BRAIN DECODING: TOWARD REAL-TIME READOUT OF VISUAL IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding visual mental imagery typically requires subject-specific calibration, raising the question of whether imagery can be decoded from representations shared across individuals. We show that a decoder trained on perception EEG from 21 participants transfers to imagery in a previously unseen participant, with no labelled imagery trials from the target participant. Per-participant standardisation is sufficient, while learned alignment provides no additional benefit. Transfer is carried by a pre-specified posterior alpha-power readout, increases with the number of source participants, and is unchanged when the standardisation is estimated on epochs that are never scored. At the single-neuron level, perception-to-imagery transfer appears within each patient in human ventral temporal cortex with no fitted transformation. Optimal-transport analysis of EEG further suggests that the perception-to-imagery transformation is described by a translation and a rescaling of a common representation. Finally, conditioning a frozen next-scale autoregressive image model on brain-predicted token prefixes enables real-time reconstruction of perceived images, while imagery reconstruction remains limited primarily by neural decoding accuracy. Together, these results show that cross-subject transfer can replace subject-specific calibration in imagery decoding and provide evidence for a common visual representation underlying perception and imagery.

Keywords: EEG, mental imagery, cross-subject transfer, brain decoding, neural representations

1 INTRODUCTION

Reconstructing seen images from brain activity has become a standard benchmark for neural decoding, from fMRI with generative priors (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023) to EEG systems that operate on consumer-grade recordings (Li et al., 2024; Kneeland et al., 2026a). Mental imagery, visual content generated without a stimulus, carries intent and meaning beyond words, yet has far less decoding evidence than perception. Within a participant, perception and imagery share a neural representation (Horikawa & Kamitani, 2017; Dijkstra et al., 2018), which in EEG is carried by posterior alpha oscillations (Xie et al., 2020). Two obstacles have limited its use. An imagery decoder must be calibrated on the participant's own imagined trials, which are slow and effortful to collect and cannot be locked to an external event, and published reconstruction methods have been validated on perception, where the target is known. The two forms of transfer that a deployable system requires, from one participant to another and from perception to imagery, have been studied separately (He & Wu, 2020; Gifford et al., 2022; Shimizu & Srinivasan, 2022; Tong & Chen, 2026).

We investigate whether both forms of transfer can be achieved together, and whether the same transfer appears in single neurons of human ventral temporal cortex. The experimental setting corresponds to the calibration procedure of a practical system: images are shown to a group of participants, and a new participant who was not part of that group is decoded with no labelled data from that participant and no imagery labels in any fit. We then examine what the decoded state can drive, and condition a frozen image generator on it under controls that distinguish reconstruction from memorisation.

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.

We decode the imagery of known images, a closed set of ten; open-ended imagination, where the imagined content is new to the decoder, is outside this work. By real-time we mean a decoder that needs no calibration session on the target participant, reads single trials and returns its output within a fixed latency; Section 4.1 reports single-trial accuracy and a strictly causal evaluation.

Contributions. We ask one question: does a decoder trained on other participants' *perception* read the *imagery* of a participant it has never seen, with no imagery label in any fit? Three results answer it. **(i) Perception reads imagery, in EEG and in single neurons** (Sections 4.1 and 4.2). In EEG, the transfer holds on two of three stimulus sets and stays at chance on a negative control fixed in advance. What carries it is per-participant standardisation of a ten-channel alpha-power readout: neither a learned alignment, nor a broad sweep of classifiers, nor a contrastive EEG–image encoder improves on it. **(ii) What separates the two states** (Section 4.3). In EEG, a translation and a rescaling account for the distance between perception and imagery responses. In single neurons, an encoding model fitted on perception alone predicts the imagery responses, and its layer-depth gradient is preserved. **(iii) Reconstruction** (Section 4.4). Forcing a brain-predicted token prefix into the coarsest scales of a frozen next-scale image decoder reconstructs seen images. Forcing half of the token positions outperforms forcing all of them. With imagery, reconstruction is bounded by decoding accuracy.

