VARIANCE IS NOT INFORMATION: WHY PRIORS OVER EEG CANNOT CLOSE THE TRIAL-AVERAGING GAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Every headline number on THINGS-EEG2, the standard benchmark for visual decoding from EEG, is computed on responses averaged over 80 repetitions of the same stimulus. The same frozen encoder scores 0.841 at 80 repetitions and 0.158 at one, and two reproduced 2026 baselines keep 10% and 16% of their headline number at one repetition. We ask whether that 0.683 gap can be closed algorithmically and answer no, in a strong sense: classical MMSE estimators recover none of it, and the *provably optimal* diffusion prior over EEG, which is available in closed form and hence immune to a training-quality objection, recovers none of it and at $R \geq 4$ actively hurts. One measurement explains why. Ordering the response's principal directions by variance, the 32 leading directions carry 43.7% of the estimated signal energy, retrieve at chance in isolation, and *deleting them is free*; keep-k accuracy does not saturate until $k \approx 4096$. The discriminative signal is high-rank and anti-aligned with variance, so any prior, estimator or likelihood-maximisation step that concentrates its estimate where the variance is allocates capacity where the information is not. Consistently, the estimator $17\times$ closer to the high-SNR response in NMSE loses four points of Top-1. Both findings replicate on THINGS-MEG. The same principle governs the image side: multi-view target augmentation is worth $+0.278$, yet an operator fitted to reproduce that target reaches cosine 0.984 on held-out images while delivering 16.7% of its benefit, making it nonlinear adds $+0.0035$ Top-1, and training it on the retrieval loss instead gives a family whose optimum is the identity. Reproducing a target and reproducing its value are different problems. Progress on this benchmark is measurement-bound rather than model-bound.

1 INTRODUCTION

A visual decoder for EEG is trained by aligning a brain response to the CLIP embedding of the stimulus that evoked it, and is scored by retrieving the correct image out of 200 candidates. On THINGS-EEG2, published Top-1 has climbed from the high forties to the low sixties over two years, and to the mid eighties under the strongest target-side pipelines. This paper asks what that number is measuring. The test set presents each of 200 images eighty times, and every published number we are aware of is computed on the average of all eighty responses. That average is not a detail of the evaluation protocol; it is most of the signal. Holding the encoder fixed and varying only how many repetitions are averaged at test time, Top-1 moves from 0.158 at one repetition to 0.841 at eighty. The 0.683 that separates these is larger than the entire span of published progress on the benchmark.

Figure 1 summarises the analysis framework. The natural response is that this is an algorithmic opportunity: a sufficiently good prior over clean EEG should recover part of the difference without paying for the scanner time. This is the idea behind a family of recent proposals, and we show that it fails: by an upper bound for the diffusion-prior family, and by measured nulls for the estimators and denoisers we test. The strength of that closure matters, because "we tried a method and it did not work" is worth very little. For the diffusion-prior family we train nothing: Chen et al. (2024) give the optimal diffusion model over a finite reference set in closed form, so the *best possible* prior over EEG responses can be evaluated directly. Sweeping its two parameters over 28 cells and selecting

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

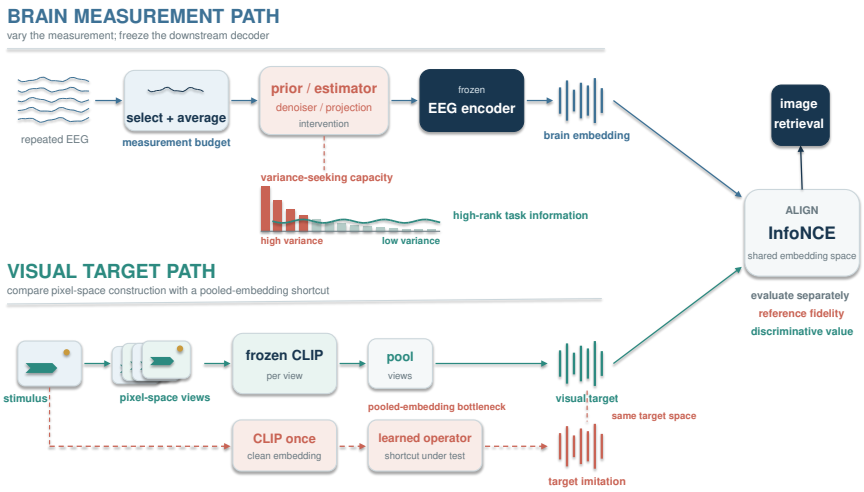

Figure 1: **Analysis framework.** The standard contrastive decoder has two paths. On the brain path, we vary the measurement budget, insert candidate priors or estimators before a frozen encoder, and probe the variance spectrum directly. On the visual path, we compare the pixel-space multi-view constructor with a shortcut that attempts to produce the same target from one pooled embedding. Both paths meet in the same contrastive objective, letting us distinguish reference fidelity from discriminative value.

the best cell *on the test metric*, the optimum sits at the degenerate corner where the update is nearly the identity; even there it loses 0.025 at $R = 8$, and cells away from that corner lose up to 0.44. When the best configuration of a method family is "do as little as possible," the family is misdirected rather than underfit. Classical estimators, learned denoisers and budget-aware retraining all concur.

One measurement explains why. Order the response's principal directions by variance and delete from the top: the leading 32 carry 43.7% of the estimated signal energy, and deleting them changes nothing (0.154 against 0.156 at one repetition), while keeping only the top k does not approach the full-basis number until $k \approx 4000$. The discriminative content is spread thinly over thousands of low-variance directions and is *anti*-aligned with the variance ordering. MMSE estimators, diffusion priors, likelihood maximisation and principal-subspace projection all concentrate an estimate in high-density regions, which here discards signal and retains nuisance; the prediction that reconstruction fidelity trades *against* retrieval accuracy holds, the estimator $17\times$ closer to the eighty-repetition response in NMSE losing four points of Top-1. Gap and dissociation both replicate on THINGS-MEG.

