REPAIR INFERENCE AFTER REPRESENTATION-ALIGNMENT FAILURE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a representation-alignment correction underperforms a task reference, which tested repair recovers more, and can the repair family reach that reference? We study this *post-failure repair inference* through matched component replacements and exact optimization over positive-scale repair paths. In THINGS-EEG2 image retrieval from electroencephalography (EEG) signals, we attribute the change in the ridge–Procrustes accuracy gap between people included in and excluded from encoder training. Direction receives greater attribution at the cohort mean. Yet for excluded people, replacing ridge's radius (centered output length) with the Procrustes radius recovers 3.46 percentage points more accuracy on average than replacing direction. The advantage is positive in all ten. In OfficeHome image–text retrieval, we scale centered outputs of fixed maps before adding back the fixed mean. Even allowing a separate label-informed positive scale for every query, the exact optimum gains 27.62 percentage points over the fitted correction but remains 9.43 points below the unmodified reference. Scale choice alone therefore cannot restore reference performance on these paths. For a second EEG encoder, the exact label-informed per-query optimum exceeds its subject-centered reference by 7.96 points, establishing attainability. Matched replacements quantify the recovery forgone by following attribution to the weaker tested repair. Exact optima determine whether the declared scale family can recover the reference.

1 INTRODUCTION

When a representation-alignment correction underperforms a task reference, what should change next? Failure alone leaves two questions: which component replacement recovers more accuracy, and whether the repair family can reach the reference. These questions arise in multimodal alignment and cross-subject transfer (Sucholutsky et al., 2025; He & Wu, 2020; Rodrigues et al., 2019). A weaker replacement forgoes recoverable accuracy; tuning can improve accuracy without reaching the reference. What evidence establishes replacement gains and the repair family's limits?

Representational similarity measures compare neural or model representations (Kriegeskorte et al., 2008; Kornblith et al., 2019; Williams et al., 2021). Several measures also quantify average agreement between optimal linear predictors across a distribution of tasks (Harvey et al., 2024). Empirically, the link to task behavior depends on evaluation data (Ciernik et al., 2025). In multimodal tasks, it also depends on the balance of shared and modality-specific task information (Tjandrasuwita et al., 2025). These relationships alone do not determine which component replacement recovers more accuracy.

Functional alignment and repair studies already evaluate such task effects. Model stitching tests whether connecting parts of different networks preserves task performance (Bansal et al., 2021; Csiszárik et al., 2021), including on ImageNet (Deng et al., 2009; Russakovsky et al., 2015; Smith et al., 2025). Retrieval repairs target hubness: a few candidates recur as nearest neighbors. Changing mapping objectives or direction can reduce hubness and improve retrieval (Lazaridou et al., 2015; Shigeto et al., 2015). Huang & Zhu (2026) improve cross-subject image retrieval from electroencephalography (EEG) signals by calibrating similarities with unlabeled test data. Sato et al. (2026) test class-score corrections for bias from reduced image–text separation. Does the component assigned greater attribution for a measured accuracy contrast also yield the larger gain when replaced?

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.

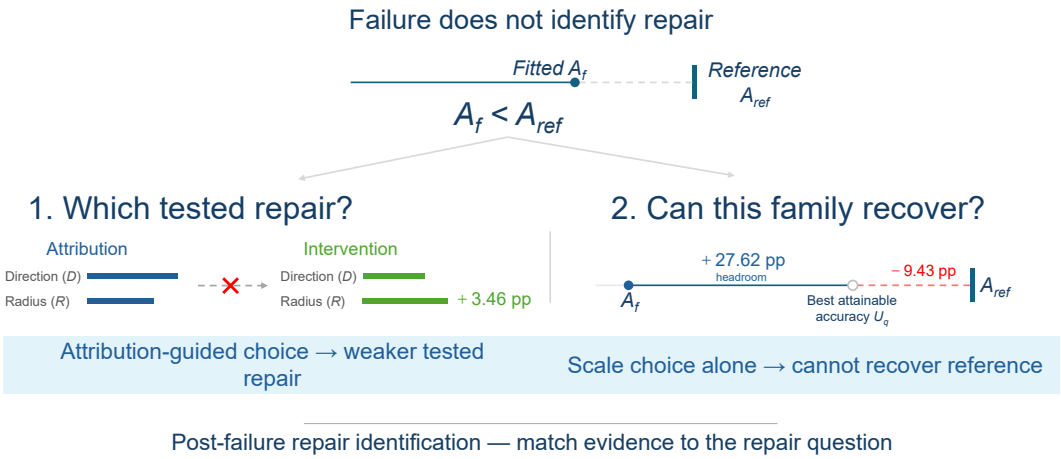


Figure 1: **Failure does not identify repair.** A_f and A_{ref} denote fitted-correction and reference accuracies. Bars and line segments are schematic; results average over people (left) and target domains (right). **Left (THINGS-EEG2 (Gifford et al., 2022)):** D and R denote the direction and radius (length) of centered map outputs. Blue bars (attribution) show component contributions to the change in the ridge–Procrustes accuracy gap between people included in and excluded from encoder training. Green bars (intervention) show accuracy gains over ridge for excluded people after replacing its radius or direction with Procrustes'. The red cross marks the opposite rankings. Following attribution would select direction, forgoing 3.46 percentage points (pp) of mean recovery relative to radius replacement. **Right (OfficeHome (Venkateswara et al., 2017)):** the family scales centered outputs before adding back the mean, keeping maps, means and retrieval candidates fixed. Using evaluation labels, we compute the exact highest attainable accuracy U_q with a separate positive scale for each query. The available improvement (headroom) is $U_q - A_f = 27.62$ pp, yet $U_q - A_{\text{ref}} = -9.43$ pp against the unmodified reference. No positive-scale choice in this family can recover the reference.

Feature importance for explaining a prediction can differ from feature importance for changing it (Mothilal et al., 2020; 2021). Recourse research evaluates which feasible actions achieve a desired outcome (Ustun et al., 2019; Karimi et al., 2021). We quantify the accuracy cost of using attribution to choose between matched alignment-component replacements. We exchange the radius (length) and direction of centered outputs from ridge regression and orthogonal Procrustes maps. The same exchange outputs support both attribution and repair comparisons. Attribution decomposes the change in the ridge–Procrustes accuracy gap between people included in and excluded from encoder training. Repair compares replacement gains for excluded people. We test whether attribution favors the weaker replacement under both orders of assigning credit to components.

Even the stronger tested repair may fall short of the reference. Can an untested member of the repair family recover it? Failed stitching leaves open whether a suitable linear mapping was not found or whether the representations are not linearly mappable (Smith et al., 2025). Unsupervised validation compares fitted candidates without target labels (You et al., 2019; Saito et al., 2021). Recourse verification instead tests feasibility over complete sets of allowable actions (Kothari et al., 2024). We compute exact capacity—the highest attainable accuracy—over all positive scales of centered map outputs. We scale before adding back the mean, keeping maps, means and galleries fixed. Using evaluation labels, we compute the optimum with one scale per map and, separately, one per query. Each optimum determines whether the reference is attainable under that scale constraint. If even the querywise optimum falls short, no positive-scale choice on these fixed paths can recover the reference.

Our first finding concerns which component to replace: attribution favors the weaker tested repair in THINGS-EEG2 EEG–image retrieval (Gifford et al., 2022). At the cohort mean, direction dominates attribution of the change in the ridge–Procrustes accuracy gap between people included in and excluded from encoder training. This holds under both allocation orders and all mixtures using the same weights in both roles. Yet for excluded people, replacing ridge's radius with Procrustes' recovers 4.43 percentage points, versus 0.97 for replacing direction. Following attribution would choose the weaker tested repair, forgoing 3.46 points of mean recovery. The radius advantage is positive in all ten people. One scalar per source fit, estimated from fitting data, retains 4.32 points of recovery without evaluation labels. In the Cogitate cohort recorded at the University of Birmingham (Liu et al., 2026), the matched analysis does not reproduce the radius-over-direction ordering (Section 3).

Our second finding is that large gains within a repair family can still leave the reference unattainable. Using evaluation labels, we compute the exact maximum accuracy over the fixed positive-scale paths with one scale per query, U_q. In OfficeHome image–text retrieval (Venkateswara et al., 2017), this optimum improves on the fitted correction by 27.62 percentage points yet remains 9.43 points below the unmodified reference. Task validation already favors the reference; exact capacity additionally excludes reference recovery by any positive-scale choice on these fixed paths. For the same excluded THINGS participants, the preselected Shared Temporal Attention Encoder (STAE) (Wu et al., 2026) instead has U_q 7.96 points above its subject-centered reference. This establishes label-informed querywise attainability; a common scale or a source-learned selection rule requires separate evidence (Section 4).

We establish two post-failure conclusions that attribution and fitted performance alone leave unresolved. Matched component replacements quantify the accuracy cost of following attribution to the weaker tested repair. Exact capacity determines whether any positive-scale choice on the declared paths can recover the reference. Together, these conclusions identify the stronger tested repair and whether scale choice alone can recover the reference (Figure 1).

2 Repair Identification: Measurements and Study Design

Measurements and repair questions. Both studies measure top-1 accuracy: how often the highest-ranked candidate is correct. A query is an item to match; its candidate set is the gallery. An evaluation panel fixes queries, candidates and correct matches. Attribution allocates an accuracy contrast across components; replacement gain measures the effect of exchanging one component. We test whether greater attribution selects the stronger replacement (Section 3). Capacity asks whether any positive output scale can reach an external reference: U_c allows one scale per fitted map and panel, and U_q one per query. Both optima use evaluation labels (Section 4).

Comparison conditions and references. Each within-role comparison fixes queries, gallery, scores, tie rules and averaging weights. Component exchanges reuse ridge and Procrustes (T_1) outputs and their shared visual mean, replacing only centered output length (radius) or direction. Attribution concerns the included–excluded role contrast; repair gains use excluded-role ridge as baseline. For the convolutional EEG encoder TSconv, excluded ridge is below T_1 but above domain centroid removal (DCR), which subtracts each person's mean representation. Capacity fixes the map and means, varying only positive scale before mean add-back and unit normalization. Its external reference T_0 is DCR for EEG and the unmodified representation for OfficeHome.

Matched roles and allowed information. THINGS uses 200-way galleries drawn from the 1,654 encoder-training concepts (object categories) of THINGS-EEG2 (Gifford et al., 2022), separately from its 200 test concepts. Ten fixed leave-one-subject-out (LOSO) encoders each exclude one person. A matched context pairs an included and excluded evaluator under the same encoder, maps and gallery. Both evaluators and the evaluation concepts are held out from the eight-person map fit and ridge selection. Each evaluator's unlabeled responses to fitting concepts supply its DCR centroid. Nine evaluator pairings crossed with five concept folds give 45 contexts per person per role. A person's two roles use different saved encoders across contexts; this comparison does not isolate a causal effect of membership.

Table 1: **Information and timing contracts.** Source denotes fitting/validation; target denotes transfer evaluation. An arm is one correction; gallery size counts candidates per query. Geometry selection scores representation agreement; task selection scores retrieval. UoB and PKU are Cogitate sites. Exchanges use fixed outputs without query correctness; evaluation labels score all arms and additionally optimize capacity. Appendices A–B give full rules.

