SEPARATING READOUT CAPACITY, FEATURE WIDTH, AND LOW-LEVEL STIMULUS STRUCTURE IN MULTIMODAL FMRI ENCODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fMRI encoders outperform unimodal ones, and the advantage is usually credited to the representation. We show that the measured multimodal gain depends on how the linear readout is restricted, not only on how much: below the free readout's operating range, restricting by ridge penalty, by random projection, or by projection structure changes the gain by nearly a factor of two at approximately matched effective dimension, and the direction depends on the restriction. Holding one frozen feature set fixed, a concatenation of three unimodal backbones, we vary only the readout, from a free ridge down to two trainable parameters per parcel, across three low-level confound arms, and measure the fusion gain at every rung. Matching width leaves the gain within 3.5% of the free readout's. In distribution, at approximately matched effective dimension, a ridge penalty retains 80% of it where a random projection retains 44%, or 50% with inputs standardized before projection. At a fixed $1{,}536$-column budget a modality-wise projection retains 86% where a joint one retains 61%; standardizing the inputs narrows the gap to 92% against 85% but does not close it, while lowering every absolute score. The penalty ladder's gain sits below the free readout's through $1{,}024$ effective dimensions and above it at $1{,}536$ and $2{,}048$. The gain is smallest in visual cortex ($+0.0095$); its inversion there under a 512-column joint projection (-0.0098) is gone at $1{,}536$ columns, a budget effect rather than a regional one. A second feature set sharing two of three backbones reproduces all four registered comparisons. A reported fusion gain is therefore partly a property of the readout that measured it.

1 INTRODUCTION

Multimodal foundation models set the state of the art on fMRI encoding benchmarks, and the standard account is representational: fusing vision, audio and language yields features that predict cortex better than any single stream. That account has not been separated from a mundane alternative. Concatenating modalities widens the design matrix, and a wider design gives a linear readout more freedom whether or not the added columns carry new information.

Three things are therefore entangled in a reported fusion gain: how much of it the readout could manufacture from the extra *width* of a concatenated design; how much depends on the readout's *capacity*; and how much of the underlying score is carried by *low-level stimulus structure* (motion energy, sound level, word rate) that every modality's features encode and that a bank of explicit covariates predicts on its own. This paper separates the three on one frozen feature set. The width account fails at once: projecting the trimodal design down to the narrowest unimodal width leaves the gain within 3.5% of the free readout's (§4). Readout dependence does not fail, and it is most visible in visual cortex, where a restricted readout turns the trimodal advantage into a deficit, the trimodal model scoring *below* the best unimodal one (the *visual-cortex inversion* of §4.5).

Published work controls this in part. Gokce et al. (2026) run their native-versus-post-hoc fusion contrast at more than one readout level; Saha et al. (2024) vary the readout mechanism on fixed backbones; Dupré la Tour et al. (2022) give each feature space its own regularization (our banded rung); Conwell et al. (2024) hold the mapping fixed across roughly 1.8 billion regressions while

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.

model properties vary. **Varying the readout at fixed features is therefore not new, and we do not claim it.**

What we are not aware of is work reporting the multimodal gain *across a graded capacity ladder* extending *below* a free readout (down to two trainable parameters per parcel) with the gain measured at every rung. Saha et al.'s readouts differ in functional form rather than in graded capacity within one form, and both of Gokce et al.'s readout points sit at or above a free ridge, the regime in which our own absorption is approximately zero; the suppression we report appears only below it (§M). That regime is where reduced-feature and heavily regularized pipelines operate, and the probing literature requires that probe capacity be controlled before structure is credited to the representation (Hewitt & Liang, 2019); fusion gains reported at one or two readout points have had no such control.

1.1 Contributions

1. **The measured multimodal gain is not a property of the representation alone.** Below the free readout's operating range, restricting the readout suppresses it: at approximately matched effective dimension a ridge penalty retains 80% of the free readout's gain where a random projection retains 44% (50% with standardized inputs), and at a fixed $1{,}536$-column budget a modality-wise projection retains 86% where a joint one retains 61%, a gap that standardizing the inputs narrows but does not close. Near the free ridge's own operating points the penalty ladder inflates the gain above the free readout's.
2. **The instruments that show it.** A graded readout-capacity ladder over one frozen feature set, measured by frozen random projections and by a fixed-feature ridge penalty at approximately matched effective dimension, paired over validation episodes; and projection controls, joint against modality-wise and raw against standardized inputs, that locate the visual-cortex inversion in the compression budget rather than in the region.
3. **Scope and diagnostics.** Common-width projection preserves the contrast (absorption within ± 0.035); suppression through $1{,}024$ effective dimensions, inflation at $1{,}536$ and $2{,}048$, and projection-structure sensitivity are three distinct outcomes; visual cortex crosses zero under projection because its baseline gain is smallest, not because its loss is larger.

1.2 Scope

We do not claim competitive absolute accuracy for the model we decompose. The encoder every cell uses ($D = 5{,}504$) scores 0.2384 in distribution and 0.1174 far-OOD officially, against 0.310 / 0.217 for MIRAGE's single model and 0.213 far-OOD for TRIBE (d'Ascoli et al., 2025); our best official standing (0.294 / 0.182, ranks 8 and 9) is a 45-member ensemble of a different configuration. The contribution is an account of *where* gains come from; whether it transfers to stronger systems is untested.

Two feature sets sharing two of three backbones, four subjects. Every registered claim rests on the first set; the second enters through the registered replication (Appendix D); every bootstrap interval conditions on the same four participants, and the only intervals across participants are the four-subject t-intervals quoted for R3 in §4, indicative at $n = 4$.

The mechanism is not established. We do not know why a penalty and a projection at approximately the same effective dimension read different gains from the same features, nor why joint compression inverts the visual-cortex gain where modality-wise compression does not; §5 names candidates, and the controls of Appendix E constrain them without identifying the mechanism.

2 Related work

Our ladder is the probing literature's capacity control (Hewitt & Liang, 2019) applied to encoding rather than classification and extended below the free-readout regime. That low-level structure carries a substantial share of an alignment score is an established genre: Hadidi et al. (2026) show that positional signals and word rate compete with trained language models and fully account for untrained ones, and Oota et al. (2024), the closest prior art, ablate a model's representation where we fit an explicit stimulus bank in competition with the encoder. We do not claim that finding; what we add is how the fusion gain's sensitivity to readout restriction interacts with confound handling. The

Algonauts 2025 lineage (Gifford et al., 2025; d'Ascoli et al., 2025; Gokce et al., 2026) ablates the readout only at or above a free ridge; we found no graded readout-capacity ladder in fMRI encoding.

3 METHODS

Data and splits. Four subjects viewing naturalistic film (Boyle et al., 2023; Gifford et al., 2025), parcellated to $1{,}000$ cortical parcels. Stimuli are partitioned *at the episode level* into 242 training, 50 validation and 17 near-OOD stimuli, with no stimulus in two partitions. Training and validation stimuli are half-episodes of the sitcom *Friends* (seasons one to five, and season six, respectively); the near-OOD stimuli are the 17 segments of two feature films (*Hidden Figures* and *Life*) viewed by the same participants in the same scanner, so the shift is in content and genre. The officially scored far-OOD films (§1) are a separate, sealed set. Delay stacking is segment-local.

The frozen representation. Every cell in this paper consumes one feature set, identical byte for byte, and no gradient reaches it. Features come from three pretrained backbones under permissive licences, taking the *final hidden state* of each and pooling over tokens or frames rather than over layers: `vjepa2-vitg-fpc64-256` for video (1,408), `w2v-bert-2.0` for audio (1,024) and `Llama-3.2-3B` for text (3,072), concatenated to $D = 5{,}504$. Pooling is a plain mean in every stream: video is encoded in clips of eight frames and each clip's tokens are averaged; audio is encoded in five-second windows and its frame states are kept; the words within a TR are encoded together and their tokens averaged. Every stream is then averaged within each 1.49 s TR and delay-stacked at one to four TRs, so the trimodal design is $5{,}504 \times 4 = 22{,}016$ columns and the unimodal designs are the corresponding sub-blocks. The root is verified against a SHA-256 manifest before every fit.

Reliability gate. Analysis is restricted to the 834 of $1{,}000$ parcels whose split-half reliability, with James–Stein shrinkage (James & Stein, 1961), clears a floor of 0.10. The estimate uses the only repeats the data contain, the two near-OOD films viewed twice by each of the four participants (8 pairs); no training or validation stimulus was repeated. Parcel selection is therefore in-sample for the near-OOD split, and the same mask is applied to every readout. §N varies the threshold down to no gate at all: ungated, R5-513 retains 82% of R1's gain against R4-512's 44% in distribution and 97% against 87% near-OOD, against 80%, 44%, 96% and 86% at the frozen threshold.

Metric. Per subject, held-out TRs are concatenated across stimuli, one Pearson r is computed per parcel, averaged over gated parcels, then across subjects.

The readout ladder. All rungs consume the same frozen features and differ only in what the readout may learn (Table 1). R3 projects each condition to a fixed dimensionality in both pre-registered directions: *down* to the narrowest unimodal width (1,024) and *up* to the trimodal width (5,504). The projector is the basis of the top D_{match} right-singular vectors of the standardized training design, fitted on training stimuli only (an economy SVD, not whitened); a unimodal design narrower than D_{match} is tiled to that width before the basis is fitted (Appendix A). Features are standardized once, *before* projection; the projected coordinates are left on their natural scale. Ridge is invariant to an orthogonal rotation only if the penalty remains isotropic in the original space, so re-standardizing the projected coordinates converts the same α into a different prior; Appendix A shows this cost 0.056 in held-out r on a rotation that discards no information. Frozen projections are dense matrices with i.i.d. standard-normal entries scaled by $1/\sqrt{\text{input width}}$, drawn by a PCG64 generator whose seed derives from SHA-256 hashes committed to a manifest of 750 matrices *before any cell ran*, and verified at load; they act on the raw feature columns (Appendix E standardizes them as a control). R4-k is ordinary least squares on its k projected columns plus an intercept (10^{-8} relative ridge for conditioning only), so its effective dimension is k up to the effect of that ridge (bounded, not measured, in §M, with the one-unit offset to R5's count). For R1, R2 and R3 the penalty is selected per parcel (R2 jointly with Dirichlet-sampled band weights) from 21 values, 10^0 to 10^{10} in log spacing, by five-fold inner cross-validation whose folds are whole training segments, never TRs (two segments of one episode may fall in different inner folds, which affects penalty selection only; outer evaluation remains on disjoint episodes), widening by two decades if more than 2% of gated parcels pin at an edge; a seed varies the fold assignment for the ridge readouts and the projection for R4; R5 sets α by construction (§M).

Table 1: The readout ladder. All rungs consume identical frozen features. R2 follows the banded-ridge construction of Nunez-Elizalde et al. (2019).

readout	trainable parameters per parcel	what it holds fixed
R1 free ridge	full design width (4,096–22,016)	nothing, the reference
R2 banded ridge	per-modality bands, Dirichlet-searched	modality-wise variance attribution
R3 capacity-matched	projection to fixed D_{match}	representation width
R4-k graded frozen	$k+1$, $k \in \{8, 32, 128, 512\}$	accessible random-feature dimension
R4 frozen	2 (scale and bias)	expressive freedom, maximally
R5 fixed-feature	penalty solved to 2–513 effective dim.	the design itself, unreduced

Confound handling. Three pre-registered arms, `excluded`, `partialled` (the bank regressed out of features *and* fMRI, coefficients fitted on train only) and `competing band` (the bank enters as its own band), plus a bank-only control. The low-level stimulus-control bank is 49 raw columns delay-stacked with the same four delays, giving 196 regressors (§F lists them). All three arms run for R1, R2, R3, R4 and R5; for R3 the bank joins *after* capacity matching, so the projector compresses features only.

Named quantities. Fusion gain is $r(\text{VAT}) - \max(r(\text{V}), r(\text{A}), r(\text{T}))$, computed within a seed. The absorption index is $1 - (\text{gain under a constrained readout})/(\text{gain under R1})$: 1.0 means the advantage vanishes when the readout is constrained, 0.0 that it survives; it is uninformative where either gain is near zero or changes sign, and raw gains are reported there. The *retention factor* $\lambda(k) = r_{\text{VAT}}(k)/r_{\text{VAT}}(\text{R1})$ is the fraction of the free ridge's trimodal score a rung keeps; if every condition shrank by this common factor the gain would too and absorption would equal $1 - \lambda$, so we call the difference between observed absorption and $1 - \lambda$ the *departure from common scaling*, a diagnostic under an assumed null rather than a variance decomposition. Excess uses a mean-of-unimodals baseline; a second definition of λ (mean unimodal retention) is a robustness check. Uncertainty is a cluster bootstrap over stimulus segments (Friends half-episodes, film segments), never TRs, 1,000 resamples, BCa (Efron, 1987), computed *per cell* and unpaired across the conditions of a contrast, because the frozen readouts did not retain per-segment statistics; R1 and R4-512 were refitted with them retained (§O). Segments of one episode or one film are not independent: the validation intervals behind the main claims are re-clustered at the episode level (23 episodes) and quoted that way, the remaining intervals resample segments and are labelled so, and, since the 17 near-OOD segments come from two films, §O reports the near-OOD contrast per film and offers no near-OOD interval as evidence of generalization across films.

4 Results

4.1 Common-width projection preserves the fusion contrast

Figure 1 and Table 2 carry the paper. Both pre-registered matching directions leave absorption within ± 0.035 of zero *in the* `excluded` *arm*; the bound is specific to that arm. A trimodal model squeezed into 1,024 dimensions (the width of the narrowest unimodal condition) retains the contrast, while its absolute score falls by 0.006–0.008. Raw design width alone therefore does not explain the fusion contrast. The converse control agrees: padding V to the trimodal width with uninformative columns doubles effective dimension yet lowers in-distribution r by 0.011 and leaves near-OOD unchanged (§P).