2 Related Work

Decoding Mental Imagery Imagery and perception share function and neural mechanism yet differ in direction of flow, cortical layer and phenomenology (Koenig-Robert & Pearson, 2021): the perceptual feed-forward cascade reverses during imagery (Dijkstra et al., 2020; Dentico et al., 2014; Hu & Yu, 2023), generative feedback yields distinct activity codes for seen and mental images (Breedlove et al., 2020), and the imagery network is heterarchical, with a domain-general node in the left fusiform gyrus and no consistent early-visual activity (Spagna et al., 2024; Bartolomeo et al., 2020; Liu et al., 2025a). Within participants, perception-trained decoders read imagery in fMRI (Horikawa & Kamitani, 2017), during sleep (Horikawa et al., 2013), and in EEG in posterior alpha power with heavy trial averaging (Xie et al., 2020; Stecher et al., 2025), at a coarser level of detail than perception (Stecher & Kaiser, 2024), and with a temporal profile that generalises more widely in time than perception (Dijkstra et al., 2018). Visual imagery as an EEG control signal is feasible but weak (Kosmyna et al., 2018; Lee et al., 2022; Kilmarx et al., 2024), and open paired perception-imagery EEG datasets remain few (Gao et al., 2026; Wilson et al., 2023; Chang et al., 2025). Shimizu & Srinivasan (2022) augment each participant's own training set with the others' perception, and Tong & Chen (2026) train one domain-adversarial network per participant on perception and imagery together. Both need labelled imagery from the tested participant, neither tests accuracy against a chance-level null, and neither corrects for multiple comparisons. Gao & Diniz (2026) frame perception-to-imagination EEG decoding as real-time: on 46 participants they classify perceived against imagined trials, a state rather than a content decision, at 96.7% leave-one-subject-out with 4 ms of inference per trial; every training participant contributes labelled imagery to the fit, so the model learns imagery from imagery, and no image content is decoded. No imagery label enters our fits, and the decision is which image was imagined. In fMRI, mental images have been reconstructed from perception-trained decoders (Kneeland et al., 2026b; Caselles-Dupré et al., 2024; Liu et al., 2025b): accuracy on seen images does not carry over to mental images, and low-dimensional features reconstruct imagery best (Kneeland et al., 2026b). At single neurons, Wadia et al. (2026) record human ventral temporal cortex while patients view and then imagine objects: about 80% of visually responsive neurons use a distributed axis code in a deep-network feature space, and imagery reactivates the same neurons proportionally to the same axis projection, while meta-analytic evidence places the imagery content code in the left fusiform gyrus rather than early visual cortex (Spagna et al., 2021).

Cross-Subject EEG Decoding Cross-subject transfer is widely studied in EEG (He & Wu, 2020); what it preserves is a coarse, low-dimensional component without functional alignment (Guntupalli et al., 2016). Riemannian recentring (Zanini et al., 2018; Congedo et al., 2017) and its learned form (Kobler et al., 2022) are the usual alignment operators; Euclidean alignment is their feature-space approximation (He & Wu, 2020). For motor imagery, Halme & Parkkonen (2018) decode across participants at within-participant parity, then find that training on others' passive movements and testing on imagery leaves all six methods short of significance. Large pretrained EEG models target

subject-invariant representations (Jiang et al., 2024; El Ouahidi et al., 2025); a benchmark across 36 tasks finds they only marginally beat task-specific models (Banville et al., 2026). Zero-shot cross-subject visual decoding from EEG has since been addressed with subject-adaptive front ends and contrastive alignment to image embeddings (Zhang et al., 2025; Huang et al., 2025; Jean Jacques et al., 2026; Chen et al., 2026). All of these evaluate perception, with leave-one-subject-out accuracy well below the within-subject figure (9.1% against 18.5% top-1 on THINGS-EEG for Zhang et al. (2025)), and none transfers from perception to imagery. Model-to-brain alignment has its own benchmarks, for perception (Schrimpf et al., 2018) and for the correspondence between network depth and cortical stage (Nonaka et al., 2021), and optimal transport has been used to align EEG sessions without retraining (Peterson et al., 2022).

Image Reconstruction from EEG Reconstruction from brain activity has moved from fMRI with deep features (Shen et al., 2019; Kamitani et al., 2025) to EEG, where contrastive EEG-image encoders (Song et al., 2024; Chen & Wei, 2024; Liu et al., 2026) feed a generator, including from low-density portable EEG (Guenther et al., 2024). Most EEG-to-image systems compress the recording into a single global vector that conditions a diffusion generator through a single vector (Li et al., 2024; Kneeland et al., 2026a; Dai et al., 2026). Yu et al. (2026) add frequency-oriented alignment and dual diffusion priors. The speed of the generator sets the latency of any closed-loop use (Benchetrit et al., 2023). Visual representations in the brain are hierarchical (Horikawa & Kamitani, 2017; Ho et al., 2023), as is next-scale autoregressive generation (Tian et al., 2024). Partial conditioning has named relatives: masked token schedules (Chang et al., 2022), noising strength (Meng et al., 2022), guidance scale (Ho & Salimans, 2022). The standard seven-metric panel (Ozcelik & VanRullen, 2023; Kneeland et al., 2026a) is known to be earnable by degenerate outputs (Shirakawa et al., 2024).