Study	Fit / evaluation contract	Allowed evidence / timing	Reporting unit
THINGS components	8-person map fit; matched included/excluded evaluation; 200-way galleries.	Same four arms; TSconv discovery. Source-scalar recipe frozen before STAE rule outcomes.	10 people; 45 contexts per person per role.
Matched UoB	44-person map fit; identity-disjoint fit/evaluation halves; 10-way galleries.	Same four arms; component/capacity forecasts frozen before target endpoint.	46 people.
STAE capacity	Original ridge maps and 200-way galleries fixed.	Source scalars fixed; all positive scales allowed. Exposed-cohort capacity analysis.	10 people; 450 contexts in total per role.
OfficeHome capacity	One OpenCLIP encoder; fixed maps and 10-way galleries.	45/10/10 geometry/source-task/target classes. Full positive-scale envelope frozen before outcomes.	4 overlapping domain routes.
PKU forecast	Same UoB model/map bank; 10-way galleries.	Forecast frozen before PKU access; preliminary evaluation (Appendix D.2).	15 interim people; 46 shared models/person.

Rule timing and prior exposure. Table 1 distinguishes when rules were fixed relative to outcomes. TSconv discovery motivated the source-scalar recipe; it was frozen before STAE rule outcomes on the same people. Matched Cogitate UoB (University of Birmingham) tests a frozen component prediction under a different task. Its later capacity forecast precedes a new endpoint on the previously analyzed cohort. OfficeHome (Venkateswara et al., 2017) also reuses analyzed routes; the full positive-scale range was fixed before new capacity outcomes. PKU forecasts precede data access; the 15-person evaluation is preliminary (Appendix D.2). Supporting score analyses use 654 concepts with prior geometry exposure but withheld task outcomes.

Empirical scope and aggregation. Contrasts are conditional on saved encoders and maps. Neural summaries average within physical person, then weight people equally. Contexts, folds, path points, queries and architectures do not increase population sample size. OfficeHome weights classes and domains equally across four routes, each holding out one domain and fitting on the other three; shared sources induce dependence. Bootstrap intervals resample fixed person summaries without refitting. They are descriptive resampling ranges, not calibrated population confidence intervals. Person counts describe cohort recurrence. Appendix A gives full protocols, including the distinct sealed-target test.

3 Across-Role Attribution Can Mis-rank Tested Repair

3.1 The larger attribution component is the weaker tested repair

Does the component explaining the larger included–excluded performance difference also offer the stronger repair for excluded people? The same four matched arms test this inference. THINGS attribution favors direction, yet radius replacement gains 4.43 points over excluded ridge, versus 0.97 for direction, and wins in all ten people. Recovery also persists with one scale per fitted map estimated from fitting data. This rule gains 4.32 percentage points over ridge for TSconv excluded people, compared with 4.43 for query-specific radius replacement (Section 3.4; Table 2).

For a given query, write the nonzero centered Procrustes and ridge outputs as $h_T = r_T u_T$ and $h_R = r_R u_R$. Their Euclidean lengths are $r_T = \|h_T\|$ and $r_R = \|h_R\|$, and their unit directions are $u_T = h_T/r_T$ and $u_R = h_R/r_R$. Within each evaluation role, let t, r, a, b denote the top-1 accuracies

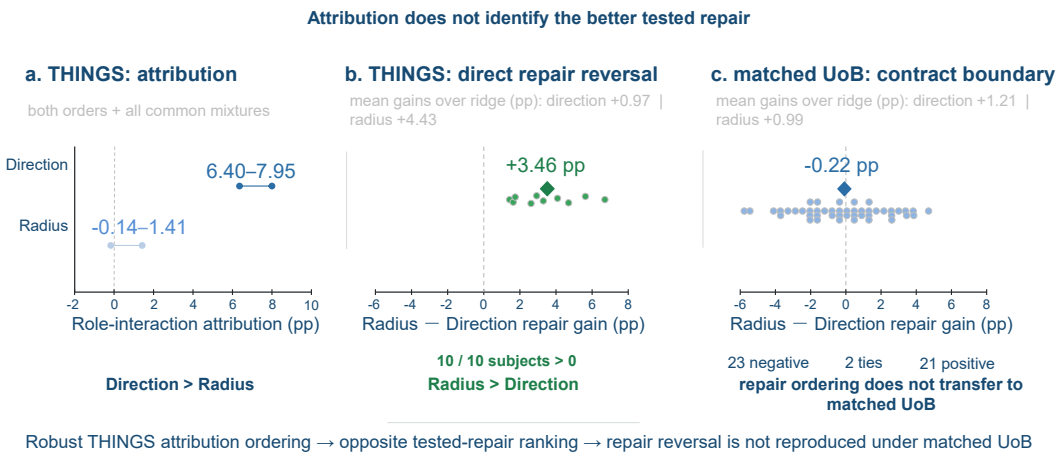


Figure 2: **Across-role attribution favors the weaker tested repair. (a)** THINGS TSconv attribution decomposes the change in the ridge–Procrustes accuracy gap between encoder-included and excluded people. At the cohort mean, direction receives 6.40–7.95 percentage points and radius −0.14–1.41. Segments span both allocation orders and every mixture with the same weights in both roles; they are not uncertainty intervals. **(b,c)** G_R and G_D are excluded-role accuracy gains over ridge from replacing its radius or direction with the Procrustes component. Their difference, $\Delta_{\text{repair}} = G_R - G_D$, averages +3.46 points in THINGS and is positive in all ten people. In matched UoB it averages −0.22 points: 21 positive, 23 negative and two tied people. Dots are people; diamonds are means. Cohorts retain their respective protocols and are not pooled.

of the four arms below. The second row gives their query vectors before normalization:

accuracy	t	r	a	b
query vector	$\mu + r_T u_T$	$\mu + r_R u_R$	$\mu + r_T u_R$	$\mu + r_R u_T$

Here μ is the shared fitted visual mean. Query vectors are normalized to unit length and scored by dot product with unit candidates (cosine similarity). Arm a replaces radius and b direction, without refitting maps or regularization. Excluded gains are $G_R = a - r$ and $G_D = b - r$. On the same four arms, choosing direction instead of radius costs $G_R - G_D = a - b$, averaging 3.46 percentage points. This is a paired, retrospective comparison of the two tested choices.

Arm a reaches 29.16% versus T_1's 29.07%, establishing neither superiority nor statistical equivalence. The reverse exchange b reduces T_1 accuracy by 3.37 points in all ten people. Removing the shared mean eliminates the ridge–Procrustes ranking reversal between roles but lowers absolute accuracy (Appendix B).

3.2 The ranking mismatch survives both attribution orders

The attributed contrast is $J = (r - t)_{\text{included}} - (r - t)_{\text{excluded}}$. Starting from T_1, direction-first replacement follows $t \to a \to r$; radius-first follows $t \to b \to r$. Let $w \in [0, 1]$ weight the first order, using the same weight in both roles. Each order, and every such mixture, allocates the within-role gap $r - t$ to direction and radius. Their included-minus-excluded differences are $J_D(w)$ and $J_R(w)$, with $J_D(w) + J_R(w) = J$. Direction dominance means $J_D(w) > |J_R(w)|$. All contributions are affine in w, so the two endpoint checks suffice (Appendix B, including the excluded-loss test). The repair difference $a - b$ is independent of this allocation.

Direction dominates the TSconv role contrast for every common weight in all ten people (Figure 2). At the mean, radius instead dominates the excluded loss: its contribution is −4.43 to −3.37 points, versus direction's −0.97 to +0.08. Both dominance properties hold jointly for every common weight in $8/10$ people. The direct repair advantage $a - b$ is positive in $10/10$.

At the cohort mean, attribution of the excluded-role loss agrees with the repair ordering: both favor radius. The mismatch appears when attribution of the role contrast is used to rank repairs within the excluded role. Radius replacement improves included accuracy by 4.57 points and excluded

Table 2: **Source-fitted recovery can leave a reference deficit.** Top-1 accuracy (%), averaged equally across ten people per role. Δ is scaled ridge minus ridge in percentage points; differences use unrounded means. DCR is centering and T1 is Procrustes.

Encoder	Role	DCR	T1	Ridge	Scaled ridge	Δ vs. ridge
TSconv	Included	46.05	57.08	60.54	65.15	+4.61
	Excluded	22.40	29.07	24.73	**29.04**	**+4.32**
STAE	Included	66.53	67.29	61.46	72.04	+10.58
	Excluded	22.32	23.56	13.77	**18.05**	**+4.28**

accuracy by 4.43 points. Under the direction-first allocation, these gains enter the role contrast with opposite signs, leaving a radius contribution of only -0.14 points. A repair gain present in both roles can therefore be weak in their contrast while remaining useful within each role.

3.3 The matched UoB prediction fails

A frozen UoB prediction requires direction-dominant interaction and radius-dominant excluded loss. It fails at the mean and holds in $0/46$ people. Radius improves excluded ridge by 0.99 points, versus 1.21 for direction. The radius-minus-direction difference averages -0.22 points, with 21 positive, 23 negative and two tied people. Both replacements improve mean accuracy; the failed prediction concerns their ordering and attribution pattern. DCR and T_1 exceed 10% chance in all 46 excluded people. UoB's partial-polar map retains only directions supported by fitting data; gallery size and fitting support also differ. These conditions limit transfer without identifying a population cause. Appendix B retains the separate 12-person scalar and radius-exchange failures.

3.4 Recovery with one scale fitted on source data

This control uses one positive scale per map. Define $\mathrm{norm}(v) = v/\|v\|$ for nonzero v. $\mathbb{E}_{\mathrm{fit}}$ averages fitting person–concept pairs equally over eight people and their concepts. The scale s is their geometric mean Procrustes-to-ridge length ratio; the repaired query is $\hat{z}$:

$$s = \exp\left\{\mathbb{E}_{\mathrm{fit}}\left[\log\frac{\|h_T\|}{\|h_R\|}\right]\right\}, \qquad \hat{z} = \mathrm{norm}(\mu + sh_R). \tag{1}$$

Both lengths must be positive and $\mu + sh_R \neq 0$. Ridge regularization uses validation within fitting people and concepts. The same s serves all queries in both roles. Scaling precedes mean add-back: for a nonzero output, $\mathrm{norm}(\mu + sh_R) = \mathrm{norm}(h_R + \mu/s)$. It changes the output-to-mean ratio and hence can change cosine rankings. Table 2 retains ridge, DCR and T_1 comparisons; TSconv excluded recovery remains just below T_1.

The recipe was frozen before testing STAE on the same ten people. Gains are positive in every person in both roles: 4.28 excluded and 10.58 included points over ridge. Excluded accuracy remains below DCR and T_1. Section 4 tests whether any scale can close that deficit.

4 Headroom Can Misstate Family Sufficiency

The STAE deficit in Table 2 leaves open whether another scale can recover the reference. We separate the improvement available from a better common scale from the additional improvement available when each query may use its own scale. We then compare these exact capacities with the external reference. This yields different conclusions on the fixed STAE and OfficeHome paths: querywise reference recovery is attainable in STAE and excluded in OfficeHome.

