal coordinate or an architecture-invariant ranking. Curvature and common-norm controls further separate local mechanism, fixed-map sensitivity, and selected-family behavior (Appendices B and E).

## 6 Discussion

**Disentanglement is relational, not absolute.** OrthoReg constrains update matrices in Euclidean geometry, whereas the tested operators transform them in different bases. Checkpoint collections should therefore be described by a training–operator compatibility matrix, not one disentanglement score. Crucially, a common movement of all operator gains, a negative individual interaction, or an unresolved trend is not evidence; the claim requires multiplicity-controlled differences between operator responses.

**Compatibility is distinct from endpoint quality.** A negative interaction need not mean a worse final model: raw vision accuracy and the language aggregate show stronger OrthoReg identity endpoints, while some structured rules add less beyond them (Tables 5 and 18). Merge operators should therefore be evaluated under multiple source geometries, with radius support, search budget, identity fallback, final utility, and incremental compatibility reported separately.

**Limitations.** The training factor compares standard fine-tuning with one orthogonality regularizer; its strength sweep is not an independent geometry. Metric relativity resolves on one backbone, while the ViT-L/14 null and three training runs limit generalization. Broader training recipes, models, tasks, and metrics remain necessary tests.

## 7 Conclusion

Task-vector disentanglement cannot be summarized by training-side orthogonality alone. After allowing an arbitrary common shift in merge-rule gains, paired operator contrasts show that source training selectively changes relative operator utility: two vision backbones exhibit preference reversals, the language suite exhibits resolved changes in operator gaps, and the larger vision backbone remains unresolved under the primary metric. Reaggregating the same held-out outputs also changes the response signature, demonstrating that the conclusion depends on the functional objective rather than weights alone. Equal-radius controls rule out global scale as the sole explanation, while the local second-order analysis shows why Euclidean orthogonality need not survive an operator's functional geometry. These findings do not identify a universally best merge rule. Instead, they motivate reporting endpoint quality and incremental compatibility separately, and evaluating training geometry, merge readout, and functional metric as a coupled system.

**AI use statement.** Generative AI assisted with literature search and synthesis, checking mathematical claims and proofs, and drafting and polishing the manuscript. The authors verified the cited sources and mathematical arguments, reviewed and revised all AI-assisted text, and take full responsibility for the final content of this work.

## REFERENCES


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

Levy Chaves, Chao Zhou, Rebekka Burkholz, Eduardo Valle, and Sandra Avila. Bridging domains through subspace-aware model merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7990–7999, 2026.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, 2024.

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*, 2025.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A benchmark for merging domain-specialized LLMs. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 38, 2025.

Oğuz Kağan Hitit, Leander Girrbach, and Zeynep Akata. A systematic study of model merging techniques in large language models. *arXiv preprint arXiv:2511.21437*, 2025.

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

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

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *International Conference on Machine Learning*, pp. 3570–3579. PMLR, 2019.

Longhua Li, Lei Qi, Qi Tian, and Xin Geng. Model merging in the essential subspace. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Shangge Liu, Yuehan Yin, Lei Wang, Qi Fan, Yinghuan Shi, Wenbin Li, Yang Gao, and Dacheng Tao. Understanding and enforcing weight disentanglement in task arithmetic. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28744–28753, 2026.

Juanwu Lu, Anand Bhaskar, Brian Axelrod, Ekaterina Tolstaya, and Tristan Emrich. Model merging on loss landscapes: A geometric perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7644–7653, 2026.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, volume 267, pp. 43177–43199. PMLR, 2025.

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

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A second-order perspective on model compositionality and incremental learning. In *International Conference on Learning Representations*, 2025.

Angelo Porrello, Pietro Buzzega, Felix Dangel, Thomas Sommariva, Riccardo Salami, Lorenzo Bonicelli, and Simone Calderara. Dataless weight disentanglement in task arithmetic via kronecker-factored approximate curvature. In *International Conference on Learning Representations*, 2026.

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*, pp. 8748–8763. PMLR, 2021.

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*, 2024.

Wenju Sun, Qingyong Li, Wen Wang, Yangli-ao Geng, and Boyang Li. Task arithmetic in trust region: A training-free model merging approach to navigate knowledge conflicts. *arXiv preprint arXiv:2501.15065*, 2025.

Anke Tang, Li Shen, Yong Luo, Enneng Yang, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. FusionBench: A unified library and comprehensive benchmark for deep model fusion. *Journal of Machine Learning Research*, 26(307):1–38, 2025.

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

Ke Wang, Nikos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations*, 2025.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In *International Conference on Machine Learning*, volume 267, pp. 66178–66193. PMLR, 2025.

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*, pp. 23965–23998. PMLR, 2022.

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*, volume 36, 2023.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal, and Tsendsuren Munkhdalai. What matters for model merging at scale? *Transactions on Machine Learning Research*, 2025.

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

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*, 2024.

Frederic Z. Zhang, Paul Albert, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Knowledge composition using task vectors with learned anisotropic scaling. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

# A PROOFS OF THE GEOMETRIC CLAIMS

*Proof of Proposition 1.* Let $\tau_1 = (1,1)^\top/\sqrt{2}$ and $\tau_2 = (1,-1)^\top/\sqrt{2}$, so that $\tau_1^\top \tau_2 = 0$. Let $H = \mathrm{diag}(h_1, h_2)$ with $h_1, h_2 > 0$. For the diagonal whitening operator $A_d = \mathrm{diag}(h_1^{-1/2}, h_2^{-1/2})$,

$$\tau_1^\top A_d^\top H A_d \tau_2 = \tau_1^\top \tau_2 = 0.$$

For the coordinate selector $A_c = \mathrm{diag}(1,0)$, however,

$$\tau_1^\top A_c^\top H A_c \tau_2 = h_1/2 > 0.$$

The vectors and functional metric are unchanged; only the operator differs. Hence Euclidean orthogonality cannot determine the induced functional interaction for both operators. □

*Proof of Proposition 2.* The reverse implication is immediate: if $M = cI$, then $x^\top M y = c x^\top y = 0$ whenever $x \perp y$. For the forward implication, let $\{e_k\}_{k=1}^d$ be the standard basis. Since $e_i \perp e_j$ for $i \neq j$, the premise gives $M_{ij} = e_i^\top M e_j = 0$; hence $M$ is diagonal. For any $i \neq j$, the vectors $x = e_i + e_j$ and $y = e_i - e_j$ are orthogonal, so

$$0 = x^\top M y = M_{ii} - M_{jj}.$$

All diagonal entries are therefore equal and $M = cI$. Finally, $M = A^\top H A \succeq 0$ implies $c \geq 0$. □

*Proof of Proposition 3.* Take $\tau_1 = e_1$, $\tau_2 = e_2$, and $A = I$. For any $0 < \rho < 1$, define

$$H_+ = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \qquad H_- = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

Both matrices are positive definite because their eigenvalues are $1 - \rho$ and $1 + \rho$. Nevertheless, $\tau_1^\top H_+ \tau_2 = \rho$ whereas $\tau_1^\top H_- \tau_2 = -\rho$. Thus a fixed operator acting on fixed Euclidean-orthogonal vectors can have opposite pairwise interactions under two functional metrics. Different backbones need not realize this exact two-dimensional construction; it establishes that architecture-invariant interaction signs do not follow from orthogonality alone. □

*Proof of Proposition 4.* With $g_t = 0$, substitute $d = \rho u$ and $d = \rho w$ into the exact quadratic loss and subtract. This gives the displayed identity. If $H_t = cI$, then $u^\top H_t u = w^\top H_t w = c$ for all unit vectors, so the contrast vanishes. Conversely, suppose it vanishes for every unit pair. Then $u^\top H_t u$ is constant on the unit sphere. Taking $u = e_i$ shows that all diagonal entries of $H_t$ equal the same scalar $c$. Taking $u = (e_i + e_j)/\sqrt{2}$ and comparing it with $e_i$ gives $(H_t)_{ij} = 0$ for each $i \neq j$. Hence $H_t = cI$. □

**Lemma 1** (Remainder of the matched-radius local contrast)**.** *Suppose $L_t$ is twice continuously differentiable and its Hessian is $L_H$-Lipschitz in a convex neighborhood containing the line segments from $\theta_0$ to $\theta_0 + d_1$ and $\theta_0 + d_2$. Let $q_t$ be the quadratic model in Equation 8. Then*

$$|L_t(\theta_0 + d_k) - L_t(\theta_0) - q_t(d_k)| \leq \frac{L_H}{6}\|d_k\|_2^3, \qquad k \in \{1,2\}, \tag{13}$$

$$|[L_t(\theta_0 + d_1) - L_t(\theta_0 + d_2)] - [q_t(d_1) - q_t(d_2)]| \leq \frac{L_H}{6}(\|d_1\|_2^3 + \|d_2\|_2^3). \tag{14}$$

*In particular, if both displacements have radius $\rho$, the contrast error is at most $L_H \rho^3/3$.*

*Proof.* Taylor's theorem with integral remainder gives

$$R_t(d) = \int_0^1 (1-s) d^\top [\nabla^2 L_t(\theta_0 + sd) - \nabla^2 L_t(\theta_0)] d \, \mathrm{d}s.$$

Hessian Lipschitzness bounds the integrand in absolute value by $L_H s(1-s)\|d\|_2^3$. Since $\int_0^1 s(1-s)\,\mathrm{d}s = 1/6$, this proves Equation 13. The triangle inequality applied to $R_t(d_1) - R_t(d_2)$ proves Equation 14; substituting $\|d_1\|_2 = \|d_2\|_2 = \rho$ gives the final bound. □

