WHERE, WHEN, AND WHAT ALIGNS: A RETRIEVAL-GROUNDED ATLAS OF EEG–VIT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The alignment between human brain activity and pretrained vision models has become both a window into biological vision and the engine of EEG-to-image decoding. In this work, we draw the first systematic atlas of the EEG–ViT interface measured in the deployment metric itself, end-to-end 200-way retrieval, resolved jointly over scalp position, epoch time, network depth, sub-layer computation, token readout, and training objective. Analyzing 21 vision models spanning contrastive, supervised, self-distillation, and reconstruction training, we evaluate approximately 21,000 probe cells on all ten THINGS-EEG2 subjects at three trial-count regimes, with every claim supported by a per-subject sign test. Our analysis reveals three key findings. (1) **Where and when**: the depth axis of the network carries a brain timestamp. Each block aligns best with a distinct, progressively later moment of the EEG epoch, rising from 177 to 201 ms across the twelve blocks in every subject, in register with the posterior-to-anterior latency gradient on the scalp. (2) **Where in the network**: the well-documented mid-depth alignment peak has a mechanism. Early blocks write EEG-visible content into the residual stream, the mid-network stream integrates those writes and overtakes every computation, and late computations fade while the stream coasts on what was already written. (3) **What**: EEG shares with a vision transformer low-level visual structure at coarse spatial grain, not fine-grained semantics. Alignment is carried by electrodes over early visual cortex, peaks at intermediate depth, deeper still under masked autoencoding, and fades exactly where representations turn semantic unless the training target is pixel reconstruction, while a coarse pooling of the token grid outperforms the field's global readouts by a wide margin. Building on these insights, the atlas configures decoders by measurement alone. It distills a five-model fusion pipeline to eight electrodes and a single backbone that surpass the published state of the art, and it composes a four-state fusion chain that reaches 97.5% final-epoch accuracy together with the first single-trial accuracy above 22%. Our results suggest that the brain–transformer interface is a lawful and largely unexplored territory, and that deployment-currency measurement offers a new instrument for studying biological vision through transformer computations and vision models through the organization of the brain.

1 INTRODUCTION

The striking parallels between deep vision models and the primate visual system have catalyzed a fundamental question at the intersection of machine learning and neuroscience: do artificial and biological vision converge on similar solutions? A decade of encoding studies has answered affirmatively for representation: features drawn from successive network layers predict neural responses at successive stages of the visual hierarchy (Yamins et al., 2014; Güçlü & van Gerven, 2015; Cichy et al., 2016; Schrimpf et al., 2018), a hierarchy that reaches object-level representations within roughly two hundred milliseconds (Thorpe et al., 1996; Felleman & Van Essen, 1991; DiCarlo et al., 2012). This correspondence has since become an engineering engine. In EEG-to-image decoding, a contrastive encoder aligns a subject's EEG response with the embedding of the viewed image under a pretrained vision model, and the seen image is retrieved zero-shot from a 200-way gallery (Gifford et al., 2022; Song et al., 2024; Li et al., 2024). Successive refinements of the encoder and of the alignment target have driven top-1 accuracy from 16% to 75% in two years (Zheng et al., 2026), and the

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.

same recipe now extends across recording modalities and toward image reconstruction (Benchetrit et al., 2024; Takagi & Nishimoto, 2023; Scotti et al., 2024).

As decoders improved, attention converged on the depth axis of the vision model: the representation that aligns best with the brain is often an intermediate layer rather than the final embedding (Benchetrit et al., 2024; Du et al., 2026; Tang et al., 2026; Xia et al., 2026). Much of the current progress accordingly concerns the choice of layer: decoding methods select a reading depth through layer-wise sweeps (Du et al., 2026), empirical depth profiles (Tang et al., 2026), or routers learned over layers (Jiang et al., 2026), and encoding studies map layer-wise brain predictivity (Naselaris et al., 2011; Horikawa et al., 2019; Yang et al., 2024), recently descending below the layer for language models and fMRI (Chen & Sivakumar, 2026). Yet a vision transformer offers more axes than its stack of layers, and an EEG recording carries more than one vector per trial. Patch tokens occupy positions in the visual field. The EEG epoch unfolds at millisecond resolution, placing every readout in brain time. Each block comprises atomic computations, from the queries and keys to the residual stream they write into, each a distinct representation that can be read. For EEG and vision transformers these axes remain unexplored, and they are where this work begins. We resolve alignment to a single electrode, a single time window, a single patch position, and a single computation inside the block, all in the deployment metric itself, and at this granularity the interface exhibits regularities that have not been reported before.

We chart these axes through a comprehensive, retrieval-grounded atlas of the EEG–ViT interface. Our instrument is a probe: a small contrastive decoder trained end-to-end on a restricted slice of the EEG (a channel group and a time window) against a restricted slice of a vision model (a depth, a component, and a token readout), scored by 200-way zero-shot retrieval on THINGS-EEG2 (Gifford et al., 2022). Spanning 21 vision models across the major training families (Radford et al., 2021; Zhai et al., 2023; Caron et al., 2021; Oquab et al., 2024; He et al., 2022), the atlas comprises approximately 21,000 logged probe cells, evaluated on all ten subjects at three trial-count regimes ($N{=}1/4/80$ averaged test repetitions), with every claim supported by a per-subject sign test. Through this atlas, we make three principal contributions:

- **Revealing that the network's depth axis carries a brain timestamp.** Each transformer block aligns best with a distinct, progressively later moment of the EEG epoch. The latency that drives retrieval rises from 177 ms at the first block to 201 ms at the last, in all ten subjects, and co-registers with the posterior-to-anterior latency gradient that the same instrument recovers on the scalp. Hierarchical correspondence in time (Cichy et al., 2016) thereby becomes a per-block, per-subject law in decoding currency (§4).
- **Uncovering the mechanism of the mid-depth alignment peak through a sub-layer neuroanatomy of the vision transformer.** Probing queries, keys, values, attention and FFN outputs, and the residual stream (Elhage et al., 2021) at matched width reveals a consistent three-act anatomy. Early blocks *write* EEG-visible content into the stream, the accumulating stream overtakes every individual computation by mid-network, and late computations fade while the stream retains what was already written, each effect holding in all ten subjects. The celebrated peak is thus the depth at which integration completes, and below the layer, keys out-align queries in all ten subjects, so EEG tracks the content that attention offers (§5).
- **Establishing what EEG shares with a vision transformer: low-level visual structure, at coarse spatial grain.** The best electrode, Oz, sits over early visual cortex. Alignment peaks at one-third to one-half of network depth in contrastive and supervised models and deeper in masked autoencoders, whose features stay low-level longer (He et al., 2022), and it fades exactly where representations turn semantic, at a cost that grows with the abstraction of the training target (Wang et al., 2023; Fang et al., 2024). The alignment is likewise coarse. A 2×2 to 4×4 pooling of the token grid outperforms the global readouts inherited from image classification (Dosovitskiy et al., 2021) by 6 to 22 points, the grain expected of 63 electrodes reading through the spatial low-pass filter of the scalp (§6).

Building on these insights, we let the atlas configure decoders by measurement alone. Eight occipital electrodes and a single backbone read at one-third depth reach $91.67 \pm 3.32\%$ final-epoch accuracy, surpassing the published state of the art of 75.7% (Zheng et al., 2026) and matching the strongest preprint report (Jiang et al., 2026) (§7.2). Greedy, sign-test-gated fusion over cached target slices

("reservoir states"), audited member by member, then yields a four-state chain that reaches 97.47 ± 1.74 final-epoch accuracy together with the first single-trial 200-way accuracy above 22%, both in all ten subjects. We report these records as evidence that the atlas is drawn accurately rather than as its purpose (§7), and we accompany them with a ledger of every hypothesis of ours that the measurements retired (Appendix J). Taken together, our results suggest that measuring the brain–transformer interface in deployment currency makes its laws visible.

2 Related Work

EEG-to-image decoding. NICE (Song et al., 2024) established the 200-way zero-shot retrieval protocol on THINGS-EEG2 (Gifford et al., 2022), training a contrastive EEG encoder against CLIP embeddings (Radford et al., 2021), and ATM (Li et al., 2024) strengthened the encoder. Brain-HIVE (Zheng et al., 2026) broadened the alignment target instead, fusing a pixel-level VAE with a semantic CLIP to reach 75.7%, the current published state of the art. Among recent preprints, SAMGA (Jiang et al., 2026) learns a per-subject soft router over the layers of a large vision backbone and reports 91.3% intra-subject with public code. These systems measure success in the same retrieval currency as our atlas, and we quote their numbers in §7.4.

