DECODE, THEN AVERAGE: REAL-TIME VISUAL EEG DECODING FROM SINGLE TRIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding perceived images from non-invasive electroencephalography (EEG) has progressed rapidly in recent years. This progress relies on averaging, a practice inherited from decades of neurophysiology. An individual EEG trial is dominated by noise, so decoders are trained on averages of repeated trials of each image, and at test time every available trial of a test image is averaged into one clean query that is decoded once. We argue that the single trial, not the average, should be the default operating point. We first build a decoder for this regime, with every design choice, from training data to hyperparameters, made for single trials. On THINGS-EEG2, the standard benchmark for this task, where a decoder must identify which of 200 candidate images evoked a response, it attains 27.39% top-1 accuracy from one trial, more than twice the single-trial accuracy of the strongest published method. From this base, the decoder scores each trial separately and averages the resulting decisions, a readout we call *decode-then-average*, in contrast to the standard *average-then-decode*. Aggregating decisions rather than signals offers three advantages. (1) **Data scale**: training on trials rather than on their averages preserves the training set at its full recorded size. (2) **EEG response timing**: averaging superimposes responses whose latencies differ from trial to trial, whereas per-trial decisions preserve each response's own timing. (3) **Real-time decoding**: the number of trials becomes a decision made during inference. A patience stop during real-time decoding, halting presentation once the answer stops changing, matches the 80-trial accuracy of the strongest published method from 2.6 trials per image, a 96.8% reduction in presentations. The averaged decisions reach 94.50% with all 80 trials and exceed the accuracy of every published method, under either readout, at every trial count. Real-time decoding also exposes a new property of the stimuli: each image has a stable *decodability*, from under four trials for the most decodable to nearly twice as many for the least. Decodability reflects how crowded an image's neighborhood is among the 200 candidates, not its pixels, and a new subject's per-image trial counts can be predicted in advance. Single-trial EEG decoding is therefore more tractable than existing results indicate, and treating it as the default turns decoding into a real-time process with a measurable, predictable cost.

1 INTRODUCTION

Decoding visual experience from non-invasive brain recordings has advanced rapidly in fMRI (Scotti et al., 2023), MEG (Benchetrit et al., 2024), and EEG. Among these modalities, EEG is the least expensive and most portable, and the one on which most practical brain-computer interfaces are built (Wolpaw et al., 2002; Värbu et al., 2022). EEG-to-image decoding is benchmarked largely on THINGS-EEG2 (Gifford et al., 2022), which records EEG responses of ten subjects viewing THINGS images (Hebart et al., 2019). The benchmark task is 200-way zero-shot retrieval: given the EEG response to a held-out image, a decoder must identify that image among 200 candidate images, none of which appears in training. Reported top-1 accuracy on this task has risen from 13.8% (Song et al., 2024) to 75.7% (Zheng et al., 2026) in two years, driven by contrastive alignment to vision-language embeddings (Radford et al., 2021), image-side priors, and stronger encoders.

A single EEG trial has an unfavorable signal-to-noise ratio, and the strongest of these accuracies are therefore obtained on averaged EEG. Averaging the responses to repeated trials of the same

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.

stimulus is a standard remedy with decades of history in event-related potential research (Blankertz et al., 2011): it raises the signal-to-noise ratio, retains the component of the response that is reproducible across trials, and dilutes the influence of occasional corrupted recordings. THINGS-EEG2 is designed to support this remedy: each training image is presented 4 times and each test image 80 times. Averaging accordingly plays a role at two separate moments. At training time, decoders are commonly fit to averages of the 4 training trials rather than to the trials individually. At test time, the 80 trials of each test image are averaged into one clean query that is decoded once, and this is the configuration in which headline results have been reported since 2024 (Song et al., 2024; Li et al., 2024; Zhang et al., 2025; Wu et al., 2025; Zhang et al., 2026; Zheng et al., 2026). Both choices are reasonable, and they are well matched to the question averaged evaluation addresses, namely what visual information the EEG response carries in principle.

We propose the single trial, and not the average, as the default operating point of EEG-to-image decoding. We assemble a decoder from standard components configured for single trials, and it reaches $27.39 \pm 2.15\%$ from single trials, where the strongest published method, run from its released code, reaches 12.0%. Our proposed decoder does not average the EEG at either moment. Trained on individual trials, it decodes every repetition of a test image separately and averages the resulting decisions, so that the averaging moves from the signal to the decisions. We call the standard readout, which averages the EEG and decodes the average once, *average-then-decode* (ATD), and ours *decode-then-average* (DTA). Under DTA the same trained model reaches $94.50 \pm 0.79\%$ when all 80 trials are used, where the strongest published method reaches 76.2%, and this decision-averaged accuracy exceeds that of every published method, under either readout, at every trial count (Figure 1). Averaging decisions rather than signals brings three advantages:

- **Data scale.** Training on averages of 4 trials leaves one sample where the recording contains 4 trials, so the training set shrinks fourfold. Accuracy grows log-linearly with the number of training trials, and a single-trial model reaches the accuracy of the averaged recipe when trained on about twice as many samples as that recipe uses, and surpasses it when all recorded trials are used (Section 3.1).
- **EEG response timing.** The latency of a visual response varies from trial to trial by about 10 ms, a quantity we measure directly on this dataset (Section 3.2). Averaging superimposes these temporally misaligned responses, blurring the timing structure of each response before the encoder receives the signal. Decoding each trial separately keeps every response aligned to its own timing, and under injected jitter, decision averaging loses less than half as much accuracy as signal averaging in relative terms.
- **Real-time decoding.** An averaged readout must fix the number of trials before its single decode. With per-trial decisions, evidence accumulates separately for each stimulus and can stop as soon as the answer stops changing, so the number of trials becomes a quantity the decoder controls in real time. A patience stop built on this rule matches the strongest published method's 80-trial accuracy from 2.6 trials per image (Section 4.1).

The first two advantages explain why decision aggregation wins; the third opens a capability that signal averaging cannot express, and it carries the scientific heart of this paper. Each stimulus turns out to have a stable, measurable *decodability*, with four properties established in Section 4: (1) the most decodable images settle in under four trials while the least decodable need nearly twice as many; (2) decodability is a property of the image, not the subject; (3) it reflects how crowded the image's neighborhood is among the 200 candidates, not its pixels; and (4) a new subject's per-image trial counts can be predicted before the subject sits down, to a median error of 3 trials.

The single-trial regime is where applications live: a stimulus appears once, and a decision must follow from that response. We offer that regime a decoder built for it, a readout aggregating its decisions, and a stopping rule deciding when enough has been seen. Our results suggest that single-trial decoding is more tractable than reported, once decisions, not signals, are aggregated.

2 TWO OPERATING POINTS ON ONE DATASET

THINGS-EEG2 (Gifford et al., 2022) is the standard benchmark for EEG-to-image decoding. Ten subjects view images from the THINGS database (Hebart et al., 2019) while 63-channel EEG is recorded. The training partition contains 16,540 images, each presented 4 times; the test partition

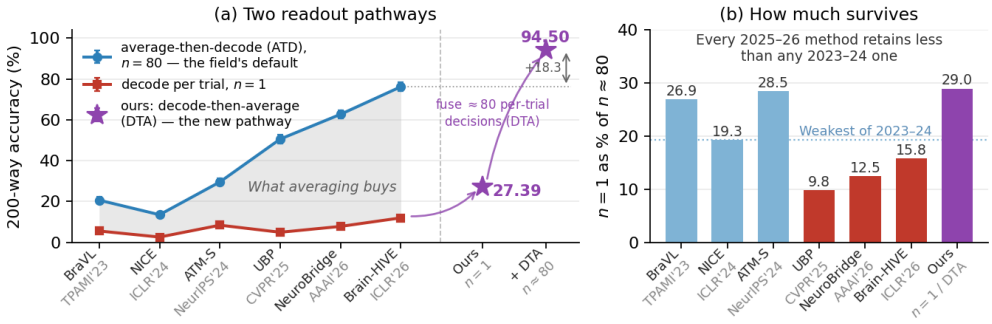

Figure 1: **Two readout pathways (average-then-decode and decode-then-average) for EEG decoding.** *(a)* Six published methods in publication order, each scored by its own released code, under the standard average-then-decode readout at $n{=}80$ and on single trials at $n{=}1$. Our single-trial decoder achieves 27.39 at $n{=}1$ and 94.50 at $n{=}80$ when applying decode-then-average. *(b)* Retention, the $n{=}1$ accuracy as a fraction of the same model's $n{=}80$ accuracy. Retention degrades in 2025–2026 compared to 2023–2024, as later methods specialize toward the averaged setting.

contains 200 held-out images, each presented 80 times, so the dataset provides 80 single-trial EEG responses for every test image. A decoder is evaluated by 200-way zero-shot retrieval: given an EEG response, it must identify which of the 200 test images was being viewed.

The 80 responses per test image are released individually rather than pre-averaged. An evaluation may therefore average any number n of these responses into a single query before decoding: at $n{=}80$ the query is the average of all 80 responses, and at $n{=}1$ every response is decoded on its own. We call a choice of n an *operating point*. Published results are generally reported at $n{=}80$ (Song et al., 2024; Li et al., 2024; Zhang et al., 2025; Wu et al., 2025; Zhang et al., 2026; Zheng et al., 2026), the operating point suited to the question those studies address, namely how much visual information an EEG response carries in principle.

This paper evaluates every method at both extremes. The $n{=}80$ operating point measures what averaging all available repetitions can recover; the $n{=}1$ operating point measures what a deployed system would obtain from a single trial of a stimulus. To our knowledge, the two operating points have not previously been compared side by side on this benchmark; Section 2.1 provides the comparison. Appendix B details the three protocol variables that differ across the literature (test-repetition averaging, train-repetition averaging, and montage) and how we hold each fixed in the comparison.

2.1 Six published methods at both operating points

We evaluate six published methods, BraVL (Du et al., 2023), NICE (Song et al., 2024), ATM-S (Li et al., 2024), UBP (Wu et al., 2025), NeuroBridge (Zhang et al., 2026), and Brain-HIVE (Zheng et al., 2026), at both operating points. Every number is produced by the method's own released code, and the only difference between the two evaluations is the number of responses averaged into a query. Each value is the mean over ten subjects and three independent runs, and all six methods reproduce their published accuracy to within one point.[1]

Figure 1a shows the outcome, with methods ordered by publication date. Under the standard $n{=}80$ readout, accuracy rises from 20.7% for BraVL to 76.2% for Brain-HIVE. When the same six trained models decode single trials, accuracy falls to between 2.6% and 12.0%. Figure 1b expresses the same result as *retention*, the ratio of a model's $n{=}1$ accuracy to its $n{=}80$ accuracy. Retention ranges from 0.098 (UBP) to 0.285 (ATM-S), and it is lower for the more recent methods, whose accuracy relies most on image-side priors. This is not a deficiency of those methods: each was designed and

[1]Our independent re-implementations of all six methods reproduce the same rankings at both operating points (Appendix D). The complete numeric grid, with published values beside reproduced values and full provenance, appears as Table 5 in Appendix C. An anonymized repository containing the full implementation, configurations, and evaluation logs of every experiment in this paper will be released during the discussion period.

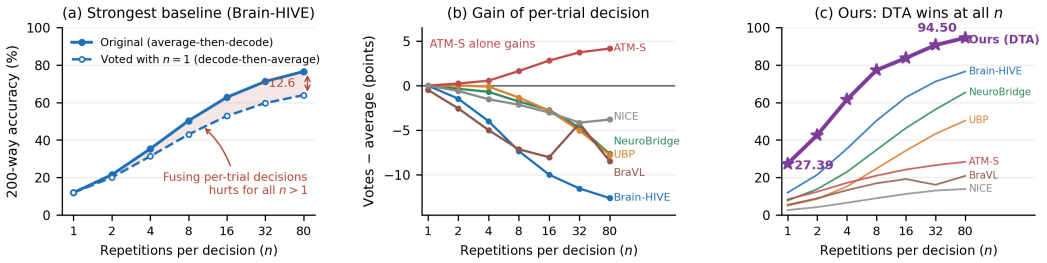

Figure 2: **Decision averaging versus signal averaging across repetition counts.** *(a)* Averaging Brain-HIVE's per-trial decisions loses to its original average-then-decode readout at every $n > 1$, by 12.6 points at n=80. *(b)* The gain of decision averaging tracks per-trial competence: it reduces accuracy for five of the six methods, and the single method it improves, ATM-S, has the highest single-trial retention as shown in Figure 1b. *(c)* Our single-trial decoder applying decode-then-average exceeds every published method at every repetition count.

validated at the averaged operating point, and components specialize toward the setting in which they are selected. The consequence, however, is that accuracy measured at one operating point does not predict accuracy at the other, and the two must be measured separately.

2.2 Decision averaging alone does not surpass the standard readout

The released repetition structure invites a question that requires no new decoder: since the 80 repetitions of each test image are available individually, their *decisions* could be fused in place of their signals. Figure 2a evaluates this substitution on Brain-HIVE, the strongest published method, using its own released code and trained model throughout (Appendix C). The comparison is between the original, unmodified Brain-HIVE under the field's default readout, which averages the n repetitions into one query and decodes it once (average-then-decode), and the same trained model reading the same n repetitions as n separate decisions whose scores are averaged (decode-then-average). Decision averaging raises Brain-HIVE's accuracy far above its single-trial value, from 12.0% at n=1 to 64.0% when all 80 decisions are fused. It nevertheless remains below the original average-then-decode readout at every repetition count, by 12.6 points at n=80. For the unmodified Brain-HIVE, replacing signal averaging with decision averaging degrades accuracy at every n greater than one.