The same principle governs the other side of the pipeline, where the field's real progress has come from. Building the CLIP-side target by averaging embeddings of many degraded views is worth $+0.278$, and because that pipeline is hand-designed with no stated principle we try to learn it. The least-squares-optimal linear map reaches cosine 0.984 on held-out images and delivers 16.7% of the benefit; a nonlinear map that starts at the linear solution and is verified to fit better out of sample buys $+0.0035$ Top-1 in 5 of 10 subjects; and training the same operator on the retrieval loss alone gives a family whose optimum is the identity (0.565 against 0.569). Both objectives available without running the augmentations fail, which moves the suspicion from the loss to the pooling. One result strains this account and we report it: on THINGS-MEG the nonlinear null replicates exactly, but a linear operator recovers 58% rather than 17%. Finally, the one place where new information is available is across subjects: ten subjects at eight repetitions reach 0.973 where one subject at eighty reaches 0.841, yet four attempts to distil this into a single-subject encoder are null, for a reason we offer as a hypothesis. Taken together, what has been improving on this benchmark is not robustness to neural noise but the quality of the visual target, measured under an averaging regime that removes most of the noise before the model is asked to handle any.

Contributions.

1. **The trial-averaging gap, on three encoders.** Top-1 falls from 0.841 to 0.158 from 80 repetitions to 1. No published method we surveyed reports this curve; on the frozen checkpoints of two faithfully reproduced 2026 baselines (62.6 vs 62.3; 59.2 vs 60.2), one repetition retains 10% and 16% of the headline number.
2. **A closure with an upper bound.** The closed-form optimal diffusion prior over EEG recovers none of the gap and at $R \geq 4$ hurts; classical estimators, learned denoisers and distillation concur, and budget-aware retraining trades rather than adds.
3. **The mechanism.** Discriminative signal is high-rank and anti-aligned with variance; the top 32 directions hold 43.7% of signal energy and are free to delete. Fidelity anti-correlates with retrieval. Both replicate on THINGS-MEG.
4. **The same principle on the target side, and one ruled-out fix.** An operator at cosine 0.984 from the multi-view target delivers 16.7% of its benefit; a nonlinear operator adds $+0.0035$; the retrieval-trained family's optimum is the identity. The nonlinear null replicates on MEG; the 16.7% does not (58%), and we report the discrepancy.
5. **Cross-subject information, and the limits of distilling it.** 0.973 vs 0.841 at matched budget; four distillation variants null, with a hypothesis for why.

2 Setup and reproductions

Task and data. THINGS-EEG2 (Gifford et al., 2022) presents 16,540 training images (4 repetitions) and 200 test images (80 repetitions) to 10 subjects, over 63 electrodes sampled at 250 Hz in a 1 s post-stimulus window. The decoder is trained to align the EEG response with a CLIP (Radford et al., 2021) embedding of the stimulus under a symmetric InfoNCE loss (van den Oord et al., 2018), and is evaluated by 200-way Top-1 retrieval. THINGS-MEG (Hebart et al., 2023) supplies 4 subjects, 271 magnetometers, 201 samples, and 12 test repetitions.

Why the strongest target matters. Recent methods build the CLIP-side target not from the clean image but from an average of embeddings of several degraded views of it. We adopt the strongest such pipeline available to us, a 24-level blur-and-noise pool, because an analysis of what limits this literature should be run at its ceiling. All encoder arms differ only in the component under study; the architecture, optimiser (AdamW, 10^{-3}, weight decay 0.01), batch size, 200-epoch schedule, and validation-selected checkpoint are held fixed.

Evaluation protocol. Throughout, an accuracy at budget R means: draw a random size-R subset of the available test repetitions, average them, and retrieve. Trained arms use 5 draws with a fixed seed so that every arm sees identical draws; the frozen-encoder curves of Section 4 use 20 independent draws per subject. We report the whole trials-to-accuracy curve for every arm so that no arm can buy low-budget accuracy by quietly surrendering the high-budget number.

Reproductions. So that we can modify published methods rather than only cite their numbers, we reproduced two 2026 baselines end to end from their released code: NeuroBridge (Zhang et al., 2026) at 62.6 against a published 62.3, and ATS (Wu et al., 2026) at 59.2 against a published 60.2. Section 7 uses them to vary the number of averaged views inside NeuroBridge's own code.

3 Related work

Contrastive visual decoding from EEG. The pipeline this paper examines was established by NICE (Song et al., 2024), which paired a temporal-spatial convolutional EEG encoder with a frozen CLIP image encoder (Radford et al., 2021) and aligned them by InfoNCE (van den Oord et al., 2018), reporting 200-way zero-shot retrieval on THINGS-EEG2 (Gifford et al., 2022). ATM (Li et al., 2024) generalised the encoder across sources, extended the framework to MEG, and added a diffusion stage for image generation. The two 2026 systems we reproduce continue this line: NeuroBridge (Zhang et al., 2026) fuses CLIP features of the stimulus with those of several perturbed variants, and ATS (Wu et al., 2026) adapts the strength of the visual teacher to the asymmetry

Table 1: Trials-to-accuracy on THINGS-EEG2, 10 subjects, three frozen encoders. Within each row only the number of averaged test repetitions changes; rows are not comparable in absolute terms because each baseline is scored in its own pipeline. Every headline number we are aware of sits at the right edge.

encoder \ repetitions R	1	2	4	8	16	32	80
ours (released checkpoint)	**0.158**	0.273	0.445	0.601	0.723	0.792	**0.841**
NeuroBridge (Zhang et al., 2026)	0.062	0.108	0.192	0.309	0.433	0.539	0.626
ATS (Wu et al., 2026)	0.095	0.163	0.259	0.372	0.467	0.537	0.592

between modalities. HyFI (Jo et al., 2026) interpolates semantic and perceptual image features along a hyperbolic geodesic, and Brain-HIVE (Zheng et al., 2026) aligns to a single token fused from several visual encoders. Every one of these works reports accuracy on the average of all 80 test repetitions, and none reports how that accuracy varies with the repetition budget; Section 4 measures what that choice hides.

Progress has been target-side. Read as a group, the recent gains on this benchmark come less from the brain encoder than from how the CLIP-side target is built. NeuroBridge's perturbed variants and ATS's shrunken teacher are both operations on the image embedding, and our view-count ladder inside NeuroBridge's own code shows its accuracy rising monotonically with the number of perturbed views. Section 7 asks what that operation *is*.

Generative priors and shrinkage estimators for noisy measurements. The proposal that a learned prior over clean measurements should substitute for repeated acquisition has a long lineage, from James–Stein shrinkage (James & Stein, 1961) through empirical-Bayes denoising (Efron, 2011) to diffusion models used discriminatively (Ho et al., 2020; Li et al., 2023). Section 5.2 leans on Chen et al. (2024), who give the diffusion model that is optimal over a finite reference set in closed form; this lets us bound an entire method family rather than report that our own training runs failed. To our knowledge the resulting negative result, that the best possible diffusion prior over EEG responses is the near-identity, has no counterpart in a literature that reports successes on natural images rather than on low-SNR neural measurements.