4.1 Exact capacity of the declared scale family

Fix a map, its mean μ, n queries and K unit candidate vectors $y_1, \ldots, y_K$. Query i has centered output h_i and correct candidate index c_i. Its path is $q_i(\alpha) = \mu + \alpha h_i$ for $\alpha > 0$. Cosine scoring is defined on $\mathcal{A}_i = \{\alpha > 0 : q_i(\alpha) \neq 0\}$; $\mathcal{A} = \bigcap_i \mathcal{A}_i$ contains scales legal for all queries. The fitted

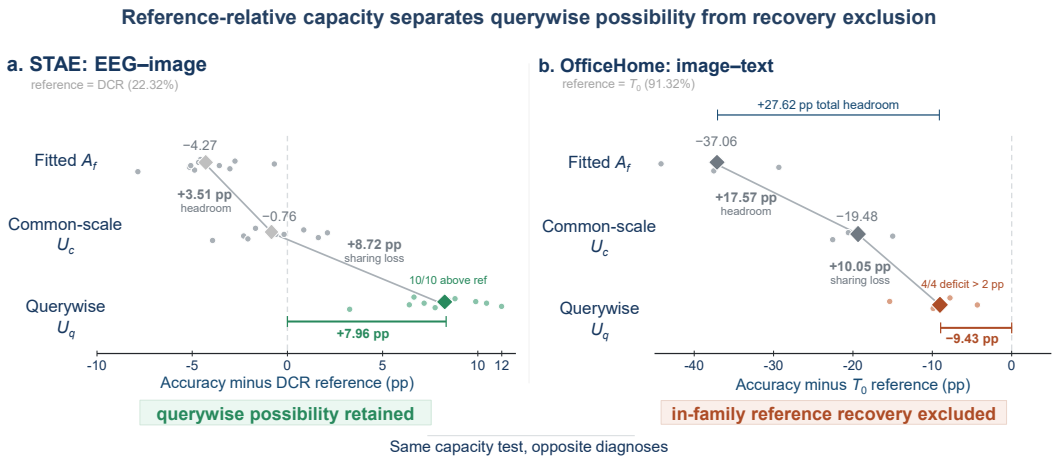


Figure 3: **Reference-relative capacity separates querywise attainability from recovery exclusion.** Accuracy is relative to each unit's reference: DCR for STAE and unmodified T_0 for OfficeHome. Points are people or target routes; diamonds are means. Panel scales differ. A_f is the source-scaled ridge accuracy in STAE and the geometry-selected accuracy in OfficeHome. Both capacities are label-informed maxima: U_c uses one common scale per map/panel; U_q permits one per query. **(a) STAE.** Common-scale headroom $U_c - A_f$ is 3.51 percentage points and the sharing gap $U_q - U_c$ is 8.72. Mean U_q exceeds DCR by 7.96 points, with a surplus in all ten people. **(b) OfficeHome.** Internal headroom $U_q - A_f$ is 27.62 points, yet U_q remains 9.43 points below T_0. All four routes have deficits above 2 points, excluding reference recovery on these fixed positive-scale paths. Differences use unrounded means.

rule has nonzero outputs, so $\mathcal{A}$ is nonempty. On a legal scale, query normalization cancels from candidate comparisons:

$$m_{ij}(\alpha) = \alpha h_i^\top (y_{c_i} - y_j) + \mu^\top (y_{c_i} - y_j), \qquad j \neq c_i.$$

Here $\top$ denotes transpose and m_{ij} is the correct candidate's score advantage over competitor j before normalization. These affine margins (linear plus constant) and the original tie rule define a winning interval. Its intersection with $\mathcal{A}_i$ is the legal winning set I_i, possibly empty or split by a zero-output scale. Valid isolated wins are retained. For fixed weights $w_i \geq 0$ summing to one, let $\mathbf{1}\{\cdot\}$ indicate a true condition. Then

$$U_c = \max_{\alpha \in \mathcal{A}} \sum_{i=1}^{n} w_i \mathbf{1}\{\alpha \in I_i\}, \qquad U_q = \sum_{i=1}^{n} w_i \mathbf{1}\{I_i \neq \varnothing\}. \tag{2}$$

U_c is the best common-scale accuracy; U_q permits a separate legal scale per query ($I_i \neq \varnothing$ means a winning scale exists). Both use evaluation labels. Optimization is separate for each map, panel and role, before fixed-weight aggregation; OfficeHome shares one common scale across the role's full route panel. Formula accuracies are fractions; reported differences are percentage points.

A finite scan checks open segments and legal endpoints partitioned by margin boundaries and zero-output scales, including the unbounded segment. Appendix C supplies tie rules, exact rational arithmetic conditional on represented float64 coefficients, and checks showing that zero exclusions preserve the reported bounds. It also verifies original fitted outcomes. Calibration paths and label-informed optima are established tools (Chao et al., 2016; Lindauer et al., 2019); here complete-path optimization tests reference recovery.

Let A_f be fitted-rule accuracy using a scale in $\mathcal{A}$ and the same weights. The nested constraints give $A_f \leq U_c \leq U_q \leq 1$ and

$$1 - A_f = \underbrace{(1 - U_q)}_{\text{unreachability}} + \underbrace{(U_q - U_c)}_{\text{sharing loss}} + \underbrace{(U_c - A_f)}_{\text{fitted-rule headroom}}. \tag{3}$$

The terms are, respectively, query weight unreachable at any scale, capacity lost by sharing a scale, and improvement available above the fitted rule. All are nonnegative: this is exact accounting, not causal mechanism attribution. For external accuracy A_{ref} on the same queries, candidates, scores and weights,

$$A_{\mathrm{ref}} - A_f = (U_c - A_f) + (U_q - U_c) + (A_{\mathrm{ref}} - U_q). \tag{4}$$

The last term is signed: positive for residual shortfall and negative for querywise surplus. The reference need not lie on the fixed path. If $U_q < A_{\mathrm{ref}}$, no positive scale choices recover the reference, even with a separate label-informed choice per query. If $U_q \geq A_{\mathrm{ref}}$, such recovery is attainable on this panel; this establishes neither common-scale recovery nor a learnable selector. Fixed-weight aggregation preserves nesting; claims about every unit use unitwise comparisons. Appendix C.1 formalizes the inference boundary.

4.2 Retained possibility: STAE crosses its reference

The original 450 STAE fits retain all source scalars, means and 200-way galleries. Excluded accuracy is 18.05% for the fitted rule, 21.56% for U_c and 30.28% for U_q, against DCR's 22.32%. The querywise surplus is positive in all ten people (Figure 3); U_q also exceeds T_1's 23.56%. Of the fitted rule's 4.27-point DCR deficit, 3.51 points are common-scale headroom. The residual 0.76-point common deficit excludes recovery at this cohort mean, with seven people below DCR and three above. It misses the prespecified practical screen of 2 points and eight people, so it does not establish a stable need for query adaptation. The included sharing gap is also 8.70 points; the large gap is not excluded-role specific.

4.3 Family exclusion: OfficeHome cannot close the deficit

Source geometry selected a correction on all four OfficeHome routes; source-task validation and the best tested action chosen with target labels both selected unmodified T_0. All routes pass reference-accuracy and mismatched-pair competence controls (Appendix A). The full positive range was declared before new capacity outcomes, keeping maps, means, galleries and class/domain weights fixed.

Mean target accuracy is 54.26% fitted, 71.84% common, 81.89% querywise and 91.32% for T_0. Common headroom is 17.57 points and the sharing gap 10.05; each exceeds 2 points on every route. Despite 27.62 points of internal headroom, the reference deficit is 9.43 points, exceeding 2 points on all four routes (range 4.42–15.46; Figure 3). The original fitting range $[0.5, 2]$ gives a smaller 2.03-point sharing gap; the 10.05-point gap concerns all $\alpha > 0$ (Appendix C). This fixed-map exclusion neither rules out another map nor identifies information loss.

What capacity adds to task selection. Source-task validation already selects T_0 from the tested fitted actions. The complete calculation additionally rules out reference recovery by any positive-scale choice on these OfficeHome paths, including untested scales. Recovering the reference from these corrected outputs therefore requires a change beyond scale. Retaining T_0 is already supported by task validation; STAE, in contrast, retains querywise recovery capacity.

4.4 Supporting recurrence of positive capacity

A source querywise surplus of at least 2 points forecasts target retention; a shortfall of at least 2 predicts insufficiency, and intermediate values abstain. On historically analyzed UoB, forecasts preceded the new capacity endpoint: all 46 routes retain querywise capacity, including 39 where the source-common forecast differs (24 negative, 15 abstentions). Mean target surplus is 16.36 points over DCR (Appendix D.1). The unchanged forecast preceded any PKU access. Preliminary results from all 15 people retain capacity, with a 17.16-point mean surplus and all 39 disagreements supported per person. Always predicting retention ties on both cohorts. PKU mean common capacity also exceeds DCR (17.07% versus 15.41%), so these results establish neither state discrimination nor a need for query adaptation. Full forecast rules, results and interim history appear in Appendix D.

5 Supporting Diagnosis: From Scores to Actual Order

Supporting score analyses motivate evaluating candidate order directly (Appendix E.1). A retrospective THINGS source comparison adds a source-fitted scale to Procrustes on fixed retrieval panels. Similarity to the correct answer increases, but mean retrieval rank worsens. The predicted average rank harm across three prespecified intermediate scales does not occur. A later test uses 654 concepts with prior geometry exposure but withheld retrieval outcomes. In all ten people, pairs where the answer initially leads by a small score gap reverse order more often than pairs with a large gap. These results support direct task evaluation; they do not establish a common mechanism for the two repair findings in Sections 3 and 4.

6 Research Consequences and Scope

Matched gains identify the stronger tested repair. A gain shared by both roles can be small in their contrast yet useful within each role. THINGS directly measures the cost of using across-role attribution for excluded-role repair choice. Its radius advantage does not recur at the UoB mean, delimiting transfer under that study's task and fitting conditions.

Complete capacity identifies what scale cannot repair. OfficeHome's deficit requires a change beyond scale on the fixed paths; it does not prescribe another correction. Task validation already supports retaining T_0. STAE instead retains querywise capacity above DCR, leaving an explicit unresolved step: realizing it without evaluation labels. Capacity alone establishes neither selector accuracy nor value from search.

Scope follows the tested information conditions. Positive capacity recurs in UoB and preliminary PKU results; OfficeHome provides the complementary exclusion case across four overlapping routes. These outcomes span EEG–image and image–text retrieval, each conditional on its map, reference and scale constraints. Component gains concern encoder-seen concepts held out from map fitting. In a separate THINGS test holding out both people and concepts from encoder training, T_2 (Procrustes with a source-fitted scale) worsens mean rank relative to DCR. The original UoB task study also favors centering (Appendices A and F). Thus geometric correction can harm retrieval under these conditions; the component gains' extension to joint subject and concept holdout remains unestablished.