Figure 2b extends the same comparison to all six methods, reporting for each the gain of decision averaging, defined as decision-averaged accuracy minus average-then-decode accuracy, at every repetition count. Five methods lose from decision averaging at nearly every count. One method gains: ATM-S improves at every count beyond a single trial, by 4.2 points at n=80. The property that distinguishes ATM-S is visible in Figure 1b: it retains 0.285 of its averaged accuracy on single trials, the highest retention of the six. These observations yield the organizing principle of this section: **decision averaging pays in proportion to the quality of the individual decisions**. Weak single-trial decisions are better averaged before decoding; strong ones are better decoded first. The principle explains why average-then-decode has remained the field's default, and it raises the question this section is built to answer:

> **Hypothesis.** Given a sufficiently strong single-trial decoder, decision averaging will not merely stop hurting but will surpass signal averaging, at every repetition count.

The remainder of this section constructs such a decoder and tests this hypothesis.

2.3 A decoder built for single trials

Our decoder builds on the contrastive-alignment pipeline that underlies the strongest published methods: EEG is encoded and aligned to pretrained image embeddings with a symmetric contrastive loss, following the architecture family of Brain-HIVE (Zheng et al., 2026) under NICE-lineage preprocessing (Song et al., 2024); the full specification is in Appendix E. Three

choices are specific to the single-trial regime. Training uses the individual responses rather than their averages. The encoder is wide and shallow, a single-hidden-layer projection, which under single-trial noise outperforms the deep transformer encoders common at n=80. Regularization is strong throughout, matched to the low signal-to-noise ratio of individual trials. The remaining components are techniques established in prior work: multi-subject pretraining with per-subject fine-tuning (Défossez et al., 2023), alignment targets drawn from intermediate layers of the vision backbone and compressed to low rank (Zheng et al., 2026), two fused analysis windows (Dietterich, 2000), and an unsupervised test-time calibration (Radovanović et al., 2010; Lample et al., 2018). One further choice, a 596 ms analysis window in place of the customary $\sim$1000 ms, improves accuracy at every repetition count: the decodable signal concentrates in the first 400 ms after stimulus onset (Appendix G). We claim none of these as contributions. What is new is their systematic tuning and deployment for the single-trial operating point.

Configuration	n=1 (%)
Base (standard recipe)	9.50
+ Multi-subject training	12.74
+ Intermediate-depth targets	15.53
+ 596 ms analysis window	17.32
+ Single-trial training	23.26
+ Regularization, montage, optimizer	25.23
+ Fusion, calibration (dev)	27.36
Frozen test (pre-registered)	**27.39**

Table 1: From the standard recipe to 27.39%: each row adds one configuration choice. Steps are path-dependent and are not isolated effects (Appendix F.1).

Table 1 is the complete record of this reconfiguration: it documents each choice's contribution on the path from the standard recipe at 9.50% to the final 27.39%, so that which techniques matter at n=1, and by how much, is reported. Hyperparameters, the protocol, and per-component profiles appear in Appendices A, E and F. The final number is protected from selection effects by pre-registration: every configuration choice was committed in writing before the frozen test data were accessed, and the frozen runs were evaluated once.

2.4 With the right decoder, decision averaging wins at every repetition count

Figure 2c completes the test of the hypothesis. The decoder of Table 1 is read out in both ways at every repetition count, and decision averaging now wins everywhere: the averaged decisions reach 94.50% at n=80 and exceed the accuracy of every published method, under either readout, at every repetition count n. At matched models, decode-then-average finishes 2.3 points above the dedicated average-then-decode pipeline at the standard operating point, a pre-registered contrast (Appendix J). The hypothesis of Section 2.2 is therefore confirmed: a decoder configured for single trials does not trade the standard operating point for the deployment one. Sections 3.1 and 3.2 identify the sources of the advantage, and Section 4.1 develops the capability the per-trial decision stream opens.

3 Why decision averaging outperforms signal averaging

Section 2.4 established that decision averaging (decode-then-average) surpasses signal averaging (average-then-decode) at every repetition count once the single-trial decoder behind it is sufficiently strong. This section establishes why, through two mechanisms measured directly on the data, one on the training side and one at test time. Their division of labor explains how one single-trial decoder serves both operating points: training on individual trials outperforms training on averages at n=1, and averaging its per-trial decisions outperforms averaging the signal at n=80.

3.1 Data scale: more samples beat better samples

Training on individual trials rather than on their averages contributes $+5.94$ points at n=1 (Table 1). The gain is not noise matching. At *equal* sample count, training on 4-trial averages surpasses training on single trials by 3.30 points even under single-trial evaluation: an average is a better training sample than a trial. What averaging sacrifices is quantity. Collapsing four trials into one average shrinks the training set fourfold, and accuracy grows log-linearly with the number of training samples: each doubling of the training set adds 3.42 ± 0.40 points, steadily across the whole measured range. Figure 3a resolves the competition between the two effects directly: as the number of retained trials grows, the single-trial curve passes the fixed 4-average recipe near twice its sample count and finishes 2.24 points ahead once every recorded trial is kept. More samples beat better samples.

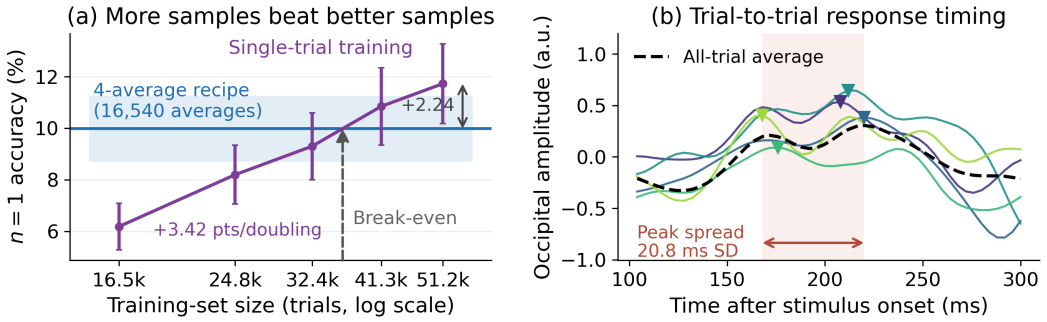

Figure 3: **The two mechanisms behind single-trial decoding.** *(a)* Accuracy of single-trial training at n=1 as the training set grows, against the standard recipe trained on the same recordings' four-trial averages. Accuracy grows log-linearly with the number of training trials and passes the averaged recipe near twice its sample count. *(b)* Five single-trial occipital responses to one test image (subject 8, channels O1/Oz/O2, smoothed; the trials most correlated with the average are shown). The peak of each smoothed response is marked, showing the misalignment that signal averaging superimposes before decoding and decision averaging avoids.

This remedy belongs to the single-trial operating point alone. Repeating the analysis at n=80 under signal averaging, three times the training data still leaves single-trial training 16.5 points behind, and no realistic data volume closes the gap. Training data repairs the single-trial operating point; the averaged operating point is won instead by the readout of Section 2.4. One decoder, two remedies, each applied where it works. The complete five-arm measurement, its per-subject break-even statistics, and the two arms whose values were predicted before being run appear in Appendix F.

3.2 EEG RESPONSE TIMING: A SINGLE TRIAL PRESERVES WHAT AVERAGING BLURS

Visual responses do not arrive on a fixed clock. Cross-correlating each trial against its subject's average response measures the natural latency jitter of this dataset: a shift of 9.9 ± 2.8 ms standard deviation across the ten subjects. Figure 3b makes the jitter visible in raw data: single-trial responses to the same image peak at visibly different times around the average response. This jitter determines where averaging does its damage. Signal averaging superimposes the misaligned responses *before* the encoder sees them, so the components that identify the image are blurred in every averaged query. Decision averaging encodes each trial in its own timing frame and averages afterward, where timing no longer matters. A causal test confirms the mechanism: we inject uniform jitter $\pm\tau$ into the test recordings and re-evaluate our committed pipeline and Brain-HIVE's released pipeline on the same shifted trials. At τ=0 both cells reproduce their reference values, so the comparison is matched by construction. At ± 8 ms, a dose close to the naturally measured 9.9 ms, decision averaging loses less than half as much as signal averaging in relative terms, as shown in Table 2. The advantage of decoding before averaging thus rests on a measurable property of the EEG response itself.

	τ=0	τ=8	Drop
Ours (DTA)	94.5	90.8	−3.7 (3.9%)
Brain-HIVE (ATD)	75.6	69.5	−6.1 (8.1%)

Table 2: Accuracy (%) at n=80 under injected latency jitter $\pm\tau$ ms.

4 THE DECODABILITY OF A STIMULUS

Decision averaging, compared to signal averaging, changes more than a benchmark score. A pipeline that averages the signal must fix the number of trials before its single decode. By contrast, a pipeline that averages the decisions receives an answer after every trial, watches that answer evolve as evidence accumulates, and can stop the moment the answer settles. The number of trials stops being a constant of the experimental protocol and becomes a quantity under the decoder's control during inference. This section develops the two consequences: an adaptive, real-time stopping rule (Section 4.1), and a stable per-stimulus property that the rule exposes (Section 4.2).

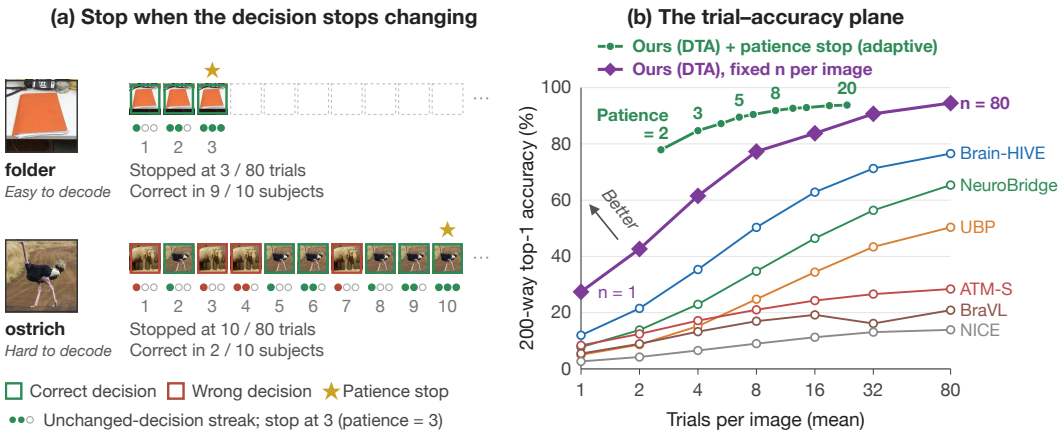

Figure 4: **Real-time decoding with a patience stop.** *(a)* The decode-then-average (DTA) accumulates the averaged decision after each new trial. The stream stops once the decision has been unchanged for three consecutive trials (patience $= 3$). An easy image, *folder*, settles after 3 of 80 available trials and is decoded correctly in 9 of 10 subjects; a hard image, *ostrich*, repeatedly loses its streak to a close competitor, *elephant*, before settling at 10, and is correct in 2 of 10 subjects. *(b)* 200-way top-1 accuracy against mean trials per image ($\log_2$ axis). DTA with a fixed n (purple) exceeds all six baselines at every budget, from 27.4% at n=1 to 94.5% at n=80. Adding the patience stop (green, labeled by patience) traces the Pareto frontier: at 2.6 trials per image it matches Brain-HIVE's n=80 accuracy (76.6%), and at 23 trials it is within 0.7 points of the DTA ceiling.

4.1 Real-time decoding: the number of trials becomes a decision

Figure 4a shows the decoder at work on two images from the test set. The same image is presented repeatedly, and after every trial the running average of the per-trial scores yields one decision. For *folder*, the first decision is correct and never changes. For *ostrich*, the running average is captured repeatedly by a close competitor, *elephant*: the decision flips four times before settling on the correct answer at the tenth trial. The two streams suggest a rule: the decoder should present an image until its answer stops changing. We call this rule a *patience* stop. Presentation halts at the first trial at which the running-average decision has been unchanged for three consecutive trials, and the stable decision is the answer. The rule is early stopping moved from training time to inference time. Training stops when the loss stops improving, and presentation stops when the answer stops changing.

Figure 4b places every method on one trial–accuracy plane, and two results are visible at once. Decode-then-average with a fixed trial count exceeds all six published baselines at every count, from 27.4% at a single trial to 94.5% at n=80. The adaptive patience stop then pushes past the fixed-count curve itself. Swept over patience values, it matches Brain-HIVE's 80-trial accuracy at a mean of 2.6 trials per image, a 96.8% reduction with respect to the 80-trial protocol, and comes within 0.7 points of the 80-trial ceiling at 23 trials, where a fixed allocation needs 44 trials for the same accuracy. The advantage holds for each of the ten subjects individually.

The saving grows with the accuracy level, from 18% of trials at a patience of three to 47% at a patience of twenty. Figure 4a shows why. A fixed allocation keeps re-presenting images that settled on their third trial, while the patience stop spends trials only where the answer is still moving. The adaptive rule pays because images differ, and the difference is a stable property of the stimulus.