4.2 Restricting the readout changes the measured gain

Under frozen projections the measured gain behaves differently. Near-OOD absorption falls as trainable capacity rises from 2 to 513 parameters per parcel, from 1.133 to 0.136 under the max convention of Table 2 and from 0.843 to 0.148 under the mean-of-unimodals baseline that Figure 1 plots (Table 9 gives both); in distribution it falls too, non-monotonically. What the measurement *means* is decided by a decomposition, given after establishing which rungs measure and which only bound.

Where the ladder measures, and where it only bounds. A rung estimates the advantage's size only where the gain exceeds its own seed spread. Out of distribution this fails at one rung, nine

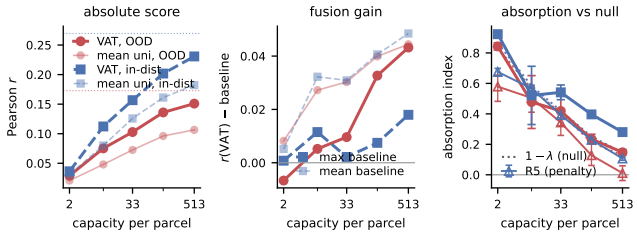


Figure 1: **Absolute scores shrink together, so the gain must be read against a null.** `Excluded` arm, subject-means over three seeds. **Left:** absolute score for the trimodal condition and the mean unimodal, with free-ridge values as dotted references. **Centre:** fusion gain under the pre-registered max baseline (solid) and the mean-unimodal baseline (faint). **Right:** observed absorption against $1-\lambda$, the value implied if every condition shrank by a common factor. Error bars: for the projection ladder, the seed standard deviation of the gain propagated onto this axis; for the fixed-feature ladder, paired 95% BCa intervals on the observed absorption itself (not on its excess over $1-\lambda$), over the 23 validation episodes in distribution and the 17 near-OOD segments, from a retained script. Open triangles are the fixed-feature ladder (horizontal axis: effective dimension; for the projection ladder, trainable parameters per parcel). Neither is variation across participants. The null is itself an estimate, so departure from it is judged by the paired interval on the *difference* (§H), not by whether a bar clears the dotted line. The projection ladder's near-OOD points track the null; its in-distribution points sit above it at the three largest rungs, and the fixed-feature series departs from the null near-OOD at the lowest measurable rung. Mean baseline where compared with the null.

Table 2: The capacity ladder measured by both instruments, `excluded` arm, subject-mean, max-over-unimodals baseline. Absorption is defined against R1 within the same arm. R4 reaches a rung by random projection with $k+1$ trainable parameters per parcel, R5 by a ridge penalty on the unreduced design solved to approximately the same effective dimension (R4's count is its column count; R5's excludes the intercept, §M).

readout	effective dim.	val VAT r	val gain	OOD gain	absorption val	absorption OOD
projection ladder: capacity removed by random projection						
R4 frozen	2	0.0364	+0.0008	−0.0067	0.981	1.133
R4-8	9	0.1123	+0.0115	+0.0052	0.719	0.896
R4-32	33	0.1567	+0.0022	+0.0096	0.947	0.808
R4-128	129	0.2017	+0.0075	+0.0327	0.818	0.346
R4-512	513	0.2307	+0.0180	+0.0432	0.561	0.136
penalty ladder: same design, capacity removed by regularization						
R5	2	0.1626	−0.0062	+0.0101	1.152	0.798
R5	9	0.1768	+0.0010	+0.0140	0.976	0.720
R5	33	0.2029	+0.0119	+0.0231	0.710	0.538
R5	129	0.2345	+0.0225	+0.0375	0.451	0.250
R5	513	0.2588	+0.0327	+0.0482	0.202	0.036
R3-up (5,504)	matched	0.2636	+0.0423	+0.0501	−0.034	−0.003
R3-down (1,024)	matched	0.2617	+0.0407	+0.0484	+0.007	+0.032
R1 free	full	0.2695	+0.0410	+0.0500		

parameters, where the gain of +0.0052 is smaller than its seed spread of 0.0111; in distribution, the noisier column, it fails at nine and 33. Below those the ladder is a scale bar rather than a set of estimates (Appendix H applies the criterion to the excess). **From 33 parameters upward the OOD gains are strictly ordered across all three projection seeds** (§H); below it the ordering is monotone in the seed mean only, one seed of three inverts the two smallest rungs, and the in-distribution non-monotonicity at the same pair of rungs is the same fact, gain below noise.

What the frozen ladder does and does not isolate, and what the missing control shows. Changing k changes two things at once, the number of trainable coefficients per parcel and the dimension and content of the accessible random subspace; R3 rules out a width artifact at 1,024 dimensions but says nothing about what a projection to 8 or 32 retains. Separating the two needs a ladder holding the feature matrix fixed and varying only the ridge effective degrees of freedom. We ran it (R5; §M), and it does not vindicate the frozen ladder: at approximately matched effective dimension,

restricting by penalty alone retains far more of the trimodal advantage than restricting by random projection, 0.0327 against R4-512's 0.0180 in distribution at 513 effective dimensions (80% versus 44% of R1's gain; paired difference $+0.0147$ $[+0.0136, +0.0158]$, resampling the 23 validation episodes) and 0.0482 against 0.0432 near-OOD (96% versus 86%; $+0.0050$ $[+0.0001, +0.0088]$ over segments, retained fractions not separated; §O), with R5 higher at every near-OOD rung and in all six arm $\times$ split cells. The two readouts also differ in input scaling, since R4 projects raw columns; with inputs standardized before projection, R4-512 retains 50% of R1's gain and the paired difference to R5-513 is $+0.0122$ $[+0.0114, +0.0132]$ in distribution and $+0.0157$ $[+0.0051, +0.0215]$ near-OOD, so what remains between the instruments is the subspace and the regularization form (§O). This is the outcome Amendment 8 named as damaging in advance (§M): *the frozen ladder overstates the effect of restricting the readout*.

Decomposing the recovery. On the projection ladder absorption must approach 1 as k approaches zero (a readout with no columns predicts a constant, scored $r = 0$); on the penalty ladder it need not, because as $\alpha \to \infty$ the held-out predictions shrink along the fixed direction $X_{\text{test}}X^\top y$ and Pearson r, which ignores rescaling, tends to that direction's correlation with y_{test}. Whether either ladder tracks anything a capacity-driven account does not give is answered by the common-scaling null of §3, measured rather than assumed because the unimodal conditions were run at every rung (Figure 1). Three statements follow.

Out of distribution, uniform shrinkage accounts for most of the recovery. Every condition recovers at nearly the same rate (at 513 parameters the trimodal score retains 87.3% and the mean unimodal 88.2%), and shrinkage alone accounts for 86% of the absorption at that rung. The residual is *small but not zero*, $+0.021 \pm 0.012$ at 513 and positive across all three projection seeds (computational replicates); at 129 it is $+0.022 \pm 0.030$, not sign-consistent. The fixed-feature ladder revises it below.

In distribution, an excess survives, on both ladders. On the projection ladder it is $+0.136 \pm 0.013$ at 513 and $+0.145 \pm 0.003$ at 129. Because that ladder overstates the capacity effect (§M), we recompute the decomposition on the fixed-feature ladder, where the excess is $+0.0668$ $[+0.0584, +0.0740]$ at 513, $+0.0998$ $[+0.0864, +0.1101]$ at 129 and $+0.1441$ $[+0.1302, +0.1572]$ at 33 (validation resampled by episode; §O): smaller, roughly one half to two thirds, but every interval excludes zero under the pre-registered max baseline as well as the mean and under both definitions of λ, and **it grows monotonically as capacity falls**, which the projection ladder's flat 129 and 513 excesses could not show.

Out of distribution the same excess is present at the lowest measurable rung and not detected at the two highest. On the fixed-feature ladder it is $+0.1333$ $[+0.0778, +0.1763]$ at 33, $+0.0435$ $[-0.0027, +0.0923]$ at 129 and $+0.0007$ $[-0.0339, +0.0349]$ at 513 (resampled by segment within the two films); the projection ladder's $+0.021 \pm 0.012$ at 513 does not reproduce on the cleaner instrument. We do not write that the near-OOD excess is absent: its intervals are about four times as wide as the in-distribution ones, because near-OOD resamples 17 segments of two films where validation resamples 23 episode clusters. Whether the splits differ is tested directly from independent resamples: the in-distribution minus near-OOD difference excludes zero at 513 ($+0.0661$ $[+0.0311, +0.1008]$) and at 129 ($+0.0563$ $[+0.0065, +0.1064]$) and not at 33 ($+0.0108$ $[-0.0342, +0.0676]$).

Uncertainty, and a result that runs against us. For the free and capacity-matched readouts seed-to-seed variation is negligible (≤ 0.0006); on the middle rungs of the frozen ladder it is not (Figure 1). For R3 we therefore quote 95% Student-t intervals over the four subject-level absorption values (§O), the unit any claim must generalize over, indicative only at $n = 4$ and shown subject by subject in Figure 4: for *up* matching, absorption is -0.034 $[-0.053, -0.016]$ in distribution and $+0.002$ $[-0.049, +0.052]$ near-OOD. The in-distribution interval excludes zero on the *negative* side: R3-up's fusion gain ($+0.0423$) exceeds R1's ($+0.0410$) in all four subjects, by 0.0013 of r, against the absorption hypothesis rather than for it. Segment-level, per-cell BCa intervals on the underlying r values are in Appendix B.

Baseline sensitivity. Fusion gain subtracts a *maximum* over three unimodal conditions, which is biased upward and biases the gain downward; we keep the pre-registered max baseline as primary (Appendix I). For R1 the argmax is V on both splits and in every seed, whereas for both R3 directions

Table 3: Fusion gain retained relative to the `excluded` arm. The first three readouts sit at or above common width; the last two sit below it.

		at or above common width			below it	
split	arm	R1 (free)	R3-down (1,024)	R3-up (5,504)	R4-512	R5-513
val	partialled	70%	71%	73%	−0%	58%
val	competing band	51%	50%	55%	−15%	42%
ood	partialled	55%	52%	54%	34%	53%
ood	competing band	45%	38%	43%	25%	42%

at near-OOD it is *T*, and R3-up's in-distribution absorption *changes sign* under the mean baseline, from -0.034 to $+0.011$, so that result is baseline-specific.

4.3 Low-level stimulus covariates account for much of the raw score

A bank of 196 low-level covariates alone reaches $r = 0.2077$ in distribution, which is 77% of R1's full trimodal score (a ratio of correlations, not a share of variance). Fusion gain is therefore reported two ways: against the best unimodal condition, the field's convention, and against the bank. Fitted jointly with the bank the learned features retain incremental predictive value, a cross-validated Δr beyond the bank (r(features+bank) minus r(bank alone)) of $+0.043$ for the best unimodal condition and $+0.064$ trimodally; Pearson r does not decompose additively, so this is incremental prediction, not a unique-variance partition. The finding is not that encoders re-derive low-level structure but that a large low-level component sits beneath every reported score.

4.4 Where the confound correction depends on the readout, and where it does not

The first three readouts sit at or above common width and R3's fusion gain is within 3.5% of R1's under the tested conditions (absorption within ± 0.035), so agreement among them is not a test of readout dependence; the last two columns are that test. Retention is stable across the three wide readouts, within five points in distribution and seven near-OOD. Below common width it is not: the penalty readout retains 58% and 42% in distribution against R1's 70% and 51%, and the projection readout retains none of the corrected gain there, its corrected gains being -0.0001 and -0.0027, which is what the -0% and -15% entries mean; near-OOD the penalty readout tracks R1 (53% against 55%, 42% against 45%) while the projection readout retains 34% and 25%. **The fraction of fusion gain removed by the confound correction is stable across ridge readouts at or above common width, and is not stable below it** (Figure 3 plots the three wide readouts).

Scope and interpretation. The decomposition is **post-hoc**: the unimodal cells were pre-registered, λ and the excess were not. We offer **no account** of the difference between splits or of R3-up's slight excess over R1, and only an untested candidate for the non-monotone ladder under partialling (§6). Three seeds are computational replicates, not experimental units, and we quote a paired interval wherever one exists.

4.5 Where in cortex the ladder bites

Every row carries the gated-parcel mean of r within each of the seven Yeo networks (Figure 2; Appendix E tabulates it); this is post-hoc. Under the free ridge the gain is smallest in visual cortex, $+0.0095$ on the primary set and $+0.0092$ on the replication, and three to five times larger in the control and default networks. Under the projection readout the visual-cortex gain does not shrink, it inverts: R4-512 gives -0.0098 on the primary set and -0.0194 on the replication, and R4-128 gives -0.0272 and -0.0371; each value sits ten or more seed standard deviations below zero, which establishes stability across projections only, and each is negative in all four participants (§E). The penalty readout gives -0.0004 and -0.0025 in visual cortex, within 0.005 of zero in every participant with mixed sign, and R5 exceeds R4 in every network in every participant (§E). The inversion is the joint compression, tested directly (§E): with each modality projected through its own 512-column matrix (1,536 columns) the visual-cortex gain is $+0.0085$ $[+0.0074, +0.0098]$, positive in all four participants, and not distinguishable from zero under a joint projection to 1,536 ($+0.0008$ $[-0.0005, +0.0021]$); at equal budget the modality-wise projection retains 86% of R1's

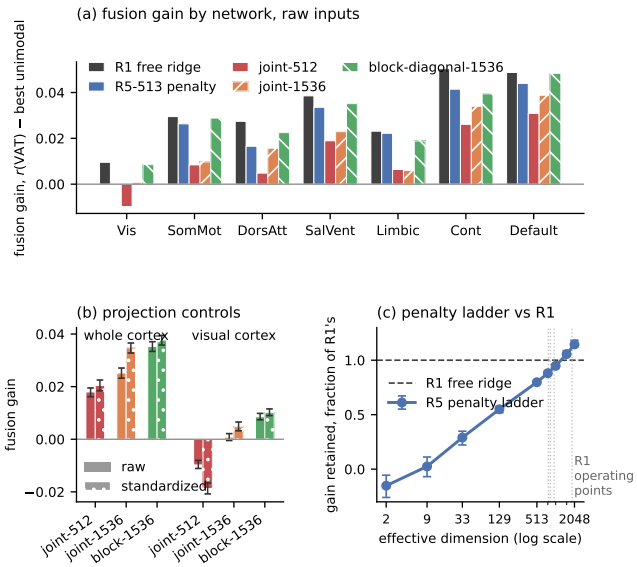


Figure 2: **Readout controls.** Validation split, `excluded` arm, subject mean. **(a)** Fusion gain by Yeo network for R1, R5-513 and the three projection readouts, raw inputs (seed spread at most 0.004). **(b)** Whole-cortex and visual-cortex gain for the projection readouts, raw and standardized inputs, paired 95% BCa intervals over the 23 validation episodes. **(c)** R5 gain as a fraction of R1's against effective dimension, paired intervals over the 23 validation episodes; dotted lines are R1's operating points (782, 844, 968, 1,876). Post-hoc; the replication's network profile is in §E.

gain against 61% joint, a paired difference of $+0.0101$ $[+0.0091, +0.0109]$ that shrinks to $+0.0029$ $[+0.0023, +0.0035]$ with inputs standardized before projection: mostly input scaling, partly subspace structure (intervals over the 23 validation episodes). The absolute loss under projection is nearly uniform across networks, so the absorption the ladder reports is concentrated where the gain was marginal to begin with (§E). We attach no mechanism; this concerns access per network, not neural integration.