Trial averaging and benchmark diagnosis. Averaging repeated presentations to raise SNR is standard in cognitive neuroscience, and both datasets we use are built around it: THINGS-EEG2 presents each test image 80 times and THINGS-MEG (Hebart et al., 2023) 12. What is specific to the machine-learning literature built on them is that the averaged number became the headline metric without an accompanying curve. This paper is a diagnosis rather than a method; its closest methodological relatives are upper-bound arguments and ablation-driven reappraisals of what a benchmark measures.

4 THE TRIAL-AVERAGING GAP

Holding the released checkpoint fixed and varying only the number of test repetitions averaged produces the first row of Table 1. The gap is not a property of our encoder: the same procedure on the frozen checkpoints of the two 2026 baselines we reproduce (Section 2), each evaluated inside its own preprocessing, channel set and target, gives the other two rows. Their $R = 80$ entries match the reproduction numbers to the third decimal, and at $R = 1$ they retain 10% and 16% of that number, against 19% for ours. Appendix F gives the draw protocol, per-subject curves for all three encoders, and why later sections' recomputed $R = 1$ baselines land a draw's width away (0.156–0.158).

This is not an accusation of error: 80 repetitions are what the dataset provides, and averaging them is a defensible protocol. It is an observation about what the number measures. A system reported at 0.841 has been evaluated under a regime that costs eighty presentations of every stimulus; its behaviour at one presentation, the regime in which any application would operate, is not reported, is not monotonically related to the headline figure across arms (Section 5.4), and turns out to be 0.158.

Table 2: The best attainable diffusion prior over EEG, evaluated in closed form. Cells are selected *on the test metric*, so these are optimistic upper bounds. The argmax sits at the largest σ and smallest γ, the degenerate corner where the update reduces to the identity.

budget	plain mean	best of 28 cells	argmax
$R = 1$	0.156	0.156 $(+0.001)$	$\sigma = 32,\ \gamma = 0.1$
$R = 4$	0.443	0.428 (-0.015)	$\sigma = 32,\ \gamma = 0.1$
$R = 8$	0.602	0.577 (-0.025)	$\sigma = 16,\ \gamma = 0.1$

5 Nothing on the EEG side closes it

If eighty repetitions carry 0.841 worth of information and one carries 0.158, a good enough prior over clean EEG should recover part of the difference. It does not, across four families. One result is an upper bound on its family (Section 5.2) and does not depend on our skill at training anything; the other three are measured nulls, and we claim for them only what a null supports.

5.1 Classical estimators

Temporal smoothing, principal-subspace projection at $k \in \{8, 32, 128, 512\}$, Wiener shrinkage (no free hyperparameter once the noise level is estimated from the repetition axis), and James–Stein shrinkage (James & Stein, 1961) toward the train grand mean: none beats the plain repetition mean at $R = 1$, 4 or 8.

5.2 The optimal diffusion prior, in closed form

The obvious objection is that a learned generative prior would do better. We answer it without training one. Chen et al. (2024) give the optimal diffusion model over a finite reference set in closed form, so the best attainable diffusion prior over EEG can be evaluated directly, with the train split's trial-averaged responses as the reference set and one likelihood-maximisation step per Tweedie denoise (Efron, 2011). We sweep 28 cells: σ from the nearest-neighbour to the grand-mean-shrinkage limit, γ from 0.1 to 1.0.

Two properties make this the paper's strongest negative result: the cells are selected on the test metric, so the numbers are optimistic upper bounds, and when the best configuration of a method class is "do as little as possible," the class is not underfit but pointed in the wrong direction.

5.3 Learned denoisers, and the capacity signature

The train split has 4 repetitions per image, so mapping an R-repetition average to the 4-repetition average is a genuinely supervised task. Fitting it in the whitened PCA basis makes the already-failed Wiener filter the diagonal linear special case, so the ladder strictly increases in capacity. Accuracy falls monotonically as capacity rises: against the plain mean's **0.156** and **0.285** at $R = 1$ and $R = 2$, the full linear map scores 0.111 and 0.206 $(-0.045, -0.079)$ and the MLP 0.074 and 0.155 $(-0.082, -0.131)$. The models are not failing to fit; they are fitting a target that is wrong, which Section 6 makes precise.

5.4 Retraining, and the budget confound

All of the above uses a checkpoint trained on 4-repetition means, so its 0.156 at $R = 1$ is partly a train/test mismatch. We remove the confound by retraining with the repetition budget as a knob, and additionally place the high-SNR target in the *objective* rather than in signal space: a teacher sees the 4-repetition mean of the student's few-trial input and the student's embedding is pulled toward the teacher's, so nothing is contracted in signal space.

Two further arms, at 10 subjects and three seeds each, sharpen the point. Distillation adds nothing: the golden-embedding teacher sits at -0.018 at $R = 80$ under a single-repetition training budget, and its apparent $R = 1$ benefit in a 3-subject pilot did not survive the larger sample. And specialising to a single budget is actively bad: training only on single trials buys the largest $R = 1$ gain and

Table 3: Removing the train/test budget mismatch, 10 subjects. The fix is a *trade*, not a gain.

arm	train budget	$R = 1$	$R = 2$	$R = 4$	$R = 8$	$R = 80$
published recipe	4	0.156	0.277	0.444	**0.605**	**0.847**
budget-randomised	1–4 random	**0.177**	**0.304**	**0.460**	0.600	0.805
delta		$+0.021$	$+0.027$	$+0.016$	-0.005	-0.042

Table 4: Is the discriminative signal where the variance is? THINGS-EEG2, 10 subjects. Deleting the 32 leading directions, which carry 43.7% of the estimated signal energy, costs nothing; in isolation they retrieve at chance. Keep-k does not reach the full-basis number until k is in the thousands.