4.2 What makes an image hard to decode

Figure 5 maps the property over the whole test set. Under a patience stop of 3, as demonstrated in Figure 4, each of the 200 test images has a characteristic cost. The mean is 5.1 trials to a stable decision; *folder* needs 3.7 ± 0.9 trials (mean $\pm$ std across the ten subjects) and *ostrich* 6.5 ± 1.8 (Figure 5b). We call the underlying property the image's *decodability*. Decodability is expressed not only in the number of trials but also in accuracy: at the same patience, the stable decision is correct

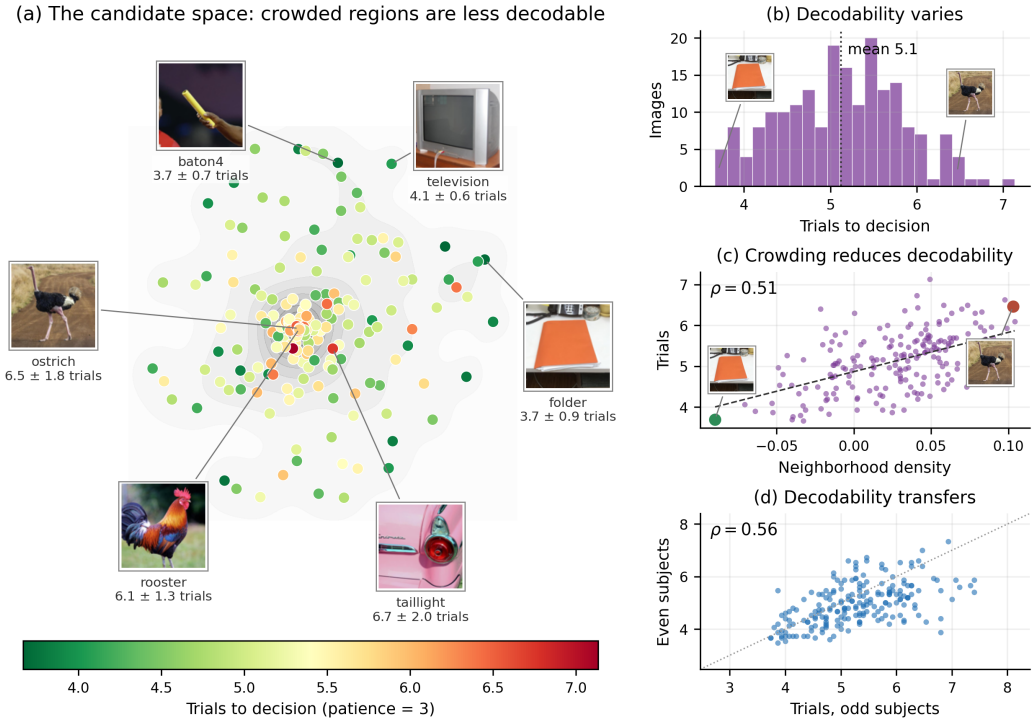

Figure 5: **What makes an image hard to decode.** *(a)* The 200 candidate images arranged so that crowding in the candidate space is visible: the angular position of each image comes from a two-dimensional embedding of pairwise similarity, and the radial position follows neighborhood-density rank. Each image is colored by the number of trials needed for a patience of three. Images in the crowded core cost the most trials; images in sparse regions are decoded from few. *(b)* The distribution of trials to decision over the 200 images. *(c)* Neighborhood density against trials to decision ($\rho = 0.51$). *(d)* Per-image trial counts measured on one half of the subjects against the other half ($\rho = 0.56$).

for *folder* in 90% of runs and for *ostrich* in 30%. A hard image costs more trials and is decided wrongly more often.

The map explains the variation. In Figure 5a each image is placed by its similarity to the other 199 candidates, with crowded images central and isolated images on the rim, and colored by its cost. Images in the crowded core cost the most trials, and neighborhood density correlates with cost at $\rho = 0.51$ (Figure 5c). The competitor that repeatedly captured *ostrich*'s running average in Figure 4a is such a neighbor: *elephant* occupies the same crowded region of the candidate space. What makes an image hard to decode is the geometry of the candidate space around it, not its pixels and not the noisiness of the response it evokes (Appendix L).

The property belongs to the image rather than to the subject. Per-image costs measured on one half of the subjects predict the other half at $\rho = 0.56$ (Figure 5d), and the ordering is not an artifact of the stopping rule: it agrees at $\rho = 0.87$ with a consensus-based measurement that lets each stream run until three single-trial decisions agree. Under that longer-stream measurement the across-subject transfer reaches $\rho = 0.81$, meeting the pre-registered criterion of $\rho \geq 0.5$; leave-one-subject-out prediction attains 0.66 ± 0.04 and anticipates a held-out subject's per-image trial counts to a median error of three trials (Figure 12, appendix).

An image's decodability is therefore known before a new subject sits down. An experimenter can budget trials per stimulus in advance rather than uniformly, and a closed-loop application can spend its session time where accuracy is still being purchased. Our results suggest that the fixed-repetition protocol, like the averaged readout it serves, is a convention rather than a necessity: once decisions

exist per trial, both the number of trials and their allocation across stimuli become choices the decoder can make at run time.

5 Related work

EEG-to-image decoding. Decoding perceived images from M/EEG has converged on contrastive alignment of neural epochs to pretrained image embeddings. BraVL couples the modalities through a trimodal generative model (Du et al., 2023), NICE formulates the CLIP-alignment recipe (Song et al., 2024), and subsequent systems add inverted-transformer encoders, modality experts, whitened flattened MLPs, and image-side priors (Li et al., 2024; Zhang et al., 2025; Wu et al., 2025; Zhang et al., 2026; Zheng et al., 2026); related lines decode from fMRI (Scotti et al., 2023) and MEG (Benchetrit et al., 2024). Our decoder is assembled from this line's own components and claims no architectural novelty. What we add is the single-trial operating point, measured for every method beside the averaged one, and a readout under which one trained model serves both.

Single-trial decoding outside this benchmark. Where repetitions are unavailable, decoding is single-trial by necessity: P300 and RSVP interfaces operate on individual epochs (Blankertz et al., 2011), speech decoding evaluates on unaveraged segments (Défossez et al., 2023), and the THINGS initiative provides designs in which test-time averaging is impossible (Grootswagers et al., 2022; Hebart et al., 2019). Two of our components come from this literature: the learned montage adapts the spatial attention of Défossez et al. (2023), whose own ablations are unaveraged, and the calibration is hubness correction (Radovanović et al., 2010), operationally CSLS (Lample et al., 2018; Bogolin et al., 2022). To our knowledge, neither had previously been applied on this benchmark.

Sequential decisions and adaptive stopping. The patience stop belongs to the tradition of sequential analysis, in which data collection halts once accumulated evidence meets a criterion rather than at a sample size fixed in advance (Wald, 1945). Brain-computer interfaces inherited this idea as dynamic stopping: P300 spellers that adapt the number of stimulus repetitions to the accumulated evidence spend few repetitions on confident decisions and more on ambiguous ones (Wolpaw et al., 2002; Mainsah et al., 2015). Our contribution to this tradition is the per-stimulus reading: the stopping time is not only a budget control but a stable property of the stimulus, transferable across subjects.

6 Conclusion

We proposed the single trial, and not the average, as the default operating point of EEG-to-image decoding. A decoder assembled from standard components reaches $27.39 \pm 2.15\%$ from single trials under a pre-registered frozen evaluation, and, read out per trial with the decisions averaged, $94.50 \pm 0.79\%$ from all 80, exceeding every published method's accuracy under either readout at every trial count. The mechanisms are measurable: single-trial training recovers the volume that averaging consumes, and per-trial decisions preserve the timing that signal averaging blurs. A patience stop then matches the strongest published method's 80-trial accuracy from 2.6 trials per image, and the per-image cost it exposes, decodability, is a stable, predictable property of the stimulus. Averaged evaluation remains informative, and the methods examined here are substantive contributions under it; our results suggest that single-trial decoding is more tractable than reported, once decisions, not signals, are aggregated.

Limitations. All measurements are on one dataset, and whether these findings hold beyond THINGS-EEG2 is the largest open question. The vote-fusion control of Section 2.2 was run on the strongest baseline only, and the per-repetition curves of Figure 2 read one run per method at its own best epoch. Decode-then-average is more robust to latency jitter than the averaged readout, not invariant to it (Table 2). The decodability analyses of Section 4.2 are exploratory and measured on development seeds, with the across-subject transfer as their one pre-registered criterion. The $n{\approx}80$ headline is a reporting-set quantity, as is every published $n{=}80$ number, since averaging all repetitions leaves nothing held out; only the $n{=}1$ headline has a genuinely frozen denominator (Appendix F).

AI USE STATEMENT

Generative AI tools were used throughout this project: for drafting and revising the manuscript text, for writing and debugging experiment and analysis code, and for literature search. All experiments were executed by the pipeline described in the paper; every number in the manuscript is read from an append-only experiment ledger and cross-checked against the logged outputs of the corresponding runs, and the single frozen-test evaluation was pre-registered before the models it scores had reported. AI-generated code paths that carry evidence (the retrieval metric, the exclusion logic, the seed-averaging) are covered by the unit and consistency checks described in Appendix C. The authors reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only the publicly released THINGS-EEG2 dataset (Gifford et al., 2022), collected by its original authors under their institutional ethics approval with informed consent; no new human-subjects data were recorded. The audit findings in Section 2 concern published evaluation protocols and released code, not the conduct of any individual; wherever a published number does not reproduce, we report the measured gap and the specific configuration ambiguity that explains it, and we distinguish protocol artifacts from method contributions explicitly. Improved EEG decoding has assistive applications (communication for locked-in patients) and, at much higher accuracy than achieved here, potential privacy implications; at 27.4% single-trial top-1 over 200 known candidate images with per-subject calibrated equipment, the latter remain remote.

REPRODUCIBILITY STATEMENT

The headline number comes from one pre-registered evaluation of a frozen test split: the configuration and decision rule were committed before the models being scored had reported (Section 2.3 and Appendix F), and every frozen-set access is logged. All baseline numbers in the main text are produced by each method's own released code, three independent runs per cell, with only the recorded environment and test-path patches applied (Section 2.1); no claim there depends on a reimplementation. An independent re-implementation of all six is reported separately, with a stated cause for every difference (Appendix D). Each baseline is an isolated module whose deletion removes exactly its own rows, enforced by an automated check. Appendix E specifies the preprocessing, splits, exclusions, and hyperparameters; Appendix C lists the reproduction subtleties of each released codebase (dead flags, silent no-op checkpoint restores, order-dependent in-batch labels) that must be handled to reproduce the grid. An anonymized repository with the source code, the experiment ledger, and the pre-registration document will be released during the discussion period, and publicly upon publication.

REFERENCES

Yohann Benchetrit, Hubert J. Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024. URL https://openreview.net/forum?id=3y1K6buO8c.

Benjamin Blankertz, Steven Lemm, Matthias Sebastian Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage*, 56 (2):814–825, 2011. doi: 10.1016/j.neuroimage.2010.06.048.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 5184–5195, 2022. doi: 10.1109/CVPR52688.2022.00513.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS)*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. doi: 10.1109/TPAMI.2023.3263181.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw Martin Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

Tijl Grootswagers, Ivy Zhou, Amanda K. Robinson, Martin N. Hebart, and Thomas A. Carlson. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. doi: 10.1038/s41597-021-01102-7.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR*, 2018. URL https://openreview.net/forum?id=H196sainb.

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 37, NeurIPS*, volume 37, pp. 102822–102864, 2024. doi: 10.52202/079017-3266.

Boyla O. Mainsah, Leslie M. Collins, Kenneth A. Colwell, Eric W. Sellers, David B. Ryan, Kevin Caves, and Chandra S. Throckmorton. Increasing BCI communication rates with dynamic stopping towards more practical use: an ALS study. *Journal of Neural Engineering*, 12(1):016013, 2015. doi: 10.1088/1741-2560/12/1/016013.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL https://jmlr.org/papers/v11/radovanovic10a.html.

Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems 36, NeurIPS*, volume 36, pp. 24705–24728, 2023.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

Kaido Värbu, Naveed Muhammad, and Yar Muhammad. Past, present, and future of EEG-based BCI applications. *Sensors*, 22(9):3331, 2022. doi: 10.3390/s22093331.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002. doi: 10.1016/S1388-2457(02)00057-3.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2246–2257, 2025.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025. doi: 10.1609/aaai.v39i13.33587.

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, pp. 18028–18036, 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 *The Fourteenth International Conference on Learning Representations, ICLR*, 2026. URL `https://openreview.net/forum?id=IEq71qS8B7`.

A A DECODER FOR ONE TRIAL: THE RECIPE IN COMPONENT DETAIL

This appendix expands the recipe paragraph of Section 2.3 into its component decisions and specifies the decoder whose behavior Appendix I profiles. Three design decisions define it: what a training sample is, how the 63 electrodes are combined, and what representation the EEG is aligned to. Each decision is validated in isolation, with the rest of the system held fixed, and the three isolated gains of $+4.24$, $+5.23$, and $+2.34$ points are each comparable to the gaps that separate entire published methods at a single trial, where the six span 2.61% to 12.02% (Table 5). Whether a component belongs to the universal core or to one regime's branch is decided by its measured profile in Appendix I.2, not assumed here. All evaluation uses the field's own test-side protocol, every hyperparameter is listed in Appendix E, and the frozen runs were evaluated once, under a single pre-registered document that fixed both operating-point configurations and every decision rule before any frozen access.

A.1 DATA: TRAIN ON THE TRIALS THE DECODER WILL FACE

The standard recipe averages the four training presentations of an image before training on them. Under averaged evaluation this is coherent, as the model is trained and tested on inputs of the same kind. At a single trial it creates a mismatch, and the model must decide about an input noisier than any it has seen. We therefore train on the individual trials, and the choice is worth $+4.24$ points at $n{=}1$ ($p{=}0.002$, ten of ten subjects), with the gain persisting, though smaller, under the field's own hyperparameters. The choice is also symmetric. A model trained on averages is specialized to averages in the same way, the two accuracy curves cross as the number of averaged repetitions grows, and each recipe wins at the noise level it is trained for (Figure 6). Section 3.1 decomposes the gain and shows that its mechanism is the recovered data volume rather than noise matching; the design rule, however, is the same under either account: a decoder that will be deployed on single trials should be trained on them, on this benchmark and under any protocol that averages repetitions its deployment will not have.