5 Discussion

What the ladder establishes is narrower than a mechanism: the measured multimodal gain depends on how the readout is restricted, and matching a scalar capacity measure, width or effective degrees of freedom, does not make readouts interchangeable. It is familiar that prediction depends on regularization, input scaling and subspace choice; what these results add is that the dependence does not cancel between the trimodal and the unimodal conditions, so a contrast built to isolate the representation inherits it, downward under restriction and upward near the free ridge's own operating points. The practical consequence below those points is that comparisons across architectures with different readouts cannot attribute the difference to representations alone, and that a weaker readout can hide a representational advantage (§4).

Two results that bear directly on ours, neither comfortable. Gokce et al.'s fusion advantage is *larger* under its weaker readout ($+0.018$ at a linear ridge baseline against $+0.009$ with its brain encoder). This is neither a contradiction nor a confirmation of ours. Their readouts differ in functional form, a linear ridge against a nonlinear encoder, not in graded capacity within one form, and both sit at or above a free ridge, the regime in which our absorption is approximately zero; and their contrast is native against post-hoc fusion rather than trimodal against best-unimodal. Whether a natively fused representation, one that aligns modalities in a shared space during pre-training (Girdhar et al., 2023), shows the suppression we find below the free ridge is therefore untested by either study, and our design, a concatenation of independent unimodal backbones, cannot speak to it.

Second, Tripuraneni et al. (2021) show overparameterized random-feature regression is more robust to covariate shift, and our frozen rungs *are* a random-feature model; their result concerns test error, not a difference between conditions, but it is the obvious candidate for why every condition degrades

faster out of distribution as capacity falls, and we do not test it. Separating that generic property of random features from a vulnerability of the learned representation needs an untrained-architecture control (the same backbones at random initialization, run down the same ladder), which was not run and is the first confirmatory experiment we would add.

Third, Scotti & Tripathy (2025) find readout architecture varies widely across top Algonauts 2025 entries, all at or above a free readout, where our absorption is near zero.

6 LIMITATIONS

1. **Generalization and statistical precision.** All four comparisons registered in Amendment 9 reproduce after changing the text backbone, layer aggregation and width: R3-up absorption is -0.033 versus -0.034, and both ladder orderings and the bank's share reproduce (Appendix D). The shared video and audio backbones limit this evidence to changes in text representation and width (an untrained-architecture control, §5, was not run). Results cover four subjects and 834 of $1{,}000$ reliability-selected parcels. Near-OOD and officially scored far-OOD are distinct test sets. Near-OOD, uniform shrinkage explains 86% of top-rung absorption, and the fixed-feature residual of $+0.0007$ has an interval spanning zero. The in-distribution excess persists across rungs. The 17 near-OOD units are segments of two films, so near-OOD intervals, which resample segments, do not establish generalization across films; the split difference in the excess is detected at the top two rungs and not at 33, and its mechanism is unknown. §O reports percentile intervals, leave-one-segment-out re-estimates and the contrast per film: no near-OOD gain or difference changes sign when any single segment is dropped, the R5-minus-R4 gain difference is positive in each film alone ($+0.0037$ in *Hidden Figures*, $+0.0093$ in *Life*), and the retained-fraction difference excludes zero under the percentile interval but not under BCa. The reliability gate is in-sample for the near-OOD split; with it removed the headline comparison is unchanged (§N). The representation concatenates independent unimodal backbones; whether a model that aligns modalities during pre-training shows the same readout dependence is untested.

2. **Interpretation and numerical robustness.** R3's systematic 0.006–0.008 score deficit relative to R1 is the standardization scope, not the projection. R1 refitted under R3's preprocessing reproduces R3-up to 10^{-10} (Appendix A). Independent implementations differ by a factor of 2.7 in partialling-leak magnitude, supporting claims about direction and order of magnitude only. Capacity ordering also depends on confound handling: under `partialled`, the fixed-feature in-distribution gains are non-monotone (0.0235, 0.0204, 0.0146, 0.0116, 0.0191), with paired intervals about 0.004 wide. This pattern is not explained by noise, and its mechanism remains unknown. One candidate, post hoc and untested: partialling flattens the design's spectrum (Appendix R; the top eigenvalue falls $7.3\times$ and the participation ratio rises from 82.8 to 549.5), so an R5 rung solved to the same effective dimension in the two arms distributes its shrinkage over different directions. The monotone ordering under `excluded` therefore does not generalize across confound arms.

3. **Exploratory analysis status.** The graded rungs were added after the analysis inclusion criteria were satisfied; they refine a pre-registered readout and introduce no new hypothesis, but were not themselves pre-registered. The post-hoc decomposition and the exploratory fixed-feature control retain their stated status; these results require confirmatory evaluation.

7 REPRODUCIBILITY AND METHOD INTEGRITY

Every result row records its configuration hash, code commit, metric version and environment, and every number here is matched mechanically against the frozen tag; the pre-registered kill condition (a sign flip on the primary contrast) is not met, and the corrections that affect published results, eight silent narrowings and one silent widening among thirty-six ledger entries, are in Appendix Q. *Reproducibility statement:* code, configurations, the analysis gate with its planted-data self-test, and the result rows behind every figure and table are released (§L); the fMRI data is third-party and obtained under its own terms.

REFERENCES

Julie Boyle, Basile Pinsard, Valentina Borghesani, François Paugam, Elizabeth DuPre, and Pierre Bellec. The Courtois NeuroMod project: Quality assessment of the initial data release (2020). In *Conference on Cognitive Computational Neuroscience (CCN)*, Oxford, United Kingdom, 2023. doi: 10.32470/ccn.2023.1602-0.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. TRIBE: TRImodal brain encoder for whole-brain fMRI response prediction. *arXiv preprint arXiv:2507.22229*, 2025.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, pp. 119728, 2022. doi: 10.1016/j.neuroimage.2022.119728.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The Algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023.

Abdulkadir Gokce, Badr AlKhamissi, and Martin Schrimpf. MIRAGE: Adaptive multimodal gating for whole-brain fMRI encoding. *arXiv preprint arXiv:2605.29850*, 2026.

Nima Hadidi, Ebrahim Feghhi, Bryan H. Song, Idan A. Blank, and Jonathan C. Kao. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications*, 17:5769, 2026. doi: 10.1038/s41467-026-72253-7.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379, 1961.

Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, 197:482–492, 2019.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528, Bangkok, Thailand, 2024.

Shreya Saha, Ishaan Chadha, and Meenakshi Khosla. Modeling the human visual system: Comparative insights from response-optimized and task-optimized vision models, language models, and different readout mechanisms. *arXiv preprint arXiv:2410.14031*, 2024.

Paul S. Scotti and Mihir Tripathy. Insights from the Algonauts 2025 winners. *arXiv preprint arXiv:2508.10784*, 2025.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

AI USE STATEMENT

(This section is required by the ICLR 2027 AI Policy for Authors and does not count toward the page limit.)

Tasks in which LLMs were used. Under the authors' direction, language models were used as coding assistants for limited software-engineering tasks, primarily to help optimize existing scripts, identify potential bugs, suggest debugging strategies, and improve code readability and maintainability. Language models were also used for minor proofreading and language refinement of text written by the authors. All LLM-assisted modifications were reviewed and approved by the authors before being incorporated into the codebase or manuscript.

Tasks in which LLMs were not used. Language models were not used to formulate the research questions, hypotheses, theoretical framework, pre-registration or its amendments, experimental design, analysis methodology, readout ladder, absorption index, or scientific conclusions. The core analysis and infrastructure code were designed and implemented by the authors, and the manuscript and its scientific arguments were written by the authors. No synthetic data or measurements were generated by language models; every numerical value derives from fitted models executed on the released fMRI data.

Verification. All LLM-assisted code changes were reviewed and tested by the authors before use. The analysis pipeline requires a planted-data self-test to pass before computing the reported contrasts. Results are frozen under a version tag. A separate checker, run on demand, compares numerical claims in the manuscript with the frozen results, including values derived through documented differences or ratios, and flags discrepancies. These checks assess implementation, transcription and arithmetic consistency between the recorded results and the manuscript; they do not establish the correctness of the underlying analyses. Appendix Q documents errors that the checks then in place did not detect.

Not applicable. Generative AI was not used to develop theoretical frameworks, prove mathematical claims, design the scientific methodology, interpret the experimental results, or reformat/clean the underlying fMRI dataset.

Responsibility. The human authors designed the study, developed the methodology, implemented the core analysis pipeline, interpreted the results, and wrote the manuscript. They reviewed and validated all LLM-assisted modifications and assume full responsibility for all code, analyses, findings, and conclusions presented in this manuscript.

A THE NORMALIZATION DIAGNOSTIC

The test case is VAT under *up* matching, where $D_{\text{match}} = 5{,}504$ is VAT's own width. The projection is the train-only right-singular basis of the standardized base features, an orthonormal matrix at full width with retained variance 1.0000, so no information is lost. R1 and R3 share the ridge solver, the segment-level inner folds, the alpha grid (10^0 to 10^{10} in 21 steps, with no widening triggered in either cell) and the selection criterion, so the selection procedure is identical. The selected penalties are not: the two designs differ (below) and per-parcel selection responds to that, landing at $1{,}876$ and $1{,}895$ effective degrees of freedom.

Table 4: VAT, in distribution, seed 0.

	held-out r	effective d.o.f.
R1 free ridge	0.2695	$1{,}876$
R3, standardize after projection	0.2132	$6{,}174$
R3, no standardization after projection	0.2636	$1{,}895$

Standardizing the projected coordinates cost 0.056 in r while inflating effective degrees of freedom $3.3\times$ on information-identical data, the signature of a changed regularization geometry, not a changed representation.

The remaining 0.0059 between R1 and the unstandardized R3 is not accounted for by the rotation, which is orthogonal. The two fits differ in one other place. R1 standardizes each of the $22{,}016$

delay-stacked columns on the training set; R3 standardizes each of the $5{,}504$ base features before delay stacking, then centers but does not rescale the projected columns, and because delayed copies are zero-padded at each segment start, the two isotropic penalties are defined in spaces that coincide only up to that edge effect. We tested this by refitting R1 under R3's preprocessing (base-feature standardization on the training stimuli, then delay stacking, then centering only), everything else unchanged. The harmonized R1 reproduces the unstandardized R3-up cell to 10^{-10}: VAT scores 0.2636 in distribution and 0.1689 near-OOD at $1{,}895$ effective degrees of freedom on both fits, so the whole 0.0059 is the standardization scope and none of it is the estimator or the rotation. The unimodal conditions agree to within 2×10^{-4} rather than exactly, because *up* matching tiles a unimodal design to $5{,}504$ columns before the basis is fitted and a tiled design is not a rotation of the original. The fusion contrast survives harmonization: under R3's preprocessing R1's gain is $+0.0424$ in distribution and $+0.0501$ near-OOD, against $+0.0410$ and $+0.0500$ canonically. The uniform 0.006–0.008 deficit Table 2 reports for R3 has the same cause. The canonical R1 refitted alongside this check reproduced its frozen rows to 6×10^{-14}.

B SEGMENT-LEVEL INTERVALS PER CELL

Segment-level BCa intervals, $1{,}000$ resamples, 50 validation and 17 near-OOD segments, seed 0, subject-mean. R1 and R3-up overlap in every cell, which is Table 2's absorption result shown at the resolution the data support.

	V	A	T	VAT
R1, val	0.2286 $[0.2213, 0.2333]$	0.2099 $[0.2038, 0.2149]$	0.1680 $[0.1629, 0.1727]$	0.2695 $[0.2614, 0.2745]$
R3-up, val	0.2212 $[0.2141, 0.2261]$	0.2020 $[0.1956, 0.2071]$	0.1677 $[0.1625, 0.1723]$	0.2636 $[0.2555, 0.2687]$
R1, OOD	0.1228 $[0.1076, 0.1402]$	0.1200 $[0.1087, 0.1297]$	0.1193 $[0.1092, 0.1298]$	0.1728 $[0.1584, 0.1875]$
R3-up, OOD	0.1180 $[0.1020, 0.1353]$	0.1127 $[0.1018, 0.1225]$	0.1188 $[0.1086, 0.1292]$	0.1689 $[0.1544, 0.1836]$

C WHAT WAS ACTUALLY RUN AT THE FREEZE

This inventory covers `results-freeze-v1`. The controls added afterwards are inventoried where they are reported: the fixed-feature ladder in §M, the width null in §P, the paired intervals in §O, and the gate-threshold sweep in §N.