7 Conclusion

We answer two questions left open by alignment failure: which tested replacement recovers more, and whether scale choice alone can reach the reference. In matched THINGS comparisons, attribution of the role contrast favors direction, yet radius replacement recovers 3.46 percentage points more accuracy for encoder-excluded people, with a positive advantage in all ten. In OfficeHome, even the label-informed per-query optimum gains 27.62 points over the fitted correction but remains 9.43 points below the unmodified reference. Scale choice alone cannot recover the reference on these fixed paths. STAE provides the complementary case: its label-informed querywise optimum exceeds the subject-centered reference by 7.96 points. Together, these results ground post-failure repair inference in the accuracy recovered by a tested replacement and the reachability of a stated reference within a declared repair family.

AI USE STATEMENT

Generative AI tools assisted in refining hypotheses and the conceptual framework, developing and checking mathematical claims and proofs, designing experiments, implementing and debugging methods and preprocessing pipelines, and interpreting results. They also assisted literature retrieval and reference checking, manuscript drafting and translation, language editing, figure preparation, and internal critical review. The authors reviewed and revised the manuscript, made the research decisions, and take responsibility for the final text, analyses, claims, references, code, and figures.

ETHICS STATEMENT

The work analyzes previously collected human neural data under their documented access and data-use terms. It involves no new recruitment and makes no clinical or individual-level deployment claim. The manuscript reports aggregate results.

REPRODUCIBILITY STATEMENT

Appendix A specifies fitting, information permissions and uncertainty limits. Appendices B–D provide component controls, exact capacity calculations, numerical checks and forecast results. Appendices E–F report supporting predictions and transfer boundaries. The supplementary verification package recomputes component and source-scalar summaries from saved outcomes, reoptimizes saved STAE winning intervals, and checks OfficeHome's saved capacity accounting. It also includes the anonymous manuscript source and figure inputs. Raw recordings, model checkpoints, and end-to-end training are not included.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004. URL https://www.jmlr.org/papers/v5/grandvalet04a.html.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pp. 52–68, 2016. doi: 10.1007/978-3-319-46475-6_4. URL https://schangpi.github.io/papers/eccv16_gzsl.pdf.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10920–10948. PMLR, 2025. URL https://proceedings.mlr.press/v267/ciernik25a.html.

Cogitate Consortium, Oscar Ferrante, Urszula Gorska-Klimowska, Simon Henin, Rony Hirschhorn, Aya Khalaf, Alex Lepauvre, Ling Liu, David Richter, Yamil Vidal, Niccolò Bonacchi, Tanya Brown, Praveen Sripad, Marcelo Armendariz, Katarina Bendtz, Tara Ghafari, Dorottya Hetenyi, Jay Jeschke, Csaba Kozma, David R. Mazumder, Stephanie Montenegro, Alia Seedat, Abdelrahman Sharafeldin, Shujun Yang, Sylvain Baillet, David J. Chalmers, Radoslaw M. Cichy, Francis Fallon, Theofanis I. Panagiotaropoulos, Hal Blumenfeld, Floris P. de Lange, Sasha Devore, Ole Jensen, Gabriel Kreiman, Huan Luo, Melanie Boly, Stanislas Dehaene, Christof Koch, Giulio Tononi, Michael Pitts, Liad Mudrik, and Lucia Melloni. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066):133–142, 2025. doi: 10.1038/s41586-025-08888-1.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9771828/.

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 140–151. PMLR, 2024.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Qunjie Huang and Weina Zhu. SATTC: Structure-aware label-free test-time calibration for cross-subject EEG-to-image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16887–16896, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_SATTC_Structure-Aware_Label-Free_Test-Time_Calibration_for_Cross-Subject_EEG-to-Image_Retrieval_CVPR_2026_paper.html.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021. doi: 10.1145/3442188.3445899. URL https://arxiv.org/abs/2002.06278.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529. PMLR, 2019.

Avni Kothari, Bogdan Kulynych, Tsui-Wei Weng, and Berk Ustun. Prediction without preclusion: Recourse verification with reachable sets. In *International Conference on Learning Representations*, pp. 20256–20278, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/591d450c9ee7670a7bc777119d046801-Abstract-Conference.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 270–280, 2015. doi: 10.3115/v1/P15-1027. URL https://aclanthology.org/P15-1027/.

Marius Lindauer, Jan N. van Rijn, and Lars Kotthoff. The algorithm selection competitions 2015 and 2017. *Artificial Intelligence*, 272:86–100, 2019. doi: 10.1016/j.artint.2018.10.004. URL https://www.cs.uwyo.edu/~larsko/papers/lindauer_algorithm_2019.pdf.

Ling Liu, Oscar Ferrante, Tara Ghafari, Dorottya Hetenyi, Shujun Yang, Rony Hirschhorn, Urszula Gorska-Klimowska, Praveen Sripad, Fatemeh Taheriyan, Tanya Brown, Diptyajit Das, Kyle Kahraman, Niccolò Bonacchi, Michael Pitts, Liad Mudrik, Ole Jensen, Huan Luo, and Lucia Melloni. An open multi-center MEG-EEG dataset for studying conscious visual perception. *Scientific Data*, 13:799, 2026. doi: 10.1038/s41597-026-07350-9.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020. doi: 10.1145/3351095.3372850.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 652–663, 2021. doi: 10.1145/3461702.3462597. URL https://arxiv.org/abs/2011.04917.

Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2019. doi: 10.1109/TBME.2018.2889705.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9184–9193, 2021.

Shota Sato, Hajime Kiyama, Tosho Hirasawa, and Mamoru Komachi. When modality gap reduction fails: Prediction-level hubness in CLIP. *arXiv preprint arXiv:2609.01103*, 2026. URL https://arxiv.org/abs/2609.01103.

Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Machine Learning and Knowledge Discovery in Databases*, volume 9284 of *Lecture Notes in Computer Science*, pp. 135–151, 2015. doi: 10.1007/978-3-319-23528-8_9. URL https://arxiv.org/abs/1507.00825.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55972–55998. PMLR, 2025. URL https://proceedings.mlr.press/v267/smith25a.html.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=Hiq7lUh4Yn.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 59723–59760. PMLR, 2025. URL https://proceedings.mlr.press/v267/tjandrasuwita25a.html.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019. doi: 10.1145/3287560.3287566.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric EEG-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17859–17867, 2026. doi: 10.1609/aaai.v40i21.38844. URL https://ojs.aaai.org/index.php/AAAI/article/view/38844.

Kaichao You, Ximei Wang, Mingsheng Long, and Michael Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7124–7133. PMLR, 2019. URL https://proceedings.mlr.press/v97/you19a.html.

APPENDIX GUIDE

The appendices answer six questions needed to interpret the main results: what information each study uses (Appendix A); whether the component contrast survives its controls (Appendix B); how complete scale capacity is computed and interpreted (Appendix C); what the capacity forecasts establish (Appendix D); which score-based predictions succeed or fail (Appendix E); and where retrieval gains do not transfer (Appendix F).

A EXPERIMENTAL INFORMATION CONTRACTS

The main studies use THINGS-EEG2 (Gifford et al., 2022), Cogitate (Cogitate Consortium et al., 2025; Liu et al., 2026), and OfficeHome (Venkateswara et al., 2017). Source means fitting or validation data; target means the data used for transfer evaluation. In this paper, "frozen" and "prespecified" mean recorded before the named endpoint was computed; they do not claim public preregistration or absence of earlier related-data exposure. Table 1 distinguishes these study-specific timing conditions.

Target information and fit identity. The supporting source score analyses use historically exposed checkpoints. The separate joint-holdout THINGS test uses source-only encoders and fixed source maps and scales. Its unlabeled target centering uses the full 200-query panel; the same panel is evaluated, making that contract transductive. Class, checkpoint, and scale choices do not use target labels or retrieval outcomes. The source score results and the joint-holdout result condition on different encoders and fits; they are not a paired membership comparison.

OfficeHome support. The fixed encoder is OpenCLIP ViT-B-32 with the OpenAI pretrained weights. Each route uses 45 fit classes and disjoint ten-class source-task and target-task roles. Centered text anchors have rank 44 in 512 dimensions. The partial-polar map acts only on the rank-44 subspace supported by the fitting anchors and discards the other 468 directions. No arbitrary orthogonal completion supplies information in those directions.

Fitting and selection in THINGS. For each of ten historical representation routes, the outer subject is absent from its fitting-data pool. Of the remaining nine subjects, one evaluates transport and eight supply fit anchors. Five deterministic concept folds cross this split. Each evaluation subject therefore contributes $9 \times 5 = 45$ contexts. This is transport cross-fitting on a frozen representation; it does not make the historical encoder an independent fit for every evaluation subject. Per-subject DCR centroids use the context's fit concepts; the normalized eight-subject mean gives each fit-concept neural anchor. Let X and Y contain paired, centered neural and visual anchors as rows, after the fixed normalization step. Write their cross-covariance singular value decomposition as $X^\top Y = U\Sigma V^\top$, where U, V contain singular vectors and Σ contains singular values. Then the Procrustes map is $R = UV^\top$, minimizing $\|XR - Y\|_F^2$ over orthogonal maps ($R^\top R = I$, where I is the identity matrix). The subscript F denotes the Frobenius norm; its inner product sums elementwise products. With mapped anchors $Q = XR$, the scale is $\hat{\alpha} = \mathrm{clip}(\langle Q, Y\rangle_F / \|Q\|_F^2, .5, 2)$. Here clip restricts the fitted value to $[.5, 2]$. These least-squares objectives differ from held-out angular evaluation. Matrix fitting uses row vectors; query formulas use the equivalent column form $R^\top x$.

Training variation and reuse. The historical source study uses a seed-42 TSconv EEG encoder and ResNet-50 (RN50) visual backbone; checkpoint exposure limits its evidence to the stated

source/forensic role. The preselected follow-up representations are the paired STAE and TSconv source-only retraining routes: seed 42, 150 epochs, batch size 1024, AdamW with learning rate and weight decay 10^{-4}, and final-epoch-only checkpoint selection. The terminal THINGS experiment reuses the paired TSconv final checkpoints. Cogitate uses separate seed-42 TSconv/RN50 LOSO fits with the same epoch, batch and AdamW settings, a tenfold learning-rate reduction every 50 epochs, and deterministic source-subject/category/relevance balancing. None is a multi-training-seed replication. Transport replay has no optimizer seed sweep; two architectures and many contexts do not estimate training-seed variability.

Paired retraining and Cogitate reuse a historical hard mean-average-precision (mAP) term computed through sorting and discrete label matches. This term supplies no gradient to the similarity scores; the remaining differentiable terms train the model. The reported results concern these actual checkpoints and do not establish behavior under a differentiable ranking-loss replacement.

OfficeHome competence controls. OfficeHome requires target T_0 mean normalized rank utility at least .55 and top-1 accuracy at least .15. For answer rank R_i and gallery size $K = 10$, query rank utility is $U_i = 1 - (R_i - 1)/(K - 1)$. These screens are above random values .5 and .1. The four observed utilities are $.982989, .944792, .991980, .997060$ and top-1 values $.898794, .811246, .962634, .980156$ (Art, Clipart, Product, Real World). The wrong-correspondence map must not match or exceed the semantic fit on both held-out source geometry and task utility. The source best nonzero task arm is T_1 in all routes, whereas geometry selects T_2 for Clipart and T_1 otherwise. On target panels, the mean normalized-rank-utility advantage of T_0 is $.257790$ over the best nonzero arm and $.264227$ over the geometry-selected arm. Their difference is entirely the Clipart action contrast.