Intermediate-layer alignment. The premise that intermediate layers can outperform final embeddings as alignment targets is now established across EEG (Du et al., 2026; Tang et al., 2026; Jiang et al., 2026), MEG (Benchetrit et al., 2024), fMRI (Xia et al., 2026), and machine learning more broadly (Skean et al., 2025; López-Cardona et al., 2025). Du et al. (2026) state the open problem our atlas answers, noting that their method "still relies on a layer-wise sweep to identify the best intermediate representation." Our atlas takes the layer finding as established and extends the axis set below the layer, across the scalp, over epoch time, and across trial budgets, with a mechanistic account of the inverted-U profile (§5).

Brain-predictivity atlases. Layer-wise encoding fits are foundational in visual neuroscience (Yamins et al., 2014; Güçlü & van Gerven, 2015; Schrimpf et al., 2018), and selection by decodability was proposed for fMRI by Horikawa et al. (2019). Closest to our sub-layer map, Chen & Sivakumar (2026) dissect thirteen intermediate states of the transformer block against fMRI during language processing and explicitly call for extending computational neuroanatomy to vision transformers, a call our decoding-side visual map answers (§5). The "Neural Visibility" score of HCF (Tang et al., 2026) is an encoding-side member of the same family, and our atlas differs from it in currency (§3).

Fusion and trial budgets. Brain-HIVE and CORTIVA (Wang & Chen, 2026) argue for fusing more teachers. Our attribution analysis measures what each teacher is worth and where (§7.1), pruning the fusion zoo by 60% with no loss. Repetition-count curves appear concurrently in Cui et al. (2026), whereas we use trial count as a *regime variable* of the atlas itself and find that optimal configurations flip between regimes (§7.1).

3 The Probe: Measuring Alignment in Deployment Currency

Why not encoding fit. The natural instrument for an alignment map is an encoding model, namely ridge R^2 from EEG to each activation. We built that instrument first, over six backbones, and it produced a clean but misleading answer. Measured R^2 is highest at the earliest blocks and decays monotonically with depth, while retrieval accuracy over the same layers follows an inverted-U that peaks mid-network. Ranking depths by R^2 therefore selects the worst-performing depth for *all ten subjects*, costing 17.3 points against a sweep. A map that is to configure a decoder must therefore be measured in the decoder's own currency.

The probe. A probe cell $A(g, w;\, m, d, c)$ trains a small contrastive decoder end-to-end on a restricted EEG slice, defined by a channel group g and a time window w, against a restricted target slice, defined by a model m, a depth d, and a component c, and reports 200-way zero-shot retrieval on the THINGS-EEG2 test partition. The host is fixed across all cells (a Brain-HIVE-style MLP

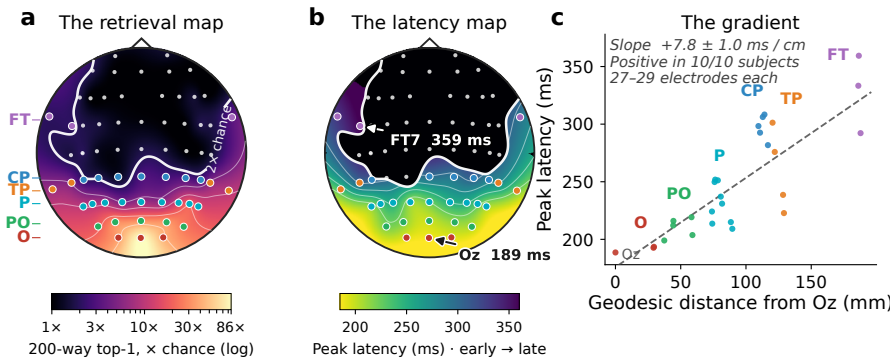

Figure 1: **The retrieval map, the latency map, and the gradient between them.** **(a)** Each of the 63 electrodes probed *alone* against a $\frac{1}{3}$-depth CLIP target (630 cells, 10-subject mean), log ×chance scale: Oz alone reaches 43% 200-way top-1 (86× chance). Bold contour: the 2×-chance inclusion threshold; colored electrodes: the six scalp groups used throughout. **(b)** Peak-latency centroid of each included electrode (10-subject mean), interpolated inside the 2× territory of (a) (black: no usable signal; thin contours every 25 ms): latency sweeps from 189 ms at Oz to 359 ms at FT7. **(c)** The same per-electrode latencies against geodesic scalp distance from Oz (head sphere $R = 96$ mm): $+7.8 \pm 1.0$ ms/cm, positive in 10/10 subjects (Wilcoxon $p = .002$; dashed: mean fit).

encoder with symmetric InfoNCE, 25 epochs), and only the slices vary, so any difference between two cells is attributable to what was read rather than to how. Chance is 0.5%.

Three currencies, fixed epochs, paired comparisons. Each cell is evaluated at three test-repetition budgets, the standard 80-repetition average (N=80) and 100 random N=4 and N=1 subsets drawn with identical seeds across all cells, so every comparison in this paper is paired down to the draw. We report *final-epoch* numbers, reserving best-epoch numbers for published-table comparability. Claims rest on per-subject sign tests over all ten subjects rather than on single cells, since the measured single-cell noise floor is three to four points, and headline cells are seed-replicated (Appendix E). Matched controls are used throughout, including matched random channel sets, width-matched targets, and GPU pairing. The full protocol appears in Appendix A.

Reading order. The three insights answer the title in turn: where and when alignment lives (Insight I), where inside the network to tap it (Insight II), and what is shared (Insight III), before the completed map configures decoders (§7).

4 Insight I: The Two Cascades Are One

The atlas opens at single-electrode resolution (Figure 1a), with a scalp map that no encoding atlas can draw: each of the 63 electrodes probed *alone*, in deployment currency. The concentration is extreme. Oz by itself retrieves at 43% 200-way top-1 (86× chance), and the useful field ends at the parieto-occipital border. The coarser geography agrees, since the eight-electrode occipital montage beats eight *matched random* channels by more than 38 points at N=80 and N=4 in all ten subjects, while the parietal and temporal groups score below their own matched-random controls.

The central finding is that this cortical geography, the epoch's chronometry, and the network's depth axis are one structure. The EEG latency that drives retrieval rises with ViT depth, from 177 ms at the first block to 201 ms at the last in all ten subjects (Figure 2a), and the same latency climbs the scalp from posterior to anterior in every subject, sweeping from 189 ms at Oz to 359 ms at FT7 (Figure 1b). The gradient survives single-electrode resolution, with each electrode's own latency climbing at $+7.8 \pm 1.0$ ms per centimeter of geodesic scalp distance from Oz in all ten subjects (Wilcoxon $p = .002$; Figure 1c). The gradient is moreover directional: posterior but lateral electrodes stay earlier than their distance alone predicts, so latency follows anteriority, which argues against a single volume-conducted source. Later cortex means later readable time and deeper usable computation, all measured in one currency.

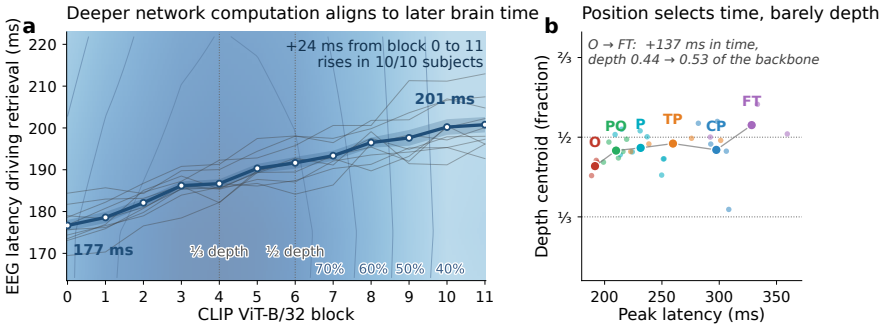

Figure 2: **Deeper network computation aligns to later brain time. (a)** The EEG latency driving retrieval rises from 177 ms at block 0 to 201 ms at block 11 (+24 ms, rising in 10/10 subjects; gray: individual subjects). Background: the measured 12-block × 7-window field; dotted guides: the $\frac{1}{3}$ and $\frac{1}{2}$ depths of panel (b). **(b)** Each included electrode placed by *when* it speaks (latency centroid, as Figure 1) and *where in depth* it reads; large markers: group means in cascade order. From O to FT, latency sweeps +137 ms while the preferred depth drifts only $0.44 \rightarrow 0.53$ (8/10 subjects, Wilcoxon $p = .02$): position selects *when* strongly, *where in depth* weakly.