		$R = 1$		$R = 4$	
k	signal energy	keep-k	drop-k	keep-k	drop-k
8	15.1%	0.003	0.154	0.004	0.454
32	43.7%	0.004	**0.154**	0.006	**0.449**
128	74.7%	0.040	0.064	0.087	0.197
512	91.8%	0.106	0.020	0.311	0.057
1024	97.7%	0.125	0.019	0.373	0.034
2048	100.0%	0.141	0.017	0.412	0.022
4096	100.0%	0.151	0.006	0.439	0.012
full	–	**0.156**	–	**0.449**	–

surrenders 0.135 at $R = 80$. Low-budget and high-budget accuracy are therefore not monotonically related, and a single-budget headline can mislead in either direction.

6 WHY: VARIANCE ORDERING IS ANTI-INFORMATIVE

Every method in Section 5 assumes that the stimulus-specific signal occupies a low-dimensional, high-variance subspace, so modelling that subspace captures the signal and discarding the rest removes noise. We test the assumption directly, with no estimator in the loop: order the principal directions of the train responses by variance, then either keep the top k or project them out. Signal energy is the eigenvalue spectrum with the trial-averaged noise floor deconvolved.

Deleting the 32 leading directions costs nothing (0.154 against 0.156; 0.449 against 0.449), and those same directions in isolation retrieve at 0.004, which is chance. The discriminative content is high-rank and *anti*-aligned with the variance ordering; what spectral methods call signal energy is here close to orthogonal to what matters.

This single fact accounts for every result in Section 5. MMSE estimators, diffusion priors, likelihood maximisation and principal-subspace projection all concentrate an estimate in high-density, high-variance regions, which on this data preferentially discards the signal. The prediction is that reconstruction fidelity should trade against retrieval, and it does. At $R = 1$ the plain repetition mean scores Top-1 **0.156** at an NMSE of 36.62 from the 80-repetition response, while the full linear map reaches NMSE **2.13** and cosine **0.413**, that is $17\times$ closer in NMSE and $2.4\times$ closer in cosine, and scores 0.115, four points worse. Across all eleven estimators we scored, NMSE and Top-1 are *positively* related (Table 13, Appendix D).

We separate two candidate mechanisms. A *scale or out-of-distribution artefact* is rejected on its first leg, since norm-matching the linear estimate to the plain mean makes it worse (0.069 against 0.115), though its second leg holds: contracting the true 80-repetition response destroys accuracy ($0.841 \to 0.615 \to 0.247$ at scale 0.7, 0.5, 0.3, all at cosine ≥ 0.87), so the encoder tolerates isotropic noise and is fragile to anisotropic contraction. The *wrong subspace* is the real cause: pure projection onto the top 512 directions with no shrinkage already costs 0.041, and variance-preserving rescaling does not help, since it acts within the same subspace.

Both findings replicate on THINGS-MEG. A different sensor modality, a different subject pool, and a test set with 12 rather than 80 repetitions, so the averaging regime the EEG benchmark reports under does not exist here. The trial gap is $+0.307$ over the available budget and the curve is still

Table 5: Replacing the augmentation pipeline with a learned operator on the clean embedding. THINGS-EEG2, 10 subjects. An operator at cosine 0.984 from the multi-view target delivers one sixth of its benefit, and making it nonlinear improves the out-of-sample fit while adding essentially nothing.

target	fit to the 24-view target	$R = 80$	fraction of $+0.278$
clean embedding	val cos 0.727	0.569	0% (reference)
scalar shrinkage	–	0.569	**0.0%**
least-squares linear operator	val R^2 0.781, cos 0.984	0.615	**16.7%**
nonlinear content-adaptive operator	val R^2 0.786, cos 0.985	0.619	**18.0%**
24-view hand-designed pool	1.000	**0.847**	100%

climbing steeply at its right edge ($0.454 \rightarrow 0.474$ from $R = 8$ to $R = 12$). The rank diagnostic (Table 7, Appendix B) is sharper than on EEG: the 8 leading directions hold 31.9% of the estimated signal energy, retrieve at 0.005 (chance) in isolation, and deleting them is free (0.154 against a 0.157 baseline at $R = 1$; 0.356 against 0.368 at $R = 4$). Keep-k has still not matched the full basis at $k = 4096$. Neither the gap nor the dissociation is an artefact of one dataset or one sensor modality. Two protocol adaptations, both forced by MEG's train split having a single repetition per image, are documented in Appendix A; neither affects the keep-k / drop-k accuracies.

7 THE TARGET SIDE, AND WHAT IT IS ACTUALLY DOING

The other side of the pipeline behaves in the opposite way. Replacing the clean CLIP embedding with the mean of 24 degraded views of the same image, with the encoder and everything else held fixed, is worth **+0.278** ($0.569 \rightarrow 0.847$ at $R = 80$). Inside NeuroBridge's own code the effect is monotone in the number of views, 39.1 at $V = 0$ rising to 63.2 at $V = 4$, so it is not specific to our pipeline. This is a hand-designed pipeline with no stated principle, so we ask what operation it performs by trying to learn it: fit an operator on the clean embedding to reproduce the multi-view target, on train images only, and then train the encoder against the operator's output instead of the augmentations.

The scalar arm tests the hypothesis suggested by a direct measurement of what averaging does to the target: it destroys roughly 73% of the variance in every direction, in the canonical basis and in CLIP's own variance-ordered basis alike. A single scalar recovers none of the benefit. The linear arm is the least-squares optimum of the entire fixed-linear family and therefore its ceiling by construction; it recovers 16.7%. The number that matters is the dissociation behind it. **The linear operator reproduces the multi-view target to within cosine 0.984 on held-out images, and delivers one sixth of its benefit.** The value of the multi-view target lives in the last 1.6% of cosine.

Is the shortfall linearity, or fitting? We lift the operator to a nonlinear map whose action depends on where in embedding space the image sits. The construction matters: an independently trained residual MLP underfits (train R^2 0.777 against the linear map's 0.809), so its shortfall would prove nothing. We therefore parameterise it as the closed-form linear solution *plus* a residual network with a zero-initialised output layer, so it starts at the linear operator and gradient descent can only improve its fit, which it does out of sample (val R^2 0.786, val cosine 0.985). The retrieval payoff of that strictly better fit is a coin flip: the nonlinear operator beats the linear one in **5 of 10** subjects, mean **+0.0035** Top-1 (per-subject numbers in Table 8, Appendix B). Recovery moves from 16.7% to 18.0%, and **82%** of the multi-view gain remains unreached.