A.2 ENCODER: CONSTRAIN THE SENSOR COMBINATION INSTEAD OF CURATING SENSORS

Recent systems hand-pick 17 occipito-parietal electrodes and flatten them into an MLP (Wu et al., 2025; Zheng et al., 2026). Given all 63 electrodes, the same architecture loses accuracy, so the curated montage is compensating for a limitation of the encoder, and identifying that limitation tells

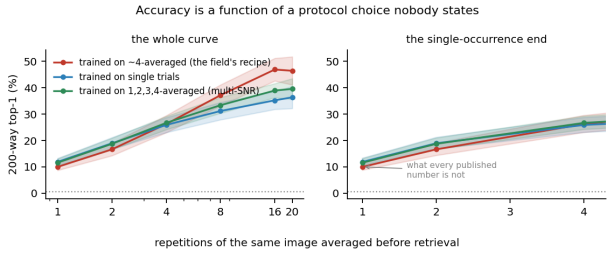

Figure 6: **Accuracy as a function of the number of averaged test repetitions is a curve, of which published work reports single points, almost always at n=80.** Three training regimes of the same architecture (single-trial, the 4-averaged trials used by several methods, and a mixture) *cross*: which model is preferable depends on an evaluation parameter that is rarely stated (Appendix H).

us what to build. Two controls isolate it. A learned per-channel gate can express exactly what the montage does, since driving a gate to zero discards a sensor, yet it recovers nothing. An unconstrained learned combination of electrodes is strictly more expressive and uses fewer parameters, yet it collapses accuracy from 15.87% to 3.15%. Neither expressiveness nor capacity is the limiting factor; what is missing is a constraint. Following Défossez et al. (2023), we therefore learn the combination as spatial attention whose weights are a truncated Fourier series in the sensor coordinates, so that they vary smoothly across the scalp. The constraint is worth $+5.23$ points on the full array (ten of ten subjects), because it reduces the function class to one that the available noisy trials can estimate. The gain is a repair rather than a general improvement: on the curated 17-channel montage the constrained combination changes nothing, and its value is that the full array becomes usable, so electrode selection is no longer a design decision. The component also illustrates the super-additivity of Appendix I.1, as it hurts when added alone to a lightly regularized system yet costs accuracy when removed from the deployed one. The construction, the full diagnosis, and the basis ablation that settles the mechanism follow below (Appendix A.3).

A.3 The montage in detail: construction, diagnosis, and the channel ladder

Each of J virtual channels is a softmax mixture over the 63 electrodes whose logits are a truncated 2-D Fourier series in scalp position $(u_c, v_c) \in [0,1]^2$:

$$a_{jc} = \frac{\exp z_{jc}}{\sum_{c'} \exp z_{jc'}}, \qquad z_{jc} = \mathrm{Re} \sum_{k,l=0}^{K-1} w_{kl}^{(j)}\, e^{\,2i\pi(ku_c + lv_c)}, \qquad \tilde{\boldsymbol{x}}_j = \sum_c a_{jc}\, \boldsymbol{x}_c, \tag{1}$$

so only the $2K^2$ real coefficients per virtual channel are learned (K=16).

The 17-channel occipital montage the strongest systems hand-pick was not information-optimal but architecture-compatible: a count-matched frontal control decodes at $1.2\times$ chance, yet including channels of that kind cost a flattened MLP 46% of its accuracy where it was first measured (27% here), because it has no way to express "ignore this sensor." One learned scalar per channel does nothing; the softmax mixture recovers the entire penalty, and with multi-subject training all-63 *overtakes* the hand-curated occ17 by $+1.42$. The mechanism ablation matters most. An unconstrained mixture with the identity basis collapses (15.87 $\rightarrow$ 3.15), so it is the Fourier smoothness constraint, not parameter count and not gating per se, that makes the montage learnable at this sample size. This overturned our initial hypothesis, which credited the mixture structure.

The channel ladder. Table 4 measures nested channel sets outward from the occipital pole under a flattened-MLP encoder, so that adjacent differences are the marginal value of the channels added. Three readings follow. The parieto-occipital row is the single largest effect in the project, $+7.12$ over O-only. The parietal row is free ($+0.59$, n.s.), so the field's 17-channel montage is a plateau from 8 to 17: half those electrodes do nothing measurable, which matters for practical systems. And every block beyond visual cortex *costs*, monotonically, in 9–10/10 subjects, even though the added channels are individually at chance (ant17). This is the flattened-MLP pathology the learned

Table 3: **Left: the montage diagnosis and fix** (single model, dev split, accuracy in %). The scalar gate recovers nothing; the spatial-attention mixture recovers the whole montage penalty ($+5.23$ on all63, p=0.0020, 10/10), bringing all 63 electrodes to parity with the hand-curated 17 (-0.13, n.s.). All values are at the K=16 the method deploys. The first three rows are span 596; the basis ablation and its reference are span 996, so that comparison is within-span. The basis-free control has *fewer* parameters than the K=16 mixture. **Right: the location control.** Frontal channels decode at chance, yet adding channels of exactly that kind to occ17 cost 46% of accuracy in the channel-ladder configuration where that was measured (27% in this table's configuration).

Encoder	occ17	all63
flattened MLP (Zheng et al., 2026)	**19.66**	14.30
+ per-channel scalar gate	19.64	14.36
+ spatial attention (ours, K=16)	19.23	**19.54**
basis ablation, span 996 throughout		
spatial attention	18.66	15.87
– positional basis (fewer params)	6.36	3.15

Montage	Accuracy
occ17 (occipito-parietal)	15.72
ant17 (frontal, count-matched)	0.58
chance	0.50

montage above repairs: with spatial attention, multi-subject training and the full target set, all-63 beats occ17 by $+1.42$.

Table 4: **Nested channel sets outward from the occipital pole** (flattened-MLP encoder, tuned recipe, PCA-256 targets, dev split). Each set contains the one above, so adjacent differences are the marginal value of the channels added. ant17 is the count-matched, location-mismatched control.

Set	n	Accuracy	vs. occ17	Marginal block
occ3 (O)	3	8.01	-7.71 (p=0.0020)	—
occ8 (+PO)	8	15.13	-0.59 (n.s.)	$+7.12$, 10/10
occ17 (+P)	17	15.72	—	$+0.59$, n.s.
occ26 (+CP/TP)	26	13.87	-1.85 (p=0.0039)	-1.85, 1/10
occ37 (+C/T)	37	12.27	-3.45 (p=0.0020)	-1.59, 1/10
all63	63	8.46	-7.26 (p=0.0020)	negative
ant17 (frontal)	17	0.58	-15.14 (p=0.0020)	—

A.4 Target: align to intermediate layers, and stop at three

Contrastive decoders align EEG to the final embedding of a vision model. A single trial, however, may say more about intermediate visual representations than about the final semantic one, so we align to both. The encoder is trained with a symmetric InfoNCE loss against a fusion of image embeddings, namely CLIP ViT-B/32, ViT-H/14 and RN50 (Radford et al., 2021; Cherti et al., 2023), a VAE latent, and mid-depth blocks of ViT-H/14, and adding the first such block is worth $+2.34$ points (ten of ten subjects). The preference for intermediate depth is an ordering, not a tuning artifact: across seven read-out depths the same block wins at both operating points, and the depth ranking is nearly identical between them, with Kendall τ=0.905 (Appendix H). This axis therefore transfers across the averaging boundary, in contrast to the encoder axis, which reverses.

The axis also saturates. Three intermediate blocks are used, a fourth changes nothing measurable in either direction, and the probes below (Appendix A.5) show that the ceiling is a property of the targets rather than of the fusion rule. Every alignment branch is further compressed to a rank-128 PCA subspace, a choice Appendix I.2 examines across the full averaging range.

A.5 The fusion probes behind the target ceiling

Our fusion is Brain-HIVE's: each branch is PCA-compressed, projected by its own linear map, and the projections are *summed* before a single residual block and a single contrastive loss (Zheng et al., 2026). The branches are summed un-normalised, so each contributes in proportion to the norm of

its projection, and the saturation reported above could in principle be an artifact of that fusion rather than a property of the targets. It is largely not. After PCA the five candidate intermediate branches have near-identical norms (within 16% of one another), so neither rejected branch is being swamped in the sum. For the shallower candidate the null result is explained outright: block 2 is 98.8% linearly predictable from blocks 4, 8 and 16, so it is close to a duplicate. The deeper candidate is not explained this way: block 20 retains about a fifth of its variance independent of all seven branches already present, a margin comparable to block 16's own, yet adds nothing. To ask whether the summed fusion is what excludes it, we gave the fusion a learned competing scale per branch, initialised so that the weighted model starts exactly at the unweighted one. It does *not* exclude the branch: block 20 is assigned a scale of 0.97, essentially the value it starts at and higher than either of the two largest branches, and accuracy still does not move significantly ($+0.34$, p=0.19, 7/10, one draw per arm). So the ceiling is not the fusion's inability to admit a fourth branch. Separately, the four non-intermediate branches are *not* balanced: their post-PCA norms span a factor of 31 (CLIP ViT-B/32 lowest, the VAE latent highest). The learned weighting pushes against this, cutting CLIP ViT-H/14 and the VAE latent to ≈ 0.6 while raising the intermediate branches to ≈ 1.2 in both runs. It only partly corrects the imbalance: effective contributions still span a factor of 18. That imbalance is inherited from Zheng et al. (2026) and we did not tune it.

A.6 Multi-subject pretraining, fusion, and calibration

The remaining components are standard, and we report them so that the final number can be decomposed. The decoder is pretrained jointly on all ten subjects through an identity-initialized per-subject channel map and then fine-tuned per subject, which is worth $+1.80$ points in isolation. At test time, two models reading the same trial through two window alignments are combined by an equal-weight mean of their standardized scores. Each candidate's fused score is then corrected by its mean similarity to its nearest queries from the same session. This is hubness correction, operationally the CSLS of Lample et al. (2018), and we claim no novelty for it. The finding is that the optimal neighborhood size is not a constant of the method but falls with the number of available queries, a regularity Appendix I.2 quantifies, and the committed schedule fixes it per operating point.

Calibration of this kind can be computed at three levels of test-time access, and the levels support different claims, so we keep them distinct. Computed from the scored query alone, it is worth nothing. Computed from unlabelled queries of the same distribution, which is the realistic deployment mode and the tier we report, it is worth $+1.20$ points. Computed transductively over the scored queries themselves, it is worth slightly more, and this tier is never a headline. Estimating the same statistic from held-out training EEG instead yields no significant gain, which identifies the mechanism as unsupervised test-time calibration rather than a learned improvement; Appendix K records the reasoning error that this three-tier reporting corrects.

B Protocol variables across the literature

Three parameters of the evaluation protocol vary across published work on this dataset. Each is a legitimate choice for the question its paper asks; we record them because a number is only comparable to another number computed the same way, and we hold all three fixed when placing methods side by side in Table 5.

Test-repetition averaging. Accuracies at n=80 describe a decision formed from $\sim$16 s of recording per image. In our measurement, averaging the 80 repetitions removes 93.6% of the trial-to-trial variance (sub-01, occ17); the residual is the component reproducible across repetitions, which is what averaging is intended to retain. This is the parameter Table 5 varies, and it is varied on a single trained model per method, so the two columns differ in grouping alone.

Train-repetition averaging. Two lineages differ here. NICE, UBP, NeuroBridge and Brain-HIVE average the four training repetitions, yielding 16,540 training samples; ATM-S and CognitionCapturer keep them separate, yielding 66,160. The choice is consequential independently of any method contribution, and its size is visible within a single published ablation: UBP's own "vanilla" control, with its novel components removed, scores 42.1 against ATM-S's 28.5. Roughly 14 of the $\sim$22-point difference between those two systems is therefore attributable to the preprocessing lineage

rather than to the components under study. Our own training-data measurement (Appendix A.1) probes this axis directly and finds it matters most at n=1.

Montage. NICE, ATM-S and CognitionCapturer use all 63 electrodes; BraVL, UBP, Brain-HIVE and NeuroBridge use 17 occipito-parietal channels. Brain-HIVE's own ablation puts the two at 75.7 and 62.5 respectively. Every method above ∼40% both averages the training repetitions and uses the 17-channel montage, so these two variables are correlated with method identity across the published record; this motivates measuring them inside one pipeline (Appendix A.2) rather than across papers.

C Reproduction notes on the released codebases

These are the defects and ambiguities that must be handled to reproduce Table 5; each is recorded with file/line evidence in the project ledger of the anonymized repository.

Table 5: **Two operating points, two readouts: the full grid** (200-way top-1 accuracy in %, mean over 10 subjects; each reproduced value is the mean of three independent draws). Every baseline number is reproduced by running the method's own released code, and the method's published value is shown beside it where one exists. The six published methods read n=80 by averaging the EEG before one decode (Average-Then-Decode); our decoder reads it by fusing its 80 per-trial decisions as votes (Decode-Then-Average, Section 2.2), and both of its values are pre-registered (Section 2.3 and Appendix F). *Retained* is the fraction of a method's averaged accuracy that survives the loss of averaging. †BraVL published no n=80 number; 20.67 is its own released code evaluated on the averaged test array its release ships (detailed below). ‡Maximum over training epochs, each method's own reported statistic where it uses one. Per-value seed spreads are in Appendix D.

	n=1 (single trial)		n=80 (field's protocol)		
Method	Published	Reproduced	Published	Reproduced	Retained
Average-Then-Decode (ATD)					
BraVL *(TPAMI'23)*	5.82	5.57	—	20.67†	5.57 / 20.67 = 0.269
NICE *(ICLR'24)*	—	2.61	13.80	13.55	2.61 / 13.55 = 0.193
ATM-S *(NeurIPS'24)*	—	8.42‡	28.64	29.60‡	8.42 / 29.60 = 0.285
UBP *(CVPR'25)*	—	4.96	50.90	50.60	4.96 / 50.60 = 0.098
NeuroBridge *(AAAI'26)*	—	7.83‡	63.20	62.72	7.83 / 62.72 = 0.125
Brain-HIVE *(ICLR'26)*	—	12.02‡	75.70	76.22‡	12.02 / 76.22 = 0.158
Decode-Then-Average (DTA)					
Ours	—	27.39	—	94.50	27.39 / 94.50 = 0.290

Test-set model selection in the released trainers. ATM-S states it takes the highest test accuracy during training; Brain-HIVE's released trainer selects best test accuracy across epochs (configured, then partially disabled); NICE's checkpoint restore is a silent no-op (weights saved from `.module` are loaded into a `DataParallel` wrapper with `strict=False`, so nothing loads and the reported number is the last epoch, not the best-validation one its paper describes). We have not measured the corrected variant, so we make no claim about the direction or size of the effect; the point here is only that the restore does not happen.