Space selects depth far more weakly than it selects time. At single-electrode grain, depth tuning is broad and nearly position-invariant. Every group's tuning curve peaks near $\frac{1}{3}$ of the backbone (Appendix Figure 6), and the depth centroid drifts deeper with scalp distance only gently, from 0.44 to 0.53 (Figure 2b; 8/10 subjects, Wilcoxon $p = .02$). The group-level parietal preference for half depth (Appendix F) emerges only when an electrode row is decoded *jointly*, so it is a property of the ensemble readout rather than of any single electrode. The latency rise has the same ensemble-level origin, since a single-electrode dissection finds every sub-layer state reading one moment at Oz (Appendix K), so the rise reflects the ensemble re-weighting electrodes of different intrinsic latencies. Anterior channels carry no usable intra-subject signal, contribute $+4.53$ points inter-subject (8/10 subjects), and *reverse* at $N{=}4$ even inter-subject, where the compact occipital montage wins. The pattern generalizes into the atlas's cross-cutting law: *the optimal montage widens with the generalization demand, while the optimal depth does not* (§7.1).

All retrieval-relevant signal lives in the first 400 ms of the epoch, and later windows score at chance in every cell across both channel sets (occipital and all-63). Within this range the peak window is 50–200 ms, stable across depth. The robust form of the depth-time interaction is the *lower edge* of the profile. The penalty for reading the shallowest blocks grows as the window moves later, in all ten subjects across three seeds, while the peak stays at one-third to one-half depth, and no window ever prefers the final block. Why the usable computation sits at that depth is a question about the inside of the network, and the next insight opens the block to answer it.

5 Insight II: Where to Tap the Network

Reading whole layers leaves the transformer a black box. We therefore probe *atomic* components, namely Q, K, V, the attention output, the FFN activations, and the residual stream at block entry and exit, width-matched at 1536 dimensions (480 cells; $N{=}4$ currency in this section). Three findings emerge.

(i) A mechanistic account of the alignment inverted-U, to our knowledge the first. Figure 3a,b traces three eras. At block 0 the block's *writes*, the attention and FFN outputs, exceed the stream at entry by $+16.8$ points in all ten subjects (the WRITING era). By mid-network the integrated stream overtakes every computation (the INTEGRATION era). Past block 6 the computations collapse while the stream coasts on what was already written, exceeding the block's own writes by $+9.9$ points at the final block, again in all ten subjects (the COASTING era). The inverted-U over depth *is* the residual stream integrating EEG-visible writes that stop arriving after mid-network.

(ii) K > Q: EEG tracks attention's content, not its queries. At early blocks the key projection out-aligns the query projection ($+4.0$ to $+2.7$ points at blocks 0 to 4, each in all ten subjects),

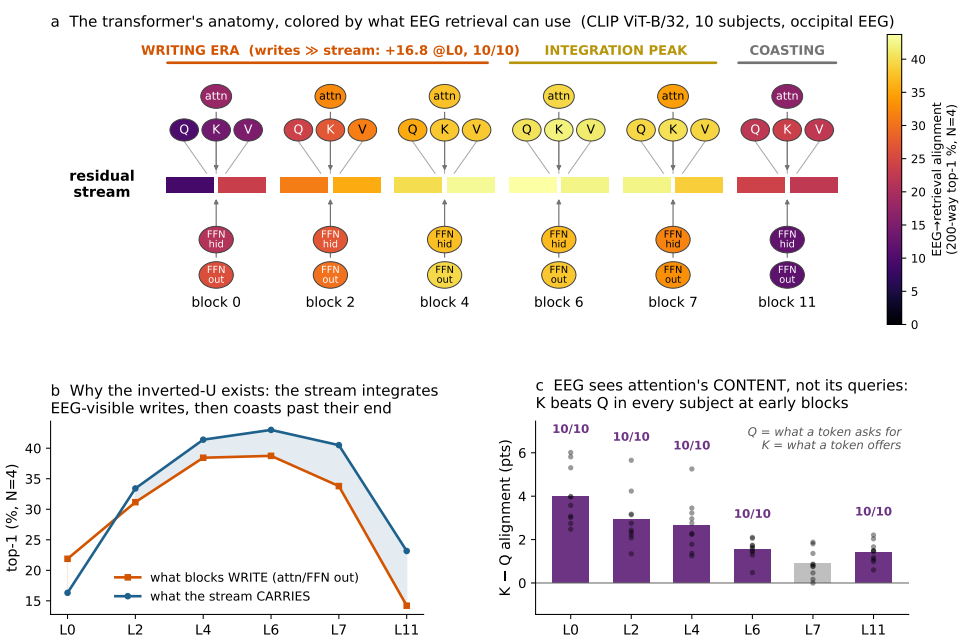

Figure 3: **Inside the block: the first sub-layer alignment map for brain decoding.** Width-matched (1536-d) probes of Q, K, V, attention output, FFN hidden/output, and the residual stream at entry/exit, per block (CLIP-B/32, 10 subjects; $N{=}4$ currency). **(a,b)** Early blocks' *writes* exceed the stream at entry ($+16.8$ at block 0, 10/10); the mid stream integrates and wins; late computations collapse while the stream coasts ($+9.9$ over the block's writes at block 11, 10/10). **(c)** K beats Q at early blocks ($+4.0/+2.9/+2.7$ at blocks 0/2/4, 10/10 each), converging by block 6.

converging by block 6 (Figure 3c). Sub-layer states have been dissected on the encoding side (Chen & Sivakumar, 2026), and the dissociation here is, to our knowledge, the first in decoding currency. §6 shows it is a property of the training objective rather than of transformers.

(iii) Stream optimality. No atomic computation beats the residual stream at matched width at the depth where the atlas reads, so block-level reading is demonstrated, rather than assumed, to be the right granularity for this backbone. §6 shows that this is a property of the objective rather than a law, since two of the five models probed at atomic resolution read best from a computation *inside* the block. What brain and network share, and why the objective decides whether it survives to the output, is the subject of the final insight.

6 Insight III: Low-Level Structure, at Coarse Spatial Grain

All backbones peak at early-to-middle depth for one reason: what EEG carries is low-level visual structure, so alignment lives wherever a model retains that structure and fades wherever training trades it for semantics. A screen over architecture $\times$ objective establishes this as a law of the training target, covering 21 models that span pixel reconstruction, masked prediction of non-pixel targets, contrastive, supervised, and self-distillation training, together with random-init controls of both families and raw pixels (Podell et al., 2024; He et al., 2022; Woo et al., 2023; Ryali et al., 2023; Fang et al., 2024; Wang et al., 2023; Assran et al., 2023; Radford et al., 2021; Zhai et al., 2023; Steiner et al., 2022; He et al., 2016; Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025), each swept over seven depths, all ten subjects, and all three currencies (1,410 cells). Five results follow.

(i) The objective does not decide which model wins at the peak. At each model's best depth, every mainstream trained encoder lands within six points, from MAE at 93.6 down to DINOv3 at 87.9 at $N{=}80$. Our own prior hypothesis, that reconstruction objectives would *rank* above contrastive ones

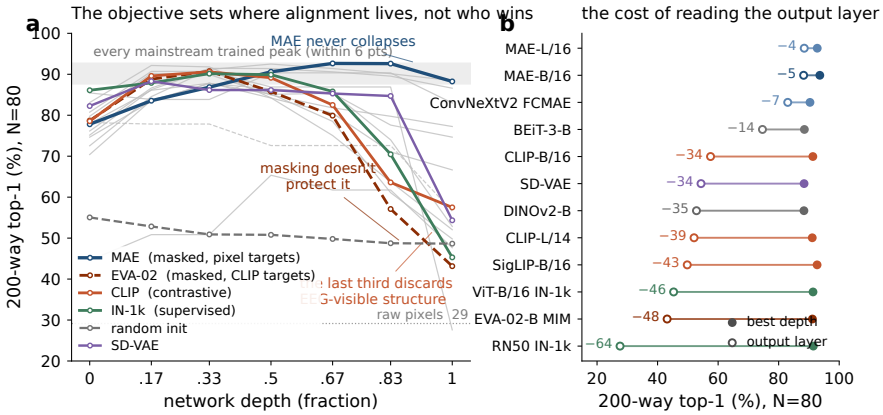

Figure 4: **The objective sets where alignment lives, not which model wins, and the protection is the target, not the masking.** **(a)** Depth-decay profiles (N=80, 10-subject mean) for the patch-matched ViT-B/16 quartet (identical architecture, four training objectives) plus the masking-matched control EVA-02 (masked prediction of CLIP features instead of pixels), over the rest of the 21-model screen (thin gray). **(b)** The cost of reading the output layer: per-model drop from best depth to the final layer, ordered by target abstraction for the masked family: pixels (MAE-B/L, ConvNeXt-V2) -4 to -7; discrete tokens (BEiT-3) -14; CLIP features (EVA-02) -48.

at the peak, failed this test. Nor is the band a law of trained networks in general, since hierarchical pooling (Hiera-MAE) and latent-target prediction (I-JEPA) fall outside it (Appendix J).