Then remove the fidelity objective. Every operator above is fitted to reproduce the multi-view target, and fidelity to a high-quality reference is precisely what Section 6 identifies as the wrong objective. The arm that closes the argument drops it entirely: we keep the nonlinear operator's architecture and zero-initialised residual parameterisation, so it again starts exactly at the clean embedding, and delete the regression loss, so its only gradient is the InfoNCE retrieval objective, optimised jointly with the encoder. The operator reads the image embedding and never the EEG, so at test time it remains a fixed function of the stimulus, like the augmentation pipeline it replaces. Sweeping its step size gives $R = 80$ of $0.575/\mathbf{0.577}/0.575/0.460/0.172$ at 10^{-5} through 10^{-3} (subjects 1–2):

Table 6: Matched-budget comparison on THINGS-EEG2. Pooling subjects dominates pooling repetitions at every budget tested, monotonically in the number of subjects.

trials spent	configuration	Top-1
10	1 subject × 10 reps	0.646
10	10 subjects × 1 rep	**0.793**
20	5 subjects × 4 reps	0.904
40	5 subjects × 8 reps	0.955
80	1 subject × 80 reps	0.841
80	10 subjects × 8 reps	**0.973**

the optimum sits at the smallest step sizes, and at 10^{-3} the operator collapses to near chance. Run at the best cell on all 10 subjects, it reaches **0.565** against the clean embedding's 0.569, **indistinguishable from the identity it was initialised at**: mean -0.004 Top-1, better in only 4 of 10 subjects, where the fidelity-fitted linear map is better in 9 of 10. This is the signature of Section 5.2 on the target side: when the best configuration of a method family is to do as little as possible, the family is misdirected rather than underfit.

What this licenses, and what it does not. Both objectives available for building a target constructor without running the augmentations fail: fidelity saturates near 18% at cosine 0.985, and the retrieval loss prefers the identity. We cannot claim that *no* function of the clean embedding suffices, but the failure is no longer attributable to the loss (Section 9).

Cross-modality: one half replicates, the other does not. Retraining all four arms on THINGS-MEG (Table 9, Appendix B; medians over five seeds, since learned-operator arms proved seed-unstable there) splits the section's two claims. The nonlinear result replicates exactly: the content-adaptive operator adds $+0.0024$ over the linear one against $+0.0035$ on EEG, i.e. nothing, twice. The recovered *fraction* does not: 58.0% on MEG against 16.7% on EEG, and within each modality the fraction is roughly flat in the budget (14–22% across R on EEG, 36–58% on MEG, Table 10), so this is not an artefact of comparing $R = 12$ with $R = 80$. The comparison has one confound, which we remove rather than argue about: EEG pools 24 views over two perturbation families (blur, and blur composed with noise) while MEG pools 12 views of blur alone, and a single-family target is a smoother function of the clean embedding that a linear map should reproduce more easily, which predicts the direction observed. Rerunning the whole EEG decomposition on MEG's exact 12-view blur-only target, the linear operator recovers 15.0% of a now-smaller gain ($+0.234$ against $+0.278$), statistically the same as 16.7% and nowhere near MEG's 58%. Whatever separates the two modalities, it is not the augmentation pipeline; it is either a property of the measurement or an artefact of MEG's four subjects and twelve repetitions, and we cannot currently tell which. The control does establish two things: the multi-view gain is monotone in view count inside our own pipeline ($+0.234$ at 12 views against $+0.278$ at 24), and the nonlinear null survives a third setting, its largest point estimate anywhere being $+0.017$ here at 6 of 10 subjects, which a paired sign test does not separate from zero ($p = 0.51$; $p = 1.00$ at 24 views).

This is Section 6 restated on the other side of the contrastive pair: there, the estimator $17\times$ closer to the high-SNR response lost four points of Top-1; here, the operator at cosine 0.985 delivers under a fifth of the benefit. On both sides, the value sits in a residual that fidelity-based fitting is structurally unable to keep.

8 Cross-subject pooling: the one real information source, and why it stays put

Repetitions of one subject average away trial noise. Different subjects also average away subject-specific bias, so pooling across subjects should be worth strictly more per trial. It is, by a wide margin, at a matched total budget (Table 6). In the canonical-correlation basis, consensus multiplies per-direction correlation by two to three times, and the count of directions with $\rho > 0.3$ rises from 8 to 22. Unlike everything in Section 5, this is genuinely new information rather than a re-weighting of what one subject already has.

It cannot, however, be moved into a single-subject encoder. We tried four variants: pulling the student's embedding toward the leave-one-subject-out consensus vector directly, and matching the consensus's pairwise similarity structure instead, at three loss weights, the latter because each subject's encoder lands in its own learned adapter frame and only the relational structure is invariant to that rotation. All are null. At 10 subjects the two surviving variants sit at -0.005 (direct) and -0.010 (relational) against the undistilled arm at $R = 80$, and the same holds for self-distillation under a single-repetition training budget (-0.018, 10 subjects $\times$ 3 seeds). That last arm is the one place where seed variance mattered (one subject collapsed to 0.155 on the default seed against a 0.712 median), so all arms in this section are medians over three seeds per subject.

We think the null is structural rather than a tuning failure, and offer the reason as a hypothesis (four null variants cannot exclude a fifth). Distillation transfers *label* information from a teacher that knows something about the answer. Here the student already has the answer: its supervision is the CLIP embedding of the actual stimulus, which is perfect. What it lacks is on the *input* side, a cleaner measurement of a brain, and no teacher can supply an input the student was never given; a method that imports cross-subject information into a single-subject decoder would have to do so through the input path rather than the objective.

9 IMPLICATIONS

Report the curve, not the point. A single 80-repetition number does not determine behaviour at lower budgets, and Section 5.4 shows the relationship is not even monotone across arms: the arm with the best $R = 1$ accuracy gives up 0.135 at $R = 80$. We suggest the trials-to-accuracy curve as the default reporting unit for this benchmark.

One mechanism, both sides of the pair. The failures in Sections 5 and 7 share one mechanism, established and replicated across modalities in Section 6: procedures that allocate capacity by fidelity to a high-quality reference allocate it away from what discriminates. Any method whose core operation is to concentrate an estimate in high-density regions of the response distribution (diffusion priors, MMSE denoisers, likelihood maximisation, principal-subspace projection) is predicted to inherit the failure on the brain side; for the diffusion-prior family Section 5.2 makes that a bound, for the others it is the mechanism's prediction, confirmed for every estimator we tested. Any operator trained to regress a hand-designed target inherits it on the image side. Nothing here is specific to EEG: any low-SNR measurement whose variance is dominated by nuisance structure should show the same signature, as THINGS-MEG already does.