The design is a full crossing of readout $\times$ modality condition $\times$ confound arm $\times$ split $\times$ seed, executed as independent cells. Every cell writes one row per subject plus a subject-mean row carrying its own provenance. The freeze contains **1,092 distinct fitted cells** and **7,470** result rows. The table below sums to 540 because it counts before splitting and omits the bank-only control; 540 plus the 6 bank-only cells is 546, and 546 evaluated on two splits is $1{,}092$.

readout	conditions	arms	directions	seeds	cells
R1 free ridge	7	3	1	3	63
R2 banded ridge	4 (multi-band only)	3	1	3	36
R3 capacity-matched	7	3	2	3	126
R4-k graded ($k \in \{8, 32, 128, 512\}$)	7	3	1	3	252
R4 frozen	7	3	1	3	63

Three properties are enforced mechanically rather than by discipline, each after a failure that cost real work.

A cell is identified by its full key (condition, readout, capacity, matching direction, confound arm, split, seed, subject) never by a label or a configuration hash. Two readouts once collided on the label `R3` and one was silently discarded; a configuration hash later proved to omit capacity and matching direction, so that 252 of 444 hashes covered more than one distinct cell while the test guarding it varied only the one axis already included.

Results merge only after proving agreement. Where the same cell was computed twice (by a restart, or on a different machine) the values are compared before either is kept, and a disagreement

raises rather than resolving by row order. Across 960 duplicated cells the largest discrepancy was 1.9×10^{-10}, including one condition recomputed under a different code commit, which we report as an incidental cross-commit reproducibility check.

Provenance is proven, not asserted. The compute machines are not repositories, so the code commit is passed explicitly at launch and the runner refuses to start without it; before any launch the machine's code is digest-compared against that commit.

D THE REPLICATION, COMPARISON BY COMPARISON

Amendment 9's four comparisons, their rules as fixed in code before any replication score existed, and both feature sets. The second feature set uses the same four participants, the same stimuli and the same video and audio backbones; it changes the text backbone (Qwen2.5-3B in place of Llama-3.2-3B), the layer aggregation (twenty layers pooled in two groups for every stream, rather than the final hidden state) and the width (2,816, 2,048 and 4,096 per stream, $D = 8{,}960$). It therefore tests whether the four comparisons are robust to feature construction, not whether they replicate on independent neural data, and "replication" is used below in that restricted sense. Retained gain is the constrained readout's fusion gain over R1's, per seed, `excluded` arm, validation split, mean over three seeds. The claim under test is qualitative agreement: absolute scores are not expected to match.

	registered rule	primary $D = 5{,}504$	replication $D = 8{,}960$	outcome
c1	R3-up absorption, $\lvert \cdot \rvert \leq 0.10$	-0.034	-0.033	holds
c2	R4-512 retains more than R4-128	43.9% vs 18.2%	39.6% vs 9.5%	holds
c3	R5-513 retains more than R4-512	79.8% vs 43.9%	80.5% vs 39.6%	holds
c4	bank share of trimodal score $> 38.5\%$	77.0%	74.8%	holds

The c1 threshold was set at three times the primary set's own ± 0.035 and the c4 floor at half its 77%, both anchored to frozen values rather than chosen freely. Twelve R3-up cells, four conditions at three seeds, back c1; the interval rule is a 95% t-interval over seeds, which is wide at $n = 3$ and therefore makes c1 harder to disconfirm and c2 and c3 likelier to carry a qualifier. None of the three needed one.

Both feature sets in full. Table 5 gives every quantity the two runs share, not only the four the amendment registered. Retained gain is the constrained readout's fusion gain as a fraction of R1's, within seed, `excluded` arm, averaged over the three seeds.

Table 5: The primary feature set and its replication, `excluded` arm, subject mean. The replication changes the text backbone, the layer aggregation and the width; the video and audio backbones are shared.

	primary $D = 5{,}504$	replication $D = 8{,}960$
R1 $r(\mathrm{V})$, val	0.2286	0.2339
R1 $r(\mathrm{A})$, val	0.2099	0.2368
R1 $r(\mathrm{T})$, val	0.1680	0.1679
R1 $r(\mathrm{VAT})$, val	0.2695	0.2777
R1 fusion gain, val	**+0.0410**	**+0.0410**
R1 fusion gain, near-OOD	+0.0500	+0.0428
R3-up retained gain, val / near-OOD	103.4% / 100.3%	103.3% / 106.5%
R3-down retained gain, val / near-OOD	99.3% / 96.8%	101.5% / 106.2%
R4-128 retained gain, val / near-OOD	18.2% / 65.4%	9.5% / 64.8%
R4-512 retained gain, val / near-OOD	43.9% / 86.4%	39.6% / 63.1%
R5-513 retained gain, val / near-OOD	79.8% / 96.4%	80.5% / 83.6%
R1 / R3-up effective dimension, VAT	1876 / 1895	2119 / 2187

Beyond the registered four. The rest of this appendix is **post-hoc**. It was computed after the branch had been selected and could not have influenced the routing, and it is labelled exploratory wherever it is cited. We report it because the pair of runs answers a question neither run could.

The gain is identical while its baseline is not. Both feature sets give a free-ridge fusion gain of $+0.0410$ in distribution. Underneath that agreement almost nothing is shared: every unimodal score moves, audio by $+0.027$, and **the best unimodal condition changes identity**, from vision on the primary set to audio on the replication. The quantity the paper decomposes is a contrast, and a contrast survived a change of the very condition it subtracts.

The two instruments do not replicate equally well, and the direction is the one §4.2 predicts. In distribution the fixed-feature ladder is nearly invariant across feature sets, R5-513 retaining 79.8% and then 80.5%, while the projection ladder moves and always downward on the wider set: R4-512 by 4.3 points and R4-128 by 8.7. The paper already argues that the projection ladder overstates the capacity effect, because a random projection destroys information beyond what the reduction in capacity accounts for. If that is right, the amount destroyed is a property of the representation and should grow with the compression ratio, which at $k = 128$ is $70\times$ on the replication against $43\times$ on the primary set. It does. **No single feature set can test that claim**, because with one representation the projection loss and the capacity restriction are confounded by construction; a second feature set that differs in text backbone, layer aggregation and width together can show whether the pattern depends on the representation at all, though not which of those changes matters, and the instrument we trust is the one that held still.

Near-OOD does not replicate, which bears on the near-OOD limitation (§6) rather than contradicting it. The near-OOD retained gains move by as much as 23 points, R4-512 falling from 86.4% to 63.1%, with no monotone pattern: one rung moves while its neighbour does not, R4-128 going from 65.4% to 64.8%. With 17 near-OOD segments of two films the paper declines to resolve that split. Two feature sets disagreeing there is consistent with a power-limited estimate; it is equally consistent with genuine dependence on the feature set, and these numbers do not separate the two.

One apparent change is arithmetic. The bank of low-level covariates contains no encoder features, so its score is the same 0.2077 for both. Its share of the trimodal score falls from 77% to 74.8% only because the denominator rises from 0.2695 to 0.2777. Nothing about the bank differs between the runs.

Width matching holds in both directions. The registered comparison read R3-up; R3-down, fitted after the branch had printed, agrees on the replication as it did on the primary set: absorption -0.015 against $+0.007$ in distribution, both inside the registered 0.10 by a wide margin, and the narrow projection's effective dimension sits below the free ridge's where the wide one's sits just above it.

The result that runs against us reproduces. R3-up retains slightly more than the whole gain on both feature sets, 103.4% and 103.3% in distribution, and its effective dimension sits just above the free ridge's on both, $+1.0\%$ and $+3.2\%$. We still have no mechanism for it (§6), and it is now unexplained on two feature sets rather than one.

E WHERE IN CORTEX THE LADDER BITES

Every result row carries the gated-parcel mean of r within each of the seven Yeo networks. Fusion gain per network is $r(\mathrm{VAT})$ minus the best unimodal, taken within seed and within network, then averaged over seeds; `excluded` arm, validation split, subject mean. R5 is deterministic and has one seed. The seed standard deviation is at most 0.004 everywhere and below 0.0001 for R1, so every pattern below is far outside seed noise. Seed noise is variation across random projections only; Table 7 gives the per-participant breakdown. **This analysis is post-hoc**: it was not pre-registered and was computed after the replication's branch was selected.

Three things hold on both feature sets, and Table 7 shows the first two hold in each participant: R1's gain is smallest in Vis for all four, R4-512 inverts Vis in all four, and R5-513 exceeds R4-512 in every network for all four.

The targeted control: is the inversion the compression? Two further trimodal readouts were fitted at the same three projection seeds and scored against the same unimodal R4-512 cells. *Joint-*

Table 6: Fusion gain by cortical network, validation split, `excluded` arm, both feature sets. Negative entries mean the trimodal model scores *below* the best unimodal in that network.

		Vis	SomMot	DorsAttn	SalVentAttn	Limbic	Cont	Default
primary	R1	+0.0095	+0.0295	+0.0274	+0.0408	+0.0237	+0.0517	+0.0487
	R5-513	−0.0004	+0.0263	+0.0166	+0.0344	+0.0232	+0.0418	+0.0452
	R4-512	−0.0098	+0.0084	+0.0048	+0.0190	+0.0064	+0.0261	+0.0309
	R4-128	−0.0272	+0.0193	−0.0064	+0.0149	+0.0067	+0.0121	+0.0184
replication	R1	+0.0092	+0.0203	+0.0272	+0.0271	+0.0119	+0.0303	+0.0278
	R5-513	−0.0025	+0.0154	+0.0156	+0.0210	+0.0105	+0.0235	+0.0219
	R4-512	−0.0194	−0.0104	−0.0029	+0.0017	−0.0057	+0.0156	+0.0082
	R4-128	−0.0371	−0.0037	−0.0125	+0.0049	−0.0026	+0.0133	+0.0111

Table 7: Per-participant fusion gain by network, primary feature set, validation split, `excluded` arm, from the refits of §O. R4-512 is the three-seed mean; R1 is seed 0. Last row: participants in which R5-513 exceeds R4-512.

		Vis	SomMot	DorsAttn	SalVentAttn	Limbic	Cont	Default
R1	sub-01	+0.0065	+0.0341	+0.0270	+0.0372	+0.0138	+0.0467	+0.0440
	sub-02	+0.0122	+0.0296	+0.0357	+0.0413	+0.0240	+0.0569	+0.0521
	sub-03	+0.0097	+0.0295	+0.0211	+0.0421	+0.0342	+0.0500	+0.0587
	sub-05	+0.0098	+0.0247	+0.0259	+0.0336	+0.0205	+0.0479	+0.0403
R4-512	sub-01	−0.0116	+0.0110	+0.0060	+0.0159	−0.0102	+0.0200	+0.0270
	sub-02	−0.0063	+0.0114	+0.0110	+0.0263	+0.0122	+0.0331	+0.0327
	sub-03	−0.0126	+0.0071	−0.0025	+0.0186	+0.0136	+0.0241	+0.0329
	sub-05	−0.0086	+0.0041	+0.0048	+0.0126	+0.0028	+0.0273	+0.0213
R5-513	sub-01	−0.0045	+0.0289	+0.0159	+0.0327	+0.0144	+0.0369	+0.0403
	sub-02	+0.0019	+0.0281	+0.0234	+0.0386	+0.0238	+0.0482	+0.0491
	sub-03	−0.0013	+0.0256	+0.0088	+0.0332	+0.0295	+0.0394	+0.0490
	sub-05	+0.0025	+0.0228	+0.0181	+0.0297	+0.0211	+0.0414	+0.0375
R5 > R4, of 4		4	4	4	4	4	4	4

1536 projects the whole design with one Gaussian matrix to 1,536 columns. *Block-diagonal-1536* projects each modality through the very matrix its unimodal R4-512 cell used and concatenates the blocks, so the trimodal design keeps exactly the 512-dimensional subspace each unimodal cell had; no new matrix is sampled. Both are ordinary least squares with an intercept, like R4-k.

Table 8: Fusion gain by network under the projection controls, validation split, `excluded` arm, subject mean over three seeds, with the whole-cortex gain and its fraction of R1's. Within each input scaling the unimodal comparators are that scaling's own V, A and T cells at 512. Last block: visual-cortex gain per participant.

	Vis	SomMot	DorsAttn	SalVentAttn	Limbic	Cont	Default	all	of R1
R1 free ridge	+0.0095	+0.0295	+0.0274	+0.0408	+0.0237	+0.0517	+0.0487	+0.0410	
R5-513	−0.0004	+0.0263	+0.0166	+0.0344	+0.0232	+0.0418	+0.0452	+0.0327	80%
R4-512, joint	−0.0098	+0.0084	+0.0048	+0.0189	+0.0064	+0.0261	+0.0309	+0.0180	44%
joint-1536	+0.0008	+0.0097	+0.0157	+0.0229	+0.0060	+0.0340	+0.0388	+0.0252	61%
block-diagonal-1536	+0.0085	+0.0289	+0.0225	+0.0352	+0.0188	+0.0394	+0.0484	+0.0353	86%
inputs standardized on the training stimuli before projection									
joint-512	−0.0187	+0.0149	+0.0014	+0.0246	+0.0141	+0.0325	+0.0355	+0.0204	50%
joint-1536	+0.0048	+0.0229	+0.0215	+0.0342	+0.0161	+0.0434	+0.0484	+0.0348	85%
block-diagonal-1536	+0.0101	+0.0285	+0.0249	+0.0376	+0.0185	+0.0425	+0.0494	+0.0377	92%
visual-cortex gain per participant (sub-01, 02, 03, 05); raw, then standardized inputs									
R4-512, joint	−0.0116, −0.0063, −0.0126, −0.0086			−0.0211, −0.0148, −0.0228, −0.0162					
joint-1536	−0.0001, +0.0044, +0.0004, −0.0016			+0.0030, +0.0087, +0.0042, +0.0032					
block-diagonal-1536	+0.0071, +0.0116, +0.0098, +0.0057			+0.0084, +0.0138, +0.0110, +0.0072					

Paired over the 23 validation episodes (the segment-level intervals, which differ by at most 0.0004 on these bounds, are in the released JSON), the visual-cortex gain is -0.0098 $[-0.0111, -0.0080]$ under the joint 512-column projection, $+0.0008$ $[-0.0005, +0.0021]$, not distinguishable from zero, under the joint 1,536-column projection and $+0.0085$ $[+0.0074, +0.0098]$ under the block-diagonal one. The inversion is therefore a property of compressing the concatenated design at fixed k, not of visual cortex: it vanishes with three times the columns and reverses when each modality keeps

its own subspace. At equal budget the modality-wise projection retains more of the gain everywhere, $+0.0101$ $[+0.0091, +0.0109]$ on the whole-cortex gain against joint-1536 and $+0.0077$ $[+0.0069, +0.0084]$ in visual cortex, so how the random subspace is drawn, not only how large it is, sets the measured gain. Near-OOD, with 17 segments, the three readouts' whole-cortex gains are indistinguishable (0.0432, 0.0437, 0.0428).