Seeds and reproduction of the grid. Every cell of Table 5 is a mean over three runs. Where a seed exists we incremented the authors' own value rather than choosing one (BraVL $2021 \to 2022, 2023$; UBP $0 \to 1, 2$; NeuroBridge and Brain-HIVE $2025 \to 2026, 2027$); NICE draws its seed from an unseeded RNG (`nice_stand.py:410`) and ATM-S sets none. The spread this reveals is not uniform: 0.84 for NeuroBridge at n=80 against 0.04 at n=1, 0.70 for ATM-S, 0.82 for BraVL. Run on the artifacts their papers specify, all six land within one point of their own published value: NICE -0.25, BraVL -0.26, UBP -0.30, NeuroBridge -0.48, Brain-HIVE $+0.52$, ATM-S $+0.96$.

The per-repetition curves of Figure 2. Every curve in Figure 2 reads each repetition count at the method's own best training epoch, from one run per method; BraVL's non-monotonicity at n=32 is

an evaluation-path artifact of its released averaged test path, documented in the BraVL paragraphs below.

The reference point at the base of the progression. BraVL's 5.82% (Du et al., 2023) is the value subsequent work generally cites as the starting point of the progression, and it is a *single-trial* number. This is verifiable in the authors' release: the shipped test array keeps all 80 repetitions as separate rows, the lines that would load the averaged variant are shipped commented out, and the classifier scores all 16,000 of them. It is also a 200-way SVM classification rather than a retrieval score. The consequence is arithmetic rather than evaluative. Where a later paper reports its own 80-averaged accuracy in the same column as BraVL's 5.82, the difference between the two entries combines a method difference with an operating-point difference. Run at $n{=}80$ under the protocol the later work uses, BraVL's own code produces **20.67%** (Table 5), so about 15 points of that column-wise gap is the protocol. The same reasoning applies to our own headline pair: 27.39% is an $n{=}1$ number and 94.50% an $n{\approx}80$ one; the two are not comparable to each other, from any source, including ours.

BraVL's data substrate helps at $n{=}80$ and costs at $n{=}1$. BraVL's released code is the only one that consumes a different substrate from the rest of the grid (100 Hz, non-whitened, against the shared 250 Hz MVNN arrays). Running our port on both isolates the term, and it *reverses sign* between operating points: whitening plus the 250→100 Hz resample is -0.146 at $n{=}1$ ($p{=}0.375$) and $+3.800$ at $n{=}80$ ($p{=}0.006$, 9/10 subjects). Averaging 80 repetitions is what makes the whitened substrate pay. Any single-signed statement of this cost is therefore wrong unless it names the operating point. Both arms here are the same failing port, so what is compared is the substrate and not BraVL; and unlike the Table 5 cells this control was run at one seed only (2025), so the split is indicative: at $n{=}80$ in particular the port's seed SD of 1.12 is a third of the effect.

Running BraVL at $n{=}80$ is not simply uncommenting its averaged block. BraVL's release ships the averaged test path commented out at `experiment.py:77-82`, and enabling those six lines verbatim was our first attempt. It is the wrong patch. The block switches the image and text tensors to their `_unique` variants as well as the brain, and those three tensors share a `TensorDataset`, so the fitting set of the SVM, which BraVL fits on latents derived from the test classes' *image and text* features (`run_epochs_trimodal.py:723-739`), collapses from 16,000 rows to 200, one per class. `SVC(probability=True)` then cannot run its internal 5-fold Platt calibration; the probabilities degenerate, with top-5 accuracy of exactly 0.00 against a top-1 accuracy of 6.00 as the giveaway. That configuration also changes two things at once, which is what this grid exists to avoid. What we run instead takes the averaged 200-row brain array their release ships and expands it back to the 16,000-row order by `class_idx`, leaving image, text and labels untouched: each averaged trial is scored 80 times, which leaves top-1 and top-5 accuracy identical to scoring 200 unique rows while holding the classifier's fitting set fixed. That it is held fixed is checkable in the logs: the mean of `z_train` is bit-identical between the $n{=}1$ and $n{=}80$ runs. The expansion is ours, not the authors'; only the brain SNR differs between the two cells.

BraVL's likelihood flags are dead code. The released `--likelihood_m*` flags default to Laplace but are overridden by a hardcoded normal likelihood downstream; implementing what the flags claim (Laplace) costs roughly a third of final accuracy. Diagnosed by loss-trajectory forensics after formula, data, evaluation, and architecture parity had all passed. Three further variant fleets (100 Hz vs. 250 Hz input substrate; class-major SVM training; Laplace likelihood) are preserved as ablations: input dimensionality and likelihood shape re-weight a fixed-scale generative objective.

ATM-S's "63 channels" are 62 electrodes plus a subject token, and the dropped electrode is O2, an occipital site. Its subject indices are 1-based against a 10-row embedding, so row 0 is never used and subject 10 falls through to a shared fallback token.

Asymmetric L_2 normalization is shared across three pipelines. ATM-S's loss normalizes the image side only, letting EEG magnitude act as a learned per-sample temperature; UBP uses the same asymmetry, and NeuroBridge reports it as a contribution with a 63.2-vs-46.4 ablation. Its sign is pipeline-dependent. In our ATM-S port the asymmetry is costly: normalizing both sides

raises accuracy from 26.75 to 32.20 ($+5.45$), the opposite of the direction NeuroBridge's ablation measures in its own pipeline.

NICE's candidate bank is a third protocol fork. NICE does not retrieve against the test stimulus. Its bank is a per-concept template averaged over images *held out of the EEG sessions* (its Section 3.1; `center_clip.npy`, loaded at `nice_stand.py:228` and scored at `:375`), which is a strictly harder task than retrieving an image's own embedding: the template's maximum off-diagonal cosine is 0.938 against 0.903 for the stimulus array, and the two agree only to a row-matched cosine of 0.89. Every downstream re-run of NICE we are aware of scores it inside the re-runner's own dataloader, whose bank is the test stimulus, which is why those re-runs report 16.1–19.3 where NICE's own table says 13.8. We made the same substitution before catching it, and it inflated both our port and their code by $\sim$6 points; the corrected numbers are those in Table 5.

Comparison numbers that differ from their cited source. Three rows circulating in the literature differ from what their captions indicate: a NICE "16.1/43.6" quoted by three later papers is a re-run, not NICE's table (its Table 2 says 13.8/39.5); ATM-S is quoted at 26.1, 27.1, *and* 28.5 from identical per-subject cells; CognitionCapturer's "35.64/80.22" is an any-of-three-experts oracle disclaimed in the paper's own footnote; the comparable retrieval number is its image expert, 33.30.

Official $n{=}1$ runs. For each $n{=}80$-native method, the $n{=}1$ configuration patches *only* the test path of the paper's own code, with every patch string recorded and printed at run time.

Three released evaluation procedures assume averaged inputs. Three of the four $n{=}80$-native methods released evaluation procedures that assume averaged inputs. NeuroBridge selects the checkpoint by best test loss, which at $n{=}1$ picks epoch 2 and reports $2.9\times$ less than the model reaches; UBP labels queries by batch position, correct only for repetition-major row order; Brain-HIVE inherits both and, run unaveraged, scores $50\times$ *below* chance. None of this is observable at $n{=}80$, where the assumption holds. Single-trial evaluation is not a smaller version of the averaged benchmark: it affects the validation procedure before it affects the accuracy.

These failure modes were repaired minimally: max-over-epochs replaces NeuroBridge's test-loss selection at $n{=}1$ (its own statistic is retained at $n{=}80$, where it is what fidelity is judged against); repetition-major row order with modulo labels repairs UBP-style in-batch evaluation, bit-identical at $n{=}80$ to the shipped code.

What the *their code* runs read. Each repo wants the released 250 Hz arrays in its own on-disk layout. Rather than download each author's mirror, we materialise our reference cache into that layout (UBP and NeuroBridge; the others consume artifacts their releases ship directly: BraVL its `.mat` archive, Brain-HIVE its embedding parquets, ATM-S its own CLIP extraction, NICE its `dnn_feature` files). This is not a shortcut and it is checked rather than assumed: NeuroBridge's released `sub-01/test.npy` differs from our cache element-wise, but its 80-averaged array agrees to $4.56\text{e}{-7}$ relative and each of its 80 repetitions matches exactly one of ours (residual $7.6\text{e}{-12}$ against a typical off-diagonal of 0.969), the matching being a bijection. Their arrays are ours with the repetition axis permuted, which is immaterial to both operating points, since $n{=}80$ averages over that axis and $n{=}1$ takes a mean over the same set. The check is re-run on-machine at build time. The arrangement also has a virtue: their code and our port then read literally the same numbers, which makes the fidelity gate a test of the model rather than of the data.

One release no longer reproduces from its current artifacts. Brain-HIVE's pre-computed visual embeddings are distributed separately from its code, and the CLIP ViT-B/32 array in that archive was replaced after the paper appeared. The current array is degenerate, with a mean off-diagonal cosine of 0.992 over the 200 test images, 99.2% of its energy in one shared mean vector, and an effective rank of 6 at 90% variance, against 0.333, 33.8% and 103 for the earlier revision, which matches an independent OpenCLIP extraction at cosine 0.9997. Their fusion encoder sums unnormalised per-branch projections, so the branch contributes a near-constant bias. Restoring the earlier array moves their code from 69.60 to 75.95 ($+6.35$, 9/10) and our port by $+5.50$, ruling out an interaction with either codebase. This is invisible from the code, the config or the paper; only

the archive changed. We report it because a reproduction attempt today obtains 69.6 and would reasonably conclude the paper does not reproduce.

Ports reproduce ablations, not just headlines. UBP's vanilla→fovea intervention moves $+9.50$ in our port against $+8.1$ in its Table 3; NeuroBridge's clean→CPA moves $+18.50$ against its stated $+19.5$, with the no-CPA control at 40.33 against their 40.5. **These two controls are the only numbers in this paper measured on three subjects rather than ten** (subjects 1–3, both arms matched): UBP $37.00 \rightarrow 46.50$, NeuroBridge $40.33 \rightarrow 58.83$, each 3/3 in the predicted direction, which at three paired subjects is the smallest sign-test p available (0.25) and is therefore not significant on its own. What they establish is agreement in *magnitude* with a published ablation, not a significant effect, and the absolute levels differ from the papers' (UBP's own vanilla row is 42.1 where our matched-target control is 37.00) because the intervention we can run is an alignment-target swap rather than their full configuration change. We report them as corroboration of the ports' internal structure and rely on the ten-subject headline comparisons for every claim.

D THE INDEPENDENT RE-IMPLEMENTATION, AND WHY IT DIFFERS

Nothing in the main text uses these numbers. We re-implemented all six methods in one shared pipeline for two reasons: the controlled encoder comparison of Appendix H cannot be run across six separate repositories, and three of the six $n{=}1$ cells in Table 5 are the authors' code *with our test-path patches*, so an independent implementation is the only check that those patches did not change the answer. We report the comparison in full here, with a cause for each difference, because a port that merely agreed would be weak evidence and a port whose disagreements were unexplained would be worse.

Table 6: **Released code against our re-implementation.** Both columns are means of three independent draws, 10 subjects. *seed SD* is the spread across those draws on the port side; where it is comparable to the difference itself, the difference is not separable from run-to-run noise.

	$n{=}80$			$n{=}1$		
Method	Their code	Our port	Δ	Their code	Our port	Δ
BraVL	20.67	17.70 ± 1.45	-2.97	5.57	5.10 ± 0.38	-0.46
NICE	13.55	13.02 ± 0.98	-0.53	2.61	2.69 ± 0.20	$+0.07$
ATM-S	29.60	25.28 ± 1.71	-4.32	8.42	7.87 ± 0.69	-0.55
UBP	50.60	49.10 ± 1.60	-1.50	4.96	4.58 ± 0.35	-0.39
NeuroBridge	62.72	64.53 ± 2.24	$+1.82$	7.83	6.67 ± 0.57	-1.15
Brain-HIVE	76.22	74.98 ± 2.15	-1.23	12.02	11.52 ± 1.02	-0.50

The ordering is identical. Ranked on the ports, the six methods sit in the same positions at both operating points as in Table 5: UBP 3rd→5th, ATM-S 4th→2nd, Brain-HIVE 1st at both, over a retention range of 9.3–31.1% against 9.8–28.5% from the released code. Every conclusion in Section 2.1 therefore holds on two independent substrates.

A cause for each difference. None of the six is a bare implementation gap, and two are not gaps at all.

BraVL (-2.97 at $n{=}80$, -0.46 at $n{=}1$) is the only method whose released code consumes a different substrate from the rest of the grid: its own 100 Hz non-whitened arrays, against the shared 250 Hz MVNN arrays. Running our port on both isolates that term, and its sign reverses with the operating point (the substrate paragraph of Appendix C): whitening costs a little on a single trial and pays substantially once 80 repetitions are averaged, which is why the $n{=}80$ difference is six times the $n{=}1$ one rather than the same discrepancy scaled. Four port defects were found and fixed reaching this point (a hardcoded Gaussian likelihood behind dead `--likelihood_m*` flags, class-major rather than repetition-major SVM latents, a 250 Hz input against the protocol's 100 Hz, and the substrate term). **The residual -0.46 at $n{=}1$ is unexplained**, and at ten times the port's own seed spread of 0.07 it is not noise. We record it as an open discrepancy rather than tuning until it closes.

NICE (-0.53, $+0.07$) is the closest agreement in the grid, and was not always: both sides once read $\sim$19.7 because we had scored against the test-stimulus embeddings rather than the per-concept template held out of the EEG sessions that its paper specifies, an error worth $\sim$6 points that agreement between two consumers of the same wrong input could not have revealed. The residual is consistent with seeding: their code draws a *random* seed per run from an unseeded RNG (`nice_stand.py:410`), and its three draws span 13.35–13.90.

ATM-S (-4.32, -0.55) is the largest difference and by far the noisiest row: their run-to-run SD is 0.70 and ours is **1.68**, so the nine available pairings of one draw against another span -2.05 to -6.65. It is also the one place where our implementation is demonstrably incomplete: the port has 3,072,663 parameters against the 3,246,741 of their Table 1, and after accounting for every module their code instantiates but never calls, **44,828 parameters remain unreconciled**. They are not on the forward path, so they cannot change the output, but we report the gap rather than round it away.