The open problem has a name, a size, and one ruled-out solution. 82% of the multi-view gain on EEG is not reachable by fitting an operator to reproduce the multi-view target, even at cosine 0.985 out of sample, and training the operator on the retrieval loss instead gives a family whose optimum is the identity. Both objectives available without running the augmentations fail, so the failure can no longer be pinned on the loss. That narrows the problem: the augmentation pipeline reenters pixel space and evaluates CLIP two dozen times, while every operator we fit acts downstream of one pooled 1024-dimensional vector. Read with the rank diagnostic, the economical explanation is that pooling discards the structure the augmentations recover, so the constructive work belongs before the bottleneck. We do not establish this, and the MEG result, where a linear map on the pooled embedding captures 58% of the gain, cautions against assuming it. Naming the problem precisely, with a measured size, a ruled-out objective, and a modality-dependence any answer must explain, is the useful first contribution.

Limitations. The closed-form diffusion prior (Section 5.2) and the construction-by-initialisation (Section 7) make a failure a property of a method class rather than of our tuning, but they are not proofs, and a different architecture or supervision for the target constructor could still succeed. The target-side decomposition only half replicates on THINGS-MEG, and we cannot tell a genuine property of the measurement from the thinness of four MEG subjects at twelve repetitions; this is the paper's least resolved claim. THINGS-MEG also shares its stimulus set with THINGS-EEG2, so the replication is not across visual paradigm.

AI USE STATEMENT

Large language model assistance was used throughout this project and we describe its role by subtask, in the categories the ICLR AI policy asks about.

Designing methodology and experiments. LLM assistance was used to propose candidate diagnostics and to argue about which experiment would discriminate between competing explanations, in particular the decision to evaluate the closed-form optimal diffusion prior rather than train one (Section 5.2), and the decision to parameterise the nonlinear target operator as the linear solution plus a zero-initialised residual so that underfitting could be excluded by construction (Section 7). The scientific claims, and the choice of which results to trust, are the authors'.

Implementing methods and creating code. The experiment scripts and analysis code were written with substantial LLM assistance. All reported numbers were produced by executing that code; no number in this paper was generated or estimated by a language model.

Retrieval and discovery. LLM assistance was used to search for prior work that might already contain our diagnostics, and to locate the closed-form result we rely on in Section 5.2. Every work cited here was retrieved and read in its original form by the authors, and its bibliographic record was checked against the publisher or arXiv listing.

Interpreting results. LLM assistance was used in discussing what the measurements imply. Every interpretation reported here was checked against the underlying per-subject data by the authors, and several earlier readings were revised when a larger sample contradicted them, notably the budget-randomisation result in Section 5.4, which appeared to be a free gain at three subjects and turned out to be a trade at ten.

Drafting and editing. The manuscript was drafted with LLM assistance and edited by the authors.

The authors take full responsibility for all content, including every claim, number and citation.

REPRODUCIBILITY STATEMENT

The evaluation protocol, including how a repetition budget R is sampled and how many draws are averaged, is specified in Section 2, and the same fixed seed is used across all arms so that every arm sees identical draws. The two protocol adaptations required to run the rank diagnostic on THINGS-MEG, which differ from the THINGS-EEG2 procedure because MEG's train split has a single repetition per image, are stated in Appendix A. Both datasets are public. Our reproductions of the two 2026 baselines (Section 2) were run in our own pipeline from the authors' released code, and land within 1.0 point of the published numbers. Source code for all diagnostics and training arms is provided as supplementary material.

REFERENCES

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning (ICML)*, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. HyFI: Hyperbolic feature interpolation for brain-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *International Conference on Computer Vision (ICCV)*, 2023.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, 2024.

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

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. NeuroBridge: Bio-inspired self-supervised EEG-to-image decoding via cognitive priors and bidirectional semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026. doi: 10.1609/aaai.v40i21.38863.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IEq71qS8B7.

A THINGS-MEG PROTOCOL ADAPTATIONS

MEG's train split has a single repetition per image, which forces two departures from the THINGS-EEG2 protocol of Section 6. Both are documented here because they are the only places where the two replications are not literally the same computation.

Whitening. On EEG we whiten by the within-image (across-repetition) noise standard deviation, which is unavailable with one repetition. On MEG we whiten by the across-image standard deviation of the train responses instead. This is train-only, so no test statistic enters the basis, and it still yields a variance ordering: it is the PCA of the correlation rather than the covariance matrix.

Noise floor. On EEG we deconvolve the noise from the eigenvalues analytically using the repetition count. On MEG we measure it: project the test single trials into the basis and take the within-image variance across the 12 repetitions, per direction, giving $\mathrm{signal}_j = \lambda_j - \mathrm{noise}_j$. This uses a test-derived quantity, but an unsupervised per-direction scalar that carries no label information, which is the standard noise-ceiling construction. It affects only the reported signal-energy column of Table 7, never the keep-k or drop-k accuracies.

Table 7: The rank diagnostic on THINGS-MEG, 4 subjects. Referenced from Section 6.

		$R = 1$		$R = 4$	
k	signal energy	keep-k	drop-k	keep-k	drop-k
8	**31.9%**	0.005	**0.154**	0.007	**0.356**
32	50.3%	0.018	0.139	0.022	0.310
128	72.4%	0.040	0.095	0.063	0.233
512	90.9%	0.087	0.041	0.183	0.100
1024	96.6%	0.115	0.023	0.263	0.043
2048	98.7%	0.147	0.011	0.327	0.018
4096	100.0%	0.146	0.008	0.355	0.011
full	–	**0.157**	–	**0.368**	–

Table 8: Per-subject retrieval at $R = 80$ for the linear and nonlinear target operators of Section 7. The nonlinear operator fits the multi-view target strictly better out of sample and wins in 5 of 10 subjects.

subject	1	2	3	4	5	6	7	8	9	10	mean
linear	0.630	0.630	0.645	0.625	0.540	0.635	0.560	0.635	0.560	0.695	0.615
nonlinear	0.610	0.630	0.610	0.645	0.535	0.645	0.540	0.665	0.580	0.730	0.619
delta	$-.020$	.000	$-.035$	$+.020$	$-.005$	$+.010$	$-.020$	$+.030$	$+.020$	$+.035$	$+.0035$

B SUPPORTING TABLES

C THE DIFFUSION SWEEP IN FULL

Table 11 gives every cell behind Table 2. σ is the noise level at which the closed-form score of Chen et al. (2024) is evaluated and γ the step size of the likelihood-maximisation update, so $\gamma \to 0$ and $\sigma \to \infty$ are both routes to the identity map. The accuracy surface is monotone in both directions towards that corner at all three budgets, which is what makes the conclusion robust to the grid: there is no interior optimum to be found with a finer sweep. Away from the corner the losses are severe rather than marginal, up to 0.151 at $R=1$ and 0.440 at $R=4$, and even the argmax fails to beat the plain mean once $R \geq 4$.