### A.1 Local compatibility and nonlinear chain-rule derivations

Let $q_t(d) = g_t^\top d + \frac{1}{2} d^\top H_t d$ denote the nonconstant part of the local model in Equation 8. For $s_G = \sum_i \tau_i^G$, direct subtraction gives

$$\begin{aligned} q_t(As_G) - q_t(s_G) &= g_t^\top (A - I) s_G + \tfrac{1}{2} \left( (As_G)^\top H_t (As_G) - s_G^\top H_t s_G \right) \\ &= g_t^\top (A - I) s_G + \tfrac{1}{2} s_G^\top (A^\top H_t A - H_t) s_G, \end{aligned}$$

which is Equation 10. Applying the same identity once with $G_{\mathrm{Ortho}}$ and once with $G_{\mathrm{Std}}$, then subtracting, yields the local difference-in-differences

$$\begin{aligned} \widehat{\Delta}_{t,A}^{\mathrm{loss}} &= g_t^\top (A - I)(s_{\mathrm{Ortho}} - s_{\mathrm{Std}}) \\ &\quad + \tfrac{1}{2} \left[ s_{\mathrm{Ortho}}^\top K_A s_{\mathrm{Ortho}} - s_{\mathrm{Std}}^\top K_A s_{\mathrm{Std}} \right], \qquad K_A = A^\top H_t A - H_t. \end{aligned} \tag{15}$$

The first term measures a change in first-order alignment; the second measures the same operator through two training-induced directions in the curvature metric. Pairwise Euclidean overlap determines neither term. The empirical estimand in Equation 4 uses accuracy rather than loss, so Equation 15 is an explanatory local surrogate, not an equality for the reported percentage-point interactions.

For the nonlinear case, write $z = (\tau_1, \ldots, \tau_T)$, $d = \mathcal{O}(z)$, and $F_t(z) = L_t(\theta_0 + \mathcal{O}(z))$. The second-order chain rule in directions $u_i$ and $u_j$ gives

$$D_{ij}^2 F_t(z)[u_i, u_j] = u_i^\top J_{\mathcal{O},i}^\top \bar{H}_t J_{\mathcal{O},j} u_j + \bar{g}_t^\top \mathcal{H}_{\mathcal{O},ij}[u_i, u_j],$$

where $\bar{g}_t = \nabla L_t(\theta_0 + \mathcal{O}(z))$ and $\bar{H}_t = \nabla^2 L_t(\theta_0 + \mathcal{O}(z))$. This form follows by differentiating $D_i F_t(z)[u_i] = \bar{g}_t^\top J_{\mathcal{O},i} u_i$ with respect to block $j$. The second term is zero if the functional gradient is zero or if the merge rule is locally affine. TIES is piecewise affine away from mask and sign boundaries, so within one such region its interaction is governed by the first term, with task-dependent Jacobians created by masking and sign election. At a boundary, the classical Hessian need not exist; our empirical finite comparison remains well-defined and does not assume differentiability there.

## B Direct Directional-Curvature Diagnostic

The main experiments establish functional training–operator interactions but do not by themselves identify a local mechanism. As a validation-only diagnostic on one paired ViT-B/32 replicate, we therefore measure the quadratic form of the task loss along each selected operator output. For a direction $d$, we compute $g_t^\top d$ and $d^\top H_t d$ at the pretrained checkpoint using reverse-over-reverse automatic differentiation, which gives an exact Hessian–vector product without materializing $H_t$ (Pearlmutter, 1994). The loss is cross-entropy under the frozen task-specific head. Operator hyperparameters are inherited from validation-only selection, and the diagnostic uses 128 fixed validation examples from each of the eight tasks.

All Standard and OrthoReg directions are rescaled to the same visual-parameter norm, $\|d\|_2 = 2$. This common absolute radius is below the smallest selected identity displacement, so comparisons across training geometries do not inherit their different deployment radii. For radius multiplier $a$, we compare the measured loss change with

$$P_1(a) = a g_t^\top d, \qquad P_2(a) = a g_t^\top d + \tfrac{1}{2} a^2 d^\top H_t d. \tag{16}$$

Table 3 reports operator-minus-identity contrasts at the smallest radius. Negative loss contrast is favorable. The depth-structured operators and TIES induce different functional responses despite identical parameter norm: depth calibration lowers local loss, whereas TIES raises it in both training geometries. The quadratic prediction also reduces mean absolute error for at least six of eight tasks in every row.

The approximation has a measurable locality boundary. At $a = 0.125$, the second-order contrast error remains lower on average than the first-order error for all three operators under both training geometries. For Standard FT, this advantage is no longer uniform by $a = 0.25$ and generally disappears at larger radii, consistent with Lemma 1. We therefore do not use the local expansion as a surrogate for deployment accuracy.

Table 3: Exact directional-curvature diagnostic at common parameter norm and radius multiplier $a = 0.0625$. $\Delta q$ is the operator-minus-identity contrast in $d^\top H_t d$; $\Delta L$ is the measured loss contrast. $E_1$ and $E_2$ are task-mean absolute errors of the first- and second-order contrast predictions. The last column is the fraction of tasks for which $E_2 < E_1$.

| Training | Operator | $\Delta q$ | $\Delta L$ | $E_1$ | $E_2$ | Fraction |
|---|---|---|---|---|---|---|
| Standard | Fixed depth | $-0.4760$ | $-0.00951$ | $0.00192$ | $0.00092$ | $0.875$ |
| Standard | LiNeS | $0.6138$ | $-0.00587$ | $0.00202$ | $0.00088$ | $0.750$ |
| Standard | TIES | $0.0625$ | $0.00648$ | $0.00084$ | $0.00042$ | $0.875$ |
| OrthoReg | Fixed depth | $-0.0037$ | $-0.00036$ | $0.00004$ | $< 10^{-5}$ | $0.875$ |
| OrthoReg | LiNeS | $-0.0090$ | $-0.00054$ | $0.00003$ | $< 10^{-5}$ | $0.875$ |
| OrthoReg | TIES | $-0.0059$ | $0.00429$ | $0.00004$ | $< 10^{-5}$ | $0.750$ |

Direct curvature does not by itself explain the observed training–operator accuracy interaction. Paired task-bootstrap intervals for the OrthoReg-minus-Standard change in $\Delta q$ are $[-0.088, 1.098]$ for fixed depth, $[-1.524, 0.239]$ for LiNeS, and $[-0.541, 0.356]$ for TIES; all include zero. The diagnostic supports the narrower claim that equal-norm operators are functionally distinct and that second-order terms improve the local loss model. It does not support a stronger claim that Hessian curvature alone causes OrthoReg's held-out operator ranking. We do not substitute the empirical Fisher for this quantity because it is a gradient second moment and is not generally a Hessian approximation (Kunstner et al., 2019).

## C Supplementary Vision Results

Table 4: Complete vision compatibility under the shared radius grid (mean $\pm$ sample standard deviation over three paired runs; pp). This is the dispersion-complete version of Table 1. Identity is normalized top-1 utility; other entries are gains over the independently validation-selected identity. $\Delta$ is the paired OrthoReg-minus-Standard gain.

| ViT-B/32 | | | | ViT-B/16 | | | | ViT-L/14 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Operator | Std. | Ortho. | $\Delta$ | Operator | Std. | Ortho. | $\Delta$ | Operator | Std. | Ortho. | $\Delta$ |
| Identity | $77.61 \pm .27$ | $96.06 \pm .16$ | – | Identity | $81.21 \pm .54$ | $96.15 \pm .69$ | – | Identity | $89.40 \pm .29$ | $109.24 \pm .63$ | – |
| Fixed depth | $1.95 \pm .36$ | $3.39 \pm .28$ | $+1.44 \pm .45$ | Fixed depth | $0.21 \pm .24$ | $2.17 \pm .28$ | $+1.96 \pm .04$ | Fixed depth | $2.15 \pm .49$ | $1.15 \pm .13$ | $-1.00 \pm .52$ |
| LiNeS | $4.48 \pm .07$ | $3.23 \pm .11$ | $-1.25 \pm .04$ | LiNeS | $2.45 \pm .32$ | $2.22 \pm .34$ | $-0.24 \pm .39$ | LiNeS | $2.30 \pm .42$ | $1.02 \pm .05$ | $-1.28 \pm .40$ |
| TIES | $1.80 \pm .10$ | $0.25 \pm .40$ | $-1.54 \pm .36$ | TIES | $1.09 \pm .11$ | $-0.33 \pm .12$ | $-1.42 \pm .20$ | TIES | $0.80 \pm .24$ | $-0.65 \pm .31$ | $-1.45 \pm .13$ |
| TSV-Merge | $8.24 \pm .40$ | $8.88 \pm .29$ | $+0.65 \pm .58$ | TSV-Merge | $6.36 \pm .34$ | $7.70 \pm .10$ | $+1.34 \pm .31$ | TSV-Merge | $4.43 \pm .59$ | $2.87 \pm .16$ | $-1.56 \pm .75$ |

Table 5 separates endpoint quality from compatibility on a denominator-invariant scale. The OrthoReg endpoint is stronger in every row, while a negative interaction only says that TSV-Merge adds less beyond that endpoint.

The raw-gain interaction preserves the normalized interaction's sign in eleven of twelve non-identity cells; B/32 TSV-Merge is near zero under either view. The headroom interaction retains the main negative TIES pattern on every backbone and the architecture-dependent depth pattern. These alternative scalings reduce concern that the descriptive interactions are purely mechanical consequences of different identity headroom; the multiplicity-controlled operator contrasts remain the inferential test. Complete per-task records and intervals accompany the evaluation artifacts.