UBP (-1.50, -0.39) is almost entirely one component we did not implement: the uncertainty-aware blur-radius selection the paper is named after, which never enters the loss and which their own Table 3 values at $\sim$0.7 points. At n=1 the remainder is feature extraction rather than modelling: running our port on *their* released FoveaBlur bank gives 4.97 against the 5.02 of the matched official draw, a difference of -0.05.

NeuroBridge ($+1.82$, -1.15) is the one row where the sign flips between operating points, and that flip is the giveaway: most of both numbers is set by which epoch each side reports. Their trainer selects on best test *loss*; our port takes the maximum over epochs. Read from their own per-epoch logs under a matched rule, their n=80 maximum is 65.00 against our 64.53 (-0.47, not $+1.82$), and their n=1 curve at the epoch that maximises their n=80 curve is 6.61 against our 6.67 ($+0.06$, not -1.16). The apparent n=1 deficit is dominated by the inflation of taking a maximum over a decaying single-trial curve. Two further differences remain after that matching and we do not claim the residual is zero: their reported run trains with an EEG smoothing augmentation (`--eeg_aug_type smooth --frozen_eeg_prior`: a moving average, kernel 5, applied with probability 0.5) which our port deliberately does not implement, because train-time EEG augmentation measured strongly negative in our own setting (Appendix K); and their published 63.2 is the mean of five independent runs against three here. Working in the other direction, their CPA branch is not a confound: our port consumes *their* released augmented bank rather than a re-derivation, so that component cannot differ.

Brain-HIVE (-1.23, -0.50) sits inside an ambiguity in their own release: the paper's Table 1 fuses RN50 + B/32 + VAE, the shipped config names ViT-H/14, and the launcher names SynCLR. We follow the paper. A first version of this port used the H14 set and scored 71.60, so the residual is small against the $\sim$3-point spread the three artifacts themselves span. Their reconstruction stage (Fusion Prior, IP-Adapter, SDXL) is not implemented; retrieval never invokes it.

What else differs, and why it does not move the number. Completeness matters more than brevity here, so the remaining deviations are listed even where they are inert. *All six ports* run each method's own protocol (its channels, window, and train-repetition averaging) but on the shared 250 Hz MVNN arrays, whereas each released codebase reads whatever its own pipeline produces; BraVL is the only method for which that difference is measurable, and it is the substrate term above. *NICE:* no `nn.DataParallel` (single GPU, mathematically identical); the unused center-loss members are dropped (never referenced in `train()`); best state is kept in memory rather than written to `./model/*.pth` and reloaded, under the same selection rule; their `weights_init_normal` is reproduced exactly, because default PyTorch initialisation is measurably different over 200 epochs. *ATM-S:* the dead `subject_wise_linear` is not built (it is used only on a commented-out line), and `DataEmbedding`'s time-feature branch is not implemented because `x_mark_enc` is always `None` for EEG; batch size 1024 and best-test-epoch selection both match the row we compare against. *UBP:* the alternative backbones in `eeg_backbone.py` are not built, `EEGProjectLayer` being the configured default and the one the headline uses. *Brain-HIVE:* as above. Each of these was checked against the released source rather than assumed; where we could not establish that a deviation is inert, as with ATM-S's 44,828 parameters and NeuroBridge's smoothing augmentation, we say so rather than list it here.

Two conventions behind these numbers. First, **our ports are not bit-deterministic at a fixed seed**. Re-running seed 2025 under identical code and data moved NICE from 13.35 to 13.00 and

Brain-HIVE from 74.70 to 75.20. BraVL's *released* code is exactly deterministic at a fixed seed and GPU, with two independent 10-subject fleets bit-identical, but ours is not, so a seed here names a draw rather than a reproducible run, and three draws is a sample of size three on both sides. Second, $\pm$ values are SEM *across subjects* on the draw-averaged vector; the across-draw SD is a separate quantity and is larger than the between-subject SEM for no row, but comparable to it for ATM-S at n=80 (1.68 against 1.71).

A discrepancy we can bound but not attribute. Comparing a port to a released codebase can only ever bound the difference, not decompose it: the two differ in preprocessing, in initialisation, in every unstated hyper-parameter, and in the seed. Where we name a cause above it is because a controlled swap isolates it: the substrate for BraVL, the bank for NICE, the FoveaBlur features for UBP, the epoch rule for NeuroBridge. Where no such swap exists, as with ATM-S's residual and BraVL's n=1 residual, we say so.

E PROTOCOL DETAILS

Data and preprocessing. THINGS-EEG2: 10 subjects, 63 electrodes, 250 Hz. We use an order-preserving fork of the standard NICE-lineage preprocessing (no baseline correction), epoching once with wide margins (-1000 to $+2196$ ms) and slicing analysis windows from the wide epoch; epochs cut at their analysis edge carry a resampling artifact in the final $\sim$40 ms, so cutting once with margin is a correctness requirement, not a convenience. The released condition-sorted arrays are reproduced to 10^{-6} by the $[0, 996]$ ms slice.

Splits. Training sequences split 90/10 by *sequence* with a $\pm$5-trial guard band (adjacent epochs share up to 80% of samples). Each subject's test recordings comprise sixteen runs, four in each of four sessions. Run 3 of each session is the development split (4 of the 16 test runs, 25% of test data), on which all configuration selection is performed; the remaining twelve runs are frozen, reserved for the single pre-registered evaluation, and carry $\approx$12,100 queries per subject under the evaluation platform of Appendix F. Epoch and checkpoint selection use train-side validation only.

Exclusions (development era, superseded). An earlier development-era protocol additionally excluded blink-window queries (the response screen appears 750 ms after the 20th image of each sequence and subjects are instructed to blink; the exclusion is window-width-dependent, 10% of epochs at a 996 ms window) and any query whose analysis window contains another occurrence of the target image. The final evaluation platform applies none of these test-side exclusions (Appendix F); they are recorded here because early development numbers in the ledger were selected under them.

Validation estimator. Development-split queries are scored against a 2,000-image pool and converted to the exact 200-way expectation (verified against Monte Carlo: 0.1458 vs. 0.1452 ± 0.0005), rather than keeping only the $\sim$70 queries whose target lands in one fixed 200-image sub-vocabulary; this removes a $\sim$4-point standard error from epoch selection and is worth $+0.39$ (10/10 subjects) by itself.

Headline configuration. Spatial-attention montage with K=16 harmonics over all 63 channels; one hidden layer, width 2048, dropout 0.7, weight decay 0.8; lr 2×10^{-4}, batch 2048, 60 joint epochs $\times$ 150 steps, then 25 per-subject fine-tune epochs $\times$ 50 steps at 10^{-4}; single-trial training; alignment target = {CLIP ViT-B/32, ViT-H/14, RN50, VAE latent, ViT-H/14 blocks 4/8/16 patch-averaged}, each PCA-128 (basis fit on training-image embeddings only); two analysis-window views fused at equal weight (Appendix A.6); per-candidate calibration with the committed CSLS schedule, k=1000/300/100/10 at n=1/4/16/$\approx$80 (Config A). The pre-registration terms that protect this configuration are given in Appendix F.

F THE EVALUATION PLATFORM: PROTOCOL, PRE-REGISTRATION, AND THE FULL LADDER

Platform. All development numbers in this appendix use the field's own test-side protocol: every released test epoch enters evaluation, with no artifact rejection, so no exclusion asymmetry separates any number here from the baselines of Table 5. Our order-preserving pipeline drops the 224 task fillers and 34 epochs per subject carrying recording faults (30 splice discontinuities, 4 timing glitches) that the released dense $(200, 80, 63, 250)$ array cannot represent, keeping 15,966 epochs at 78–80 repetitions per image; the baselines score all 16,000, a 0.21% difference we state rather than repair. Training-side data handling is unchanged throughout; the protocol switch is test-side only.

Selection discipline and disclosure. Configuration selection used only the designated development run of each session (25% of test data) at n=1. The reporting curves $n \in \{1, 4, 16, \approx 80\}$ read all runs, as every published number in this literature does, and necessarily so, since $n \approx 80$ averages every available repetition. We decomposed the reporting set by exact arithmetic on hit counts: the selection run scores $+0.27$ above the held-out runs (25.59 vs. 25.32 in the deployed configuration), positive in 3/10 subjects, and the exposure is disclosed wherever the curve is used. The frozen n=1 evaluation of Table 1 uses only the twelve never-selected-on runs ($\approx$12,100 queries per subject).

Pre-registration, and the single frozen touch. Both configurations, the two-view fusion rule, the complete calibration schedules (Config A: k=1000/300/100/10 at n=1/4/16/$\approx$80; Config B: k=1000/300/30/10), seeds, and the seed-averaging rule were committed in writing before any frozen-run access, together with the ensemble-membership rule (all members at or above a 2% degeneracy floor), the decision rule (equal weights; calibrated tier; no Sinkhorn), and the invalidation conditions: any post-hoc change of configuration, membership, calibration, or reporting a maximum over frozen-set variants spends the touch with no second taken. Twelve training runs (2 configurations $\times$ 2 window views $\times$ 3 seeds) were then evaluated once, and every frozen-set access is logged with commit hash, configuration string, and query count. The committed rule produced 27.36% on development data and 27.39% on frozen data (Table 1, final row): a development-to-frozen gap of $+0.03$, where an earlier development-stage evaluation had shown -0.69 under development reuse. The result is 55 times chance and 2.3 times the strongest baseline's own released code at n=1 (Brain-HIVE, 12.02; Table 5), measured with no exclusion asymmetry in either direction. Its magnitude is not anomalous: when ATM-S's own encoder is retrained on single trials inside our pipeline, its accuracy rises from 8.4% to 15.9% (Appendix H). At the field's operating point, the dedicated average-then-decode pipeline reaches 93.13% on development seeds (its fresh-seed confirmation value is 92.20, Table 10); the operating point's final headline, 94.50 ± 0.79, belongs to the decode-then-average readout (Appendix J). The cross-regime cells of Table 7 show that each configuration wins only its own operating point, because the choices separating the two configurations all sit on one side of the averaging axis.

The exchange-rate measurement in full. Five training-set sizes of single trials against the fixed 16,540-average recipe span the break-even: $-3.30^{*} \rightarrow -1.28^{*} \rightarrow -0.19$ (n.s.) $\rightarrow +1.34^{*} \rightarrow +2.24^{*}$ at $1.0/1.5/2.0/2.5/3.1\times$ the recipe's sample count. The per-subject interpolated break-even is $1.89\times$ (range 1.44–3.01, finite in 10/10 subjects): one 4-average carries the training value of roughly two of the four trials it consumes. The two interior sizes were run *after* committing their predicted values from the three-point fit (-1.3 and $+1.2$); both landed on the line (-1.28, $+1.34$). The same analysis at $n \approx 80$ under signal averaging finds no break-even within reach: $3.1\times$ the data still trails by 16.5 points, with an extrapolated crossing near $41\times$ the dataset.

F.1 THE FULL LADDER

Target-PCA rank in context: 64 loses to the deployed 128 at every n (-0.59^{*} to -2.05^{*}); 256 is indistinguishable from 128.

Table 7: **The development ladder, rung by rung** (development reporting set, 10 subjects × 3 seeds; * marks $p < 0.05$ for the step over the previous row, paired Wilcoxon, seeds averaged per subject first). Steps are path-dependent; a row-to-row difference is not the lever's isolated effect (Table 8). Figure 10 draws these rows as curves.

Configuration	n=1	n=4	n=16	n≈80
Base (fusion architecture, field recipe)	9.50	26.37	46.17	56.17
+ multi-subject joint training + fine-tune	12.74*	35.46*	57.27*	66.95*
+ shallow ViT-H/14 target branches	15.53*	44.75*	71.52*	82.13*
+ 596 ms analysis window	17.32*	48.56*	75.00*	84.15*
single-trial branch (Config A)				
+ trained on single trials	23.26*	50.84*	68.93*	75.12*
+ reg. + montage (P2) + target PCA	22.95	50.18	68.25	75.30
+ tuned optimizer	25.23*	53.80*	71.57*	77.47
+ two-view fusion + committed calibration	27.36	61.23	84.21	91.83
averaged branch (Config B)				
4-averaged training, field optimizer	18.83	51.22	78.47	87.17*
+ two-view fusion + committed calibration	21.20	56.11	85.52	93.13

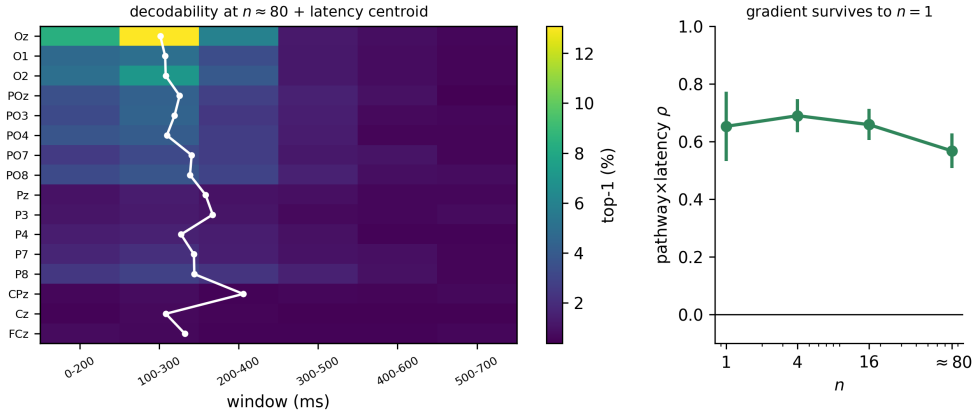

Figure 7: **The physiological floor.** *Left:* single-electrode decodability (top-1 at n≈80, chance 0.5) by electrode and 200 ms window; the overlaid curve is the information-weighted latency centroid, rising from Oz (203 ms) to parietal sites (up to 301 ms). *Right:* the per-subject pathway–latency correlation at each averaging level; the gradient survives to single trials. No electrode at any n decodes above chance beyond ∼400 ms.