D CLASSICAL ESTIMATORS AND FIDELITY, IN FULL

Table 13 gives the fidelity-against-accuracy comparison of Section 6 in full, with the two controls that rule out the obvious alternative explanation. A supervised estimator shrinks its output norm (the plain mean sits at $5.6\times$ the target norm, the linear map at $1.4\times$), so one could argue the accuracy loss is a scale artefact rather than a subspace error. It is not. Rescaling each estimate to match the target norm (+NM) leaves accuracy at or below the unrescaled value in all three families while returning NMSE to roughly the plain mean's, and equalising per-direction variance (+VP) does not recover it either. The dissociation survives both: the linear map is $17\times$ closer to the 80-repetition response in NMSE than the plain mean and four points worse in Top-1, at any norm.

E THE TARGET-RICHNESS CONTROL

F PER-SUBJECT CURVES AND THE SHRINKAGE GRID

The scalar shrinkage grid. Section 7 reports one α; the grid behind it is $\alpha = 0.20/0.35/0.52/0.70$ giving $0.568/0.557/0.557/0.567$ at $R = 80$ against a clean baseline of 0.550 and a 24-view target of 0.857 (3 subjects). The best cell recovers 6.0% of the gain and the grid is non-monotone, i.e. within noise of doing nothing; at 10 subjects $\alpha = 0.52$ recovers 0.0% (0.569 against clean 0.569). This matters because the α was not guessed: a direct measurement of the augmentation pipeline showed it destroys $\approx 73\%$ of the variance in every direction, in the canonical

Table 9: Target-side decomposition retrained on THINGS-MEG, 4 subjects. Learned-operator arms are seed-unstable on MEG (subject 3 collapsed to near chance on two of five seeds), so every cell is the median over five seeds, a protocol fixed before the aggregate was inspected and applied uniformly to all four arms. The nonlinear-buys-nothing result replicates; the recovered fraction does not.

target (cells: median of 5 seeds)	sub1	sub2	sub3	sub4	median	fraction of $+0.172$
clean embedding	0.185	0.590	0.375	0.255	0.315	0% (reference)
least-squares linear	0.215	0.680	0.470	0.360	0.415	**58.0%**
nonlinear content-adaptive	0.260	0.680	0.490	0.345	0.417	**59.4%**
12-view hand-designed pool	0.335	0.775	0.585	0.390	**0.487**	100%

Table 10: The recovered fraction is not a working-point artefact: within each modality it is roughly flat in the test-time repetition budget, so the EEG–MEG discrepancy of Section 7 is a modality difference rather than a consequence of comparing $R = 80$ with $R = 12$. The last column is each modality's maximum budget, 80 for EEG and 12 for MEG.

R	1	2	4	8	max
EEG, fraction recovered by the linear operator	14.1%	18.3%	8.1%	17.1%	21.8%
MEG, fraction recovered by the linear operator	44.9%	36.1%	44.8%	45.3%	58.0%

basis and in CLIP's variance-ordered basis alike, which predicts $\alpha = \sqrt{1 - 0.73} \approx 0.52$. The prediction is quantitatively right about what the pipeline does to the target and buys nothing, which is the section's point in miniature.

Table 11: All 28 cells of the closed-form optimal diffusion prior, at three test-time budgets. Baselines (plain repetition mean) are 0.156, 0.443 and 0.602. Bold marks each budget's argmax.

	$R = 1$ (base 0.156)				$R = 4$ (base 0.443)				$R = 8$ (base 0.602)			
σ	γ.1	.25	.5	1	γ.1	.25	.5	1	γ.1	.25	.5	1
0.5	.146	.070	.015	.007	.376	.117	.014	.008	.493	.135	.017	.007
1	.147	.076	.014	.007	.383	.128	.016	.008	.506	.148	.019	.008
2	.147	.080	.015	.007	.405	.157	.017	.008	.545	.199	.019	.007
4	.152	.105	.018	.007	.424	.258	.025	.009	.565	.340	.025	.008
8	.154	.119	.029	.007	.428	.278	.028	.006	.576	.356	.031	.005
16	.155	.120	.030	.006	.428	.290	.031	.005	**.577**	.374	.034	.006
32	**.156**	.124	.029	.005	**.428**	.291	.032	.004	.576	.368	.032	.005

Table 12: The full classical ladder of Section 5. No estimator beats the plain repetition mean at more than one budget, and temporal smoothing's small $R = 1$ edge reverses by $R = 8$. Subspace projection degrades monotonically as k falls, which is the same ordering the rank diagnostic predicts.

estimate	$R = 1$	$R = 4$	$R = 8$
plain repetition mean	0.155	**0.443**	0.612
James–Stein shrinkage	0.158	0.442	**0.613**
temporal smoothing, width 1	**0.162**	0.442	0.602
temporal smoothing, width 2	0.149	0.381	0.511
temporal smoothing, width 4	0.103	0.236	0.304
temporal smoothing, width 8	0.058	0.113	0.137
Wiener shrinkage	0.104	0.305	0.431
top-512 projection	0.115	0.317	0.442
top-128 projection	0.052	0.126	0.166
top-32 projection	0.006	0.008	0.007
top-8 projection	0.004	0.005	0.005

Table 13: Fidelity against accuracy at $R = 1$, with norm-matched (+NM) and variance-preserving (+VP) controls. The *golden* rows interpolate the ground-truth 80-repetition response toward the origin and are the reference for what fidelity ought to buy if fidelity were the right objective.