Input scaling versus subspace structure. R4 projects the raw feature columns, whose scales differ across modalities; in the joint trimodal subspace the audio block accounts for 0.2% of the entering variance. To separate that from the subspace structure, every cell of the comparison was refitted with each base feature standardized on the training stimuli before delay stacking and projection, the matrices unchanged and the unimodal comparators refitted the same way (Table 8, lower block). Standardization raises every gain and lowers every absolute score: the trimodal joint-$1{,}536$ score falls from 0.2379 to 0.2355 while its gain rises from $+0.0252$ to $+0.0348$ $[+0.0328, +0.0366]$, because the unimodal scores fall more. The equal-budget difference between modality-wise and joint projection shrinks from $+0.0101$ $[+0.0091, +0.0109]$ with raw inputs to $+0.0029$ $[+0.0023, +0.0035]$ with standardized ones, and stays positive in visual cortex ($+0.0053$ $[+0.0046, +0.0061]$). So most of the joint projection's penalty at this budget is input scaling, which standardization removes (joint-$1{,}536$ gains $+0.0096$ $[+0.0089, +0.0104]$ from it, the block-diagonal only $+0.0024$ $[+0.0020, +0.0030]$), and a smaller part is the subspace structure, which survives it. Neither is the whole mechanism. At 512 columns standardization deepens the visual-cortex inversion, -0.0187 $[-0.0208, -0.0166]$ against -0.0098, negative in all four participants, while raising the whole-cortex gain to $+0.0204$, so the inversion is a budget phenomenon under either scaling and its depth depends on the scaling. One candidate explanation, untested: with equal column variance the widest modality contributes the most columns to the joint subspace, which need not mean it carries the most predictive information. Near-OOD the standardized set's gains are smaller (0.033 to 0.039) and the within-set differences are within noise.

Network by readout, paired. From the same episode-level resamples, R5-513 exceeds R4-512 in every network with every interval excluding zero (Vis $+0.0094$ $[+0.0086, +0.0104]$), and the advantage is larger in association cortex (Cont and Default) than in visual cortex by $+0.0056$ $[+0.0040, +0.0073]$. The loss from projection itself, R4-512 minus R1, is nearly uniform across networks, from -0.017 in Limbic to -0.026 in Cont, and is *larger* in association cortex than in visual cortex by 0.0025 $[0.0006, 0.0044]$: visual cortex inverts because its free-ridge gain is the smallest, not because projection removes more of it there.

The free-ridge fusion gain is smallest in visual cortex and largest in association cortex. Vis is $+0.0095$ on the primary set and $+0.0092$ on the replication, near-identical; Cont and Default are three to five times larger. The gain is largest where the three streams are individually weak and near-equal and absent where a single stream already carries the region, as vision does in visual cortex; the network profile itself is not new, and what is new here is its dependence on the readout.

Under a restricted readout the visual-cortex gain does not shrink, it inverts. R4-512 gives -0.0098 and -0.0194 in Vis, R4-128 -0.0272 and -0.0371, each ten or more seed standard deviations below zero: with capacity withdrawn, the trimodal model is *worse* than the best unimodal there, and the sign is stable across random projections, which is all that seed spread establishes. The instrument reading, that at a fixed k the trimodal design is compressed harder than a unimodal one, is tested below with a modality-wise projection and holds. The fixed-feature ladder has no such handicap and gives -0.0004 and -0.0025 in Vis on both sets, values within 0.005 of zero in every participant and of mixed sign across them (Table 7). Limbic rests on the fewest gated parcels of the seven and its row is the least reliable.

Association cortex keeps the gain. Under R5-513, Default retains 93% and 79% of R1's gain on the two sets and Cont 81% and 78%; under R4-512 the best-preserved network is Default on the primary set (63%, Cont 50%) and Cont on the replication (51%, Default 29%). The absorption the main text reports is therefore not spread evenly across cortex. It is concentrated where the gain was marginal to begin with, and the association networks where fusion matters keep most of it under the penalty ladder and, on the primary set, half to two thirds under the projection ladder (on the replication, half in Cont and under a third in Default). The replication's projection ladder is harsher in every network (§D); its penalty ladder tracks the primary set's network by network. The near-OOD split is not

tabulated: its per-network gains move between feature sets by more than their seed spread, which is the instability §D reports.

F THE LOW-LEVEL STIMULUS BANK, AND WHY R2 BORROWS UNIMODAL SCORES

The 49 columns. Five text rates (word rate, character rate, mean word length, speech present, sentence ends); ten positional and drift terms (linear, quadratic, eight low-frequency cosines), which are not stimulus features and are included so that slow within-episode structure is charged to the bank rather than to the encoder; fourteen audio descriptors (RMS, log RMS, envelope change, zero-crossing rate, spectral centroid, flux and 85% roll-off, and seven octave-band energies from 0 to 8 kHz); and twenty motion-energy descriptors (sixteen principal components of a motion-energy pyramid plus luminance, luminance change, RMS contrast and scene cuts). Each is delay-stacked at the same four delays as the features, giving 196 regressors, so confounds get the same temporal reach as the signal they compete with. A bank-only condition serves as a control and reaches $r = 0.2077$ in distribution.

R2 is fitted only on the four conditions with more than one band, because a banded ridge with *one* band reduces **exactly** to a free ridge, an invariant asserted in our test suite (identical selected α, weights agreeing to 1.1×10^{-7}). Its fusion gain therefore borrows R1's unimodal scores, which are R2's own by construction rather than by approximation. The substitution is performed by a named function that documents the equivalence it depends on, so the figure carries the same provenance as every other number here; the general fusion-gain routine returns *nothing* for R2 rather than inventing a baseline.

G A FIGURE WE WITHDREW

An earlier version of this paper defended the absorption index against the tautology objection by quoting 86% score retention against 44% gain retention at 513 parameters in distribution. The 44% used a max-over-unimodals baseline. Under the mean-of-unimodals baseline, which our own sensitivity analysis reports alongside the pre-registered one, gain retention is 72%. The dissociation is real, an excess of $+0.136 \pm 0.013$ over uniform shrinkage, but roughly half the size we first reported, and the larger figure carried max-selection bias that grows as capacity falls. We record the withdrawn number rather than silently replacing it.

H FIGURE 1, RIGHT PANEL, IN NUMBERS

The near-OOD gains from 33 parameters upward are strictly ordered in each projection seed: $+0.0159 < +0.0324 < +0.0430$; $+0.0074 < +0.0337 < +0.0433$; $+0.0056 < +0.0320 < +0.0433$.

The values plotted in the right panel of Figure 1, so the panel carrying the paper's second result can be checked against the frozen release like every table. Mean-of-unimodals baseline, $\lambda = r_{\mathrm{VAT}}(k)/r_{\mathrm{VAT}}(\mathrm{R1})$, `excluded` arm, subject-mean over three seeds. The spread column is the seed standard deviation of the excess.

The rule that selects the cells the claims rest on has two steps, and both were used. First, a rung's excess is assessed only where the gain itself exceeds its seed spread (§4: in distribution that fails at 9 and 33, near-OOD at 9). Second, the excess counts as measured where its magnitude exceeds its own seed standard deviation and its sign agrees across the three seeds. This yields val 129, val 513 and ood 513; ood 33 and 129 pass the first step and fail the second ($+0.014$ against 0.066; $+0.022$ against 0.030). Two rungs need stating separately. Val 33 passes the second test ($+0.123$ against 0.041) and fails the first. Val 2 passes both numeric tests ($+0.058$ against 0.006; gain $+0.0008$ against a seed sd of 0.0006) and is set aside by an interpretive decision that rests on §3's definition: its gain is near zero and the projection ladder's absorption must approach 1 as $k \to 0$, so the index carries no information there. "Near zero" is a judgement, not a numeric criterion, so no single numerical rule selects exactly the three retained cells; we record the exclusion as interpretive. The

split	params	observed absorption	$1-\lambda$	excess	seed sd
ood	2	0.843	0.832	+0.011	0.021
ood	9	0.478	0.565	−0.087	0.129
ood	33	0.419	0.405	+0.014	0.066
ood	129	0.235	0.213	+0.022	0.030
ood	513	0.148	0.127	+0.021	0.012
val	2	0.923	0.865	+0.058	0.006
val	9	0.521	0.583	−0.062	0.159
val	33	0.541	0.419	+0.123	0.041
val	129	0.397	0.251	+0.145	0.003
val	513	0.280	0.144	+0.136	0.013

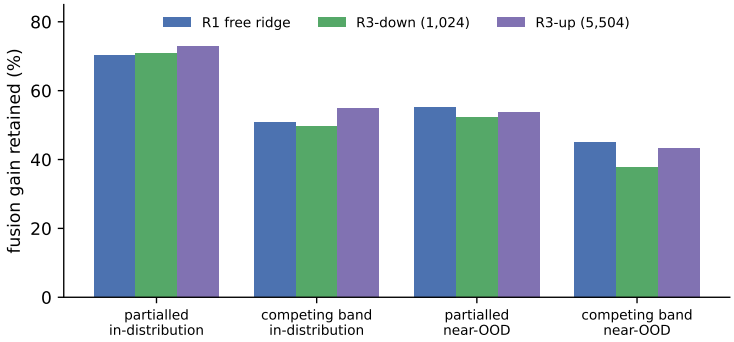


Figure 3: **Confound-correction retention agrees across readouts at or above common width.** Fusion gain retained relative to the `excluded` arm, at a free ridge and at the capacity-matched readout in both directions. These three agree to within five points in distribution; the exception is visible rather than averaged away, at near-OOD under a competing band, where R3-down retains 38% against R1's 45%. This is agreement among readouts whose fusion gains we separately show to agree within 3.5%, so it is not a test of readout dependence. The two readouts below common width, which are that test and which disagree, are in Table 3.

rungs at 9 are noise-dominated in both columns, which is why the residual there is negative without that signifying anything.

I BASELINE SENSITIVITY IN FULL

Table 9: Gain and absorption under three baselines, `excluded` arm, subject-mean. The pre-registered estimand is the max baseline.

quantity	split	max (pre-registered)	fixed V	mean unimodal
R1 gain	val	0.0410	0.0410	0.0674
R1 gain	near-OOD	0.0500	0.0500	0.0521
R3-up absorption	val	−0.034	−0.034	+0.011
R4-512 absorption	val	0.561	0.561	0.280
R4-512 absorption	near-OOD	0.136	0.126	0.148

Fusion gain subtracts a *maximum* over three unimodal conditions, and a maximum over noisy estimates is biased upward, which biases the gain downward. Two checks. First, *for R1* V is the argmax on both splits and in every seed, so the max-based baseline and a fixed V baseline are numerically identical there (0.0410 val, 0.0500 OOD). This does not generalize: for both R3 directions at OOD the argmax is *T* (0.1188 against V's 0.1180), again in every seed, which is precisely why a V baseline and a max baseline diverge for R3 and not for R1. The unimodal scores at OOD are close (0.1228, 0.1200, 0.1193 for R1), so which one wins is not robust across readouts and we do not treat it as a fact about the data. Second, absorption itself is baseline-sensitive: R4-512's is 0.136, 0.126 and 0.148 under max, V-alone and mean baselines, and in distribution it moves from 0.561 to 0.280 across the same three, while R3-up stays within 0.035 of zero under all of them. Which

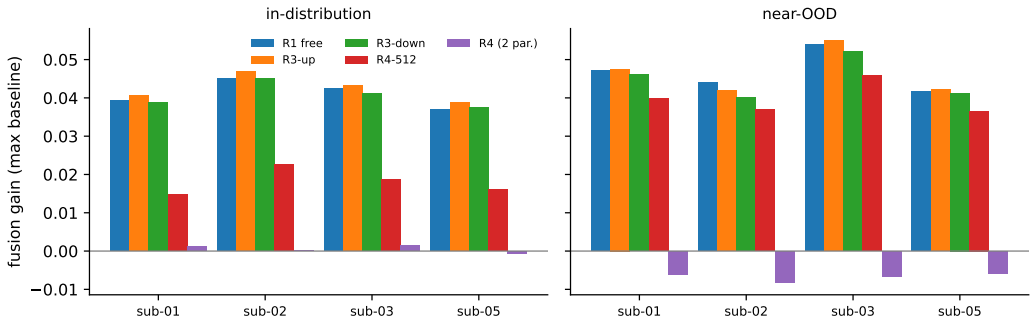


Figure 4: **The primary contrast, per subject.** Fusion gain under the pre-registered max baseline, `excluded` arm, seed-averaged. Four subjects support only the indicative across-subject interval of §4, so we show them individually. The free ridge and both capacity-matched readouts are nearly identical in every subject; the 513-parameter frozen readout is lower but positive in every subject; the two-parameter readout is negative in all four at near-OOD and near zero in distribution, where one subject is slightly negative.

column is reliable therefore depends on the quantity, OOD for the recovery curve, whose ordering is seed-stable from 33 parameters up; in distribution for the excess, whose two well-measured rungs are tight under every baseline. We report each where it is measured rather than declaring one column reliable throughout.