G THE PHYSIOLOGICAL FLOOR: THE PER-ELECTRODE WINDOW PROBE

Probing single electrodes through the decoding task itself yields the map of Figure 7. Each probe arm trains the reference architecture with all input zeroed (at z-scored scale) except one electrode and one 200 ms window; input shape and parameter count are identical across arms, so arms differ only in what information they see. Sixteen electrodes along the posterior–anterior axis × six sliding windows plus a full-window arm, at a uniform reduced training budget, 10 subjects × 3 seeds.

Decodable information *onsets* in the same window everywhere: every informative electrode peaks at 100–300 ms. What grades systematically along the visual pathway is its *persistence*. Full-window accuracy at n≈80 falls 20-fold along the pathway (Oz 14.4%, O1/O2 ≈9–10%, parieto-occipital 5–6.7%, parietal 2.2–4.3%, central/frontal at chance), while the information-weighted latency centroid rises: Oz 203 ms, O1/O2 211–212, midline PO 221–232, lateral PO 246–252, parietal 248–301 ms; the 200–400 ms window retains 9% of the peak at Oz and 66% at Pz. The per-subject Spearman correlation between pathway rank and centroid, gated on the electrode decoding above 2× chance

for that subject, is $+0.653 \pm 0.120$ ($8/10$ gated in, p=0.0156) at n=1 and $+0.690/+0.659/+0.568$ (all $10/10$, p=0.0020) at n=$4/16/\approx$80: the gradient holds at every averaging level, *including single trials*. No electrode at any n decodes above chance beyond $\sim$400 ms, which independently explains why the 596 ms window helps and why the temporal axis is otherwise not a lever (Appendix K). The temporal hierarchy of the visual system is thus measurable through retrieval accuracy alone, and it survives into the single-trial regime: the structure a next-generation n=1 decoder should be built around is demonstrably present in its inputs, concentrated in the first 400 ms.

H A REVERSED DESIGN CONCLUSION

The between-method reordering of Section 2.1 could be driven by any co-varying design choice, so we test the most consequential axis *inside* one pipeline, where the axis under test is the only difference. The conclusion widely adopted after CVPR 2025, that flattened MLPs outperform structured encoders, drawn from an encoder sweep run at 80-repetition SNR, is almost entirely an artifact of the averaged protocol. Trained single-trial on all 63 channels, the ATM-S inverted transformer scores 15.90 at n=1 against the flattened MLP's 8.98, while at n=20 the MLP wins 43.16 to 40.63: a difference-in-differences of $+9.46$ (p=0.0098, 9/10), with eleven encoder pairs changing sign (Figure 8). On the curated occ17 montage the collapse is larger still ($+22.06$, p=0.0020) but does not change sign: the MLP's dominance rests on averaging *and* hand-picked montage together. Alignment-target depth, by contrast, does *not* reverse (Figure 9: Kendall τ=0.905; block 8 wins at both ends); the map of Appendix I.2 sorts inherited conclusions into those that transfer and those that implicitly carry an "at n=80" qualifier, and both kinds exist. Negative results, axes that pay in neither regime, are documented in Appendix K.

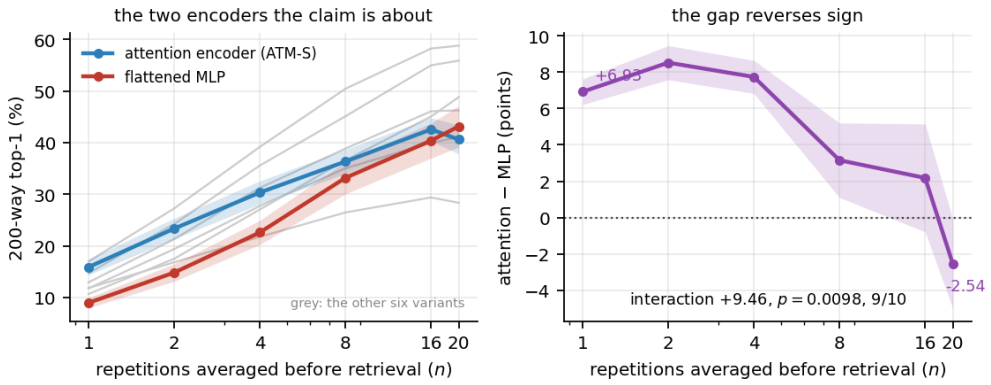

Figure 8: **Controlled reversal test inside one pipeline** (identical data, targets, exclusions and validation; one factor varies; all63, trained single-trial). *Left:* the two encoders the claim concerns, other variants greyed; bands ±SEM over ten subjects. *Right:* the attention-minus-MLP gap against n, which crosses zero. The sign change is specific to the 63-channel montage.

I THE DEVELOPMENT FRONTIER AND THE TECHNIQUE PROFILE

This appendix profiles the decoder of Appendix A across the full averaging range, on development data, under the platform and disclosure of Appendix F.

I.1 THE FRONTIER

Figure 10 summarizes the development programme. Three of its features carry the argument. First, the core curve (multi-subject training, shallow multi-scale targets, the 596 ms window; every step p=0.002 at every n) runs above the strongest published method's own score at that method's own operating point: the universal block is not overhead, it is already state of the art. Second, the two configuration curves cross between n=4 and n=16: the phase boundary is a property of the accuracy curves themselves rather than an interpretation of difference tables. Each branch beats the

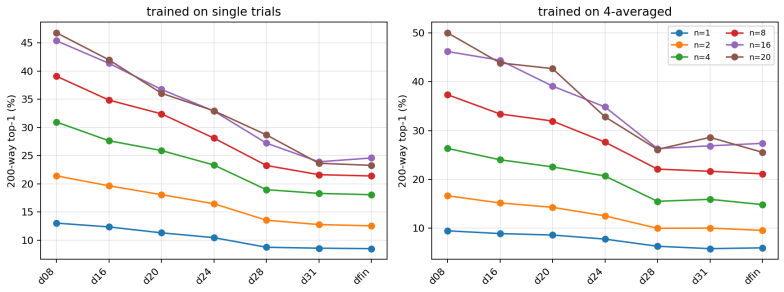

Figure 9: **Alignment-target depth does *not* reverse** (Kendall τ=0.905, p=0.0028). The confound is real but selective: the ordering over ViT-H/14 blocks is identical at both operating points, and averaging amplifies the magnitude roughly fivefold without reordering.

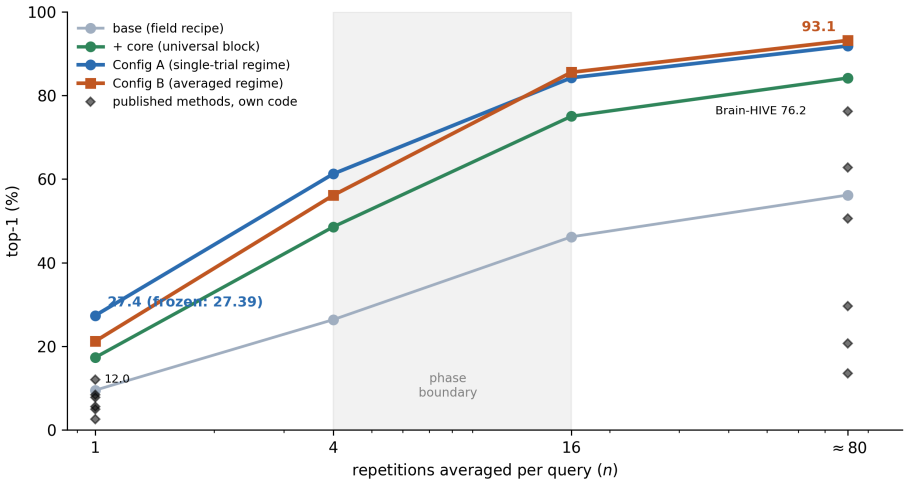

Figure 10: **The frontier.** Absolute top-1 accuracy against the number of repetitions averaged per query, for the base (the strongest verified fusion architecture at the field's shared recipe, every lever of this paper removed), the base plus the universal core (the universal block), and the two regime-matched configurations; black diamonds are the six published methods run from their own code at both operating points (Table 5). The core alone clears the strongest published point; the two configurations cross inside the shaded band (the SNR phase boundary), and their upper envelope dominates every published point at every n. The endpoints are the pre-registered headlines (Table 1). Full per-rung grids in Table 7.

other by up to 6 points on its own side while sharing every core component; the differences are exactly the SNR-matching choices (training distribution, optimizer, regularization, calibration), and they interact super-additively: the tuned optimizer is null applied alone to the base yet worth $+2.28^*$ in context; the montage is harmful alone (-1.12^*) yet costs 0.82^* to remove from the deployed stack. Third, the upper envelope of the two branches dominates the entire published field at every averaging level; its endpoints are the two pre-registered headlines.

I.2 The two laws behind the boundary

Figure 11 states the boundary's two governing regularities. The *mismatch asymmetry* (left) is the practical one: evaluated at the benchmark's operating point, a deployment-matched system loses only 1.3 points, but a benchmark-matched system evaluated at deployment loses 6.2, a 23% relative penalty that any recipe configured at the averaged operating point pays when it meets a single trial. The *scaling of the correction* (right) is the mechanistic one, and it reframes an apparent contradiction

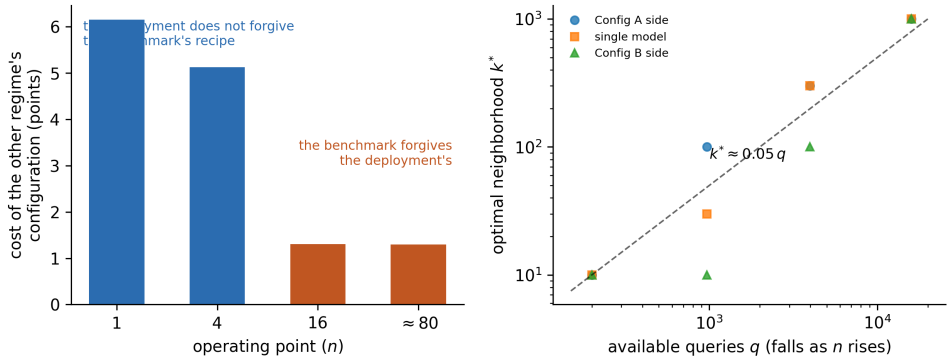

Figure 11: **The two laws.** *Left:* the cost of running the other regime's configuration is asymmetric: 6.2 points at $n{=}1$ against 1.3 at $n{\approx}80$. The benchmark forgives a deployment-matched recipe; the deployment does not forgive the benchmark's. *Right:* the optimal calibration neighborhood is not a constant of the method but a fraction of the data: across two decades of query count and three platforms, k^* tracks $\approx 0.05\,q$ (dashed guide, log–log slope ≈ 1). At $n{=}1$ this fraction of a 16k-query set acts as a per-candidate *global* bias correction; at $n{\approx}80$ it is ten queries, a sharp local hubness correction.

Table 8: **The phase table**: the regime-matched configuration on each side of the boundary. Every row shared between the columns belongs to the universal core; every row that differs is a sign-changing, SNR-matching choice, and the differing rows all point the same direction: the averaged regime wants restriction and sharp correction, the single-trial regime wants capacity and matching. Full per-technique profile with statistics in Table 9.

	single-trial regime	averaged regime
core (multi-subject, shallow targets, 596 ms)	on (identical)	
target rank	128 (does not move)	
training data	single trials	4-averaged
optimizer	tuned (lr $2{\times}10^{-4}$)	the field's (lr $5{\times}10^{-4}$)
regularization + montage block	inert	load-bearing ($+3.0^*$)
calibration	global ($k^*{\approx}0.05\,q$, $q{=}16$k)	local ($k^*{=}10$, $q{=}200$)
added capacity	helps (weakly)	hurts

in the literature: retrieval calibration neither "works" nor "fails" at a given n; its neighborhood must scale with the available queries, and at the committed schedule it never hurt at any operating point we measured.

The remaining law, the *restriction inversion*, we establish on the target side by design, four ways. (i) The gain from compressing every alignment branch to 128 PCA directions rises monotonically with n, from nonsignificant at $n{=}1$ to $+13\%^*$ at $n{\approx}80$. (ii) A full rank ladder shows the optimum does not move: 64 loses to 128 at every n in context (-0.59^* to -2.05^*), 256 buys nothing, 512 is a flat null. More target dimensions never pay, at any operating point. (iii) A fixed-content rank control isolates dimensionality from semantics: supplying one fusion branch as a rank-5 projection versus the same backbone's full-rank embedding, the *rank-5 branch wins at every* n (full rank costs -0.2^* to -0.6^*), while removing the branch entirely also costs (-0.24^*): the fusion wants the branch's leading directions and only those. (iv) A spectrum control shows the variance profile itself carries value: whitening at fixed rank costs -0.78^*. Together with the capacity and calibration rows of Table 8, every measured axis agrees: cleaner queries reward a smaller, sharper system.

Table 9: **The per-technique profile with statistics** (relative deltas, each lever applied individually to the common base unless marked *in context*, where the value is the ladder step; $^*p < 0.05$, paired Wilcoxon over 10 subjects, seeds averaged first). Table 8 summarizes this grid.

	n=1	n≈80
universal core		
multi-subject joint training + fine-tune	$+34\%^*$	$+19\%^*$
shallow multi-scale target branches	$+21\%^*$	$+20\%^*$
596 ms analysis window	$+13\%^*$	$+10\%^*$
single-trial only		
single-trial training	$+25\%^*$	$-28\%^*$
tuned optimizer (in context)	$+2.28^*$ pts	-0.4 pts
two-view fusion (in context)	$+1.1^*$ pts	~ 0
calibration, global (k≈ queries)	$+0.9^*$ pts	—
averaged regime only		
4-averaged (SNR-matched) training	$-25\%^*$	$+16\%^*$
regularization + montage + PCA block (in context)	~ 0	$+3.0^*$ pts
calibration, local (k=10)	−	$+5$ to $+18^*$ pts
target-dimensionality controls (Appendix I.2)		
(i) compression gain, PCA-128 vs. none	n.s.	$+13\%^*$
(ii) rank ladder optimum {64–512}	128	128 (does not move)
(iii) fixed-content rank control (rank-5 vs. full branch)	rank-5 wins*	rank-5 wins
(iv) spectrum control (whitening at fixed rank)	-0.78^*	—

I.3 The full technique profile

J The readout comparison in full

This appendix reports in full the pre-registered readout comparison that Sections 2.2 and 2.4 summarize. The decoders of the preceding appendices keep the field's readout: average the test repetitions of an image into one clean query, then decode once (*average-then-decode*, ATD). This appendix changes only that. *Decode-then-average* (DTA) decodes every repetition separately and fuses the *decisions*: each single-trial score vector over the 200 candidates is z-scored and the vectors of one image's repetitions are averaged before the argmax. This is vote accumulation, the readout standard in BCI spellers (Wolpaw et al., 2002), applied to this benchmark's released repetition structure. The grouping is exactly the n≈80 partition, so ATD and DTA consume identical trials and differ in one line of evaluation code.