Uncertainty under shared source fits. Normalized mean rank averages $(R_i - 1)/(K - 1)$; lower values are better. The primary empirical object is the mean contrast across the observed physical subjects, conditional on the saved representations and transport fits. Subject-level aggregation prevents pair/context pseudoreplication; it does not make these estimates independent. The reserve study's 10,000-draw percentile bootstrap uses PCG64 seed 2026090413, resampling ten fixed subject summaries with replacement; no training or transport fitting is repeated. The nominal 2^{-10} sign probability assumes independent Bernoulli-.5 null signs. Neither calculation establishes new-population coverage under arbitrary shared-fit dependence. We retain the recorded numerical summaries but interpret their intervals as fixed-summary resampling ranges, not population-calibrated confidence guarantees. Such overlap is a known obstacle to naive cross-validation uncertainty (Bengio & Grandvalet, 2004); that work does not provide a correction specific to our design. A valid pipeline-level interval would require a justified dependence model or independently sampled cohorts with refitting. Deleting saved blocks cannot reproduce either.

Validation scope. Separately implemented checks verify artifact identity and aggregation at their declared scopes. In particular, the prospective representation/path and reserve validators reconstruct persisted context-to-subject statistics, identities and decisions; they do not perform a second raw-source transport replay. The reserve validator additionally has an independently implemented context reference fixture. These implementation checks do not provide independent population replication.

B MATCHED ENCODER-MEMBERSHIP AND MAP CONTRASTS

This separate discovery study uses the ten fixed, final-epoch TSconv LOSO encoders and their 1,654 training concepts. Each of 450 contexts compares an encoder-training-included person and the encoder-excluded person under the same checkpoint, source-fitted maps, and 200-way gallery. Neither evaluated person enters the eight-person transport fit; evaluation concepts are also excluded from that fit. The excluded person's unlabeled fit-concept support supplies the DCR centroid. Concepts were seen by the encoder on other people: this is not the sealed joint unseen-subject/unseen-concept test. Each physical person contributes 45 contexts per role, averaged separately before equal-person aggregation; shared encoders and fits prevent treating contexts as independent.

For the centered fitting-anchor matrix X_c, ordinary affine ridge uses $\lambda = c\,\mathrm{tr}(X_c^\top X_c)/768$, $c \in \{.001, .01, .1, 1, 10\}$, where tr sums diagonal entries and 768 is the output dimension. Inner

validation holds out one of the eight fitting people and 200 fitting concepts, selecting accuracy with ties favoring larger c before refitting on all eight people. The later one-standard-error-style rule chooses the largest c within $\sqrt{p_*(1-p_*)/200}$ of the best inner accuracy p_*; this is a selection heuristic, not a population uncertainty estimate. Here T1 denotes the Procrustes arm T_1. Ridge's top-1 accuracy exceeds T1 on encoder-included people (60.54% versus 57.08%), but falls below T1 on excluded people (24.73% versus 29.07%). Excluded-role DCR is 22.40%: T1 and ridge gains are respectively 6.67 and 2.33 percentage points, positive in $10/10$ and $8/10$ people. The prespecified map-by-role interaction is 7.81 points, positive $10/10$; individual method-preference reversal occurs in $7/10$, not $10/10$. A subsequent source-only one-standard-error-style ridge selector improves excluded accuracy by only 0.14 points ($5/10$), while reducing included accuracy by 1.77 points ($10/10$). This rejects that fixed repair, not every regularization explanation.

Removing only the fitted visual-mean add-back from both original maps gives the following exact accounting of their accuracy difference. All values are percentage points; the final column is included minus excluded.

Ridge minus T1	Included	Excluded	Difference
Without add-back	$+10.27$	$+2.22$	$+8.06$
Differential add-back effect	-6.81	-6.56	-0.25
Complete maps	$+3.46$	-4.35	$+7.81$

Without add-back, ridge wins in $10/10$ people in each role, with no individual preference reversals. The interaction therefore persists while the preference reversal disappears. Adding the center benefits T1 relative to ridge by similarly sized amounts in both roles, exceeding only the smaller excluded-role ridge advantage. A component may nearly cancel from the interaction yet decide the winner in one role. This is controlled functional accounting, not unique causal mediation, evidence that the center is irrelevant, or a novel theorem. Removing it also harms absolute accuracy and is not a proposed repair.

To localize this effect further, a subsequent fixed counterfactual writes the nonzero centered Procrustes and ridge outputs as $h_T = r_T u_T$ and $h_R = r_R u_R$. Their lengths are r_T, r_R and their unit directions are u_T, u_R. It exchanges only the lengths before adding the shared fitted visual mean μ. Fits, ridge regularization, directions, and galleries remain unchanged. The table includes the fitting-data scalar s from Equation (1). Let t, r, a, b denote the accuracies of its first four rows, respectively, within each role:

Prediction before normalization	Included (%)	Excluded (%)
$r_T u_T + \mu$ (T1)	57.08	29.07
$r_R u_R + \mu$ (ridge)	60.54	24.73
$r_T u_R + \mu$	65.11	29.16
$r_R u_T + \mu$	55.11	25.70
$s h_R + \mu$ (source-fixed)	65.15	29.04

Replacing ridge's radius improves excluded accuracy by 4.43 points and included accuracy by 4.57 points, both in $10/10$ people. The reverse exchange reduces excluded T1 accuracy by 3.37 points, also $10/10$. Thus the disadvantage can be transferred by radius exchange, and its top-1 recovery does not require a source-accuracy sacrifice. The recovering hybrid differs from excluded T1 by only $+0.08$ points, with five positive and five negative people; $8/10$ meet the prespecified lower bound of -1 point. This is practical deficit closure under that screen, not statistical equivalence or superiority. Normalized mean rank remains worse than T1 (0.04515 versus 0.03988). These query-specific diagnostic exchanges alone do not establish a deployable calibration rule or a novel normalization operator.

Dependence on intervention order. We next test whether the attribution ranking depends on which component receives credit first. The attributed contrast is $J = (r-t)_{\text{included}} - (r-t)_{\text{excluded}}$, the change in the ridge–T_1 accuracy gap between roles. Within each role, start from T_1 and replace its two components with ridge's. The direction-first path $t \to a \to r$ assigns direction and radius contributions $(D_1, R_1) = (a - t, r - a)$; the radius-first path $t \to b \to r$ assigns $(D_0, R_0) = (r - b, b - t)$. Let $w \in [0, 1]$ be the weight on the direction-first allocation. The mixed

within-role contributions are $D(w) = (1-w)D_0 + wD_1$ and $R(w) = (1-w)R_0 + wR_1$, with $D(w) + R(w) = r - t$. Using the same w in both roles gives

$$J_D(w) = D_{\text{included}}(w) - D_{\text{excluded}}(w), \qquad J_R(w) = R_{\text{included}}(w) - R_{\text{excluded}}(w).$$

These signed interaction contributions satisfy $J_D(w) + J_R(w) = J$.

The two dominance checks are

$$J_D(w) > |J_R(w)|, \qquad -R_{\text{excluded}}(w) > |D_{\text{excluded}}(w)|.$$

The first requires a positive direction contribution larger than the radius contribution in magnitude; the second requires a negative radius contribution larger in magnitude than the direction contribution. Every contribution is affine in w (a constant plus a term proportional to w). Each magnitude inequality is equivalent to two signed affine inequalities, so checking $w = 0$ and $w = 1$ establishes it for every common weight. The direct repair difference $a - b$ is independent of this credit allocation.

At cohort means, direction's role interaction ranges from 6.4011 to 7.9478 points, versus -0.1400 to 1.4067 for radius. Excluded direction contributions range from -0.9711 to $+0.0833$, versus -4.4289 to -3.3744 for radius. Direction dominates interaction magnitude and radius dominates excluded-loss magnitude throughout. The first property holds for all weights in $10/10$ people; both jointly hold in $8/10$. This is a bound over these common path mixtures, not all attribution schemes or uniquely identified causal shares.

Directly comparing repairs on ridge, radius replacement gives $a - r = 4.4289$ excluded points, versus $b - r = 0.9711$ for direction replacement. Their 3.4578-point difference is positive in $10/10$ people; the included difference is 9.9989 points, also $10/10$. These paired consequences require no attribution allocation. They are retrospective consequences of the original four arms, not independent replication or prospective policy regret. The general distinction between attribution importance and outcome-changing interventions already appears in recourse research (Mothilal et al., 2021); the present result concerns these matched retrieval maps and their subsequent repair test.

A subsequent control replaces querywise radii with one scalar per source fit, $s = \exp\{\mathbb{E}_{\text{fit}}[\log(\|h_T\|/\|h_R\|)]\}$, computed over all eight fitting people and their fitting concepts. Neither evaluation person nor outer task concepts enter this calibration. Applying $sh_R + \mu$ improves excluded ridge accuracy by 4.32 points and included accuracy by 4.61 points, both $10/10$. Its differences from the querywise hybrid are -0.11 and $+0.04$ points respectively, passing the prespecified -1-point mean and $8/10$-person preservation screens in both roles. Source scalars range from 1.52 to 2.08 across fits; no query-dependent scale or scale grid is used. Excluded normalized mean rank remains worse than T1 (0.04549 versus 0.03988). The rule was motivated by outcomes on this exposed TSconv cohort, although its coefficients use source data only. It supports a simple source-fixed account of the top-1 recovery within this contract, not a novel scaling operator or independent confirmation of generalization.

The next experiment froze the complete source-fitting, ridge-selection and scalar rule for the second preselected EEG architecture, STAE, before its new rule outcomes. All 450 contexts completed; the saved-rank scalar check had zero error. Table 2 gives all four arms. Excluded/included improvements over ridge are 4.2800/10.5811 points, each positive $10/10$; competence and source preservation pass. Scaled ridge gains over DCR are $-4.2678/+5.5100$ points respectively, with all ten signs agreeing in each role. Thus the prespecified gain over ridge recurs, while the planned excluded-role gain over DCR does not. Excluded normalized mean rank is 0.09085, worse than DCR 0.05526 and T1 0.05821. Source-selected $c = .1$ in 442 contexts and .01 in 8; scalars range from 1.32727 to 1.60123. The primary material-improvement screen requires mean gain at least 2 points and positive gain in at least 8 people. Source preservation permits at most 1 point loss in the mean and requires at least 8 people above that lower bound; excluded DCR competence requires 2.5% accuracy in at least 8 people.

Original STAE ridge loses to T1 in both roles in all ten people, so it does not reproduce TSconv's original preference reversal. After scaling, it wins on included people in $9/10$ and loses excluded in $10/10$. This latter reversal is a descriptive companion, not an additional prospective prediction. The original/scaled interactions are 3.9556/10.2567 points. The similar excluded correction means in TSconv/STAE (4.3156/4.2800 points) are not an established effect equivalence. In particular, partial recovery of a large initial deficit remains compatible with harm relative to fallback.