Figure 4 makes three levels of variation explicit: training geometry within a backbone, operator family within a geometry, and independent repetitions within each cell. The metric-sensitivity panel is a robustness check, not a second selection criterion: all configurations remain those selected by normalized validation accuracy. Its sign agreement shows that the reported pattern is not created by normalizing each task by its single-task reference, while the nonunit slope confirms that metric choice still affects effect size.

Table 5: Absolute vision endpoints on a common scale. Entries are raw mean top-1 accuracy with hierarchical run–task bootstrap 95% intervals. The final column is the paired OrthoReg-minus-Standard endpoint difference. Operator configurations remain those selected by normalized validation accuracy.

| Backbone | Endpoint | Standard FT | OrthoReg FT | Difference |
|---|---|---|---|---|
| ViT-B/32 | Identity | 70.14 [64.84, 75.55] | 74.56 [69.97, 79.33] | +4.42 [2.40, 6.44] |
| | TSV-Merge | 77.22 [72.84, 81.60] | 81.41 [76.81, 86.05] | +4.19 [2.14, 6.40] |
| ViT-B/16 | Identity | 75.24 [70.02, 80.37] | 79.46 [74.44, 84.41] | +4.22 [1.92, 6.72] |
| | TSV-Merge | 80.95 [76.30, 85.46] | 85.41 [81.07, 89.66] | +4.46 [2.10, 7.07] |
| ViT-L/14 | Identity | 84.07 [79.63, 88.27] | 89.03 [85.71, 92.17] | +4.96 [3.15, 6.94] |
| | TSV-Merge | 88.15 [84.12, 91.86] | 91.20 [88.18, 94.03] | +3.05 [1.62, 4.68] |

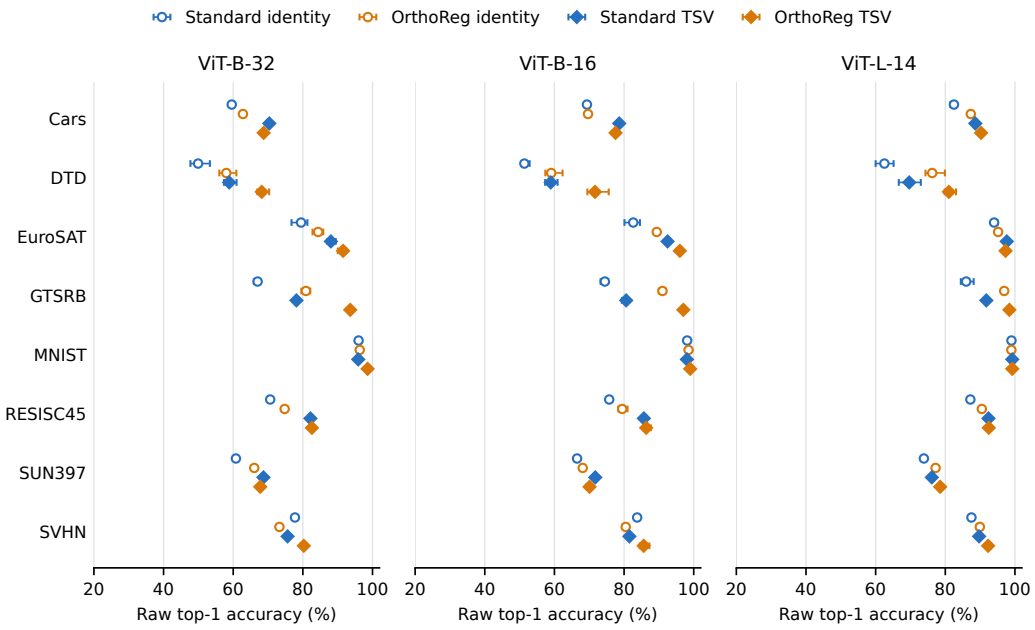


Figure 3: Per-task absolute accuracy for identity and TSV-Merge. Markers are three-run means and bars are run-bootstrap 95% intervals. Open circles denote identity; filled diamonds denote TSV-Merge. This common-scale view supports endpoint claims without using geometry-dependent single-task denominators.

## D Training-Strength Sweep

We repeat the ViT-B/32 factorial for $\lambda \in \{0, 1, 3, 10, 30\}$ with three paired runs per strength. The training diagnostic is the scale-free isotropy error

$$E_{\text{iso}} = \frac{\|G(W) - \bar{d}I\|_F}{\|G(W)\|_F}, \qquad \bar{d} = \frac{\operatorname{tr}(G(W))}{\dim(G(W))},$$

where $G(W) = WW^\top$ when $W$ has fewer rows than columns and $G(W) = W^\top W$ otherwise. We average this quantity over all matrix-valued residual-block updates and tasks within each run; zero denotes isotropy up to scale. The statistic measures the internal update geometry encouraged by OrthoReg. It is not a measure of cross-task orthogonality and is not treated as a causal mediator.

The ordered changes in $E_{\text{iso}}$ and identity accuracy verify that the training intervention spans more than two isolated checkpoint recipes. The operator response does not follow the same monotone path: fixed depth remains positive, LiNeS and TIES remain negative, and TSV-Merge crosses from positive to negative. This argues against a single-strength accident and shows that this isotropy statistic alone does not order the observed compatibility responses. It does not establish that changing $\lambda$ alone

Table 6: Common-scale robustness audit. Error reduction is raw accuracy gain over identity. Interactions compare that gain after OrthoReg versus Standard FT; brackets are hierarchical bootstrap 95% intervals. Headroom divides each gain by the identity error rate before taking the same interaction.

| Backbone | Operator | Standard error reduction | OrthoReg error reduction | Raw interaction | Headroom interaction |
|---|---|---|---|---|---|
| ViT-B/32 | Fixed depth | 1.70 | 2.69 | $+0.99\ [0.03, 2.00]$ | $+4.86\ [1.53, 9.03]$ |
| | LiNeS | 4.05 | 2.63 | $-1.42\ [-1.84, -0.99]$ | $-3.22\ [-4.91, -1.34]$ |
| | TIES | 1.70 | 0.59 | $-1.11\ [-1.71, -0.49]$ | $-3.37\ [-5.30, -1.02]$ |
| | TSV-Merge | 7.09 | 6.85 | $-0.23\ [-2.08, 1.71]$ | $+3.22\ [-3.06, 10.49]$ |
| ViT-B/16 | Fixed depth | 0.10 | 1.61 | $+1.51\ [0.52, 2.42]$ | $+7.44\ [2.91, 12.06]$ |
| | LiNeS | 2.25 | 1.74 | $-0.50\ [-1.28, 0.20]$ | $-0.58\ [-4.06, 2.29]$ |
| | TIES | 1.05 | $-0.06$ | $-1.11\ [-1.75, -0.50]$ | $-4.55\ [-7.17, -1.83]$ |
| | TSV-Merge | 5.71 | 5.95 | $+0.23\ [-1.33, 1.84]$ | $+5.89\ [-0.74, 12.87]$ |
| ViT-L/14 | Fixed depth | 1.97 | 0.71 | $-1.25\ [-1.90, -0.59]$ | $-5.84\ [-11.21, -1.30]$ |
| | LiNeS | 2.12 | 0.67 | $-1.45\ [-2.02, -0.88]$ | $-7.15\ [-12.10, -3.25]$ |
| | TIES | 0.75 | $-0.30$ | $-1.05\ [-1.71, -0.43]$ | $-7.37\ [-11.75, -2.26]$ |
| | TSV-Merge | 4.08 | 2.17 | $-1.91\ [-2.83, -1.03]$ | $-5.81\ [-10.54, -0.76]$ |

Table 7: ViT-B/32 strength sweep over three paired runs. Identity is raw mean top-1 accuracy; operator columns are raw-accuracy interactions relative to the paired $\lambda = 0$ training condition (pp). Values are mean $\pm$ sample standard deviation. Operators and radii are selected on validation data under the primary family-wise protocol.

| $\lambda$ | $E_{\mathrm{iso}}$ | Identity | Fixed depth | LiNeS | TIES | TSV-Merge |
|---|---|---|---|---|---|---|
| 0 | $.9561 \pm .0007$ | $70.14 \pm .30$ | – | – | – | – |
| 1 | $.8720 \pm .0003$ | $72.32 \pm .42$ | $+.31 \pm .36$ | $-2.38 \pm .10$ | $-1.78 \pm .12$ | $+.42 \pm .89$ |
| 3 | $.8110 \pm .0002$ | $73.55 \pm .05$ | $+.52 \pm .34$ | $-1.91 \pm .20$ | $-1.66 \pm .11$ | $+.23 \pm .41$ |
| 10 | $.7492 \pm .0093$ | $74.56 \pm .04$ | $+.99 \pm .36$ | $-1.42 \pm .03$ | $-1.11 \pm .25$ | $-.23 \pm .40$ |
| 30 | $.7209 \pm .0002$ | $76.23 \pm .13$ | $+.66 \pm .40$ | $-1.93 \pm .40$ | $-1.69 \pm .18$ | $-1.69 \pm .46$ |

causes the measured interaction, because checkpoint quality and other aspects of the learned geometry change jointly.

## E Fixed-Configuration Absolute-Norm Audit

The primary factorial gives every family the same relative-radius support, but independent validation selection can produce different endpoint radii and each training geometry has a different identity-vector norm. We therefore add a stricter ViT-B/32 audit. Within each paired run, the Standard-FT identity displacement at relative radius one defines an absolute-norm anchor $\rho_s$. Every operator under both training geometries is evaluated at $\{0.25, 0.5, 0.75, 1\}\rho_s$. Thus each paired comparison uses exactly the same parameter-space norm, not merely the same relative grid.