estimate at $R = 1$	Top-1	NMSE to 80-rep	cos to it	norm ratio
plain mean	**0.156**	36.62	0.174	5.60
top-512 projection	0.115	25.88	0.181	4.56
Wiener	0.103	21.69	0.190	4.12
+NM	0.101	36.48	0.190	5.60
+VP	0.092	6.83	0.274	2.38
full linear	0.115	**2.12**	**0.413**	1.38
+NM	0.069	34.14	0.413	5.60
+VP	0.119	6.66	0.284	2.37
MLP	0.077	5.89	0.250	2.29
+NM	0.064	35.89	0.250	5.60
+VP	0.089	8.75	0.220	2.80
golden $\times 1.0$	0.841	0.00	1.000	1.00
golden $\times 0.7$	0.782	0.05	0.985	0.84
golden $\times 0.5$	0.615	0.14	0.948	0.76
golden $\times 0.3$	0.247	0.27	0.872	0.69
golden $\times 0.1$	0.021	0.45	0.745	0.66

Table 14: The control that removes the target-richness confound from the EEG–MEG comparison of Section 7. All four EEG target arms rerun on MEG's exact 12-view blur-only target, 10 subjects, $R = 80$. Holding modality fixed and simplifying the target leaves the recovered fraction essentially unchanged, so the discrepancy with MEG is not attributable to the augmentation pipeline. The gain itself is monotone in view count, independently reproducing the view-count ladder.

target set (statistic)	clean	linear	nonlinear	views	gain	linear %
24-view blur + noise, EEG (mean, 10 subj.)	0.569	0.615	0.619	0.847	+0.278	**16.7%**
12-view blur only, EEG (mean, 10 subj.)	0.569	0.604	0.620	0.803	+0.234	**15.0%**
12-view blur only, MEG (median, 4 subj. × 5 seeds)	0.315	0.415	0.417	0.487	+0.172	**58.0%**

Table 15: The nonlinear operator's advantage over the linear one, per subject, in all three settings. No setting separates from zero by a paired sign test and the largest point estimate is $+0.017$; this is the form in which we prefer the claim of Section 7 to be read.

setting	mean	median	subjects won	sign test	$\lvert t \rvert$
EEG, 24-view blur + noise	$+0.0035$	$+0.0050$	5/10	$p = 1.00$	0.47
EEG, 12-view blur only	$+0.0165$	$+0.0125$	6/10	$p = 0.51$	1.65
MEG, 12-view blur only	$+0.0024$	–	–	–	–

Table 16: The trial-averaging gap per subject, on the frozen published encoder; the mean row is Table 1. The effect is not carried by a few subjects: every one of the ten loses at least 0.635 Top-1 between $R = 80$ and $R = 1$, and the weakest subject at $R = 80$ (0.755) still outscores the strongest subject at $R = 8$ (0.699). For $R < 80$ each cell averages 20 size-R repetition subsets drawn without replacement, independently per subject; $R = 80$ uses all repetitions and is deterministic. The draw-to-draw standard deviation is 0.028 at $R = 1$, peaks at 0.038 at $R = 4$, and falls to zero at $R = 80$, which is the scale on which this row differs from other sections' recomputed baselines.

subject	$R = 1$	$R = 2$	$R = 4$	$R = 8$	$R = 16$	$R = 32$	$R = 80$
1	0.157	0.241	0.404	0.574	0.702	0.821	0.880
2	0.141	0.252	0.424	0.590	0.725	0.797	0.845
3	0.171	0.297	0.480	0.643	0.758	0.812	0.845
4	0.127	0.235	0.399	0.570	0.700	0.768	0.820
5	0.096	0.184	0.316	0.459	0.601	0.686	0.755
6	0.190	0.298	0.483	0.650	0.766	0.838	0.885
7	0.165	0.296	0.478	0.612	0.709	0.760	0.800
8	0.208	0.383	0.564	0.699	0.790	0.830	0.865
9	0.117	0.204	0.360	0.520	0.665	0.753	0.840
10	0.209	0.344	0.537	0.698	0.811	0.858	0.875
mean	**0.158**	0.273	0.445	0.601	0.723	0.792	**0.841**

Table 17: The trial-averaging gap per subject on the two reproduced 2026 baselines, evaluated on their frozen checkpoints under the same draw protocol as Table 16 (20 size-R subsets per cell, deterministic at $R = 80$). Each baseline is scored inside its own pipeline: its 17 occipito-parietal channels, its own preprocessing, its own test-time target and projector. The $R = 80$ column reproduces the per-subject numbers of our reproduction runs exactly, which is the check that the loader is theirs. Every subject on both encoders loses at least 0.386 between $R = 80$ and $R = 1$; the draw-to-draw standard deviation is 0.018–0.031 (NeuroBridge) and 0.023–0.029 (ATS) for $R < 80$.

subject	$R = 1$	$R = 2$	$R = 4$	$R = 8$	$R = 16$	$R = 32$	$R = 80$
NeuroBridge (Zhang et al., 2026), final-epoch checkpoint							
1	0.059	0.087	0.133	0.228	0.296	0.415	0.445
2	0.054	0.105	0.180	0.289	0.412	0.531	0.640
3	0.071	0.118	0.211	0.320	0.456	0.535	0.595
4	0.061	0.099	0.181	0.296	0.414	0.507	0.590
5	0.035	0.055	0.115	0.206	0.311	0.420	0.565
6	0.073	0.129	0.220	0.372	0.506	0.614	0.685
7	0.069	0.115	0.222	0.341	0.465	0.569	0.640
8	0.087	0.158	0.261	0.402	0.537	0.623	0.695
9	0.039	0.073	0.137	0.229	0.377	0.505	0.645
10	0.076	0.142	0.263	0.409	0.560	0.673	0.760
mean	**0.062**	0.108	0.192	0.309	0.433	0.539	**0.626**
ATS (Wu et al., 2026), final-epoch checkpoint							
1	0.088	0.134	0.185	0.292	0.360	0.444	0.480
2	0.089	0.156	0.253	0.377	0.474	0.545	0.615
3	0.103	0.173	0.278	0.386	0.499	0.562	0.610
4	0.089	0.158	0.255	0.357	0.454	0.531	0.585
5	0.061	0.103	0.177	0.282	0.386	0.474	0.550
6	0.121	0.187	0.322	0.452	0.561	0.620	0.675
7	0.095	0.167	0.263	0.367	0.437	0.483	0.510
8	0.123	0.230	0.341	0.472	0.568	0.622	0.665
9	0.061	0.106	0.183	0.274	0.386	0.468	0.555
10	0.120	0.214	0.331	0.455	0.547	0.620	0.670
mean	**0.095**	0.163	0.259	0.372	0.467	0.537	**0.592**