(ii) It decides where alignment lives, and whether it survives. Collapse onset orders exactly by how aggressively the objective discards low-level structure (Figure 4a), from depth 0.67 for CLIP-L/14 and SigLIP, through 0.83 for CLIP-B/16, supervised ViT, and the DINO family, to the output for the VAE and supervised ResNet, while **MAE never collapses** at any scale. At the output layer, the one the literature actually reads, the spread across the patch-matched quartet runs from MAE at 88.2 down to supervised ViT at 45.3 (Figure 4b). The design rule for the field follows directly. *Never align to a contrastive or supervised output embedding*. Depth-optimize the read, or use a masked autoencoder, for which depth ceases to matter.

(iii) The protection is the reconstruction target, not the masking. The decisive control is EVA-02, which follows the identical masked-image-modeling recipe but predicts CLIP features instead of pixels, and it collapses *harder than supervised ViT*. Collapse severity is monotone in target abstraction, moving from pixel targets (MAE, -4 to -6 across its three scales) through discrete visual tokens (BEiT-3, -14) to CLIP features (EVA-02, -48). Nor is the law a transformer artifact, since a masked-reconstruction *convnet* (ConvNeXt-V2-FCMAE) never collapses either, while Hiera shows that hierarchical pooling relocates the profile without breaking it. *The objective sets whether alignment survives, and the architecture sets where it sits.*

(iv) The VAE's advantage is location, not quality. The SD-VAE's celebrated usefulness for brain decoding reflects not a superior representation but a convolutional encoder whose *whole depth range* stays reconstruction-preserved, readable at a shallow state with no depth search. That is exactly the low-level territory its fusion marginal occupies (§7.1).

(v) Controls, and one screen-level observation. Every trained model beats its matched random-init control in all ten subjects, and raw pixels are the floor. One observation carries forward: SigLIP posts the best low-trial numbers of the screen, and the composed pipeline later confirms it (§7.3).

Cross-objective anatomy. Repeating the atomic probe of §5 for MAE, DINOv2, EVA-02, and ConvNeXt-V2 (1,740 cells) shows an opening that is universal, block-0 writes exceeding the stream at entry in all five models (CLIP included) and all ten subjects, while the endings, and even the K>Q ordering, split along the target families of Figure 4 (full anatomy in Appendix I).

The readout law (Figure 9). The depth results are one half of the picture, and the spatial half comes from the token-readout axis, measured in the same currency on concept-disjoint validation over patch-grid reservoirs (design, figure, and the full retinotopic map with its pre-registered control in Appendix G). Coarse spatial pooling beats the literature's default readouts everywhere, with the largest effect sizes in the atlas. A 4×4 average pool beats the class token by $+22.4$ points and global averaging by $+16.7$ on the pixel-reconstruction backbone, and a 2×2 pool beats global averaging by $+5.7$ on the semantic backbone, each in all ten subjects. The optimal grain follows the objective, and global max pooling is uniformly destructive. EEG carries the image's coarse layout, finer than a global summary yet far coarser than the token grid. Every published EEG–ViT alignment we are aware of reads the class token or a global average, and this axis alone is worth more than any model swap within the trained band of result (i).

7 The Constructive Reading, and the Records

7.1 Territory, not count

The three insights are drawn on one map, and a map is tested by navigating with it. Fusion pipelines such as Brain-HIVE report a single number for a five-model target zoo. Leave-one-out gain maps over model $\times$ input condition localize each model's marginal and show a double dissociation. The VAE's marginal is worth $+2.2$ to $+6.5$ points at $N{=}4$ in all ten subjects, peaking at occipital electrodes in the 100–300 ms window, exactly the low-level territory, while CLIP-B/32's marginal is significant everywhere but there. The other three zoo members contribute marginals near zero. The atlas thus turns "fuse more models" into "fuse models that cover uncovered territory."

Territory is regime-bound, and gates cannot see it. Channel restriction, window restriction, and the anterior stream all *reverse sign* between the low-N intra-subject and averaged inter-subject regimes, three independent instances of the cross-cutting law of §4, invisible to any evaluation reporting only the 80-repetition average. Nor can learning recover the arbitration: learned per-stream gates either *open on the harmful stream* or never open at all, because the contrastive gradient sees only the training regime. Regime arbitration must come from measurement across regimes.

The constructive proof. A greedy state-fusion program takes the territory lesson literally, running forward selection over 27 reservoir states (model $\times$ depth $\times$ component targets), acceptances sign-test-gated on concept-disjoint validation and confirmed on the official test before adoption, stopping when five rounds of shrinking marginals saturate. The finished chain (MAE-L at 0.5, VAE at 0.17, SigLIP at 0.5, BEiT-3 at 0.33) samples the target-abstraction spectrum one member per tier, and a leave-one-out audit certifies every member, since dropping any one costs accuracy in all ten subjects at $N{=}4$. Ported into the dual-window host of §7.3 below, the chain sets the records of Figure 5. The audit returned two structural lessons. First, the greedy seed itself is evicted, because a deeper read of the same model absorbed its territory, so construction order is not dependency structure. Second, the semantic member's optimal depth climbs as its partners cover more low-level territory (SigLIP reads at 0.17 alone, 0.33 in a duo, 0.5 in the chain). The chain is a certified local optimum, not a proven global one.

7.2 Pruning: eight electrodes, one window, one backbone

The attribution map says the B/32+VAE duo covers the full zoo's territory, and it does, matching the five-model target within noise. The spatial and temporal maps say eight electrodes and one window suffice, and they do: the occipital-8 montage with a *single* backbone read at one-third depth reaches $\mathbf{91.67 \pm 3.32}\%$ final-epoch accuracy, surpassing the published state of the art of 75.7 (Zheng et al., 2026) and matching, under our stricter final-epoch protocol, the best-epoch 91.3 of the SAMGA preprint.

7.3 Composition: the atlas as the configuration tool

The atlas not only prunes but composes: the regime flips above prescribe different pipelines for different deployments.

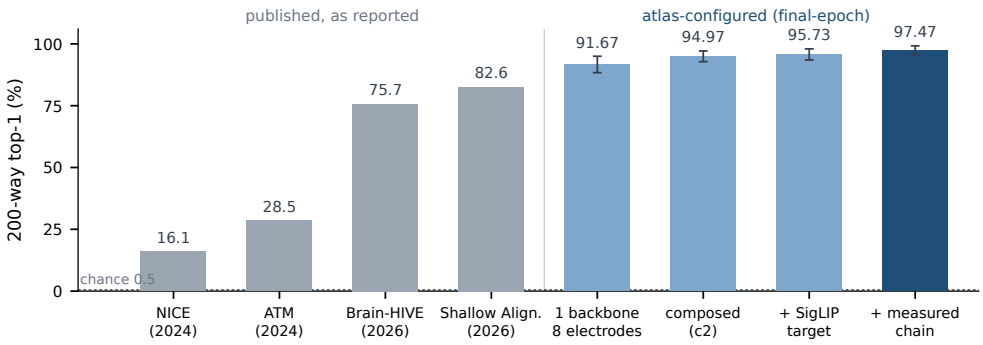

Figure 5: **THINGS-EEG2 intra-subject 200-way top-1 (%): the published landscape vs. the atlas-configured decoders.** Gray bars show published baselines as reported by their authors (Song et al., 2024; Li et al., 2024; Zheng et al., 2026; Du et al., 2026), best-epoch where stated; preprint numbers are quoted in the text. Blue bars are the atlas-configured decoders at the stricter final-epoch protocol (mean±sd over the ten subjects; the measured chain reaches 98.12 ± 1.33 at best-epoch). Inter-subject and trial-budget results appear in the text (§7.3).

Intra-subject, low-trial (pipeline c2). Two occipital streams, one on the 100–300 ms window and one on the full epoch, fused on the duo target, reach 94.97 ± 2.17 at $N{=}80$, 59.76 at $N{=}4$, and 20.16 at $N{=}1$ (final-epoch, %). These configurations achieve, to our knowledge, the first single-trial 200-way top-1 above 20% (the concurrent NEAR preprint reports 19.8 (Cui et al., 2026)); the measured chain below lifts it above 22%. Against the matched single-stream control, the dual-window design wins exactly where the time atlas said it would, at the low trial budgets, in all ten subjects.