Configurations are also fixed across geometries and runs. The fixed-depth profile and TSV-Merge definition are unchanged; LiNeS uses $\beta = 1$ and TIES uses density $0.6$, the modal development-selected settings in the primary ViT-B/32 factorial. All prespecified curve points are reported. We summarize each curve by trapezoidal area over anchor fraction, including the common zero-gain origin; no held-out point is selected by its audit score. Specifically, if $a \in \{0, 0.25, 0.5, 0.75, 1\}$ is the anchor fraction and $g_{G,\mathcal{O}}(a)$ is raw mean-top-1 gain over identity at exactly $a\rho_s$, the summary is the trapezoidal estimate of $\int_0^1 g_{G,\mathcal{O}}(a)\,\mathrm{d}a$. We compute the OrthoReg-minus-Standard area difference within run and then report its mean and sample standard deviation. The anchor may vary across runs, but never within a paired training–operator comparison.

Table 8: Fixed-configuration ViT-B/32 audit at common absolute norms. AUC is the raw mean-top-1 gain over identity integrated across anchor fraction; $\Delta$ columns are paired OrthoReg-minus-Standard differences. Values are mean $\pm$ sample standard deviation over three paired runs (pp).

| Operator | Standard AUC | OrthoReg AUC | $\Delta$ AUC | $\Delta$ endpoint |
|---|---|---|---|---|
| Fixed depth | $0.736 \pm 0.065$ | $-0.403 \pm 0.023$ | $-1.139 \pm 0.042$ | $-2.451 \pm 0.242$ |
| LiNeS | $1.450 \pm 0.040$ | $1.265 \pm 0.060$ | $-0.185 \pm 0.091$ | $-1.865 \pm 0.312$ |
| TIES | $-0.031 \pm 0.086$ | $-0.073 \pm 0.036$ | $-0.041 \pm 0.117$ | $-1.747 \pm 0.144$ |
| TSV-Merge | $-3.272 \pm 0.148$ | $0.131 \pm 0.013$ | $+3.402 \pm 0.136$ | $+0.106 \pm 0.457$ |

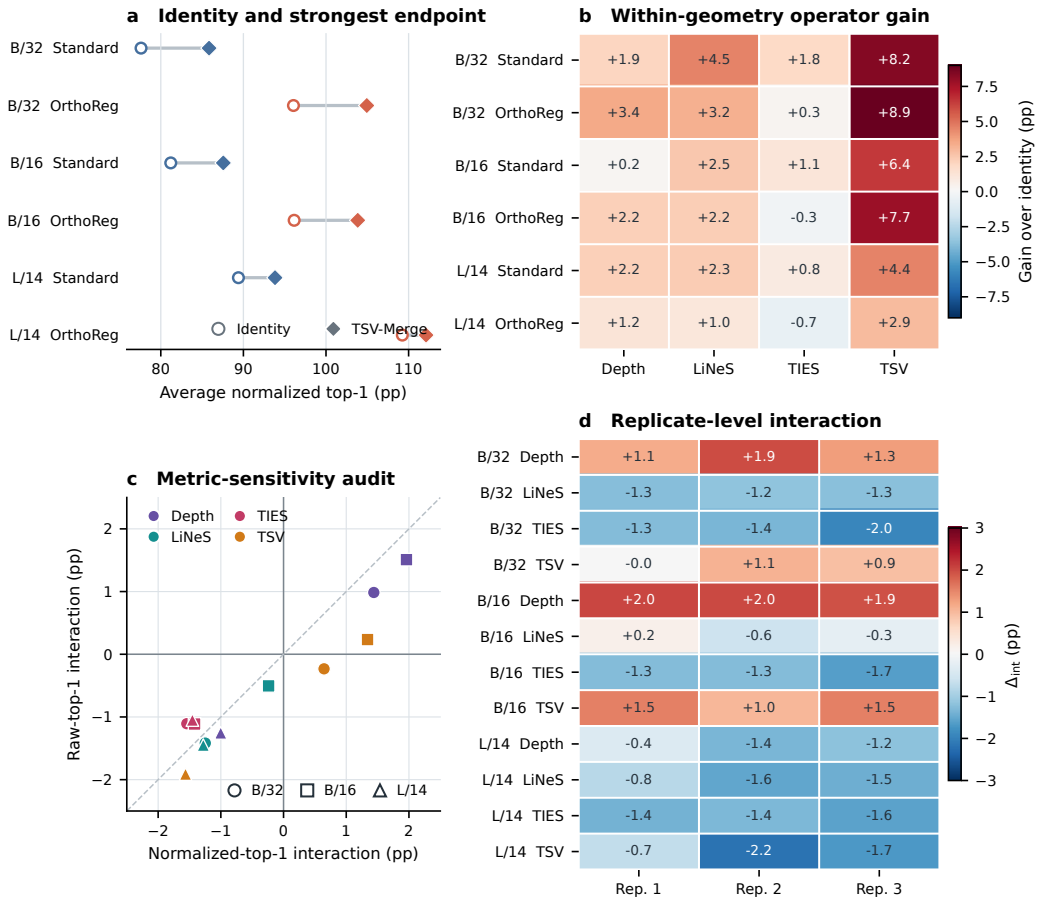


Figure 4: Multilevel vision audit from the frozen paired records. (a) Identity and the strongest operator endpoint separate source quality from post-hoc gain. (b) The complete within-geometry gain matrix shows that endpoint strength and compatibility are distinct. (c) Eleven of twelve training–operator interactions retain their sign when average normalized top-1 is replaced by unnormalized mean top-1; distance from the diagonal reflects the different task weighting. Colors denote operators and markers denote backbones. (d) Replicate-level interactions expose sign stability and the near-zero B/32 TSV-Merge boundary case without averaging it away.

Table 8 and Figure 6 separate two estimands that the primary factorial intentionally combines. LiNeS retains a negative interaction, while TSV-Merge has a strongly positive curve-average interaction because it harms Standard FT at small radii but is nearly neutral under OrthoReg. TIES is near zero on average over the curve yet negative at the common endpoint. Most importantly, fixed depth changes from a positive validation-selected family interaction in the primary experiment to a negative fixed-member interaction at common absolute norms. The primary result is therefore not evidence that one fixed depth direction is intrinsically favored by OrthoReg. Rather, training geometry changes both the response of a fixed map and which scale is useful. Family-level compatibility and fixed-member sensitivity answer different questions and should not be interchanged.

## F Fixed Relative-Radius and Profile-Strength Audit

We next fix the relative radius to one for every family. Within a training geometry, each structured displacement is rescaled to the norm of that geometry's identity sum, so the comparison removes operator-specific norm differences without validation-selecting a radius. It is not a common absolute-norm comparison across training geometries: averaged over runs, the OrthoReg-to-Standard identity-norm ratios are 7.90, 7.26, and 6.97 on ViT-B/16, ViT-B/32, and ViT-L/14, respectively. Table 9

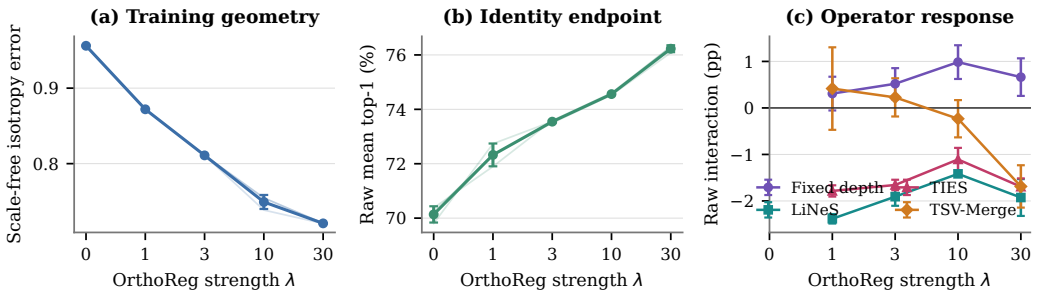


Figure 5: Replicated ViT-B/32 training-strength audit. Thin lines in (a–b) are individual runs; markers and bars are means and sample standard deviations. (a) The internal update-matrix isotropy error falls with regularization strength. (b) The raw direct-addition endpoint improves. (c) Raw-accuracy interactions relative to paired Standard FT diverge by operator. The x-axis is categorical; no linear dependence on $\lambda$ is assumed.