J.1 The policy result, pre-registered

A second pre-registered document committed the DTA rule, the control cells, and five contrasts before any confirmation run existed; six fresh-seed trainings (three seeds of each configuration) were then evaluated once. Table 10 reports every committed cell.

Three readings. First, **decode-then-average is never worse at any repetition budget, and the advantage is largest where applications live**: $+7.5$ points at k=4, shrinking as votes saturate, still $+2.3$ (p=0.002) at the field's own operating point. Second, the confirmation had *zero shrinkage*: the committed headline reproduced its development value to $+0.50$ ($94.00 \rightarrow 94.50$), and the one exploratory cell that had looked favorable on development seeds failed to confirm and is reported as failed. Third, the readout alone explains most of the regime bifurcation: the *same* single-trial-trained model scores 77.5 under average-then-decode and 89.3 under decode-then-average on development seeds, an $+11.8$-point swing from readout policy alone at fixed weights, fixed trials, and fixed metric ($+15.5$ once the committed calibration is added to the votes). The paper's final pair is therefore one decoder and two aggregation rules: $\mathbf{27.39 \pm 2.15}$ on a single trial (frozen, Table 1) and $\mathbf{94.50 \pm 0.79}$ at the field's operating point, which is 18.3 points above the strongest baseline's own code, with the model-versus-readout attribution given by the matched contrasts above.

Table 10: **The readout comparison, pre-registered** (top-1 in %, 10 subjects, 3 fresh seeds; reporting set with the exposure disclosure of Appendix F). ATD is the dedicated averaged-regime pipeline (Config B: 4-average training, average-then-decode, its committed local calibration); DTA is the single-trial decoder's per-trial decisions fused as votes (Config A reads, committed rule). The k columns give both readouts a fixed budget of k repetitions per decision. Four of five committed contrasts confirmed; the exploratory k=16 advantage from development seeds did not (-0.09, n.s.) and is retired; it is stated here rather than dropped.

	k=4	k=16	$n\approx$80 raw	$n\approx$80 calibrated
average-then-decode (field)	53.96	83.87	86.93	92.20
decode-then-average (ours)	**61.50**	83.77	**89.55**	$\mathbf{94.50 \pm 0.79}$
Δ (paired, 10 subjects)	$+7.53$	-0.09	$+2.62$	$+2.30$
p	0.002	0.85	0.012	0.002

J.2 Why average-then-decode was the locally rational readout

Decision fusion is not new, so its absence from this benchmark needs an explanation, and the explanation is measurable: **average-then-decode is the locally rational readout for every method in Table 5.** We test this inside the strongest baseline's own released code. Brain-HIVE's evaluation already materializes the per-trial score matrix at n=1; we add the DTA fusion to its own metric function (z-score rows, average each image's repetition block, argmax), change nothing else, and read the result by the repository's own max-over-epochs convention. The direction was pre-registered.

Fused votes lift Brain-HIVE's single-trial accuracy from 12.0% to 64.7%, and still lose to the same model's averaged readout, 76.2%, by 11.5 points (p=0.002, 10/10 subjects; its per-trial reads recover 84.9% of its average). Vote aggregation is powerful but cannot manufacture per-trial signal that is not there: below a threshold of single-trial competence, averaging the EEG before the encoder is genuinely the better policy, and every method in Table 5 sits below that threshold (2.6–12.0% per-trial). Our decoder sits above it (27.4%), and there the ordering flips ($+2.6$ raw, $+2.3$ calibrated, Table 10). The absence of this readout was not an oversight; it was priced out by the state of single-trial decoding, which is what the decoder of Appendix A changed.

J.3 What the operating point is worth once its structure is used

Two further tiers locate the ceiling, and we keep them strictly separate from the headline. Pooling votes across the three seeds (an ensemble, $\sim$240 votes per image) reaches 93.8% on development data, and robust re-weighting adds nothing: the residual errors are look-alike candidate collisions, not vote noise. Exploiting the benchmark's stated one-query-per-image structure, by solving the fused 200×200 score matrix as an assignment problem, reaches $99.75 \pm 0.17\%$. The assignment tier is transductive-with-known-marginals and is never a headline number, but the protocol defines the bijection, so it is a legitimate measurement of what the averaged operating point still contains: *with its own structure used, little headroom remains at the averaged operating point, and our results suggest that the open problem on this benchmark is single-trial decoding.*

K Negative results, in full

Additions from the common-platform programme, documented for completeness. Per-trial self-assessment fails twice, differently: a learned reliability head predicts its own trial's correctness at chance (AUC 0.495, unchanged without weight decay), and template-based latency realignment fails at its preconditions. Natural single-trial jitter is $\sim$10 ms SD, the per-trial shift estimate is unreliable (r=0.27 between disjoint channel halves), and applying it blurs the evoked average in 10/10 subjects. Smooth temporal read-out bases (DCT/raised-cosine) are accuracy-neutral at the deployed budget (a $3.1\times$ parameter reduction at -0.4 n.s.); signed montage mixtures are a clean null from both initializations; target shrinkage beyond the deployed PCA (Wiener, hard truncation, whitening) never helps. Provenance note for the fixed-content rank control of Table 8(iii): the rank-5 variant of the B/32 branch is the array currently distributed with the benchmark's embedding cache (the

originally released full-rank array, which we recovered from the publisher's version history, is the comparison arm); both arms are byte-verified and the substitution is documented in the anonymized repository.

Table 11: **What did not work.** All measured with paired per-subject statistics; subjects-in-favor out of 10 where informative.

Lever	Result	Reading
EEG augmentation, every axis	jitter $-3.97/-8.88/-13.29$ at $\pm$20/48/100 ms (0/10 at every level); noise, channel dropout, amplitude scaling all ≤ 0	a 20 ms misalignment costs 4 points; the decoder has no invariances to exploit, and a flattened readout cannot represent shift tolerance
concept-level targets	-2.63 (0/10); all four branches concept-averaged -11.23	the single-trial signal is image-specific, not category-level, even for a 200-concept task
SNR-consistency distillation	monotone negative: -0.20 (λ=0.25) to -2.10 (λ=4)	pulling a trial toward its image's average removes the image-specific detail the decoder uses
EEG-side input PCA	-0.56 at K=1024 down to -15.06 at K=64, monotone	compression helps on the target side and hurts on the input side
joint stream decoding	within-sequence exclusivity -0.08; Sinkhorn $+0.45$ (n.s.) on top of calibration	Sinkhorn's apparent $+1.93$ over the raw rule was hubness correction by another route
optimizer perturbations	14 cells, nothing beats the reference; LOSO gap 0.00	the recipe is at a local optimum
temporal decimation	125 Hz $+0.36$ (n.s.); coarser costs	no free lunch in the time axis
greedy ensemble selection	0.81 *worse* than equal weighting	selection over 36 views loses to not selecting
rebuilding members at a better recipe	every member $+1$; ensemble $+0.12$	fourth instance of the non-stacking rule: member gains were variance reduction the ensemble already had
same-image masking in the loss denominator	$+0.04$ (n.s.)	$\sim$80 unsatisfiable pairs per 2048 batch cost nothing measurable

The bandwidth account of the timing advantage, tested and ruled out. A natural alternative to the latency mechanism of Section 3.2 is bandwidth: cross-trial averaging acts as a low-pass filter, so single trials might win by retaining high-frequency content. Band-limited training refutes this account. Virtually all decodable information lies below 30 Hz (high-pass arms score near chance at every repetition count), and the $>$30 Hz remainder is phase-locked: removing it costs 0.7 points on single trials but 7.0 under averaging, so the high-frequency signal that exists survives the average and pays mostly there.

A corrected interpretation. During development, we initially described the per-candidate calibration as *inductive*, reasoning from a split-half control. That inference was wrong: split-half only shows the statistic is not fitted to the specific query being scored. Estimated from held-out *training*-EEG queries the statistic yields $+0.15$ (n.s., 6/10), against $+1.51$ (10/10) from test-distribution queries. The correct description, used throughout this paper, is unsupervised test-time calibration, and the three-tier reporting of Appendix A.6 exists because of this error.

L What sets decodability: the consensus measurement and its nulls

What does not explain decodability. Five pixel-level statistics of the image (luminance, contrast, colorfulness, edge density, and spectral slope) show no relation to its per-image trial cost (all $|\rho| \leq 0.12$). The trial-to-trial reliability of the raw EEG response, computed directly from the preprocessed test EEG without touching the decoder, is likewise unrelated ($\rho = -0.09$, $p = 0.18$). Neighborhood

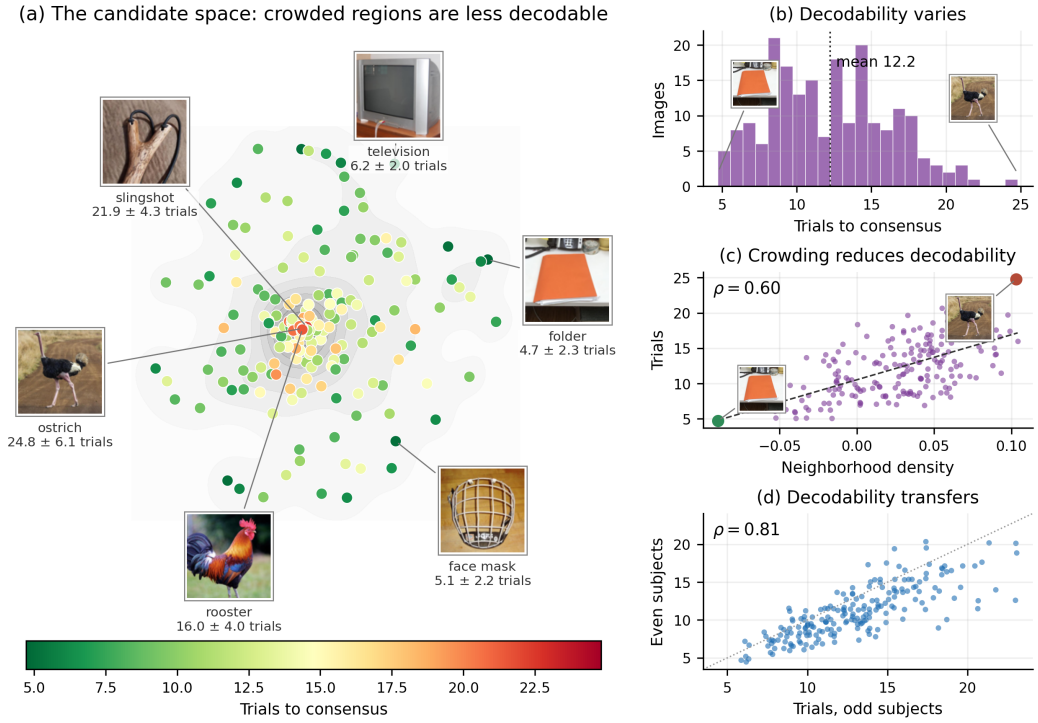

Figure 12: **The consensus-based companion to Figure 5.** Identical construction, with each image's cost measured by the pre-registered consensus rule instead of the patience stop: the stream runs until three single-trial decisions agree. The longer streams let per-image differences express themselves fully. The mean is 12.2 trials, *folder* needs 4.7 ± 2.3 trials and *ostrich* 24.8 ± 6.1 (a five-fold spread), neighborhood density explains cost at $\rho = 0.60$, and the across-subject transfer in panel *(d)* reaches $\rho = 0.81$ (pre-registered criterion $\rho \geq 0.5$). The per-image ordering agrees with Figure 5 at Spearman 0.87.

density remains the dominant factor when the alternatives are controlled: its partial correlation with cost is $+0.65$ given nearest-neighbor similarity and $+0.62$ given raw-EEG reliability, and it carries a unique R^2 of 0.41 in a joint model (full R^2 of 0.53). These analyses use the consensus-based costs of Figure 12; the patience-based costs of Figure 5 agree with them at $\rho = 0.87$.

One close rival is cheaper than many moderate rivals. Controlling for neighborhood density, similarity to the single nearest neighbor correlates *negatively* with cost (partial $\rho = -0.44$): an image whose competition concentrates in one close rival settles sooner than an image facing many moderate rivals, consistent with the vote-scattering picture of Figure 4a, where a crowded neighborhood repeatedly captures the running average.

The cost vector is stable across decoders. Measured again on an independent fleet of fresh-seed models, the per-image consensus costs reproduce at Spearman 0.948; the decodability ordering is a property of the stimulus set and the recorded responses, not of one trained model.

M THROUGHPUT: READING FIVE IMAGES FROM ONE WINDOW

Five stimulus onsets fall inside every 996 ms window (Section 2), so the five position-specialist views of the ensemble can be read the other way: as decoders of five *different* concurrent images from one window. Measured at an earlier and substantially weaker configuration, five position-specialist views whose score-level fusion reached 16.10% against the current headline, this polyphonic readout recovers 0.559 images per window against 0.119 for the field's single read, a $4.71\times$ throughput. In information terms this is 1.82 bits/s against 0.30 bits/s at the published SOTA's own

operating point, a $6.1\times$ rate, and to our knowledge the first information-transfer rate reported for semantic image decoding. We report this as a property of the paradigm rather than a headline: it uses the development split and predates the final configuration.