These studies use an exposed cohort and differ in checkpoint across a person's roles; they identify neither a causal membership effect nor independent-population scope. Source-effective ordinary maps, beneficial excluded-subject action on encoder-seen concepts, and unresolved joint-shift utility must remain separate. The fixed interventions deepen diagnosis and support source-derived recovery within this contract, without establishing a new method or predictive validity under independent-population or joint subject–concept shift.

A different-cohort radius prediction and functional control. A distinct UoB TSconv encoder was trained on 23 people (seed 42, 150 epochs, final checkpoint). On 11 other source people we fit maps, select ridge regularization and compute the same log-radius scalar, then evaluate 12 other people. This historically exposed cohort uses 240 stimulus–orientation identities shared across roles and 20-way category–orientation galleries. Each query averages its available cached neural vectors before normalization and source centering. A metadata-only amendment, frozen before new outcomes, retains all observed identities and all people: 2,629 source and 2,858 evaluation queries. People receive equal weight, with equal observed-query weights within person. This is a distinct task and instrument, not an exact THINGS replication or an unseen-concept test. Source selection gives ridge $c = 1$ and scalar 3.1493; the Procrustes reference uses its identified rank-239 partial-polar operator, without arbitrary null-space completion.

Mean evaluation accuracy is 7.276% for DCR, 7.730% for T1, 6.649% for ridge and 6.579% for scaled ridge. The fixed-scalar improvement prediction fails: -0.070 points, with five positive, five negative and two tied people. A subsequent prespecified control holds these maps and queries fixed and exchanges radii bidirectionally. Ridge direction with T1 radius gives 6.545% accuracy, a -0.103-point change from ridge (four positive, six negative, two tied). The reverse exchange gives 6.403%, a 1.327-point loss from T1 (eight positive losses, two negative, two tied). Neither passes its separate screen of mean effect at least 2 points and positive effect in at least ten people. All 11,432 original comparator ranks match exactly.

Rank sensitivity remains: the forward exchange improves normalized mean rank by .00932 in nine people, and the reverse worsens it by .02035 in all twelve. These companions do not replace the frozen top-1 prediction. Thus this task does not reproduce THINGS' top-1 radius recovery, even when the fixed scalar is replaced by query-specific exchange. Modest competence (best 7.73% against 5% chance), shared fitting, prior exposure and protocol differences limit broader mechanism or population conclusions.

Matched-role UoB test of the interaction/repair prediction. A separate frozen test uses the existing 46 leave-one-person-out encoders, one cyclically assigned encoder-included evaluator per route, and 44 fitting people excluding both evaluators. Two identity-disjoint halves each contain 120 stimulus–orientation identities; galleries contain ten candidates within category and orientation. Transport fits and evaluation identities are disjoint, but identities are encoder-seen and the cohort is historically exposed. Per-person centering uses only fitting-identity support. Ridge selection uses the original five-value source-only grid, and T1 uses the identified rank-119 partial-polar operator. Each physical person's observed queries are pooled across the two folds before equal-person aggregation; each role contains 10,965 observed query positions. This tests the same component-intervention estimands under a different task contract, not an exact full-rank, 200-way THINGS replication or an identified population cause.

Matched UoB arm	Included (%)	Excluded (%)
DCR	24.094	16.478
T1	20.941	16.597
Ridge	17.467	13.864
Ridge direction, T1 radius	19.699	14.858
T1 direction, ridge radius	18.589	15.075

The frozen direction-interaction/radius-deficit conjunction fails at the cohort mean and holds in $0/46$ people. Across the two allocation orders, direction's interaction term ranges from $+0.088$ to $+0.498$ points and radius's from -1.238 to -0.829; the total interaction is -0.741 points. Excluded radius replacement improves ridge by 0.994 points, but direction replacement improves it by 1.211: their difference is -0.217 points (21 positive, 23 negative, two tied). Included radius improvement is

2.233 points, retained as secondary evidence. DCR and T1 exceed 10% chance in all 46 excluded people; this is a valid failed prediction, not a no-signal instrument. The bounded THINGS finding is preserved without assuming its component ordering transfers.

C COMPLETE POSITIVE-SCALE CAPACITY: STAE AND OFFICEHOME

Exact positive-scale capability on the fixed STAE maps. The earlier STAE rule's failure does not establish that every positive scale fails. A subsequent discovery fixes its original 450 contexts, ridge fits, source scalars and galleries and considers $q_i(\alpha) = \mu + \alpha h_i$, $\alpha > 0$, for centered ridge output h_i. Use the legal domains $\mathcal{A}_i$ and $\mathcal{A}$ from Section 4. For unit candidates, query i with correct candidate index c_i wins at $\alpha \in \mathcal{A}_i$ exactly when

$$\alpha h_i^\top (y_{c_i} - y_j) + \mu^\top (y_{c_i} - y_j) \geq 0 \quad \text{for all } j \neq c_i,$$

with strict inequality for competitors $j < c_i$ in the stable tie order. The inequalities alone define an interval. Its intersection with $\mathcal{A}_i$ is the legal winning set I_i, which can be split by an excluded zero-output point. We compute represented float64 slope/intercept coefficients, then use exact rational differences and comparisons. A complete scan partitions the positive axis at all finite interval endpoints and zero-output scales. It checks open segments, valid endpoints and the unbounded segment. The shared maximum excludes every scale outside $\mathcal{A}$; querywise attainability tests $I_i \neq \varnothing$. Original ridge and source-scalar correctness match saved outcomes in all 360,000 comparisons across both roles.

Each context has 200 equally weighted queries. We use U_c, U_q and fitted-rule accuracy A_f exactly as defined in Section 4. The common optimum is separate for each map and role and uses evaluation labels; it is more privileged than the source-fitted scalar. Averaging the original 45 contexts per physical person in each role gives:

Fixed STAE capability	Included (%)	Excluded (%)
DCR	66.528	22.320
Original ridge	61.457	13.772
Source-fixed scalar	72.038	18.052
U_c	75.268	21.564
U_q	83.964	30.283

Excluded U_c falls below DCR by 0.756 points, with seven negative and three positive person differences. This misses the frozen material/stable deficit screen of 2 points and eight negative people. Nevertheless, $U_q - U_c = 8.719$ points and $U_q - \mathrm{DCR} = 7.963$ points are positive in all ten people. The sharing gap passes its separate 2-point/eight-person screen; included sharing loss is also 8.697 points, so this does not identify an excluded-specific mechanism. The original 4.268-point source-rule deficit comprises 3.512 points below the common oracle and a residual 0.756-point class deficit. Thus the finite-rule loss cannot be attributed wholly to shared-scale insufficiency. The query oracle uses the answer to identify a legal winning set: its advantage is available capacity, not a learned adaptive rule or evidence of source-only attainability. This STAE result alone supplies no new population or cross-application evidence.

Why zero-output exclusions leave the STAE bounds unchanged. The saved records contain 180,000 query intervals across 900 map/role panels. Every nonempty interval contains an open segment; none consists of one scale. In all 900 panels, the saved common optimum also holds on an open segment adjacent to its stored witness. Each affine path is nonzero at the fitted scale and can vanish at most once. Removing finitely many such scales from an open segment leaves legal points. Hence each previously reachable query remains reachable, even if scales must belong to $\mathcal{A}$, and each common optimum remains attainable. Exclusion cannot increase either capacity, so both reported values are unchanged. This argument uses the saved interval geometry; it does not assert that every STAE path is nonzero at every scale.

Cross-application capability test on OfficeHome. A subsequently frozen discovery applies the same nested capability functional to the original four OfficeHome maps and 45/10/10 class roles, without refitting or feature extraction. Here μ is the saved fit-text mean and $h_i = R^\top(\mathrm{norm}(x_i) -$

μ_x). Here x_i is the image embedding, μ_x the fitted mean of normalized image anchors, and R the fitted rank-44 partial-polar map in row-vector convention. This map is fixed for every scale. Candidates are the original ten unit text anchors. OfficeHome's original top-1 counts a correct co-maximizer, so all interval inequalities are closed, unlike STAE's stable-first tie rule. The saved minimum output norm over the tested positive rays is $0.8889207369817566 > 0$. Thus $\mathcal{A}_i = \mathcal{A} = (0, \infty)$ on these panels, and no scale is removed by the nonzero condition. Exact rational comparison conditions on the represented float64 score coefficients as above.

Each route's target panel has one held-out domain; its source panel contains the three remaining domains. Let D be the number of panel domains and $n_{d,c}$ the image count for domain d and class c. We preserve weights $w_i = 1/(D \cdot 10 \cdot n_{d,c})$ and require one common scale across the complete panel. Thus $U_c = \max_{\alpha \in \mathcal{A}} \sum_i w_i \mathbf{1}\{\alpha \in I_i\}$ and $U_q = \sum_i w_i \mathbf{1}\{I_i \neq \varnothing\}$. Fitted accuracy A_f is that of the previously recorded source-geometry action: T_1 in Art/Product/Real World and T_2 in Clipart. It is not the source-task selector, which already chose T_0 in all routes. Here T_0 is the unmodified normalized image embedding and generally lies outside the fixed-map ray.

Held-out domain	A_f	U_c	U_q	T_0
Art	52.975	70.343	82.015	89.879
Clipart	36.853	60.457	65.664	81.125
Product	66.838	81.182	91.846	96.263
Real World	60.393	75.373	88.031	98.016
Target macro	54.265	71.839	81.889	91.321
Source macro	32.747	56.650	67.060	89.073

Values are class/domain-balanced top-1 percentages. Across target routes, fitted-rule headroom averages 17.574 points and sharing loss 10.050 points. Each exceeds 2 points in 4/4 routes, passing the frozen mean/three-route rule. The error components are 18.111 points unreachable, 10.050 sharing loss and 17.574 headroom; they sum to 45.735 points of A_f error, not causal fractions. Querywise oracle scaling remains 9.432 points below T_0, with route deficits 4.417–15.461 points, passing the separately frozen reference-deficit test. Thus large calibration headroom coexists with insufficient capability even after removing the requirement to share a scale. This contrasts with the STAE query-oracle reference crossing without identifying its upstream cause.

The original OfficeHome fitted scale was bounded to $[.5, 2]$; the full positive envelope is a prospectively declared diagnostic extension, not an original deployment search. Within $[.5, 2]$, target common/query bounds are 63.170/65.201%, and the sharing gap is 2.031 points, exceeding 2 points in only 2/4 routes. The larger 10.050-point gap therefore belongs specifically to the full positive envelope. All 480 original class/action top-1 summaries match within 1.12×10^{-16}; weighted witnesses and error identities agree exactly. No per-image outcomes are exported. The four overlapping routes are descriptive cross-application evidence, not independent populations. Labels inform bounds only; no deployable query selector, new alignment operator or positive UoB component transfer follows.