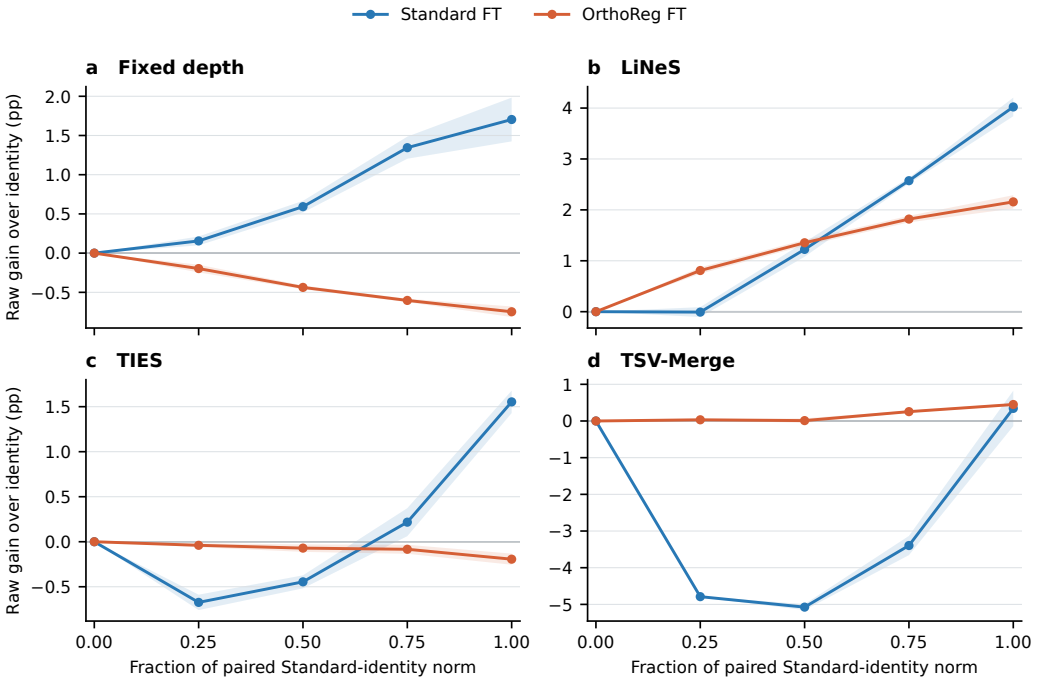


Figure 6: Fixed operators at identical absolute displacement norms. Curves show raw gain over identity; lines and bands are the mean and sample standard deviation over paired runs. Every point is prespecified and reported. The response depends on operator, training geometry, and radius even after removing configuration and norm differences.

therefore estimates a prespecified relative-radius deployment rule, whereas Table 8 isolates fixed-map behavior at identical parameter-space norms.

At this relative operating point, LiNeS and TIES have negative interactions on all three backbones, while TSV-Merge has a positive interaction on all three. Fixed depth is positive on both base models but unresolved on ViT-L/14. The TSV-Merge sign on ViT-L/14 differs from its validation-selected family interaction in Table 1, further showing that the operating radius is part of the training–merging relation.

Table 9: Interaction at fixed relative radius one. Entries are paired raw mean-top-1 interactions, reported as mean ± sample standard deviation over three runs (pp). Every operator has the identity-sum norm within its training geometry; absolute norms differ across geometries.

| Backbone | Fixed depth | LiNeS | TIES | TSV-Merge |
|---|---|---|---|---|
| ViT-B/16 | $+3.158 \pm 0.378$ | $-1.172 \pm 0.080$ | $-0.999 \pm 0.396$ | $+7.940 \pm 0.698$ |
| ViT-B/32 | $+1.897 \pm 0.482$ | $-1.812 \pm 0.171$ | $-0.762 \pm 0.177$ | $+7.536 \pm 0.717$ |
| ViT-L/14 | $-0.222 \pm 0.591$ | $-1.247 \pm 0.477$ | $-0.319 \pm 0.089$ | $+3.154 \pm 0.634$ |

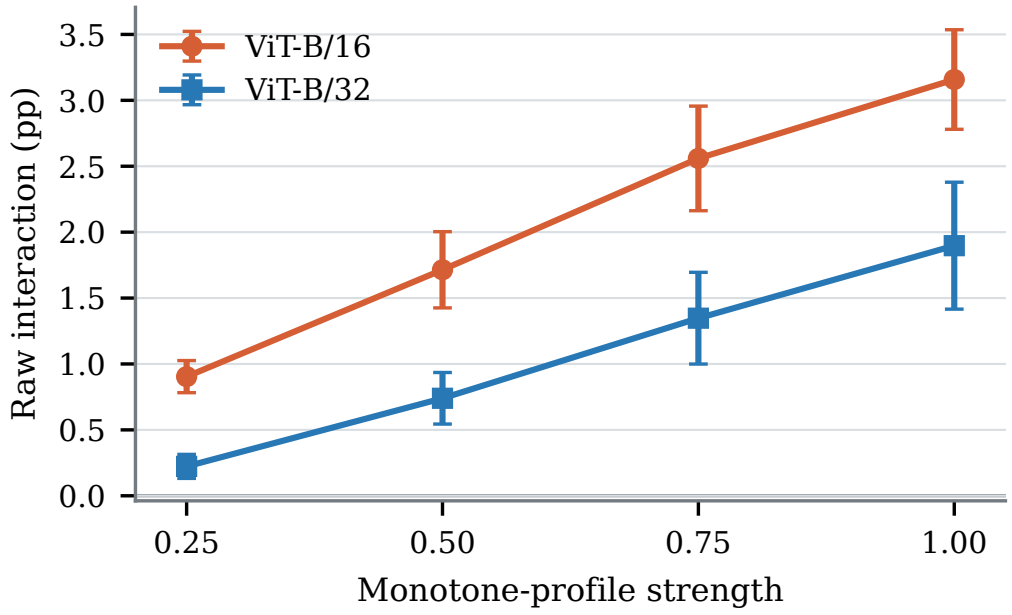


Figure 7: Fixed-relative-radius interaction across prespecified monotone profile strengths. Markers and bars are the paired mean and sample standard deviation over three runs. The interaction is positive for every tested profile on ViT-B/16 and for strengths $s \geq 0.5$ on ViT-B/32; the shallowest B/32 profile is a disclosed sign boundary. The response generally increases with stronger depth reweighting. This robustness does not imply a fixed-map advantage at common absolute norm; that distinct estimand is reported in Table 8.

To test whether the base-model fixed-depth result depends on the single profile used in the primary factorial, we prespecify four monotone profiles $(0.30^s, 0.75^s, 1.25^s)$ for $s \in \{0.25, 0.50, 0.75, 1.00\}$. We apply every profile at relative radius one and report all strengths without selecting among them.

## G Complete Operator-Response and Metric Audits

Table 10 reports every operator pair entering the common-shift test. The simultaneous intervals use the maximum absolute bootstrap error over all six pairs within each backbone or the language suite. Consequently, the main text does not infer heterogeneity from isolated pointwise intervals. Here D, L, T, and V denote fixed depth, LiNeS, TIES, and TSV-Merge.

Table 11 reports every individual interaction used in the fixed-output metric audit. Macro raw and normalized utility give broadly similar signatures. Worst-quartile utility changes the ViT-L/14 signature most strongly: TIES moves above its common shift while the depth and subspace readouts move below it. We test all three metric pairs and six operator pairs; the three familywise-resolved normalized-macro versus worst-quartile contrasts are reported in Table 12. Complete intervals for all 18 comparisons are included in the machine-readable audit.

Table 10: Complete operator-response contrasts (pp). Pointwise and simultaneous 95% intervals use paired run–task hierarchical bootstrap. "Rev." marks an observed reversal in within-geometry operator-gain ordering; it is descriptive unless the simultaneous contrast also excludes zero.

| Setting | Pair | $\mathcal{H}$ | Pointwise CI | Simultaneous CI | Rev. |
|---|---|---|---|---|---|
| B/32 | D−L | $+2.69$ | $[+1.63, +3.81]$ | $[+0.14, +5.23]$ | yes |
| B/32 | D−T | $+2.98$ | $[+1.50, +4.45]$ | $[+0.44, +5.53]$ | no |
| B/32 | D−V | $+0.79$ | $[-1.04, +2.65]$ | $[-1.75, +3.34]$ | no |
| B/32 | L−T | $+0.29$ | $[-0.56, +1.23]$ | $[-2.25, +2.84]$ | no |
| B/32 | L−V | $-1.89$ | $[-4.09, +0.28]$ | $[-4.44, +0.65]$ | no |
| B/32 | T−V | $-2.19$ | $[-4.71, +0.28]$ | $[-4.74, +0.36]$ | no |
| B/16 | D−L | $+2.20$ | $[+1.27, +3.23]$ | $[-0.33, +4.72]$ | no |
| B/16 | D−T | $+3.38$ | $[+1.77, +5.04]$ | $[+0.85, +5.91]$ | yes |
| B/16 | D−V | $+0.62$ | $[-1.42, +2.75]$ | $[-1.90, +3.15]$ | no |
| B/16 | L−T | $+1.18$ | $[+0.04, +2.55]$ | $[-1.35, +3.71]$ | no |
| B/16 | L−V | $-1.58$ | $[-3.38, +0.26]$ | $[-4.10, +0.95]$ | no |
| B/16 | T−V | $-2.76$ | $[-5.20, -0.43]$ | $[-5.28, -0.23]$ | no |
| L/14 | D−L | $+0.28$ | $[+0.05, +0.54]$ | $[-1.14, +1.70]$ | yes |
| L/14 | D−T | $+0.45$ | $[-0.72, +1.92]$ | $[-0.97, +1.87]$ | no |
| L/14 | D−V | $+0.56$ | $[+0.07, +1.08]$ | $[-0.86, +1.98]$ | no |
| L/14 | L−T | $+0.17$ | $[-0.94, +1.53]$ | $[-1.25, +1.60]$ | no |
| L/14 | L−V | $+0.28$ | $[-0.27, +0.79]$ | $[-1.14, +1.71]$ | no |
| L/14 | T−V | $+0.11$ | $[-1.38, +1.34]$ | $[-1.31, +1.53]$ | no |
| Language | D−L | $+0.32$ | $[-0.08, +1.09]$ | $[-5.43, +6.08]$ | no |
| Language | D−T | $+6.61$ | $[+1.60, +13.00]$ | $[+0.86, +12.36]$ | no |
| Language | D−V | $+5.28$ | $[+1.50, +11.05]$ | $[-0.47, +11.03]$ | no |
| Language | L−T | $+6.29$ | $[+1.56, +12.04]$ | $[+0.53, +12.04]$ | no |
| Language | L−V | $+4.96$ | $[+1.42, +10.01]$ | $[-0.80, +10.71]$ | no |
| Language | T−V | $-1.33$ | $[-3.77, +0.62]$ | $[-7.08, +4.42]$ | no |