The screen upgrades the target. Swapping CLIP-B/32 for SigLIP at depth 0.33, the screen's best low-trial model, lifts the same pipeline to $\mathbf{95.73 \pm 2.25}\%$ final-epoch accuracy, $\mathbf{61.08}$ at $N{=}4$, and $\mathbf{20.62 \pm 4.72}$ at $N{=}1$, with each low-trial gain holding in 9/10 subjects over three seeds. One honest note: the screen predicts the *model*, while the in-pipeline depth came from a small pilot, landing one step deeper than the solo optimum because the VAE partner covers the low-level territory.

Inter-subject (pipeline c3). Adding the anterior stream as its own fused branch, unusable intra-subject yet informative inter-subject, yields $\mathbf{28.85 \pm 6.1}\%$ final-epoch accuracy (33.05 ± 6.7 best-epoch, above the published state of the art of 22.4 (Du et al., 2026)). The stream is worth $+9.50$ points at $N{=}80$ in all ten subjects (vs. $+4.53$ in the single-stream arm of §4) and almost nothing at $N{=}4$, the regime flip once more.

7.4 The records, in context

Figure 5 places the composed pipeline against the published landscape, and every architectural choice in it is traceable to a specific atlas finding. The final bar puts the whole atlas to work end to end: the measured four-state chain, inside the atlas-prescribed dual-window host, reaches $\mathbf{97.47 \pm 1.74}\%$ final-epoch accuracy (98.12 ± 1.33 best-epoch), $\mathbf{63.98}$ at $N{=}4$, and $\mathbf{22.06 \pm 5.12}$ at $N{=}1$, with both low-trial gains holding in all ten subjects over three seeds, and with eight electrodes throughout.

8 Conclusion

An atlas earns trust by publishing its dead ends; the full ledger of retired hypotheses is Appendix J. Where the field established *that* the best-aligned depth is often intermediate, this paper maps *where, when, and what* actually aligns, in the currency that matters. The map configures, pruning a five-model fusion pipeline by 60% at parity, setting records under a stricter protocol, and performing arbitration that the training loss demonstrably cannot. Its limits are one benchmark, one montage family, and a posterior-dominated recording. The axes charted here exist for every brain-model pairing, and deployment currency is the way to measure them. The territory is open.

ETHICS STATEMENT

This work analyzes THINGS-EEG2 (Gifford et al., 2022), a public, de-identified EEG dataset collected under its authors' institutional ethics approval, and no new human-subjects data are collected. The methods decode which of a fixed set of images a consenting participant viewed and require cooperative per-subject recording. They do not read unconstrained mental content. Improving trial efficiency moves such systems closer to practical use, so mental privacy should govern any deployment, with informed, revocable consent and no repurposing of brain data. The vision backbones probed are public pretrained models and carry whatever biases their training corpora contain.

REPRODUCIBILITY STATEMENT

Every number in this paper is the output of a logged probe run ($\sim$21,000 result files with fixed seeds, paired draw seeds, and recorded GPU types); the full result set will be released upon acceptance together with the probe trainer, extractors, and aggregation scripts. Appendix A specifies the probe protocol; Appendix B inventories every cell family; the atomic reservoir specification is in Appendix D. All comparisons against published methods quote their authors' own numbers with protocol annotations rather than reimplementations.

AI USE STATEMENT

In this work, we used generative AI tools for writing and editing the manuscript text, for drafting and refactoring the analysis and figure code, and for literature search. We have not used generative AI tools for research ideation or for producing or altering experimental results, and the generation of datasets or experimental stimuli is not applicable to this work, which analyzes a public benchmark. Additionally, we used generative AI tools for aggregating logged results into tables and figures and for copy-editing. We have reviewed all AI-assisted work. AI-drafted code was verified by the authors and re-run end to end under the fixed-seed protocol of Appendix A; every number reported in this paper is the output of a logged probe run, never an AI-generated value; and references surfaced by AI-assisted literature search were checked against the original papers before citation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *International Conference on Learning Representations (ICLR)*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Cheng-Yeh Chen and Raghupathy Sivakumar. The mind's transformer: Computational neuroanatomy of LLM-brain alignment. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=PgIlCCNxdB.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016. doi: 10.1038/srep27755.

Zhenyao Cui, Siyuan Kan, Dingkun Liu, and Dongrui Wu. Beyond trial averaging: Anchoring neural and visual representations for few-repetition brain-to-image retrieval. *arXiv preprint arXiv:2608.19128*, 2026.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. doi: 10.1016/j.neuron.2012.01.010.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Du, Siyuan Dai, Yonghao Song, Paul M. Thompson, Haoteng Tang, and Liang Zhan. Deep models, shallow alignment: Uncovering the granularity mismatch in neural decoding. *Transactions on Machine Learning Research (TMLR)*, August 2026. arXiv:2601.21948.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. `https://transformer-circuits.pub/2021/framework/index.html`.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. doi: 10.1093/cercor/1.1.1.

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

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, 2018. doi: 10.1016/j.neuroimage.2018.02.044.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Tomoyasu Horikawa, Shuntaro C. Aoki, Mitsuaki Tsukamoto, and Yukiyasu Kamitani. Characterization of deep neural network features by decodability from human brain activity. *Scientific Data*, 6:190012, 2019. doi: 10.1038/sdata.2019.12.

Jonathan C. Horton and William F. Hoyt. The representation of the visual field in human striate cortex: a revision of the classic Holmes map. *Archives of Ophthalmology*, 109(6):816–824, 1991. doi: 10.1001/archopht.1991.01080060080030.

Lin Jiang, Qingshan She, Jiale Xu, Haiqi Xu, Duanpo Wu, and Zhenzhong Kuang. Subject-aware multi-granularity alignment for zero-shot EEG-to-image retrieval. *arXiv preprint arXiv:2604.17782*, 2026. v2; code at `https://github.com/LinJiang8/SAMGA`.

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 (NeurIPS)*, volume 37, 2024.

Ángela López-Cardona, Sebastián Idesis, Mireia Masias-Bruns, Sergi Abadal, and Ioannis Arapakis. Brain–language model alignment: Insights into the platonic hypothesis and intermediate-layer advantage. In *Proceedings of UniReps: the Third Edition of the Workshop on Unifying Representations in Neural Models (NeurIPS 2025 Workshop)*, volume 322 of *Proceedings of Machine Learning Research*, pp. 81–101. PMLR, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29441–29454, 2023.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint 407007*, 2018. doi: 10.1101/407007.

Paul S. Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875, 2025.

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

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research (TMLR)*, 2022.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14453–14463, 2023.

Jingyi Tang, Shuai Jiang, Fei Su, and Zhicheng Zhao. Aligning what EEG can see: Structural representations for brain-vision matching. *arXiv preprint arXiv:2603.07077*, 2026.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. doi: 10.1038/381520a0.

Junhan Wang and Kani Chen. CORTIVA: Candidate-score fusion of complementary visual teachers for EEG- and MEG-to-image retrieval. *arXiv preprint arXiv:2608.01355*, 2026.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186, 2023.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, 2023.

Tian Xia, Zihan Ma, Yawen Zhang, Xinlong Wang, Qing Liu, Xiaowei He, and Yudan Ren. Brain-MCLIP: Brain image decoding with multi-layer feature fusion of CLIP. *Meta-Radiology*, 4(3): 100219, 2026. doi: 10.1016/j.metrad.2026.100219. arXiv:2510.19332.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences (PNAS)*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Huzheng Yang, James Gee, and Jianbo Shi. Brain decodes deep nets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23030–23040, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

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

A Probe Protocol

Data. THINGS-EEG2 (Gifford et al., 2022): 10 subjects, 16,540 training images (4 repetitions each), 200-image zero-shot test partition (80 repetitions). EEG preprocessed at 250 Hz with multivariate noise normalization (Guggenmos et al., 2018). The montage carries no Fz (it is the reference); anterior-row analyses use FCz.

Probe trainer. Each cell trains the Brain-HIVE-style host (MLP brain encoder, per-branch linear projections, symmetric InfoNCE with learned temperature, AdamW (Loshchilov & Hutter, 2019), 25 epochs, batch 1024) on the specified EEG slice (channel group $\times$ time window) against the specified target slice (model $\times$ depth $\times$ component). Fixed seed 2025; the full per-epoch test curve is logged and the *final* epoch is primary.

Evaluation. $N{=}80$ uses the standard 80-repetition average. $N{=}4$ and $N{=}1$ average random test-repetition subsets, 100 draws each, drawn with a fixed draw seed (777) shared across *all* cells, so low-N comparisons are paired at the draw level. Sign tests are exact binomial over the 10 subjects; slope and drift claims additionally report a Wilcoxon signed-rank test over the ten per-subject values.