3 METHODS

3.1 DATASETS

The imagery dataset of Gao et al. (2026) records 22 participants (20 with two sessions) on 32 channels at 1000 Hz, resampled to 250 Hz, with interleaved classes and no baseline correction (van Driel et al., 2021), while they view an image for 4 s and, after a mask, imagine it for 4 s; the three runs present animals (AVI), figures (FVI) and objects (OVI), about 80 trials per class, each class being one fixed image, ten in all, so imagery decoding is closed-set classification among known images. THINGS-EEG2 (Gifford et al., 2022) provides perception EEG of 10 subjects on 63 channels for 16,540 training and 200 test images, averaged over repetitions as the benchmark prescribes, with a 400-image validation split for all model selection. The single-neuron release of Wadia et al. (2026) provides 367 screening neurons from 16 patients viewing 500 stimuli and 43 neurons from 6 of them imagining 6 to 8 objects, trial-averaged and pre-selected by the authors of the release (Appendix F). Preprocessing and the label audit are in Appendix A.

3.2 EVALUATION PROTOCOL

All transfer analyses were pre-registered. The evaluated unit never enters a fit: on the EEG datasets it is the participant, who contributes no labelled data and whose class grouping serves only to form pseudo-trials; on the Gao dataset, where each class is one image, classification is pairwise between classes; where a class has several stimuli (THINGS-EEG2, Wadia) evaluation is also leave-one-stimulus-out. Cross-patient transfer on the single-neuron data is evaluated leave-one-patient-out. Every comparison is repeated with the source labels permuted, and the reported effect is the difference $\Delta_p = A_p - \mathbb{E}_\pi[A_p^\pi]$. A negative control is designated in advance on published grounds (Prasad & Jena, 2013), the animal set on the Gao dataset (Appendix B). Correspondence and artefact controls comprise a cyclic derangement of the class correspondence, which preserves any geometry mappable onto arbitrary labels and abolishes a genuine correspondence, a shared presentation-order control, an ocular-surrogate control with a topographic test, and a simulated guard band between each readout window and the following event. Primary comparisons are enumerated and corrected (Bonferroni over 37 on the Gao dataset), descriptive readouts support no claim, and a detection floor accompanies every null. Its split, its own null, its analysis windows, its designated negative control and its detection floors are implemented as a task family in an open benchmarking framework (Appendix H).

3.3 Cross-Subject Readout

A trial $x_i \in \mathbb{R}^{C\times T}$ is represented by its log alpha power on $C_{\text{post}} = 10$ posterior electrodes, averaged over the alpha bins of a fixed Morlet grid within 8 to 13 Hz and over a readout window W; the electrodes and the band follow Xie et al. (2020) and were fixed in advance:

$$z_{i,c} = \frac{1}{|W|}\sum_{t\in W} \log\Big(\frac{1}{|\mathcal{F}_\alpha|}\sum_{f\in\mathcal{F}_\alpha} \big|(x_{i,c} * \psi_f)(t)\big|^2\Big), \qquad \tilde{z}_i^{(m)} = \big(z_i^{(m)} - \mu_p^{(m)}\big) \oslash \sigma_p^{(m)}, \quad (1)$$

where the second expression standardises each participant p and modality m to its own mean and standard deviation, without labels and without rotation (the standardisation hereafter). Trials are averaged within class into pseudo-trials ($k = 40$ for the pooled source, $k = 10$ for the target, $R = 40$ redraws), and a linear SVM trained on the pooled perception pseudo-trials of all participants except p classifies each pair of classes, so chance is $1/2$ for every class count and the held-out participant contributes no labelled data to any fit. The moments are estimated on the target participant's own unlabelled epochs, so the procedure needs no label and no calibration session.

3.4 Single-Neuron Analyses

Within a patient and recording block, a map from perception to imagery (identity, scalar gain, per-neuron gain; a general linear map is under-determined at 6 to 8 stimuli) is fitted on the paired responses of the $S - 1$ training stimuli, the held-out stimulus's perception vector is mapped, and the block's imagery vectors are ranked against it; the held-out imagery never enters a fit, the null permutes the imagery labels within block, and blocks with fewer than three neurons are excluded (7 blocks, 36 neurons, 4 patients). An axis model is fitted on each neuron's perception responses to 499 screening stimuli and applied to its imagery responses, per layer of two networks. Cross-patient transfer uses the stimulus as the shared axis, with representational similarity and a fitted ridge map (Appendix F).