J WHERE THE CONTRAST SITS ACROSS CORTEX

An obvious deflationary reading of any encoding contrast is that it tracks signal-to-noise: the gain appears wherever prediction is already good. It does not here.

Using the per-parcel values of the capacity-matched readout (`excluded` arm, seed 0, subject-mean), the correlation between a parcel's trimodal predictability and its fusion gain is -0.038 in distribution and -0.020 at near-OOD, effectively zero. The gain is also broadly distributed rather than carried by a few parcels: 97% of parcels show a positive gain in distribution and 79% at near-OOD. Binning parcels by predictability gives a mildly inverted-U profile (in distribution, $+0.0175$, $+0.0283$, $+0.0320$, $+0.0196$ from worst- to best-predicted quartile), which we report without an account.

Two limitations bound this. At the freeze, per-parcel values were persisted only for the capacity-matched and frozen readouts; the R1 and R4-512 refits of §O now retain per-episode, per-parcel statistics, so the free ridge could be analysed this way and per participant, which we have not done, while the bank-only control remains unavailable. And this analysis is exploratory and post-hoc. It rules out the simplest SNR reading; it does not establish a positive account of where the contrast lives.

K CONFOUND ARMS ACROSS READOUTS

Table 10: Fusion gain under each confound arm. R2's gain uses R1's unimodal scores (Appendix F).

readout	split	excluded	partialled	competing band
R1 free ridge	val	0.0410	0.0288	0.0208
R2 banded	val	0.0432	0.0327	0.0228
R1 free ridge	ood	0.0500	0.0276	0.0226
R2 banded	ood	0.0507	0.0286	0.0237

A readout that tunes regularization per band and one that does not are affected almost identically: retention agrees to six points in distribution and one point out of it under partialling, and to two points on both splits under a competing band.

L CODE AND DATA AVAILABILITY

All analysis code, the pre-registration and its ten amendments with their dates, the self-audit ledger and the record of corrections, the frozen result rows behind every figure and table, the per-episode statistics and the scripts that reproduce every paired contrast and every figure from them, and the projection manifest are in the anonymised repository (its index is `README.md`):

`https://anonymous.4open.science/r/readout-capacity-fmri-review-07EB/`

The stimuli and fMRI responses are the Algonauts 2025 release (Gifford et al., 2025) of the Courtois NeuroMod data (Boyle et al., 2023) and are obtained from the original providers under their own terms; we redistribute neither. The frozen feature root is identified by a manifest of per-file hashes, which the repository includes so a reader can confirm they hold the same inputs without our having to ship them.

M THE FIXED-FEATURE EFFECTIVE-DIMENSION LADDER (R5)

R4-k confounds trainable parameters with the accessible random subspace. R5 separates them: the full feature matrix is held fixed (no projection) and only the ridge penalty varies, solved per subject so that the effective degrees of freedom $\sum_i s_i/(s_i+\alpha)$ land on the R4 rungs. α is therefore fixed by construction rather than cross-validated, a departure confined to R5. The sum runs over the eigenvalues s_i of the centered design's Gram matrix and excludes the unpenalized intercept, whereas R4-k's $k{+}1$ counts its intercept, so an R5 rung solved to 513 matches R4-512's 512 projected columns to within one degree of freedom. The runner records the achieved value beside each target; at every rung in Tables 2 and 12 it equals the target to four decimals. R4-k's own effective dimension is k less the effect of its 10^{-8} conditioning ridge, which lowers it by at most $k \cdot 10^{-8} \cdot \bar{s}/s_{\min}$, the ratio being that of the mean to the smallest eigenvalue of the projected Gram; we did not compute that ratio, so the count is exact to within 0.05 of a degree of freedom only if the ratio is below 10^4. We did not re-solve R5 at 512. The match is therefore approximate, and it is a match in one scalar: the two readouts still differ in subspace, in shrinkage and in the scale of their inputs, which is what the comparison is about. R5 is deterministic given the design, so it runs at one seed; the single-seed entries below carry no spread, which is not the same as a measured spread of zero.

This control was specified after the frozen results were seen, in response to a reviewer naming the objection, and is **exploratory throughout**. It names three outcomes in advance, of which "R4 more absorbing than R5" is the one that damages the ladder's reading.

Table 11: Fusion gain, and in parentheses the fraction of R1's gain it retains, at approximately matched effective dimension, `excluded` arm, subject-mean. Percentages are of R1's gain **in the same split and the same arm** (0.0410 val, 0.0500 near-OOD under `excluded`). Table 3 normalises differently, against the same readout's own `excluded` gain, so its entries for these readouts differ; both conventions are stated in §4.4. R4 is a three-seed mean.

	in distribution		near-OOD	
edof	R4 (projection)	R5 (penalty)	R4 (projection)	R5 (penalty)
2	$+0.0008$ (2%)	-0.0062 (-15%)	-0.0067 (-13%)	$+0.0101$ (20%)
9	$+0.0115$ (28%)	$+0.0010$ (2%)	$+0.0052$ (10%)	$+0.0140$ (28%)
33	$+0.0022$ (5%)	$+0.0119$ (29%)	$+0.0096$ (19%)	$+0.0231$ (46%)
129	$+0.0075$ (18%)	$+0.0225$ (55%)	$+0.0327$ (65%)	$+0.0375$ (75%)
513	$+0.0180$ (44%)	$+0.0327$ (80%)	$+0.0432$ (86%)	$+0.0482$ (96%)

Two normalisations, both used. Table 3 divides retained gain by the same readout's own `excluded` gain; the table above divides by R1's gain within the arm. The same cell is therefore 58% there and 66% here (0.0191 over 0.0327 against 0.0191 over 0.0288); both are correct, and both conventions are stated in §4.4.

Paired intervals on the R5 rungs. R5's runner retains the per-episode sufficient statistics, so its gains admit a bootstrap that resamples the same units (validation episodes; near-OOD segments) across all four conditions rather than combining intervals from separately fitted cells (§O). In the `excluded` arm the paired 95% BCa intervals exclude zero at rungs 33, 129 and 513 in both splits, and contain zero exactly at rung 9 in distribution and rung 2 near-OOD – the two rungs we already decline to interpret. Pairing matters: at 513 in distribution the interval on the *contrast* is 0.0040 wide against 0.0131 for the per-cell interval on the trimodal score alone.

The comparison holds under every confound arm. At the matched 513-parameter rung, R5 retains more gain than R4 in **all six arm × split cells**: 80% versus 44% and 96% versus 86% of R1's gain under `excluded`, 66% versus 0% and 93% versus 54% under `partialled`, and 65% versus −13% and 88% versus 48% under `competing band`. Under both confound corrections the projection readout's in-distribution gain reaches zero or below while the penalty readout retains about two thirds.

What the comparison rests on, and what it does not. Rungs 33, 129 and 513 carry it. Rungs 2 and 9 carry nothing: both gains sit near zero against R1's, R5's is negative at 2 in distribution, and at rung 9 R4's seed spread exceeds the R4–R5 difference. At 129 and 513, where R4's seed spread is smallest, the separation is far outside it. One thing does *not* generalize: R5's own capacity trend rises monotonically under `excluded` but not under `partialled`. The R4-versus-R5 direction is arm-invariant; the shape of the R5 ladder is not, and we claim only the former.

A mis-specified check, and what it revealed. A sixth rung at 4,000 was run first as a wiring check, on the stated expectation that a ridge at high effective dimension reproduces R1. It does not, and the expectation was wrong rather than the readout: the rung was chosen against the *nominal* design width, whereas R1's own effective dimension **at seed** 0 is 781.76 (V), 968.35 (A), 844.01 (T) and 1876.35 (VAT), which is 6.9% to 23.6% of nominal. Elsewhere we quote the mean over the three seeds, 782.78 for V, and the two figures differ only by that aggregation: the frozen and re-derived runs agree on this cell exactly at every seed. A penalty pinned at 4,000 is under-regularized relative to R1, and it scored below it: VAT 0.2626 against 0.2695 in distribution and 0.1644 against 0.1728 near-OOD. The corrected check is that the curve rises to the rung nearest R1's effective dimension in log spacing and recovers R1's score there; at 513 the unimodal conditions reach 0.972–1.000 of R1. We report the wiring rung separately and compute no absorption index from it. That R1 operates at 782–1,876 effective dimensions (seed 0) also means the R4 ladder, topping out at 513, stops below the range R1 actually uses; the free-ridge operating points themselves lie above its top rung.

The penalty ladder through the free-ridge range. Four further R5 rungs were solved at 768, 1,024, 1,536 and 2,048 effective dimensions on all four conditions (`excluded` arm, per-episode statistics retained), bracketing R1's cross-validated operating points, 782 (V), 968 (A), 844 (T) and 1,876 (VAT) on average at seed 0, with per-parcel values from 17 to 6,834. Paired with the R1 reference over the same units (validation episodes; near-OOD segments):

Table 12: R5 near and beyond the free-ridge operating range, validation split, `excluded` arm, paired with R1 over the 23 validation episodes (segment-level intervals in the release). All rows share one resample stream; the headline stream of Table 15 gives $[0.781, 0.814]$ for the 513 row, a difference of 0.001 from the draws alone. Unimodal scores are the subject mean.

edof	V	A	T	VAT	gain	retained of R1's +0.0410
513	0.2261	0.2041	0.1657	0.2588	+0.0327	0.798 [0.780, 0.815]
768	0.2274	0.2075	0.1669	0.2635	+0.0361	0.881 [0.865, 0.898]
1,024	0.2272	0.2087	0.1669	0.2660	+0.0388	0.948 [0.933, 0.967]
1,536	0.2248	0.2081	0.1653	0.2681	+0.0433	1.056 [1.039, 1.088]
2,048	0.2213	0.2056	0.1628	0.2682	+0.0469	1.145 [1.123, 1.184]
R1	0.2286	0.2099	0.1680	0.2695	+0.0410	

The R5 gain reaches R1's between 1,024 and 1,536 effective dimensions and exceeds it above, because the unimodal scores peak near 1,000 and decline while the trimodal score has not yet declined at 2,048. At the unimodal free-ridge operating points (782 to 968) R5 retains 88% to 95% of R1's gain, and at the trimodal one (1,876) it exceeds it. The penalty ladder therefore suppresses the gain

through $1{,}024$ effective dimensions and inflates it at $1{,}536$ and $2{,}048$; it matches R1 at no tested rung. The claims of §4.4 are claims about the suppression regime; the inflation above $1{,}024$ is a separate outcome, and the projection-structure difference of §E persists at the one budget tested there, $1{,}536$ columns, so the three are not one effect with one cutoff. Near-OOD, with 17 segments, R5 already matches R1 at 513 (0.964 $[0.886, 1.043]$, the headline stream of Table 15; this appendix's own stream gives $[0.890, 1.036]$) and exceeds it from $1{,}024$.

N DOES THE RELIABILITY GATE DECIDE THE RESULT?

The gate keeps the 834 of $1{,}000$ parcels clearing a split-half reliability floor of 0.10, estimated from eight repeat pairs that lie inside the near-OOD films. It is therefore in-sample for that split, and a reader is entitled to ask whether the contrast is a property of parcel selection rather than of the readouts.

It is not. Result rows store per-parcel scores for all $1{,}000$ parcels ungated (the gate is applied when the summary is computed, not when the row is written) so the threshold can be varied without refitting anything. At the frozen threshold the re-mask reproduces every stored value to 3.65×10^{-8}, so the sweep differs from the frozen pipeline in the threshold alone.

Table 13: Fusion gain by reliability threshold, `excluded` arm, subject-mean. Parcels retained in parentheses. The frozen threshold is 0.10. The R1 and R4-512 rows come from the refits of §O (R1 seed 0; R4-512 the three-seed mean), which retain per-parcel scores for all $1{,}000$ parcels; the last four rows are the fraction of R1's gain each readout retains at each threshold.

readout	split	0.00 (1000)	0.05 (980)	**0.10 (834)**	0.15 (653)	0.20 (487)
R1	val	0.0376	0.0380	**0.0410**	0.0443	0.0455
R1	near-OOD	0.0440	0.0447	**0.0500**	0.0567	0.0628
R3-up	val	0.0391	0.0395	**0.0423**	0.0455	0.0465
R3-down	val	0.0376	0.0380	**0.0407**	0.0435	0.0442
R3-up	near-OOD	0.0440	0.0447	**0.0501**	0.0577	0.0636
R3-down	near-OOD	0.0425	0.0432	**0.0484**	0.0555	0.0608
R4 (2 par.)	val	0.0008	0.0008	**0.0008**	0.0006	0.0013
R4 (2 par.)	near-OOD	−0.0060	−0.0060	**−0.0067**	−0.0080	−0.0085
R4-512	val	0.0165	0.0166	**0.0180**	0.0197	0.0203
R4-512	near-OOD	0.0382	0.0388	**0.0432**	0.0492	0.0543
R5-513	val	0.0307	0.0309	**0.0327**	0.0345	0.0348
R5-513	near-OOD	0.0428	0.0434	**0.0482**	0.0541	0.0590
R4-512 retained	val	0.438	0.437	**0.439**	0.444	0.446
R5-513 retained	val	0.816	0.814	**0.798**	0.779	0.763
R4-512 retained	near-OOD	0.867	0.868	**0.865**	0.868	0.865
R5-513 retained	near-OOD	0.971	0.970	**0.964**	0.953	0.940

With no gate at all the primary contrast remains clearly positive in both splits, and every R5 rung keeps its sign and its ordering. The comparison the paper rests on is unmoved: ungated, R5-513 retains 0.816 of R1's gain against R4-512's 0.438 in distribution and 0.971 against 0.867 near-OOD, and the R5-minus-R4 gain difference is $+0.0142$ and $+0.0046$ against $+0.0147$ and $+0.0050$ at the frozen threshold. The retained fractions move by at most 0.053 across the whole sweep, in distribution and near-OOD, while the absolute near-OOD gains rise by more than a third from no gate to the strictest threshold: the gate changes how much signal there is to fit, not how the readouts divide it. Gating monotonically *increases* the gain, which is the direction expected if it enriches for reliable signal rather than manufacturing an effect. Magnitudes do move (roughly 8% in distribution and 12% near-OOD between no gate and the frozen threshold) so what survives is the direction and the ordering, not the exact values.