Controls. Channel-count-matched random electrode sets; width-matched (1536-d) targets for all atomic comparisons; fixed 200-ms windows for temporal comparisons; cells of a paired contrast run on the same GPU model (identical-seed runs differ by several points across GPU types on this benchmark).

Montage. The occipital-8 montage used throughout the composed pipelines is PO7, PO3, POz, PO4, PO8, O1, Oz, O2.

Selection hygiene. Target-side feature selection is brain-free (activation-variance based); depth and configuration choices were made on concept-disjoint validation splits, and no test data enters any selection or hyperparameter choice, with one stated exception: each acceptance of the state-fusion chain (§7.1) was gated on validation and then confirmed on the official test partition before adoption.

B CELL INVENTORY

Table 1: Probe-cell families (21,138 logged runs: 20,888 intra-subject, 250 inter-subject). Each cell = one trained probe with full three-currency evaluation. Counts are of result files on disk; two smoke-test runs belonging to no family are excluded.

family	resolves	cells
space: channel groups × depth, + matched-random controls	§4	840
space, inter-subject arm	§4	180
single-electrode (63 electrodes, alone)	§4	630
single-electrode × windows (29 electrodes × 6 windows)	§4	1,740
single-electrode × depths (29 electrodes × 6 depths)	§4	1,740
electrode rows (8 rows × depth × windows)	§4	1,040
time: windows × depth, + fine grid, + seed replications	§4	1,220
HD cascade grid (12 blocks × 7 windows)	§4	840
atomic components, CLIP-B/32 (8 comp. × 6 blocks)	§5	480
atomic components, MAE-B/16 + DINOv2-B	§6	960
atomic components, EVA-02 + ConvNeXt-V2 (conv set)	§6	780
objective screen, round 1 (11 models × $\leq$7 depths)	§6	710
objective screen, round 2 (10 models × 7 depths)	§6	700
fusion attribution (LOO × input conditions) + duo	§7.1	500
composed pipelines (c1–c4) + learned gates	§7.1,7.3	346
target-swap pilots + 3-seed confirm (SigLIP)	§7.3	50
greedy state-fusion: selection rounds, refinement, pilots, norm-ablation arm	§7.1	2,772
state-chain test cells: checkpoints, LOO audits, 3-seed, dual-window port	§7.1,7.4	280
retinotopy + readout: patch cells, coarse grids, pooling ladders	App. G	3,170
single-electrode sub-layer anatomy + chronometry (Oz)	App. K	2,160

C SINGLE-ELECTRODE DEPTH TUNING BY SCALP GROUP

This appendix resolves the depth-tuning claim of §4 at single-electrode grain. Figure 6 plots one tuning curve per scalp group: the curves stack in the cascade order of Figure 1, peak together near one-third depth, and drift deeper with scalp distance only weakly, which is the single-electrode form of the law that position selects *when* strongly but *where in depth* only weakly.

D ATOMIC RESERVOIR SPECIFICATION

For each probed block we extract, per image: the residual stream at block entry and exit; the Q, K, V projections; the attention output (post W_O); and the FFN hidden and output activations. Token grids are read through the class token and global average pooling and reduced to a common 1536-d width by variance-ranked column selection, so no component wins on dimensionality. Blocks probed: $\{0, 2, 4, 6, 7, 11\}$ for the 12-block models, tagged by nominal depth fractions $0/.17/.33/.5/.67/1.0$ (block index = fraction × 11, rounded). Extraction scripts and the per-component `npz` layouts will be included in the code release upon acceptance.

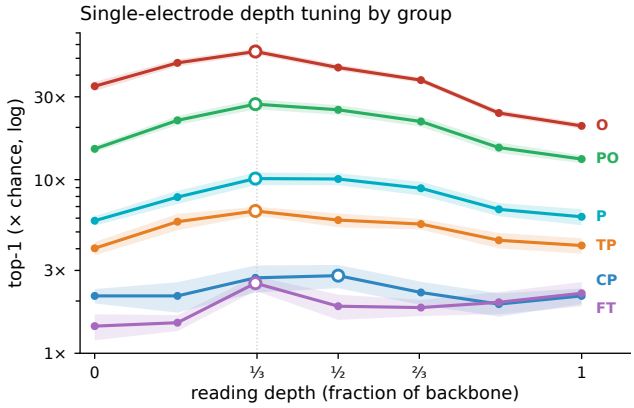

Figure 6: Single-electrode depth tuning by the six scalp groups (1,740 electrode×depth cells; mean±SEM, ×chance, log; open circles: peaks). The groups stack in the cascade order of Figure 1; tuning is broad and peaks near $\frac{1}{3}$ in every group, and the depth centroid drifts deeper with scalp distance only weakly (O $0.44 \rightarrow$ FT 0.53; slope positive in 8/10 subjects, Wilcoxon $p = .02$): position selects *when* strongly but *where in depth* only weakly (§4).

E Seeds and the Noise Floor

Repeating headline cells across 3 seeds gives a within-subject seed-sd of 0.54 points at $N{=}80$, and single-cell differences of ± 3–4 points occur across GPU models at identical seeds, which is why (i) no claim in this paper rests on a single cell, (ii) all contrasts are GPU-paired, and (iii) significance is by per-subject sign test rather than by cell-level magnitude. Best-epoch reporting inflates $N{=}80$ numbers by $+2$–3 points on this benchmark; we report it only for published-table comparability.

F Full Atlas Matrices

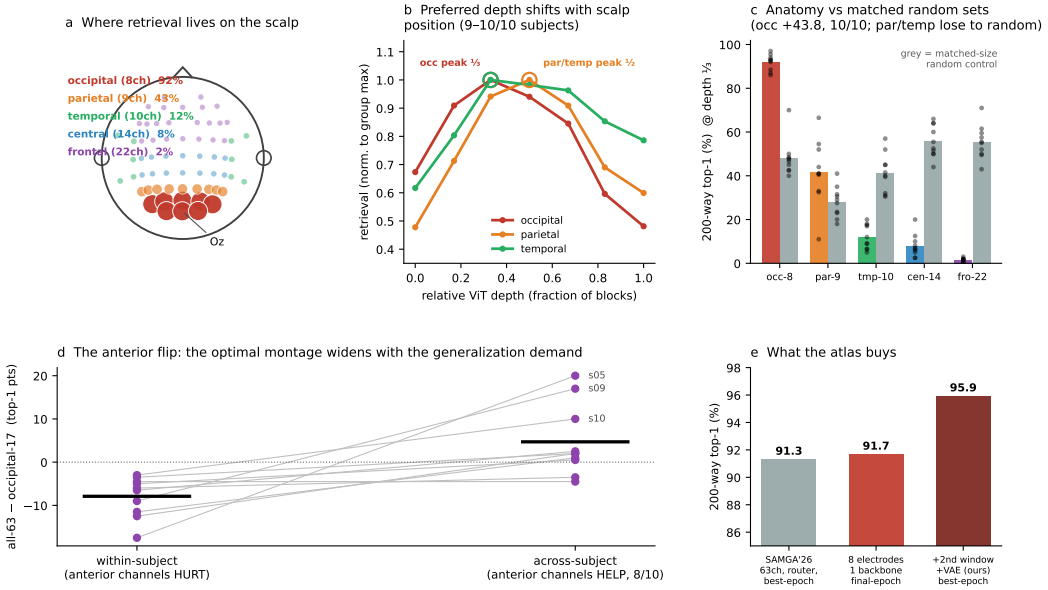

Figure 7: Full channel-group × depth matrix with matched-random controls and scalp insets (summary form of §4).

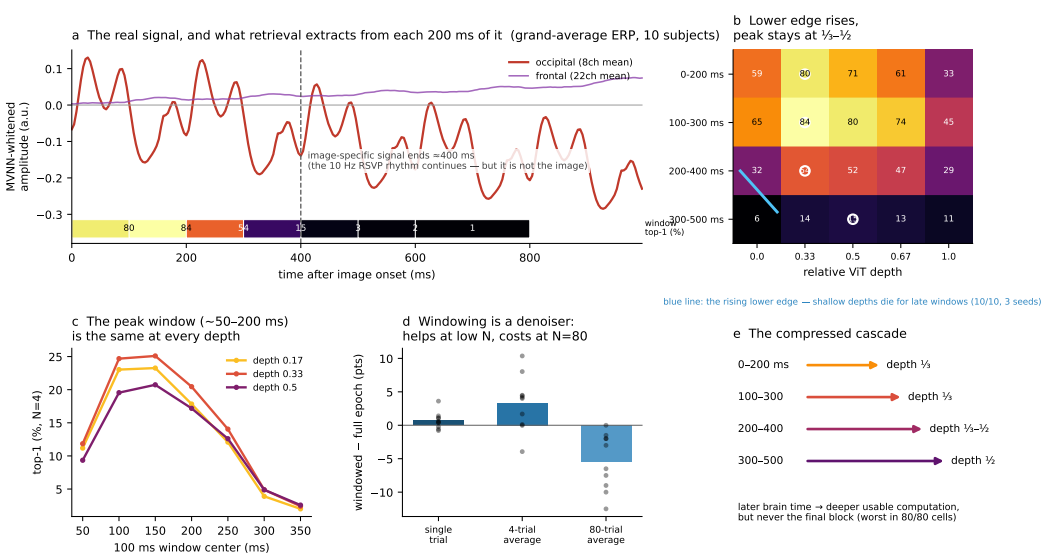

Figure 8: Full window × depth matrix with the lower-edge statistic (summary form of §4).