Table 11: Complete fixed-output vision interaction signatures under three task aggregations (pp). WQ is worst-quartile specialist-normalized utility.

| Backbone | Metric | Fixed depth | LiNeS | TIES | TSV-Merge |
|---|---|---|---|---|---|
| B/32 | Normalized macro | $+1.44$ | $-1.25$ | $-1.54$ | $+0.65$ |
| | Raw macro | $+0.99$ | $-1.42$ | $-1.11$ | $-0.23$ |
| | Normalized WQ | $+0.22$ | $-4.44$ | $-0.28$ | $-2.77$ |
| B/16 | Normalized macro | $+1.96$ | $-0.24$ | $-1.42$ | $+1.34$ |
| | Raw macro | $+1.51$ | $-0.50$ | $-1.11$ | $+0.23$ |
| | Normalized WQ | $+1.29$ | $-2.00$ | $-1.19$ | $-1.38$ |
| L/14 | Normalized macro | $-1.00$ | $-1.28$ | $-1.45$ | $-1.56$ |
| | Raw macro | $-1.25$ | $-1.45$ | $-1.05$ | $-1.91$ |
| | Normalized WQ | $-3.68$ | $-3.54$ | $+0.77$ | $-4.62$ |

## H Evaluation Protocol

**Vision training.** Both source-training regimes use AdamW with learning rate $10^{-5}$, weight decay $0.1$, cosine decay, 500 optimizer-step warmup, and no label smoothing. The dataset-specific epoch counts are 35 (Cars), 76 (DTD), 12 (EuroSAT), 11 (GTSRB), 5 (MNIST), 15 (RESISC45), 14 (SUN397), and 4 (SVHN), following the OrthoReg training code. ViT-B/32 and ViT-B/16 use batch size 128. ViT-L/14 uses microbatch size 64 with two-step gradient accumulation, giving the same effective batch size. Standard FT and OrthoReg differ only by the added regularizer within each paired run; the data order, optimizer schedule, and task-specific heads are otherwise matched.

**Implementation environment.** Experiments use NVIDIA RTX A6000 accelerators, PyTorch 2.0 with CUDA 11.8, NumPy 1.24, and timm 0.9. The language extension uses Transformers 4.44 and Datasets 1.18. All aggregate tables are produced from machine-readable frozen records by the supplied analysis scripts; test records are not read by any selection routine.

For a raw operator vector $v$, the coefficient at relative radius $r$ is $r\alpha_{\mathrm{ref}}\|v_{\mathcal{I}}\|_2/\|v\|_2$. The reference coefficient is obtained from a validation-selected identity run for the same model, training mode, and paired run. Candidate radii are $\{0.5, 0.625, 0.75, 0.875, 1.0, 1.125, 1.25, 1.375, 1.5\}$. The identity baseline is reselected over this same grid. For each non-identity family, the best joint family-hyperparameter/radius configuration is selected by average normalized accuracy on task validation splits. Exactly that configuration is then evaluated once on test data.

Table 12: Resolved fixed-output metric contrasts on ViT-L/14. $\mathcal{H}$ is the operator-response contrast under normalized macro or worst-quartile normalized utility; $\mathcal{Q}$ is their difference. Intervals simultaneously cover all 18 metric-by-operator-pair comparisons for this backbone (pp).

| Operator contrast | $\mathcal{H}_{\text{macro}}$ | $\mathcal{H}_{\text{worst}}$ | $\mathcal{Q}$ simultaneous 95% CI |
|---|---|---|---|
| Fixed depth − TIES | $+0.45$ | $-4.45$ | $[+1.20, +8.60]$ |
| LiNeS − TIES | $+0.17$ | $-4.31$ | $[+0.79, +8.18]$ |
| TIES − TSV-Merge | $+0.11$ | $+5.40$ | $[-8.99, -1.59]$ |

The fixed-depth operator has no shape hyperparameter in the primary experiment. Its early/middle/late factors are $(0.30, 0.75, 1.25)$ and each transformed task vector is normalized to its original $\ell_2$ norm before summation. LiNeS uses the official multi-task definition; $\beta$ ranges from $0.1$ to $1.0$ in increments of $0.1$. TIES uses summed reduction and densities $\{0.2, 0.4, 0.6\}$. TSV-Merge follows layerwise memory-reduced orthogonalization. For eight tasks, each matrix update contributes its leading one-eighth singular subspace; non-matrix parameters are averaged. Global TSV scale is selected by the same validation radius protocol. The complete validation curves and selected configurations are included in the machine-readable supplementary archive.

For the added placement diagnostic, the post-merge tiered profile raises each base factor to a strength selected from $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 2\}$ and preserves the merged-vector norm. The per-task linear profile selects $\beta$ from the same numeric grid and renormalizes each task vector before summation. These broader grids are why Table 14 is interpreted by interaction sign rather than by small differences in magnitude.

### H.1 Additional vision controls

Table 13 is a one-run audit of the validation-to-test transition. It is not used in the three-run main estimate, but verifies that the reported operator settings do not reverse their interaction sign at test on these paired runs. Table 14 isolates the two components that differ between the fixed-depth probe and LiNeS: profile shape and whether the profile is applied before or after summation. It is an explanatory, single-run diagnostic and is therefore not used to support a population claim.

Table 13: Validation–test interaction audit (pp) for one paired replicate per backbone. All signs agree between validation and test.

| | B/32 | | B/16 | | L/14 | |
|---|---|---|---|---|---|---|
| Operator | Val | Test | Val | Test | Val | Test |
| Fixed depth | $+2.006$ | $+1.253$ | $+2.737$ | $+1.997$ | $-1.420$ | $-1.228$ |
| LiNeS | $-1.080$ | $-1.256$ | $+0.042$ | $+0.157$ | $-1.166$ | $-1.468$ |
| TIES | $-2.029$ | $-1.955$ | $-1.227$ | $-1.295$ | $-1.621$ | $-1.596$ |
| TSV-Merge | $+0.925$ | $+0.898$ | $+1.926$ | $+1.515$ | $-1.281$ | $-1.730$ |

Table 14: Depth-shape/application-placement diagnostic on ViT-B/32 (one paired run). Entries are validation-selected test gains over independently selected identity on the shared radius grid (pp). Tiered and linear shapes retain opposite interaction signs after moving across the summation boundary.

| Shape | Placement | Standard | OrthoReg | Interaction |
|---|---|---|---|---|
| Tiered | Per task | 2.309 | 3.562 | $+1.253$ |
| Tiered | Post merge | 2.459 | 3.665 | $+1.206$ |
| Linear | Per task | 4.445 | 3.634 | $-0.812$ |
| Linear | Post merge | 4.461 | 3.205 | $-1.256$ |

The fixed probe in the main factorial is the $s = 1$ member of $(0.30^s, 0.75^s, 1.25^s)$, fixed before this robustness readout. The interaction is positive in all runs for all four B/16 shapes. On B/32, its mean is positive for $s \geq 0.5$, but all runs are positive only at $s \in \{0.75, 1.00\}$; the shallowest transform is a disclosed negative boundary. Thus the result is not evidence that every monotone profile is compatible. It supports the narrower, graded claim that compatibility changes with the shape of the depth intervention.

Table 15: Fixed monotone-profile robustness over three paired runs. Each profile is reported rather than selected on test; its radius is selected on validation data. Values are normalized test gains over identity (pp), mean ± sample standard deviation. The final column is the paired interaction.

| Backbone / strength $s$ | Standard | OrthoReg | Interaction | Positive runs |
|---|---|---|---|---|
| ViT-B/32, 0.25 | $1.651 \pm 0.009$ | $1.464 \pm 0.041$ | $-0.187 \pm 0.038$ | $0/3$ |
| ViT-B/32, 0.50 | $2.328 \pm 0.123$ | $2.488 \pm 0.218$ | $+0.160 \pm 0.341$ | $2/3$ |
| ViT-B/32, 0.75 | $2.318 \pm 0.313$ | $3.174 \pm 0.275$ | $+0.856 \pm 0.481$ | $3/3$ |
| ViT-B/32, 1.00 | $1.946 \pm 0.361$ | $3.386 \pm 0.283$ | $+1.440 \pm 0.447$ | $3/3$ |
| ViT-B/16, 0.25 | $0.720 \pm 0.111$ | $1.144 \pm 0.286$ | $+0.425 \pm 0.308$ | $3/3$ |
| ViT-B/16, 0.50 | $0.987 \pm 0.236$ | $1.757 \pm 0.305$ | $+0.770 \pm 0.402$ | $3/3$ |
| ViT-B/16, 0.75 | $0.598 \pm 0.184$ | $2.100 \pm 0.390$ | $+1.502 \pm 0.246$ | $3/3$ |
| ViT-B/16, 1.00 | $0.209 \pm 0.243$ | $2.169 \pm 0.283$ | $+1.961 \pm 0.043$ | $3/3$ |

# I Selection-Budget and Validation-Curve Audit

Table 16 states the complete family-wise search budget. Every family sees the same validation examples and radius grid within a domain, but shape-parameter counts differ. Consequently, Tables 1 and 18 compare validation-selected family envelopes; they do not estimate performance under equal numbers of hyperparameter trials.