C.1 What the capacity comparison identifies

This subsection states elementary implications used to interpret the empirical capacity results. It supplies no new oracle or general decision theorem. Fix an evaluated panel $D = \{(x_i, c_i, w_i)\}_{i=1}^n$, with nonnegative weights summing to one, query inputs x_i, correct indices c_i, a gallery, a tie rule, and a nonempty correction family $\mathcal{F}_0$ whose shared actions are defined on every query. Let $s_i(a) \in \{0, 1\}$ denote correctness of action a, $A_D(a) = \sum_i w_i s_i(a)$, and $A_{\text{ref}} = A_D(r)$ for reference r. Define shared-action capacity and its querywise relaxation by

$$C_D(\mathcal{F}_0) = \sup_{a \in \mathcal{F}_0} A_D(a), \qquad U_{q,D}(\mathcal{F}_0) = \sum_i w_i \sup_{a \in \mathcal{F}_0} s_i(a).$$

Every $a \in \mathcal{F}_0$ satisfies

$$A_D(a) \leq C_D(\mathcal{F}_0) \leq U_{q,D}(\mathcal{F}_0). \tag{5}$$

For the positive-scale family, shared actions use $\mathcal{A}$, giving $C_D(\mathcal{F}_0) = U_c$. Section 4 allows each query's scale to lie in its own domain $\mathcal{A}_i$. This can further relax legality, so $U_{q,D}(\mathcal{F}_0) \leq U_q$. A deficit in either querywise bound therefore excludes shared recovery. The main text's positive U_q

result licenses query-specific legal choices; it need not give choices valid for all other queries. In the reported STAE and OfficeHome panels, both domain conventions give the same querywise capacity (Appendix C). Fixed nonnegative averaging weights preserve these bounds.

Fitted failure versus family exclusion. For a fitted action $a_f \in \mathcal{F}_0$, an observed $A_D(a_f) < A_{\mathrm{ref}}$ rejects recovery by a_f. A finite nonempty evaluated subset $S_0 \subset \mathcal{F}_0$ provides only a lower bound $\max_{a \in S_0} A_D(a) \leq C_D(\mathcal{F}_0)$; without additional structure, its failure does not exclude untested members. In contrast, $U_{q,D}(\mathcal{F}_0) < A_{\mathrm{ref}}$ excludes reference recovery by every shared action and even by a separate choice for every query. The comparison concerns this panel and family; it says nothing about a different map, representation or population. Verification of a complete reachable set, rather than failure to find one action, has established precedents in recourse (Kothari et al., 2024).

Positive capacity and the quantifier boundary. If $U_{q,D} \geq A_{\mathrm{ref}}$, querywise recovery is not excluded. Because the panel is finite and success is binary, label-informed per-query choices can attain $U_{q,D}$. This does *not* imply $C_D \geq A_{\mathrm{ref}}$: exchanging a shared supremum for a sum of per-query suprema relaxes the requirement of one common action. Nor does it supply a selector learnable from legal source information or guaranteed to transfer to new queries. This is the familiar distinction between a per-instance oracle and a learned selector (Lindauer et al., 2019). Here, "retained capacity" means retained querywise possibility, not demonstrated shared-action sufficiency or a recommendation to continue searching.

What exclusion adds to fitted-action validation. If the reference already outperforms every evaluated family member, removing those members leaves the selected action unchanged under the same scoring and tie rules. When $U_q < A_{\mathrm{ref}}$, complete capacity additionally excludes recovery by untested family members. When calculated from the same scores and labels, this is an additional conclusion from those data, not an independent observation or a demonstrated search benefit.

Exclusion does not prescribe replacement. No inequality above orders an action in a different family $\mathcal{F}_1$ against r or against the best legal $\mathcal{F}_0$ action. Changing the family may be necessary for reference recovery yet still fail to achieve it. The capacity deficit therefore supplies neither an $\mathcal{F}_1$-success theorem, a search-cost guarantee, a source-to-target transport guarantee, nor a deployable repair policy. The empirical contribution in the main text is the reference-crossing contrast on the specified alignment paths.

D Supporting capacity forecasts and interim recurrence

D.1 Source-informed capacity-retention prediction

This test retains the 46 matched UoB routes described above, including the original encoders, source-selected regularization, ridge-map factors, DCR centering, observed-identity masks, two identity-disjoint halves and ten-way galleries. For each outer person, the cyclic encoder-included source validator and encoder-excluded target are both excluded from the 44-person map fit. The validator is not excluded from encoder training. Evaluator DCR calibration uses fitting identities only. The positive-scale family and first-index tie convention are identical on source and target. All common bounds optimize one scale per fixed map/panel; counts are pooled across the two folds within a route, then routes are averaged equally.

Forecast, amendment and outcome timing. Superscript s denotes source evaluation. Source surplus $V^s = U_q^s - A_{\mathrm{ref}}^s$ is compared with $B_c^s = U_c^s - A_{\mathrm{ref}}^s$ and fitted-action $B_f^s = A_f^s - A_{\mathrm{ref}}^s$. For each quantity, a value at least $+.02$ predicts retention, at most $-.02$ predicts insufficiency, and the remainder abstains. The source rule predicted retention everywhere, failing an initial qualification requiring nonconstant within-cohort predictions. That result was preserved. A subsequent source-informed amendment specified the one-sided target-capacity consequence before calculating that endpoint. It retained all source predictions, the 2-point margin, all 46 routes and all 92 maps; no source map was refitted or target-driven rule selected.

The predeclared primary was the equal-route mean target surplus. The 39 route disagreements and 75 map disagreements with common-scale forecasts were also declared before target capacity and

are reported in full, alongside the remaining units. The disagreement count is not a retrospectively substituted primary endpoint. Source predictions and this analysis plan were recorded before the new capacity calculation, but the cohort, representations and old action outcomes had been analyzed previously. The test has prospective ordering relative to this endpoint; it is not a sealed-cohort confirmation or evidence that prior related outcomes could not influence its design.

Fixed-path measurement	Source (%)	Target (%)
DCR reference	24.094	16.478
Fitted ridge	17.467	13.864
Common-scale capacity U_c	22.208	17.602
Querywise capacity U_q	39.261	32.840

Prediction consequences and constant control. The source common mean is 1.886 points below DCR, inside the abstention band; source query surplus is 15.167 points. Target query surplus is 16.362 points (range 9.205–22.917 across routes), exceeding the margin in all 46 routes. Target common surplus is only 1.124 points; it is a different endpoint.

Source forecast	Retain	Insufficient	Abstain	Matched target state
Fitted-action surplus	0	43	3	0/46
Common-scale surplus	7	24	15	7/46
Query-capacity surplus	46	0	0	46/46
Constant retention	46	0	0	46/46

All 24 negative-common and all 15 abstaining-common routes instead have materially retained target query capacity: 39/39 source-defined disagreements support the capacity forecast. Abstention is unresolved evidence, not an incorrect negative prediction. At map level, query forecasts match 92/92 target states, common forecasts 17/92 and fitted forecasts 1/92; all 75 disagreements support retention. Target map surpluses range 5.042–27.119 points. The maps share people and fitting data; even the LOSO routes overlap in training. These counts describe recurrence, not independent Bernoulli trials. Original DCR and ridge ranks match exactly in 21,930 source and 21,930 target comparisons; the original failed component prediction is unchanged.

The querywise forecast correctly anticipates the target capacity state where the common-scale forecast is negative or abstains. Constant retention ties, so neither within-cohort two-sided discrimination nor superiority to application-identity predictions is established. The ordinary surplus labels here are explicit capacity-forecast heuristics, not logically valid whole-family exclusion certificates or a demonstrated failure of deployment model selection. No learnable selector or guarantee that further search pays off follows.

D.2 PKU INTERIM CAPACITY EVALUATION

We evaluated the first 15 completed PKU participants using the same 46 UoB encoders and 92 cached maps per participant. Source capacity forecasts were fixed before any PKU samples or task outcomes were accessed; the authors report no prior PKU model training or selection. The original target plan specified 47 participants from the official EEG inclusion lists and their equal-participant mean capacity surplus. Two roster entries were ineligible before outcomes: CB036 has no EEG and CB082 is absent from the official inclusion list.

Two interim looks and sample inclusion. The time required for CPU preprocessing prompted an initial pause and a three-person analysis specified before opening those participants' retrieval or capacity outcomes. Four participants had completed preprocessing then; only the first three, CB001–CB003, were evaluated. After observing their results, we resumed preprocessing. CPU preprocessing cost and an interruption limited the second analysis to 15 completed participants. We included that entire completed prefix, fixing its membership before computing the additional 12 participants' outcomes. The earlier three results were reused unchanged; no participant was excluded on the basis of a capacity outcome. This is an interim, availability-limited sample of the original roster, not a new 47-person confirmatory plan or a randomly selected 15-person cohort.

Every participant uses all 46 models, the original two identity-disjoint fitting/evaluation halves, ten-way galleries, positive-scale paths and 2-point state margin. DCR calibration uses the participant's fitting identities; there is no PKU encoder training, source-map refitting or target-based model selection. We pool observed evaluation-query counts across the two folds within each route, average routes within a participant, then average participants equally. The same rule separately summarizes the initial three and the newly evaluated twelve; the latter is a descriptive extension check, not a replacement primary endpoint.

PKU participant	DCR (%)	U_c (%)	U_q (%)	$U_q - \mathrm{DCR}$ (pp)
CB001	17.237	16.336	30.535	+13.298
CB002	16.585	15.716	33.822	+17.237
CB003	16.205	17.364	33.967	+17.763
CB006	12.970	17.832	33.168	+20.198
CB008	15.336	18.092	34.091	+18.756
CB009	18.375	17.618	31.979	+13.604
CB011	10.981	14.588	29.408	+18.427
CB012	10.943	16.131	29.585	+18.643
CB013	18.379	18.678	34.040	+15.661
CB015	20.899	19.583	35.047	+14.149
CB016	14.408	17.755	35.137	+20.729
CB019	16.182	17.500	33.400	+17.218
CB020	15.372	15.681	30.453	+15.081
CB022	10.615	14.262	30.098	+19.483
CB023	16.676	18.859	33.804	+17.129
Initial 3 mean	16.676	16.472	32.775	+16.099
New 12 mean	15.094	17.215	32.518	+17.423
All 15 mean	15.411	17.066	32.569	+17.158

Capacity and reference consequences. All 15 participants exceed the 2-point query-capacity margin, with mean surpluses ranging from 13.298 to 20.729 points. All 690 routes retain capacity; the smallest route surplus is 8.597 points. In each participant, all 39 source-defined query/common disagreements support retention: 24 routes had source-common insufficiency and 15 had abstention. Abstention is not an incorrect negative prediction. Constant retention ties on every route. These are 15 target participants conditional on a shared source model bank, not 690 independent route replications or 585 independent disagreement tests.

The fixed ridge action averages 13.803%, below DCR's 15.411%, whereas U_c averages 17.066%. Common capacity exceeds DCR by 1.656 points overall (11 participants positive, four negative); the new twelve's mean surplus is 2.120 points. Thus PKU does not establish common-scale incapacity or a need for query adaptation. The new twelve's U_q is similar to the initial three's (32.518 versus 32.775%); their larger query surplus accompanies a lower DCR reference, not improved model performance. Every participant's route-mean DCR exceeds the nominal 10% chance level, but three are only 10.615–10.981%. At route level, 23/690 DCR accuracies are at or below 10% (minimum 7.950%). These are descriptive competence checks, not a claim that every route is significantly above chance; all routes are retained.