Absorption is gate-robust too. R1 originally stored no per-parcel scores, so the ratio could not be re-gated. We re-ran R1 under its frozen configuration retaining them; it reproduces the frozen rows to 2.1×10^{-10} across all 120 cells, so it is the same result with one extra column. Absorption then sweeps as follows in the `excluded` arm: R3-down moves from -0.002 (no gate) to $+0.030$ (strictest) in distribution and stays within $[+0.023, +0.036]$ near-OOD; R3-up stays between -0.040

and -0.020 in distribution; R4's two-parameter rung stays near or above 1 throughout (0.971–0.987 in distribution, 1.133–1.142 near-OOD); and R5-513 stays far below it (0.184–0.237 against R4's 0.971–0.987 in distribution, 0.029–0.060 against 1.133–1.142 near-OOD). All three headline results therefore survive removing the gate entirely.

What this sweep cannot cover. The graded R4 rungs at 8, 32 and 128 parameters stored no per-parcel scores at the freeze and were not refitted, so the middle of the projection ladder is absent from the table; its top rung enters through the refit of §O. Its absorption sweeps as narrowly as R5-513's: 0.554–0.563 in distribution and 0.132–0.135 near-OOD. Leave-one-stimulus-out at the fitting stage would need refits and was not run; the scoring-stage version, every paired estimate recomputed with one segment dropped, is in §O. Absorption magnitudes drift with the threshold even where the ordering does not, and R4's near-OOD ratio exceeds one because its gain is negative there, which is arithmetic rather than a stronger effect. Retaining per-parcel predictions is a requirement we place on any confirmatory run.

O PAIRED INTERVALS ON THE FUSION GAIN

The per-cell intervals of this paper (Appendix B) are cluster bootstraps over stimulus segments computed independently for each cell. A fusion gain is a contrast between cells evaluated on the *same* segments, so an interval assembled from two of them treats shared segment-level variation as independent and overstates the uncertainty on the contrast.

The contrast can be resampled jointly without refitting anything, provided the per-segment sufficient statistics are kept: Pearson r over any subset of segments is a function of $(n, \sum x, \sum y, \sum x^2, \sum y^2, \sum xy)$, so one multiplicity vector applied to all four conditions is a genuinely paired resample. **The frozen readouts discarded these statistics**, so R1, R2, R3 and R4 cannot be paired retrospectively from the frozen rows; R5's runner retains them (gated parcels, `float32`), and R1 and R4-512 were later refitted with them retained (last paragraph). The arithmetic is otherwise a transcription of the per-cell estimator: the same per-subject Pearson from summed statistics, the same guard below three samples, the same order of averaging, the same BCa. The paired point estimates reproduce the reported gains to 10^{-11}, so each interval surrounds precisely the quantity beside it.

Table 14: Paired 95% BCa intervals on the R5 fusion gain, `excluded` arm, 1,000 resamples with the four subjects held fixed, over the 23 validation episodes in distribution and the 17 segments near-OOD. Intervals containing zero in bold.

edof	in distribution		near-OOD	
	gain	paired interval	gain	paired interval
2	-0.0062	$[-0.0101, -0.0023]$	$+0.0101$	$[\mathbf{-0.0015}, \mathbf{+0.0186}]$
9	$+0.0010$	$[\mathbf{-0.0023}, \mathbf{+0.0046}]$	$+0.0140$	$[+0.0025, +0.0224]$
33	$+0.0119$	$[+0.0093, +0.0145]$	$+0.0231$	$[+0.0121, +0.0312]$
129	$+0.0225$	$[+0.0204, +0.0244]$	$+0.0375$	$[+0.0280, +0.0447]$
513	$+0.0327$	$[+0.0306, +0.0346]$	$+0.0482$	$[+0.0428, +0.0534]$

The two intervals containing zero fall at rung 9 in distribution and rung 2 near-OOD, the rungs we already decline to interpret, here confirmed by a paired interval rather than by an argument about seed spread. Rungs 33, 129 and 513 exclude zero in both splits.

The same pattern holds under the other two arms: rungs 33, 129 and 513 exclude zero everywhere, while rungs 2 and 9 contain zero in `competing band` as well. Under `partialled` every rung excludes zero, but the ladder there is *non-monotone* in distribution (0.0235, 0.0204, 0.0146, 0.0116, 0.0191) and the paired intervals are about 0.004 wide, so noise does not explain it. The capacity ordering that is clean in the `excluded` arm does not survive partialling, and we report that without an account.

Pairing is not cosmetic. At 513 in distribution the interval on the contrast is 0.0040 wide, against 0.0131 for the per-cell interval on the trimodal score alone: shared segment-level variation cancels in a contrast, and does not cancel when two independent intervals are combined. The resample is

over validation episodes, or near-OOD segments, with the four subjects held fixed, so these intervals are conditional on those subjects exactly as the others are.

The R4-versus-R5 difference, paired. R1 and R4-512 discarded their per-segment statistics at the freeze. For this comparison both were refitted with the statistics retained, under the frozen code path and the frozen projection matrices. The R4-512 refit reproduces its frozen rows to 2×10^{-16} in all 24 cells and the R1 refit to 6×10^{-14}, so the intervals below are on the published fits. One multiplicity vector per resample is applied to R1's four conditions, R4-512's four conditions at three seeds (the gain averaged over seeds within the resample) and R5-513's four conditions; `excluded` arm, 1,000 resamples, BCa.

Table 15: Headline comparison, paired. In distribution the resample is over the 23 validation episodes (the segment-level intervals, a sensitivity check, move no boundary by more than 0.009; see the paragraph on segments below); near-OOD it is over the 17 segments of two films. Retained fraction is the readout's gain over R1's within the same resample. R1 is seed 0 (its seed spread is below 0.0001).

	in distribution (23 episodes)	near-OOD (17 segments, two films)
gain, R1	$+0.0410$ $[+0.0387, +0.0430]$	$+0.0500$ $[+0.0470, +0.0539]$
gain, R4-512	$+0.0180$ $[+0.0163, +0.0196]$	$+0.0432$ $[+0.0410, +0.0469]$
gain, R5-513	$+0.0327$ $[+0.0306, +0.0346]$	$+0.0482$ $[+0.0428, +0.0534]$
retained, R4-512	0.439 $[0.415, 0.464]$	0.865 $[0.826, 0.950]$
retained, R5-513	0.798 $[0.781, 0.814]$	0.964 $[0.886, 1.043]$
gain difference, R5 − R4	$+0.0147$ $[+0.0136, +0.0158]$	$+0.0050$ $[+0.0001, +0.0088]$
retained difference, R5 − R4	$+0.359$ $[+0.336, +0.380]$	$\mathbf{+0.100}$ $[\mathbf{-0.005}, \mathbf{+0.181}]$

In distribution the difference excludes zero by a wide margin on both scales. Near-OOD the gain difference just excludes zero and the retained-fraction difference does not, so the near-OOD ordering of the two instruments is not established at 17 segments; the paper's near-OOD claims already stop short of it.

How much the near-OOD boundaries owe to the interval construction. With 17 segments a BCa interval's bias and acceleration terms are fitted from a jackknife of 17 re-estimates, and a reader is entitled to ask whether a boundary that just clears or just misses zero is a property of the data or of the correction. Table 16 answers it from the same 1,000 paired resamples summarised two ways, BCa and plain percentile, beside the 17 leave-one-segment-out re-estimates themselves (the BCa jackknife, reported rather than folded into a constant). No gain or difference changes sign when any single segment is dropped; the R5-minus-R4 gain difference stays between $+0.0040$ and $+0.0071$. The two constructions disagree on the one boundary the paper already declines to lean on: the retained-fraction difference excludes zero under the percentile interval and not under BCa, and the percentile intervals sit lower than their BCa counterparts for the three gains and the two retained fractions and higher for the two differences, which is the skew of the bootstrap distributions that BCa exists to correct. We keep BCa, the pre-specified construction, and read the disagreement as the resolution the split has. In distribution the two constructions agree to 0.0003 on every gain boundary, and the leave-one-out range of the R5-minus-R4 difference is a tenth as wide (it stays between $+0.0146$ and $+0.0149$).

Table 16: Near-OOD headline comparison under two interval constructions and leave-one-segment-out, `excluded` arm, 17 segments of two films, the same 1,000 paired resamples as Table 15. LOO range is the minimum and maximum of the 17 re-estimates.

	point	BCa 95%	percentile 95%	LOO range
gain, R1	$+0.0500$	$[+0.0470, +0.0539]$	$[+0.0435, +0.0524]$	$[+0.0489, +0.0511]$
gain, R4-512	$+0.0432$	$[+0.0410, +0.0469]$	$[+0.0352, +0.0448]$	$[+0.0416, +0.0439]$
gain, R5-513	$+0.0482$	$[+0.0428, +0.0534]$	$[+0.0386, +0.0525]$	$[+0.0465, +0.0498]$
gain difference, R5 − R4	$+0.0050$	$[+0.0001, +0.0088]$	$[+0.0016, +0.0102]$	$[+0.0040, +0.0071]$
retained, R4-512	0.865	$[0.826, 0.950]$	$[0.775, 0.908]$	$[0.835, 0.880]$
retained, R5-513	0.964	$[0.886, 1.043]$	$[0.879, 1.036]$	$[0.947, 0.975]$
retained difference, R5 − R4	$+0.100$	$[\mathbf{-0.005}, \mathbf{+0.181}]$	$[+0.035, +0.208]$	$[+0.081, +0.140]$

Segments are not episodes, and near-OOD segments are not films. The 50 validation units are half- and quarter-episodes of 23 Friends episodes, and the 17 near-OOD units are segments of two films (12 of *Hidden Figures*, 5 of *Life*). Segments of one episode or one film share structure, so a segment-level resample treats correlated units as independent and can understate uncertainty. Table 15 therefore resamples the 23 validation episodes (every segment of a resampled episode enters with its episode's multiplicity, and the jackknife leaves one episode out) and that interval is the one quoted in the main text. The segment-level resample is the sensitivity check: it moves each gain boundary by at most 0.0008 (the R5-minus-R4 gain difference is $[+0.0135, +0.0158]$ over segments against $[+0.0136, +0.0158]$ over episodes, and the retained-fraction difference $[+0.335, +0.383]$ against $[+0.336, +0.380]$), and no boundary by more than 0.009 (R4-512's retained fraction). The R5 rungs, the R5 range, the projection controls and the network contrasts were re-clustered the same way from their retained statistics and are quoted at the episode level in distribution; no interval changed its position relative to zero, and the largest move of any bound was 0.0008 on a gain and 0.0166 on a retained fraction of the R5 range. The fixed-feature excess intervals of §4 were recomputed by a retained script and re-clustered the same way (below); only the per-cell intervals of Appendix B remain segment-level. Two films do not support a film-level interval, so near-OOD we report the contrast within each film (Table 17): R5-513 retains more of R1's gain than R4-512 in *Hidden Figures* and in *Life* taken alone. No near-OOD interval in this paper should be read as evidence about films the participants did not watch.

Table 17: The near-OOD headline comparison per film, `excluded` arm, subject mean; each film's column uses every segment of that film and nothing else. Point estimates only: two films do not support a film-level interval.

	all 17 segments	*Hidden Figures* (12)	*Life* (5)
gain, R1	$+0.0500$	$+0.0428$	$+0.0474$
gain, R4-512	$+0.0432$	$+0.0316$	$+0.0372$
gain, R5-513	$+0.0482$	$+0.0353$	$+0.0465$
gain difference, R5 − R4	$+0.0050$	$+0.0037$	$+0.0093$
retained, R4-512	0.865	0.739	0.784
retained, R5-513	0.964	0.824	0.980
retained difference, R5 − R4	$+0.100$	$+0.086$	$+0.196$

Harmonizing the preprocessing. The headline comparison differs in three things at once: R4 projects raw feature columns where R5 works on standardized ones (input scaling), a random subspace against the full design (subspace), and ordinary least squares against a ridge (regularization form). The standardized-input R4-512 cells of §E remove the first, and their retained per-segment statistics pair with R1 and R5-513 over the same resamples (Table 18). Standardization raises R4-512's in-distribution gain by $+0.0025$ $[+0.0016, +0.0033]$ and lowers its near-OOD gain by 0.0107 $[-0.0148, -0.0028]$, so the harmonized R4-512 retains 0.499 of R1's gain in distribution and 0.650 near-OOD. The difference to R5-513 survives on both splits and on both scales, and near-OOD it is larger than the raw comparison's, not smaller. The subspace and the regularization form remain confounded with each other in this comparison; separating them needs a ridge fitted on the random projection, which we have not run.