Table 16: Disclosed validation-search budgets. "Configs" excludes radius; evaluations are configs times radii for each training geometry and selection run. Language selection is performed once and transferred to all held-out runs.

| Domain | Family | Configs | Radii | Evaluations |
|---|---|---|---|---|
| Vision | Fixed depth / TSV-Merge | 1 | 9 | 9 |
| | LiNeS | 10 | 9 | 90 |
| | TIES | 3 | 9 | 27 |
| Language | Identity / TSV-Merge | 1 | 7 | 7 |
| | Fixed depth / LiNeS | 4 | 7 | 28 |
| | TIES | 3 | 7 | 21 |

Across the three vision runs, the selected radii are concentrated rather than arbitrarily dispersed. For B/32, Standard/OrthoReg select radius $1.00/1.00$ for fixed depth and $1.125/0.75$ for LiNeS in every run; TIES selects $1.125$–$1.25/0.75$, and TSV-Merge selects $1.50/0.875$–$1.125$. On B/16 the corresponding ranges are $1.00$–$1.125/0.875$, $1.125/0.625$, $1.25/0.625$–$0.75$, and $1.50/0.875$–$1.00$. On L/14 they are $1.125$–$1.25/0.75$, $1.125$–$1.25/0.625$, $1.50/0.625$–$0.75$, and $1.50/0.625$–$0.875$. Complete run-level counts and selected configurations are provided in the machine-readable audit.

For language, the transferred Standard/OrthoReg settings are respectively: fixed-depth value $-0.5/-0.5$ at radius $0.2/0.2$; LiNeS value $-0.5/-0.5$ at $0.2/0.2$; TIES density $0.2/0.1$ at $0.2/0.4$; and TSV-Merge at $0.2/0.2$. Figure 8 reports the full validation envelopes rather than only their maxima.

# J Three-Run Interaction Audit

The complete result files also contain validation gains, selected radii, per-task test metrics, and identity-inclusive deployment outcomes.

# K Language Extension

**Protocol.** We use the local FusionBench task formatting and evaluator. The eight source checkpoints are independently fine-tuned Flan-T5-base models for CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST-2, and STS-B. Operator settings and radii are selected once on the first run's development data and transferred unchanged to the other two runs. The operator-development split is a deterministic 10% subset of each task's training data, capped at 5,000 examples; task fine-tuning uses only the remainder. Because official GLUE test labels are unavailable, the labeled validation split (validation-matched for MNLI) is frozen as the held-out split and evaluated once per merged checkpoint. Generation

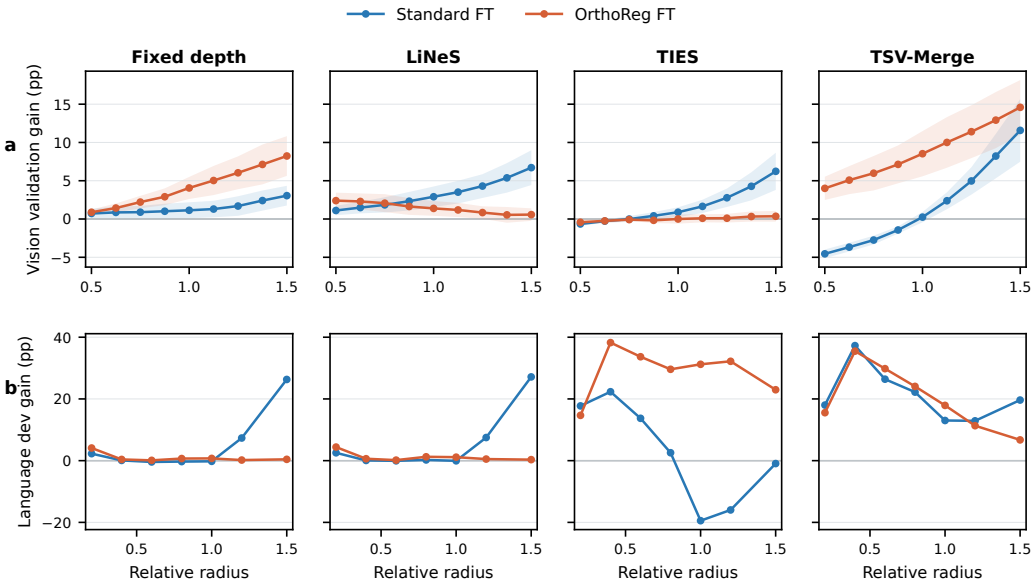


Figure 8: Validation selection audit. (a) Vision curves are family envelopes over the disclosed shape grids; lines and bands show the mean and sample standard deviation across backbones and runs. (b) Language curves use the single development selection that is transferred to all held-out runs. Every point is computed relative to identity at the same radius. Curves are reported for transparency and are not used to revise the frozen test choices.

Table 17: Paired vision interactions (pp) for every completed independent run. These are the values aggregated in Table 1. Fixed depth and TIES have the same interaction sign across all runs within every backbone. TSV-Merge is stable on B/16 and L/14; B/32 contains one near-zero sign change.

| Backbone | Operator | Replicate 1 | Replicate 2 | Replicate 3 |
|---|---|---|---|---|
| B/32 | Fixed depth | +1.116 | +1.950 | +1.253 |
| | LiNeS | −1.283 | −1.204 | −1.256 |
| | TIES | −1.300 | −1.373 | −1.955 |
| | TSV-Merge | −0.012 | +1.052 | +0.898 |
| B/16 | Fixed depth | +1.997 | +1.971 | +1.913 |
| | LiNeS | +0.157 | −0.611 | −0.260 |
| | TIES | −1.295 | −1.308 | −1.654 |
| | TSV-Merge | +1.515 | +0.985 | +1.518 |
| L/14 | Fixed depth | −0.405 | −1.373 | −1.228 |
| | LiNeS | −0.821 | −1.550 | −1.468 |
| | TIES | −1.409 | −1.355 | −1.596 |
| | TSV-Merge | −0.747 | −2.212 | −1.730 |

inputs explicitly exclude labels and `decoder_input_ids`, preventing target-token leakage during decoding. The task score is exact accuracy for CoLA, MNLI, MRPC, QNLI, QQP, RTE, and SST-2, and Spearman correlation for STS-B; Table 18 uses their unweighted mean. This is the FusionBench local evaluation protocol, not a claim of direct comparability to the official GLUE leaderboard.

All language checkpoints use AdamW with learning rate $10^{-4}$, zero weight decay, linear decay, a 6% warmup fraction, microbatch size 8, and four-step gradient accumulation. Inputs and targets are truncated to 256 and 16 tokens, respectively; bfloat16 is used when supported, and the gradient norm is clipped to 1.0. MNLI, QNLI, QQP, and SST-2 are trained for one epoch; CoLA, MRPC, RTE, and STS-B are trained for three. The OrthoReg coefficient is 0.1. Its penalty is the mean Frobenius orthogonality penalty over trainable two-dimensional attention and feed-forward kernels. Shared embeddings, token embeddings, the output head, layer-normalization weights, and relative-attention

Table 18: Complete frozen language compatibility results. Values are held-out mean ± sample standard deviation over three independent runs (pp). Identity is the aggregate FusionBench score; other rows are gains over the corresponding frozen-policy identity merge. Brackets give the hierarchical bootstrap 95% interval for each individual interaction. Pairwise simultaneous inference is reported in Table 10.

| Operator | Standard FT | OrthoReg FT | $\Delta_{\text{int}}$ |
|---|---|---|---|
| Identity utility | $62.31 \pm 0.37$ | $65.15 \pm 0.71$ | – |
| Fixed depth | $+3.04 \pm 0.70$ | $+5.46 \pm 2.16$ | $+2.42\ [-1.27, 7.56]$ |
| LiNeS | $+3.47 \pm 0.98$ | $+5.57 \pm 2.03$ | $+2.10\ [-1.38, 6.62]$ |
| TIES | $+17.47 \pm 0.66$ | $+13.28 \pm 0.92$ | $-4.19\ [-7.84, -0.80]$ |
| TSV-Merge | $+18.58 \pm 0.72$ | $+15.72 \pm 0.51$ | $-2.86\ [-5.33, -0.63]$ |

biases are excluded. Because this architecture-normalized implementation averages over eligible matrices, its coefficient is not numerically comparable to the sum-aggregated visual coefficient. Standard FT uses the identical training pipeline without this term.

For both encoder and decoder blocks, let normalized depth be $z_\ell = \ell/(L-1)$. The fixed-depth family applies $\exp[\beta(2z_\ell - 1)]$ to each task vector before summation, preserves each task-vector norm, and then matches the summed identity norm. LiNeS first sums the vectors, applies $\max\{0, 1+\beta(2z_\ell-1)\}$, and matches the identity norm. Non-block parameters receive unit weight. Both families search $\beta \in \{-0.5, -0.25, 0.25, 0.5\}$. TIES uses disjoint-mean reduction at densities $\{0.1, 0.2, 0.5\}$ and is norm-matched after merging. TSV-Merge keeps the leading $1/T$ singular subspace per eligible matrix; vocabulary, embedding, output-head, relative-position, and normalization parameters are merged without SVD. All families use the seven radii listed in Table 16 and are selected only on the development split.