The results support the original positive-state forecast at a new site, including the twelve newly evaluated participants. Prediction-before-access ordering is preserved, but continuation followed the first interim result and the second sample size reflects preprocessing availability and cost. The planned 47-participant confirmation is incomplete; the remaining 32 capacity outcomes have not been inspected. We make no sequentially calibrated significance claim from these interim looks. Neither superiority to constant retention, a two-sided predictive rule, a deployable query selector, nor extension of the component-ordering claim follows.

E SUPPORTING SCORE-TO-ORDER TESTS

E.1 SCORE-TO-ORDER DIAGNOSIS AND PREDICTION TESTS

Sections 3 and 4 evaluate repair with top-1 accuracy. A separate THINGS source study explains why changes in similarity scores must also be checked against candidate order. On fixed 200-way retrieval panels, it compares Procrustes T_1 with T_2, which uses the same map and centering means but adds a positive scale fitted to reduce squared distances between paired source representations. The scale multiplies the centered output before the mean is added back; T_1 uses scale one. Each metric selects the correction with its best observed value; only exact ties favor the simpler correction. Angular scoring averages cosine similarity to the correct candidate and favors T_2. Normalized rank utility rescales the correct candidate's rank to $[0, 1]$, with higher values better; it favors T_1. Both preferences hold in all ten people.

To separate score changes from order changes, let s_{ij} be the cosine score of query i against candidate j in a gallery of $K > 1$ candidates, and let c_i index its correct candidate, whose score is s_{ic_i}. Its mean competitor score is $\bar{s}_{i,\mathrm{comp}} = (K-1)^{-1} \sum_{j \neq c_i} s_{ij}$, and its mean correct-minus-competitor margin is $M_i = s_{ic_i} - \bar{s}_{i,\mathrm{comp}}$. For this T_1–T_2 comparison, Δ denotes T_2 minus T_1. A larger correct score need not increase M_i: the competitor scores can rise by more.

Order depends on which competitors outrank the answer. A violation means that a competitor scores higher, or ties and precedes the answer in the fixed candidate order. Let R_i be one plus the number of such violations, so rank one is best. Define normalized rank utility $U_i = 1 - (R_i - 1)/(K - 1)$, which ranges from zero to one and increases as rank improves. For $T_1 \to T_2$, let C_i^+ count violations removed and C_i^- count violations introduced. These superscripts label favorable and adverse crossings. The denominator $K - 1$ is the number of incorrect candidates. Using the same gallery and tie rule at both endpoints gives

$$\Delta M_i = \Delta s_{ic_i} - \Delta \bar{s}_{i,\mathrm{comp}}, \qquad \Delta U_i = \frac{C_i^+ - C_i^-}{K-1}. \tag{6}$$

These standard identities distinguish the average score margin from the net change in candidate order. U_i measures rank utility, not top-1 accuracy or the scale-family capacities U_c, U_q of Section 4.

Decomposing the already observed T_1–T_2 difference makes this distinction measurable. Correct scores rise by 0.08325, but mean competitor scores rise by 0.17784, giving a mean-margin change of -0.09459. Favorable and adverse crossing rates are 0.005026 and 0.007352 per incorrect candidate pair; their difference recovers the -0.002326 change in normalized rank utility. For 200 candidates, this corresponds to a mean worsening of about 0.463 rank positions. The summaries average queries within each context, then contexts within each person and people equally. All ten people exhibit the correct-score rise, larger competitor rise, margin decline and net adverse order. The rates account for the observed rank change; mean-margin erosion alone does not imply it.

A prespecified test examines intermediate scales between T_1 and T_2. Let $\hat{\alpha} > 0$ be the source-fitted scale of T_2, and define $\alpha(\lambda) = 1 + \lambda(\hat{\alpha} - 1)$, where $\lambda \in [0, 1]$ moves from T_1 at zero to T_2 at one. At the three prespecified interior points $\lambda \in \{0.25, 0.5, 0.75\}$, mean margin deteriorates, yet normalized rank utility averaged over those points improves slightly in all ten people. The frozen prediction that this score pattern would accompany net rank harm on average over the interior therefore fails. The adverse cohort-mean rank response occurs at the fitted endpoint on this grid; individual adverse crossings can occur earlier.

The next test asks where adverse crossings accumulate. It uses the remaining 654 concepts outside the earlier task panels. Their representation agreement had already been examined, but their retrieval outcomes were withheld when the prediction was fixed. Baseline margin reserve is the correct-minus-competitor score difference for one query–competitor pair at T_1. Within each context, pairs with strictly positive reserve are sorted into three groups of approximately equal size: low, middle and high reserve. Ties and already violating pairs are excluded from the groups, and group membership is fixed before later path scores are evaluated. Crossing burden is the fraction of each group that becomes a violation, averaged over the same three interior points. It is a sampled mean rate, not a continuous area under a curve.

All ten people have greater crossing burden in the low-reserve than the high-reserve group. The middle and high groups both have zero observed burden, so the result does not support a strict three-

group ordering. This comparison is empirical: baseline reserve alone does not fix the size of each pair's subsequent margin change. It supports a prediction on a withheld task panel within the same source population. Together, these studies support checking actual candidate order when interpreting score improvement; they do not establish a common mechanism for the two repair headlines. The following subsections report the full path outcomes and the reserve prediction, including the failed stronger predictions.

E.2 Representation recurrence and the failed interior prediction

The representation prediction and scale path were frozen before their outcomes. Both preselected representations satisfy the full four-direction conjunction in every physical subject. The scale-path test instead fails its mean-interior conjunction in every subject: continuous score and mean-margin directions persist, but the net rank direction is favorable on average over the interior. Adverse crossings need not be absent when net rank improves. In Table 3, s^+ and s^- abbreviate the panel-mean correct and competitor scores. Their differences use the same averaging; ΔM and ΔU follow Equation (6). The table reports person-level aggregates; contexts and representations are not treated as population replicates.

Table 3: Prespecified representation recurrence and controlled-path boundary. All changes are from the scale-one baseline.

Condition	Δs^+	$\Delta(s^- - s^+)$	ΔM	ΔU	Directional subjects
STAE	.083775	.082578	-.082578	-.010095	10/10
TSconv	.076767	.090840	-.090840	-.002958	10/10
Path $\lambda = .25$	.023492	.015794	-.015794	.000461	0/10 adverse rank
Path $\lambda = .50$	.046643	.035990	-.035990	.000684	0/10 adverse rank
Path $\lambda = .75$	.067564	.061819	-.061819	.000210	4/10 adverse rank
Path $\lambda = 1$	.083246	.094594	-.094594	-.002326	10/10 adverse rank

E.3 Reserve prediction on withheld task outcomes

NEAR, MID and FAR denote the low-, middle- and high-reserve groups, respectively. The reserve study uses the deterministic complement of the source-task panels, containing $131/131/131/131/130$ concepts across the five folds. Candidate-relative task outcomes for this complement were withheld before the prediction was frozen. All 450 contexts are competent, and inference remains at the ten physical subjects. Table 4 reports the prespecified reserve contrast and its descriptive controls.

The risk set includes only strictly positive baseline margins; ties and already violating pairs are excluded. Within each context, pairs are ordered by (margin, query index, candidate index), then split into three balanced groups. A competent context requires at least 300 risk pairs; a competent subject requires at least 40 of 45 contexts. Contexts receive equal weight, not weight proportional to their pair count. The primary branch requires all subjects competent, at least nine with erosion in every band, at least nine with positive NEAR-minus-FAR contrast, and a positive lower endpoint of the 10,000-draw subject bootstrap interval. A fixed hash of pair indices supplies a control partition of matched sizes without using margins. This partition, the MID band and endpoint are descriptive and cannot rescue the primary prediction. Each path-point crossing rate records violations at that point relative to the baseline; it does not count an unobserved crossing and return between sampled points. This support rule tests the availability of pairs, not above-chance retrieval. No practical-action margin is imposed on the crossing contrast.

Table 4: Prospective margin-reserve prediction. Sampled mean crossing rates and margin changes are first aggregated within context and then within physical subject.

Quantity	NEAR	MID	FAR
Interior crossing rate	.005102	0	0
Interior mean margin change	-.017158	-.038019	-.061720
Endpoint crossing rate	.019587	0	0

The primary NEAR-minus-FAR contrast is positive in $10/10$ subjects, with mean .005102, 95% bootstrap interval $[.004195, .005848]$, and nominal one-sided sign $p = .0009766$ conditional on independent null signs. Shared fitting prevents treating this as calibrated population significance. The fixed size-matched hash contrast has mean $-.0000213$, bootstrap interval $[-.0000715, .0000261]$, with five positive and five negative person contrasts. The descriptive strict three-band order is not supported $(0/10)$, because MID and FAR are tied at zero. This distinction is important: the reserve test supports low-reserve localization, not a smooth monotone tertile law.

F RETRIEVAL BOUNDARIES UNDER DIFFERENT INFORMATION CONDITIONS

Joint person and concept holdout in THINGS. The separate terminal THINGS test excludes both the target person and 200 target concepts from encoder training and selection. It reuses the source-only TSconv training recipe in Appendix A and fixes the map and scale before target retrieval outcomes. Its one unlabeled centroid uses the full target query panel, so the evaluation is transductive. It does not establish streaming adaptation without that panel.

The geometry-selected correction T_2 worsens mean rank relative to DCR by 1.2190 positions, with improvement in only one of ten people. Source rank validation had selected T_1 over T_2. This result therefore tests the geometry-selected action; it does not identify a failure to transfer an action chosen by source-task validation. Nor does it extend the matched radius gains to concepts unseen by the encoder.

Original UoB retrieval favors centering. A separate 46-person LOSO study uses four category-specific 20-way identity galleries and two states: task-irrelevant trials and task-relevant non-target trials. All six action/state combinations exceed chance in normalized rank utility: $U_i = 1 - (R_i - 1)/(K - 1)$ for answer rank R_i and $K = 20$ candidates, with U its panel mean. Their mean excess over chance $(U - .5)$ ranges from .048820 to .059930. Nevertheless, both corrections underperform DCR in each state:

Utility difference	Task irrelevant	Task relevant, non-target
$T_1 - \mathrm{DCR}$	$-.008111$	$-.005206$
$T_2 - \mathrm{DCR}$	$-.004063$	$-.002167$
$T_2 - T_1$	$+.004048$	$+.003039$

The last row shows that the source-fitted scale improves on Procrustes without restoring the centering reference. Its gains remain below the prespecified normalized-utility margin $2/199$. This numerical margin was inherited from the original THINGS screen; it is not two rank positions in the 20-way UoB galleries. All UoB scales reach the fitted upper bound 2, whereas the THINGS source scales lie in $[.5, .512398]$. These are different representations, galleries and information conditions, so the comparison identifies neither a causal scale-direction effect nor a population cause. The matched 10-way UoB component and capacity studies retain their separate protocols (Appendices B and D).