Table 18: R4-512 with inputs standardized before projection, paired with R1 and R5-513, `excluded` arm; three seeds averaged within each resample. Validation carries the segment-level and the episode-level (23 clusters) interval.

	in distribution (50 segments, 23 episodes)			near-OOD (17 segments)	
	point	segment BCa	episode BCa	point	segment BCa
gain, R4-512 std.	$+0.0204$	$[+0.0184, +0.0231]$	$[+0.0185, +0.0224]$	$+0.0325$	$[+0.0251, +0.0378]$
retained, R4-512 std.	0.499	$[0.464, 0.534]$	$[0.468, 0.526]$	0.650	$[0.535, 0.784]$
gain diff., R5 − R4-512 std.	$+0.0122$	$[+0.0113, +0.0131]$	$[+0.0114, +0.0132]$	$+0.0157$	$[+0.0051, +0.0215]$
retained diff., R5 − R4-512 std.	$+0.299$	$[+0.274, +0.322]$	$[+0.276, +0.323]$	$+0.315$	$[+0.112, +0.457]$
gain, R4-512 std. − raw	$+0.0025$	$[+0.0016, +0.0033]$	$[+0.0016, +0.0034]$	-0.0107	$[-0.0148, -0.0028]$

The excess over common scaling, recomputed. The script that produced the excess intervals of earlier drafts was not retained. The excess (§3: absorption minus $1-\lambda$, mean-of-unimodals baseline,

$\lambda = r_{\mathrm{VAT}}(k)/r_{\mathrm{VAT}}(\mathrm{R1})$) was therefore transcribed onto the paired machinery of this appendix. Its point estimates reproduce the ledger's to within 0.0005 (the ledger assembled each excess from four-decimal intermediate values; the retained script works at full precision, and §4 now quotes its values). Its intervals supersede the earlier ones, which are recorded in the ledger: in distribution the retained computation gives narrower intervals at 129 and 513 (Table 19), every in-distribution interval still excludes zero under the max baseline and under λ defined on the mean unimodal score, and near-OOD the verdicts at 33, 129 and 513 are unchanged (under the max baseline alone the near-OOD 129 interval excludes zero, $[+0.0064, +0.2816]$). One verdict changes: the in-distribution minus near-OOD difference, from independent resamples of the two splits with a jackknife over the union of their units, now excludes zero at 129 as well as at 513, where the earlier interval at 129 spanned zero. The four-subject R3 intervals of §4 are two-sided 95% Student-t intervals over the four subject-level absorption values, each the mean over three seeds; the retained script that reproduces them from the frozen per-subject rows is in the release.

Table 19: The excess over common scaling on the fixed-feature ladder, `excluded` arm, mean-of-unimodals baseline. Validation: BCa over the 23 episodes, with the segment-level interval in the next column; near-OOD: BCa over the 17 segments of two films. Last column: in-distribution minus near-OOD.

edof	val excess	val, segment-level	near-OOD excess	val − near-OOD
33	$+0.1441$ $[+0.1302, +0.1572]$	$[+0.1294, +0.1580]$	$+0.1333$ $[+0.0778, +0.1763]$	$+0.0108$ $[-0.0342, +0.0676]$
129	$+0.0998$ $[+0.0864, +0.1101]$	$[+0.0867, +0.1118]$	$+0.0435$ $[-0.0027, +0.0923]$	$+0.0563$ $[+0.0065, +0.1064]$
513	$+0.0668$ $[+0.0584, +0.0740]$	$[+0.0591, +0.0752]$	$+0.0007$ $[-0.0339, +0.0349]$	$+0.0661$ $[+0.0311, +0.1008]$

P THE WIDTH NULL, TESTED DIRECTLY

R3 compresses an informative design; it does not widen an uninformative one, so on its own it tests only half the width null. We test the other half by padding the best unimodal stream to the trimodal width with columns carrying the right marginals and no stimulus alignment, fitting R1 exactly as the primary grid fits it, with alpha **re-selected** on the padded design by the same inner segment-level cross-validation over the same grid. Two paddings: `perm` takes the A and T blocks of a different episode under a per-seed derangement, preserving column marginals and within-episode autocorrelation and destroying only alignment; `noise` uses Gaussian columns matched to each stimulus's own A and T column mean and variance. `perm` is primary because it is the harder null.

Table 20: The width null, `excluded` arm, three seeds. Effective dimension is divided by *that subject's own* V value at the *same seed*, never by a subject mean.

condition	val r	near-OOD r	edof ratio to own V	max pinned
V alone	0.2286	0.1228	1.00	
V padded, `perm`	0.2177	0.1222	1.96 $[1.90, 2.00]$	0.12%
V padded, `noise`	0.2169	0.1233	3.27 $[3.20, 3.37]$	0.00%
VAT	0.2695	0.1728		

The null's premise holds and its conclusion fails. Padding does buy the readout freedom: effective dimension rises to 1.90 to 2.00 times each subject's own V value under permutation, and 3.20 to 3.37 under Gaussian noise. The score does not follow, falling 0.011 in distribution and moving within seed noise near-OOD. This is what a ridge readout does when columns carry real spectral mass and no signal: they raise effective dimension at any fixed penalty and add variance, and cross-validation offsets that only partly. Edge pinning never exceeded 0.12% against a 2% tolerance, with zero widenings, so the result does not depend on the grid ceiling.

Near-OOD is unchanged in the strict sense. The two paddings straddle V: `perm` sits 0.0006 below it and `noise` 0.0005 above, against seed ranges 0.0004 wide. The near-OOD change is at most 0.0006 and its sign differs between the two paddings, so no effect in either direction is supported.

The two paddings differ in how much freedom they buy, which is arithmetic rather than a finding. Real feature blocks have concentrated spectra, so few directions carry the mass; Gaussian columns

spread it flatly across many. Permutation preserves that concentration, which is why the harder null adds less effective dimension than the easier one.

Q THE TWO FAILURE CLASSES, IN FULL

An entry made after the freeze: the replication's R3 cells were fitted twice. The first pass omitted the normalisation key Amendment 7 fixed (§A) and so ran the retired estimator, and a verdict was printed on those cells. It was voided the same day: the signature was unmistakable once looked for, effective dimension roughly tripled against the free ridge rather than sitting beside it, and the per-cell records state the setting, which is how it was caught. All twelve cells were refitted under the registered estimator and the verdict was run a second time, for that recorded cause and under unchanged rules; both runs are in the released ledger. The retired rows are kept as an audit trail and enter nothing, exactly as the primary set's pre-amendment R3 rows were. The defect was a configuration that omitted a key and inherited a default, so the conformance check now asserts the field on every R3 configuration rather than only the cell set: a check that verifies which cells ran cannot see which estimator ran them.

Table 21: The ledger at `results-freeze-v1`, by class. A reader who stops here has the account; the rest of this appendix is the detail behind it.

class	entries	remedy, and why it is incomplete
Silent narrowing	8	scope recorded per value; composition of scoped values still unchecked
Silent widening	1	the checker reports every use of a scoped value instead of certifying it
Cause asserted, later disproved	4	measure before attributing; three were our own corrections
Other incidents	20	per-cell checkpointing, cgroup-aware budgets, hash manifests
Later corrections	3	entries are appended, never rewritten
total	36	

At `results-freeze-v1` the ledger holds thirty-six entries: thirty-three incidents and three later corrections to earlier ones. The fixed-feature control (§M) added four more, including a fourth member of the collision family below, found by a guard rather than a reader, and changing no frozen number. Eight were **silent narrowing**, and a ninth was its mirror, a claim that silently widened, described in §7. The eight narrowings: an entire confound arm dropped from the analysis by a substring match that also matched the arm's own name; two distinct readouts colliding on one label; a configuration hash omitting capacity and matching direction; a partialling guard covering one code path and not the adjacent one, which reversed a prior internal conclusion once fixed; container tooling reporting host rather than cgroup resources, overstating capacity 4–$8\times$ for eight days; a memory guard sampling once against a peer whose footprint oscillated fivefold; and a configuration generator whose substring guard matched a comment it had itself just written, so that 28 cells ran under the setting an amendment existed to change; and, after the freeze, the verified-facts file that governs every number in this paper grouping on `readout` and thereby pooling all five graded rungs into one row, the same collision recorded two incidents earlier, reproduced inside the artifact built to prevent it. It was found by an automated agent reconciling printed gains against the score table, and initially misclassified by a human as a fabrication. The recompute against the frozen tag vindicated every affected row; the incident is retained regardless.

Four further entries record a **cause asserted and later disproved**. A per-cell cost overrun was attributed to band count; measurement showed per-attempt cost is identical at two bands (2982.8, 2973.4, 4220.5 s) and three (2981.2, 3074.8, 2984.4 s), and the real causes were an alpha-grid widening and an unnoticed change of GPU model. A truncated binary was attributed to filesystem corruption; the identical signature later appeared on a machine that had never rolled back. A data transfer was reported as having moved nothing, on the strength of a recency search that could not have detected it because the copy preserved source timestamps.

The mirror case in full, and why its remedy is incomplete. A figure caption in an earlier version of this paper asserted a seed spread of ≤ 0.0006 across the frozen ladder. That number is correct and appears verbatim in the frozen source, it was computed on the free and capacity-matched readouts, where it holds. Quoted over the frozen ladder it is false by a factor of thirty: the middle rungs reach 0.0181, larger than the gain the figure qualifies. Our mechanical check passed it, and could not have done otherwise, because the check verifies that a number exists in the frozen source and jurisdiction is not a property a string match can see. An external reader caught it by comparing the claim against the scope that licensed it. The remedy is not a better matcher: the frozen source now records, for each such value, what it is true of and what it must not be quoted for, and the checker reports every use of one instead of certifying it. **A verification tool inherits the scope of its source, and we had not been recording scope.** That remedy is itself incomplete, and we say so because our own next incident defeated it: a later draft paired an observed value computed under one baseline with a prediction computed under another, both individually in scope. **Composition of licensed values is not itself licensed**, and a per-value scope table does not see pairings.

We record the second class beside the first because an audit that catches every anomaly and misexplains a third of them is not obviously better than one that catches fewer. The shared shape is *attributing a residual to the most salient recent change, and treating the absence of a signature the event does not produce as evidence the event did not occur.*

R Does partialling change the conditioning of the design?

Section 4.4 reports that the fraction of gain the confound correction removes differs across readouts below common width. That descriptive claim needs no assumption about conditioning. Reading the removed fraction as *low-level signal* rather than as capacity does: it assumes the correction removes variance without changing the *conditioning* of the design. If partialling reshaped the spectrum, a fixed ridge penalty would buy a different effective capacity before and after, and part of what we charge to the confound would be a capacity change. This appendix tests that assumption rather than asserting it.

The design's Gram is decomposed before and after partialling, for V and VAT, over the three distinct training row sets: two subjects contribute identical rows as a multiset, and the other two differ from them by one and by three episodes respectively, which is the whole of the between-subject variation the design has. The reported quantity is effective degrees of freedom at a *fixed* penalty, which asks the question in the units the concern is about. The interpretation rule, including its thresholds and its disconfirming direction, was registered and tagged before any VAT cell completed, and the verdict below is computed by the script that transcribes it rather than read off the table.

Table 22: Effective degrees of freedom at a fixed ridge penalty, before and after partialling, VAT condition. Entries are the relative change $(\text{partialled} - \text{excluded})/\text{excluded}$. Designs are the three distinct training row sets; `sub-01` stands for the design it shares with `sub-02`.

design	$\alpha = 10^2$	$\alpha = 10^3$	$\alpha = 10^4$	$\alpha = 10^5$
`sub-01`	$+0.009$	$+0.045$	$+0.126$	$+0.202$
`sub-03`	$+0.009$	$+0.045$	$+0.126$	$+0.202$
`sub-05`	$+0.009$	$+0.046$	$+0.126$	$+0.202$

The registered verdict is *unresolved*. The rule called an effect material if the relative change exceeded 0.10 at a majority of the four penalties with the sign agreeing across designs, and immaterial if it stayed below 0.05 throughout. The sign agrees everywhere, but the change exceeds 0.10 at two of four penalties, which is not a majority, so neither branch is satisfied. We report the table and make no claim in either direction.

Two things we deliberately do not do with this table. The change grows with the penalty, from $+0.009$ to $+0.202$; that pattern is a post-hoc observation the rule neither anticipated nor adjudicates, and converting it into a finding by selecting the penalties where it is large is exactly the choice the advance registration existed to remove. And the participation ratio moves much more than the effective dimension does under VAT (82.8 to 549.5); it is a shape descriptor, not the registered quantity, and it cannot carry a conclusion the registered test declined to support. The V condition

is reported as exploratory, because its first two cells had printed when the rule was filed, which the registration discloses.

What this leaves is an assumption checked and not vindicated, rather than an assumption checked and confirmed: the preregistered decision rule returned *unresolved* at its stated thresholds. The descriptive readout dependence of Section 4.4 stands on its own; the reading of the removed fraction as low-level signal rests on a conditioning assumption the registered test did not settle.

The fixed-penalty objection, answered post hoc. A reader may object that a fixed absolute penalty cannot test conditioning: residualising on the bank removes variance, lowers the Gram's eigenvalues, and so must lower the effective dimension at any fixed α for reasons of scale alone. The objection does not apply to this table, and the table shows it. First, the diagnostic standardizes every column *after* the arm is applied, so the trace of the Gram agrees between the arms to 2.5×10^{-5} relative and the comparison is trace-matched by construction. Second, the effective dimension *rises* under partialling, which a pure residualisation, a projection that can only lower each eigenvalue, could not produce; what the table records is the re-standardised design, in which the variance removed along the bank's directions is returned to the columns and spread over the rest. That is a change of shape, not of scale: for VAT the top eigenvalue falls $7.3\times$ (2.17×10^8 to 2.95×10^7), the share of the trace in the top 33 directions falls from 0.355 to 0.188, and the participation ratio rises from 82.8 to 549.5. It is the change the registered rule was designed to detect and was too coarse to call. The mirror question is answered in the same file: the penalty that places a ridge at 513 effective dimensions is $1.32\times$ larger in the `partialled` arm, so an R5 rung solved to the same effective dimension in the two arms is not the same shrinkage profile, which is the candidate account of the non-monotone `partialled` ladder (§6) that we do not test. R5 itself is untouched by the objection, since it solves for a target effective dimension per design and per arm rather than fixing α. The registered verdict stands as printed; this paragraph is post hoc.