**Task decomposition and uncertainty.** Let $x_{G,\mathcal{O},s,t}$ be the held-out score for training geometry $G$, operator $\mathcal{O}$, run $s$, and task $t$. The paired task-level interaction is

$$\delta_{s,t}(\mathcal{O}) = (x_{\text{Ortho},\mathcal{O},s,t} - x_{\text{Ortho},\mathcal{I},s,t}) - (x_{\text{Std},\mathcal{O},s,t} - x_{\text{Std},\mathcal{I},s,t}). \tag{17}$$

Because the reported language score is an unweighted task mean, its interaction decomposes exactly as

$$\Delta_{\text{int}}(\mathcal{O}) = \frac{1}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} \delta_{s,t}(\mathcal{O}). \tag{18}$$

For each of 20,000 hierarchical bootstrap draws, we sample runs with replacement and then sample task metrics with replacement inside each sampled run. The 2.5th and 97.5th percentiles form the reported interval. This keeps the paired four-cell contrast intact while representing uncertainty across both training repetitions and tasks. The target is the mean interaction over the empirical task suite and training randomness; task resampling is used only to quantify sensitivity to task composition. With three independent runs, these intervals are descriptive finite-sample summaries, not asymptotic guarantees over unseen architectures or task distributions. We therefore report whether an interval excludes zero under this prespecified resampling scheme rather than assigning parametric significance levels.

Table 19: Frozen language interactions (pp) for every paired independent run. These values are the OrthoReg-minus-Standard difference in gains over identity and are aggregated in Table 18.

| Operator | Replicate 1 | Replicate 2 | Replicate 3 |
|---|---|---|---|
| Fixed depth | $+1.54$ | $+4.28$ | $+1.43$ |
| LiNeS | $+1.56$ | $+3.37$ | $+1.36$ |
| TIES | $-2.37$ | $-4.96$ | $-5.24$ |
| TSV-Merge | $-2.38$ | $-3.71$ | $-2.49$ |

The signed pattern is descriptive: TIES and TSV-Merge are negative in all three runs, while the two depth operators are positive but individually unresolved. The inferential result comes from the complete pairwise test in Table 10, where only the fixed-depth–TIES and LiNeS–TIES simultaneous contrasts exclude zero.

**Radius and geometry controls.** The selected TIES radius is $0.2$ for Standard FT and $0.4$ for OrthoReg. To exclude that difference as the sole explanation, we evaluate OrthoReg TIES at the Standard radius $0.2$ for all three frozen runs while retaining each geometry's selected density. The resulting interaction is $-5.05$ pp with a hierarchical bootstrap 95% interval of $[-8.73, -1.66]$. Mean pairwise global task-vector cosine is much lower after OrthoReg in every run: the respective three-run means are $0.1347 \pm 0.0046$ and $0.0195 \pm 0.0006$. This audit is descriptive of geometry rather than a causal decomposition, but it is consistent with reduced residual overlap in the coordinate and singular-subspace readouts.

**Fixed-policy strength diagnostic.** The main language comparison estimates validation-selected family envelopes. To separate that estimand from a fixed operator member, we additionally hold each of two exact TIES policies constant while varying the OrthoReg coefficient in one training run. The Standard-selected policy uses density $0.2$ and radius $0.2$; the OrthoReg-selected policy uses density $0.1$ and radius $0.4$. Table 20 reports every prespecified strength.

Table 20: One-run language TIES policy-by-strength diagnostic. Gains are over the identity merge trained at the same strength; interactions subtract the $\lambda = 0$ gain under the same fixed policy (pp). No configuration is reselected across strengths.

| | Standard-selected policy | | OrthoReg-selected policy | |
|---|---|---|---|---|
| $\lambda$ | Gain | Interaction | Gain | Interaction |
| 0 | 16.701 | 0 | 0.360 | 0 |
| 0.03 | 13.563 | −3.138 | 13.799 | +13.440 |
| 0.10 | 14.045 | −2.655 | 14.327 | +13.968 |
| 0.30 | 14.132 | −2.569 | 14.000 | +13.641 |

Under the Standard-selected member, the negative TIES interaction persists at all three nonzero strengths, so it is not specific to $\lambda = 0.1$. Under the OrthoReg-selected member, however, the sign reverses because that sparse, larger-radius policy transfers poorly to Standard FT. Hence the exact configuration is part of the operator: a family-level selected envelope and a fixed-member contrast answer different questions. This diagnostic uses one run and is not a substitute for the replicated family-level result in Table 18.

Figure 9 shows that the aggregate effects are not a uniform shift over tasks. CoLA is negative for every operator, whereas STS-B provides most of the positive depth-operator interaction. The negative TIES and TSV-Merge aggregates receive large contributions from MNLI and CoLA, with smaller offsets on other tasks. This heterogeneity supports an operator-specific task decomposition, but is not by itself evidence of metric relativity. That claim is tested by the fixed-output audit in Table 12. Since STS-B uses Spearman correlation and the other rows use accuracy, the cells are components of the local FusionBench aggregate, not official GLUE scores to compare across tasks.

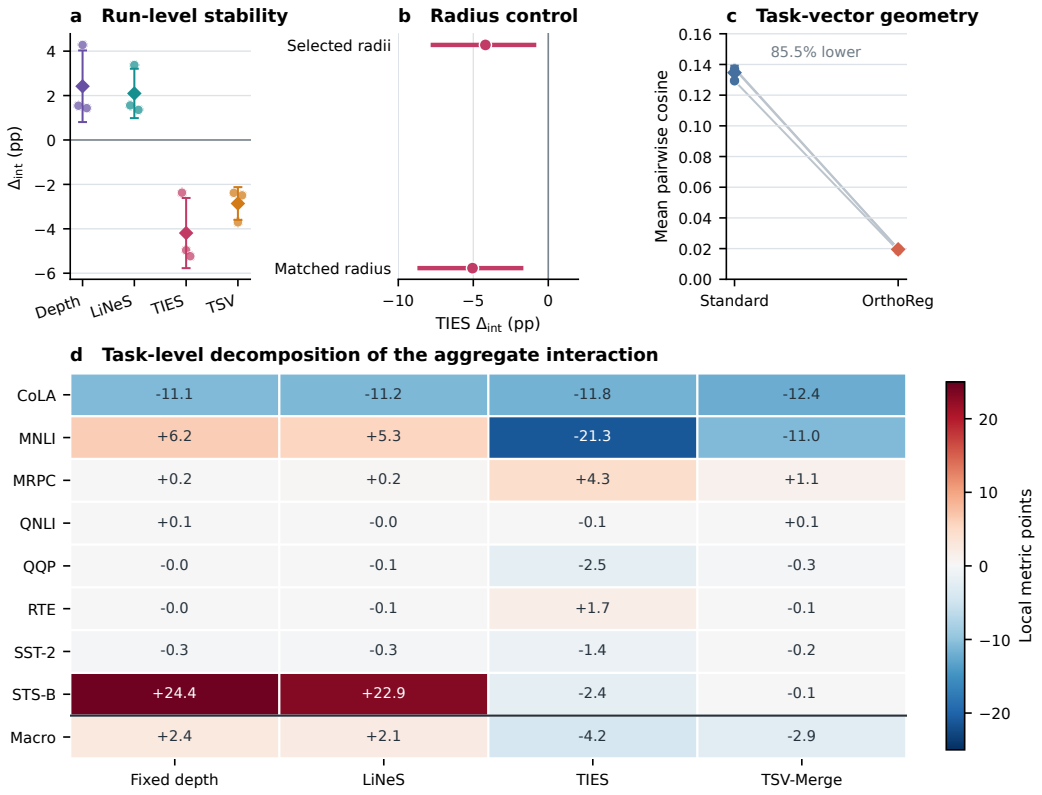


Figure 9: Multilevel language audit. (a) Run-level interactions show stable signs for all four operator families. Diamonds and bars denote the mean and sample standard deviation. (b) The negative TIES interaction persists when both training geometries use the same radius. Bars are hierarchical bootstrap 95% intervals. (c) OrthoReg lowers mean pairwise global task-vector cosine in every paired run; this is descriptive, not a causal mediation result. (d) Task-level interactions expose heterogeneous contributions to the macro effect. Cells use each task's local FusionBench metric; the Macro row is their exact unweighted mean.

## L Identity-Inclusive Deployment

The main analysis measures family compatibility and therefore applies the validation-selected member of every operator family, even when its validation gain is negative. In a deployment comparison, identity can be included as a fallback by deploying the better validation setting between identity and the selected family. This safeguard changes deployment utility but not the estimand in Equation 4: the latter asks how much a fixed operator family adds conditional on its source geometry.

The fallback removes negative deployed gains but does not remove the training–operator interaction. All language families beat identity during the frozen development selection, so their deployment and family-estimand outcomes coincide. Complete run-level decisions are included in the supplementary archive.

Table 21: Identity-inclusive vision deployment gain (pp), averaged over three runs. A family is deployed only when it beats identity on validation.

| Backbone | Family | Standard | OrthoReg |
|---|---|---|---|
| B/32 | Fixed depth | 1.95 | 3.39 |
| | LiNeS | 4.48 | 3.23 |
| | TIES | 1.80 | 0.18 |
| | TSV-Merge | 8.24 | 8.88 |
| B/16 | Fixed depth | 0.06 | 2.17 |
| | LiNeS | 2.45 | 2.22 |
| | TIES | 1.09 | 0.00 |
| | TSV-Merge | 6.36 | 7.70 |
| L/14 | Fixed depth | 2.15 | 1.15 |
| | LiNeS | 2.30 | 1.02 |
| | TIES | 0.80 | 0.00 |
| | TSV-Merge | 4.43 | 2.87 |

## M SCOPE OF EXPLORATORY LARIS RESULTS

Our initial experiments combined sequential projection, residual mixing, and depth rescaling. Those experiments motivated the present study but are not used as confirmatory evidence: the original sweeps evaluated multiple profiles on test data and did not match global displacement radius. Moreover, rescale-only ablations explained most of the addition gain, while projection effects varied by training mode. We therefore replaced the method-centric LARIS claim with the operator-relative analysis and reran the principal comparisons under the protocol in Appendix H. This separation prevents exploratory model selection from entering the primary tables.