The complete numeric matrices for every family in Appendix B, including the 21-model objective screen and the three-objective atomic matrices, will be regenerable from the result files by the aggregation scripts, both part of the release planned upon acceptance.

G THE RETINOTOPY OF EEG–VIT ALIGNMENT

Every axis of the main-text atlas treats the image as a whole, whereas the visual system does not. Cortical magnification devotes a disproportionate share of early visual cortex to the central visual field (Horton & Hoyt, 1991), and the occipital montage that carries all of our signal sits directly over it. This appendix asks whether EEG–ViT alignment is itself retinotopic, that is, whether some parts of the visual field align more than others beyond what image statistics alone predict.

The design runs the same frozen probe trainer over *patch-grid* reservoirs from two objective-matched backbones, a contrastive-semantic member and a pixel-reconstruction control, with three cell families, each scored by the same 200-way retrieval computed over a concept-disjoint validation set of 200 held-out concepts (the selection split of Appendix A), so the official test partition stays untouched. The first is the retinotopic map, in which each of the 14×14 patch positions is probed *alone*, complemented by a coarse 7×7 grid for the cross-model contrast and by eccentricity rings for the radial profile. The second is an honesty control, built in as a design rule rather than an afterthought. Per-patch *image-only* concept discriminability normalizes out the dataset's object-centered statistics, so any claim of foveal bias attaches to the *ratio* map rather than the raw one. The third is the token-pooling axis, whose results form the readout law of the main text (§6) and supersede the encoding-side screen of Appendix H.

The two-map result (Figure 10). Raw EEG alignment is strongly foveal. Center patches beat edge patches by $+7.3$ points under SigLIP and by $+12.6$ under MAE, in all ten subjects for both encoders. The pre-registered control, however, shows that the *images* are more foveal still. The THINGS stimuli are object-centered, and the image-only ceiling falls off faster with eccentricity than the EEG map does. The EEG/image ratio therefore *inverts*, with peripheral patches outperforming central ones per unit of image information, again in all ten subjects for both encoders. The claim this licenses is deliberately double-edged. EEG–ViT alignment is retinotopic, but its foveal concentration is inherited from the image statistics themselves, and relative to what each location affords, the periphery over-delivers. The control is a nearest-class-mean reference rather than a capacity-matched ceiling, so ratio magnitudes should be read qualitatively. One classical signature does survive normalization directionally. Alignment is higher in the *lower* visual field (raw $+3.2$

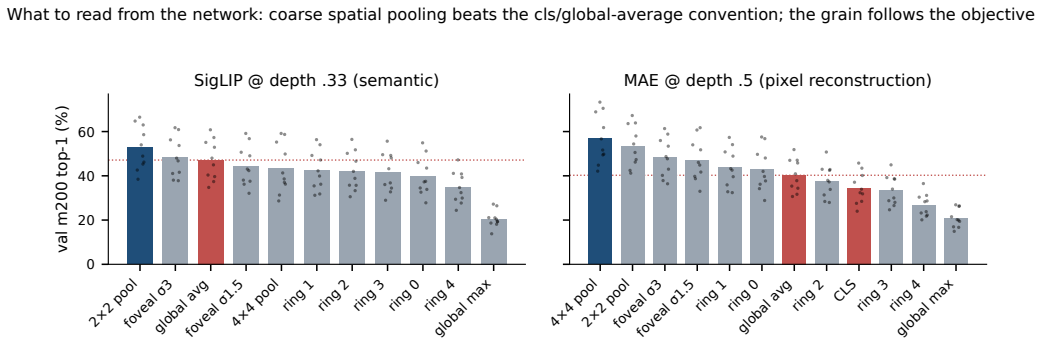

Figure 9: **The readout law.** Mean retrieval (concept-disjoint 200-way validation top-1, 10 subjects, dots = subjects) for every token readout of a semantic (left) and a pixel-reconstruction (right) backbone at the depths used by the record pipelines (§7.3). The field's default readouts (red: CLS, global average) are not competitive; the winning grain follows the objective (§6).

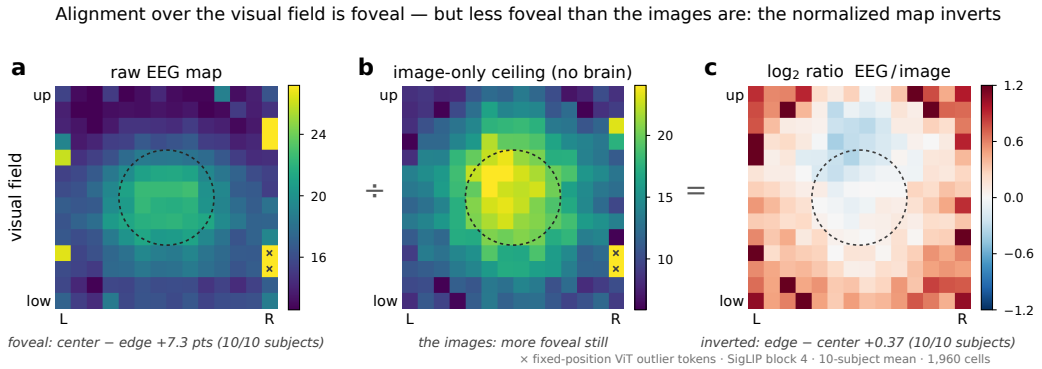

Figure 10: **The two-map retinotopy. (a)** EEG alignment for each of the 14×14 patch positions probed alone (SigLIP block 4, 10-subject mean, 1,960 cells). **(b)** The image-only ceiling: concept discriminability of the same patch features, no brain in the loop. **(c)** Their $\log_2$ ratio. The raw map is foveal, the image ceiling more foveal still, so the ratio *inverts* (blue center, red periphery). Dotted circle: the fovea of the pre-registered center/edge statistics; $\times$: fixed peripheral positions hosting ViT register/outlier tokens (global information, hot in both maps, cancelling in the ratio); color ranges percentile-clipped.

points, 9/10 subjects, p=.011; normalized 8/10 with the same sign in both encoders), the lower-field advantage of early human vision, measured here for the first time in decoding currency.

H THE ENCODING-CURRENCY READOUT SCREEN (SUPERSEDED)

Before the deployment-currency campaign of Appendix G, we had screened a token-readout axis (CLS, global average, global max, 2×2 pool, projection head) in held-out encoding R^2, the currency this paper argues against, over a different backbone set (the CLIP family, DINOv2, ResNet-CLIP). We keep the pointer here for the record and for one caution. That screen and the retrieval ladder of Figure 9 are measured in different currencies on different models and are *not* interchangeable, and by the dissociation of §3, the deployment-currency version is the one a decoder builder should use. The full R^2 readout tables will be part of the release planned upon acceptance.

I Cross-Objective Sub-Layer Anatomy

Repeating the atomic probe of §5 for MAE, DINOv2, EVA-02, and ConvNeXt-V2, the last with a convolutional analogue of the component set (1,740 cells), shows that the anatomy's opening is universal while its ending diagnoses the objective's target family. Block-0 writes exceed the stream at entry by $+25$ to $+44$ points across the four transformers and the convnet, each in all ten subjects, so this early writing era is both objective- and architecture-universal. The endings split by family. CLIP's computations collapse while its stream coasts. MAE never stops integrating, with writes comparable to the stream through block 7 and the stream still strong entering the final block, which is the depth profile of Figure 4 seen from inside. DINOv2's final block actively *damages* its own stream, whose exit state falls below the block's own Q, K, and V inputs in all ten subjects. EVA-02, masked like MAE but predicting CLIP features, ends like the self-distillation family rather than the contrastive one. Its final block drops its own stream, its computations beat its exit stream in all ten subjects, and its **attention output beats the residual stream at the model's own best depth** (9/10 subjects, p=.011), a lesson the convnet repeats through its depthwise-conv output. Stream optimality, result (iii) of §5, is therefore a property of the original trio rather than a law. For EVA-02-style targets one should read the attention output, and for ConvNeXt-V2 the depthwise-conv output. Even the K>Q ordering follows the family, strengthening under MAE, inverting late under DINOv2, and appearing only at mid-depths under EVA-02. EEG sees attention's content under contrastive and reconstruction training and attention's addressing under self-distillation, and a model's target family, not its input recipe, decides which.

J The Ledger of Retired Hypotheses

Each entry records a hypothesis of ours that the atlas retired, together with the evidence that retired it.

1. **Encoding fit as a selection criterion**, our own earlier instrument. It is anti-correlated with retrieval over early depth, so we retired it as a configuration tool and kept it as motivation (§3).
2. **"Frontal electrodes align to late layers."** Frontal electrodes alone are near chance in both settings. What survives is posterior-anterior *synergy*, and only inter-subject (§4).
3. **Cascade "peak migration."** It held in only 6/10 subjects under seed replication and was replaced by the lower-edge form, which holds in 10/10 (§4).
4. **"Anterior gain grows with depth."** It was a two-subject artifact that vanished at ten subjects.
5. **Learned gates.** They fail in both directions, and we report the failure rather than hide it (§7.1).
6. **The low-level-retention ordering as a peak ranking**, our hypothesis entering the §6 screen. It failed. At best depth all trained models sit within a narrow band, and the VAE loses to a supervised ResNet in nine of ten subjects. The principle survives only in relocated form, as depth decay rather than peak height.
7. **Universal K>Q.** The universal form died with DINOv2, which inverts the ordering late. Only the objective-conditional form stands (§6).
8. **The trained-peak band as a universal.** Extending the screen from 11 to 21 models broke it, since hierarchical pooling (Hiera-MAE, 82.3) and latent-target prediction (I-JEPA, 69.3) fall below the band, which we restate as a property of mainstream single-stream training (§6). I-JEPA's profile is the atlas's strangest object, with early and mid blocks scoring *below the raw-pixel floor* and rising only near the output, and we report it without a theory (huge-only checkpoint, not patch-matched).
9. **The greedy seed as a chain member.** The state that *started* the winning fusion chain does not belong in it. The finished chain's leave-one-out audit evicts it, because a deeper read of the same model absorbed its territory (§7.1). The lesson is constructive, in that the greedy path was correct at every step given what existed, and the audit is what turns a construction into a structure. The companion caveat is stated where the chain is, namely that greedy selection plus audit certifies a local optimum, not a global one.

10. **The naive foveal claim.** The raw retinotopy map shows exactly the foveal bias one would headline, yet our own pre-registered image-statistics control shows the stimuli are *more* foveal than the EEG, inverting the normalized map (Appendix G). The raw map alone would have been a publishable artifact. The control's own limitation, a nearest-class-mean reference rather than a capacity-matched ceiling, is stated alongside it.

K ONE ELECTRODE, END TO END

The single-electrode map of §4, in which Oz alone supports 43% 200-way retrieval, invites a fair question: how can one channel of voltage carry that much? This appendix traces the answer from the raw signal to the sub-layer states of the network, using the raw public EEG and 2,160 dedicated probe cells (Figure 11; subject 02, the strongest single-electrode subject, is shown for the signal panels).

The code is temporal, and it needs averaging. A single electrode is not a single number: the probe reads its full epoch, 250 samples at 250 Hz. Averaged over independent halves of the 80 test repetitions, each image evokes a reproducible, image-specific waveform at Oz (Figure 11a), and plain correlation between half-1 and half-2 waveforms, with no learning at all, identifies the correct concept among 200 at 72.5% top-1 (median rank 1; chance 0.5%). A principal-component analysis of the concept-average waveforms shows roughly eight temporal components carry 90% of the reliable variance, so the electrode is best understood as an approximately eight-dimensional temporal code. The code is exposed by repetition averaging: the same identification falls to 1.7% on single trials, which is why the atlas treats trial count as a regime variable throughout. The information is confined to 100–400 ms (Figure 11b), in register with the window law of §4.

The sub-layer anatomy survives at one electrode. Repeating the atomic probe of §5 with Oz as the only input (480 cells) reproduces the full three-act anatomy: block-0 writes exceed the stream at entry ($+17.0$ at $N{=}80$, $+6.5$ at $N{=}4$), the stream overtakes every computation by mid-network, and at the final block the stream retains what the collapsed computations no longer offer ($+6.9$ / $+2.4$); K exceeds Q at block 0 ($+4.1$ / $+0.7$); every one of the six statistics holds in 10/10 subjects (Figure 11c).

One timestamp per electrode. A component $\times$ depth $\times$ window grid (1,680 cells, the window grid of the $177 \rightarrow 201$ ms analysis) asked a pre-registered question: do the block's writes draw on an earlier EEG latency than the residual stream at the same depth? They do not. Every computation at every depth reads the same moment at Oz, with all latency centroids inside a 15 ms band around 180 ms, and the depth-linked rise itself is absent at the single electrode ($+4.6$ ms from block 0 to 11, rising in 5/10 subjects) while the eight-electrode ensemble rises $+24$ ms in 10/10 (Figure 11d). The $177 \rightarrow 201$ ms law of §4 is therefore a property of the ensemble read-out, which re-weights electrodes with different intrinsic latencies as depth grows, rather than of any latency shift within a single electrode; this mirrors the ensemble-level origin of the parietal depth preference in §4.

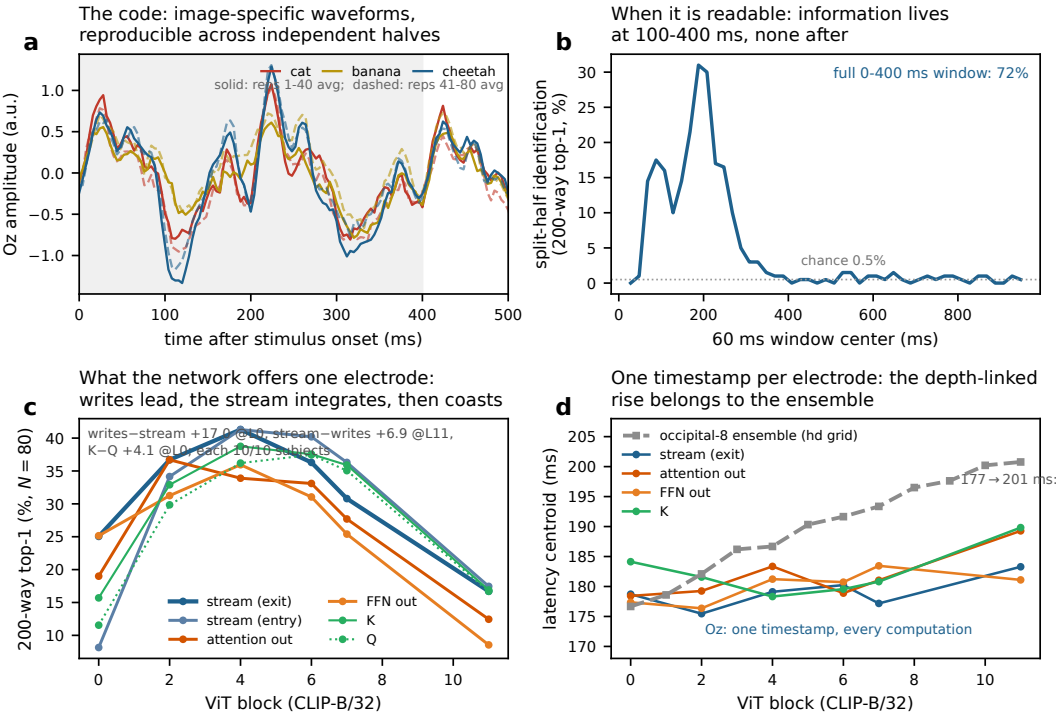

Figure 11: **One electrode, end to end.** **(a)** Real Oz waveforms (subject 02): three test concepts, each averaged over two independent halves of its 80 repetitions; the waveform is image-specific and reproducible. **(b)** Split-half correlation identification of the 200 test concepts from Oz alone in sliding 60 ms windows: the code lives at 100–400 ms and nowhere else (0–400 ms window: 72.5%). **(c)** The sub-layer anatomy of §5 probed through Oz alone (480 cells, N=80): writes lead at block 0, the stream integrates and peaks mid-network, computations collapse late while the stream coasts. **(d)** Latency centroids per computation (1,680 cells): at one electrode every computation at every depth reads ≈180 ms, and the 177 → 201 ms depth-linked rise (gray dashed: the occipital-8 ensemble) emerges only at the ensemble read-out.