3.5 Shared Readout Space

The hypothesis is that the same low-complexity mapping relates perception and imagery in EEG and in single neurons, measured in a common unit. The perception response of each modality is mapped by ridge regression into the same vision-model feature space (AlexNet fc6 reduced to its first 50 principal components, shared by both modalities). It serves three analyses: retrieval in EEG (Section 4.1), retrieval in single neurons (Section 4.2) and the distance between perception and imagery (Section 4.3). The perception-fitted readout is applied to imagery responses, and three quantities are computed identically for both modalities: the retrieval rank of the true stimulus among the unit's candidates (the task's 3 or 4 images on Gao, the 6 to 8 imagined objects of a patient block on Wadia) against a source-label-permutation null; the readout's own held-out perception retrieval, reported as its ceiling next to every transfer number; and the distance between the two states as point clouds. Each stimulus is one point of the readout space, so a modality is a cloud of points, and the 2-Wasserstein distance measures how far the perception cloud sits from the imagery cloud as distributions. Two halves of the perception data alone are separated by a distance of the same kind, and that split-half value is subtracted, so what remains is the part of the gap that noise does not explain. The remainder is then measured after each map, in order of increasing complexity, applied to the perception cloud, from the identity through recentring to diagonal scaling, each map fitted on one half of the trials and scored on the other; rotation is not estimable at these ranks. The rung at which the remainder stops falling is the complexity the map requires (Figure 7).

3.6 Reconstruction with a Frozen Decoder

We use a VAR next-scale autoregressive transformer (Tian et al., 2024) with a VQ tokenizer and scales $(1, 2, 3, 4, 5, 6, 8, 10, 13, 16)$, at depth 16 (420 M parameters) and depth 30 (2.1 B), adapted once on image embeddings alone to accept a ViT-H image embedding (Radford et al., 2021) in place of its class label, then frozen and shared across subjects; no gradient from EEG reaches it. Two closed-form ridge regressions, fitted on training images only, map the EEG embedding of the contrastive encoder of Liu et al. (2026) to a global image embedding that conditions the decoder through adaptive normalisation and to a spatial band map that the tokenizer's own quantiser turns into token maps at the coarse scales (Figure 3; Appendix G).

Table 1: Cross-subject decoding, in points above each comparison's own permutation null, mean ± SE over 22 participants and eight seeds, each row scored in its own target window. The same decoder across the rest of the trial is Figure 2.

	OVI	FVI	AVI (neg.)
perception → perception	$+7.5 \pm 1.5$	$+13.6 \pm 2.8$	$+13.5 \pm 2.0$
perception → imagery	$\mathbf{+5.3 \pm 1.2}$	$\mathbf{+3.5 \pm 1.6}$	-1.8 ± 1.9
own perception → imagery	$+4.9 \pm 1.4$	$+2.1 \pm 1.4$	$+0.9 \pm 1.6$

Prefix forcing. Let $\hat{t}_s$ denote the EEG-predicted token map at scale s, t_s^{model} the decoder's own map, and M_s a binary position mask; the conditioned map is $t_s = M_s \odot \hat{t}_s + (1 - M_s) \odot t_s^{\text{model}}$ for $s < K$, where K is the number of forced scales and q the fraction of positions forced within them, chosen per subject and scale as those of highest ridge confidence on the training images (45 of 91 positions at $K = 6$, $q = 0.5$; a random choice of the same number changes one metric by 0.007). Full prefix forcing, $q = 1$, is the only setting considered in prior hierarchical conditioning; $K = 0$ is flat conditioning.

4 Results

4.1 Cross-Subject Transfer from Perception to Imagery

A decoder trained on the pooled perception of the other participants classifies a held-out participant's imagery epochs on OVI ($+5.3 \pm 1.2$ pp above the null, $p = 0.0002$, 18 of 22 participants) and FVI ($+3.5 \pm 1.6$, uncorrected $p = 0.038$, 15 of 22), and not on AVI, the negative control (-1.8 ± 1.9; Table 1). The OVI result holds after Bonferroni correction over the thirty-seven primary comparisons and under two non-parametric tests at the participant level; FVI does not survive that correction and is reported as convergent support only. It is present when the evaluation unit is a single trial ($+1.9$ pp, 18 of 22) and grows to $+7.0$ as trials are averaged. Scored as retrieval in the shared readout space, the same transfer finds the imagined object on OVI in 30.1% of cases against a permutation null of 25.0% (Holm-corrected $p = 0.011$, 15 of 22 participants), while it is at chance on the negative control and inside its floor on FVI. The transfer increases with the number of source participants on OVI ($+0.90$ pp per doubling, $p = 0.0001$) and FVI, and not on AVI (Figure 1a).

Controls. A cyclic derangement of the class correspondence abolishes the transfer ($+5.85$ pp above the deranged condition, $p = 0.0027$). Source participants who share no presentation order with the target reproduce it, and among six combinations of frequency band and scalp region it is confined to posterior alpha (Figure 1b). Ocular-component removal leaves it unchanged while abolishing the negative control's frontal perception decoding (Appendix C). Standardising the scored trials to the moments of a never-scored half of the participant's imagery leaves the transfer unchanged, whereas moments taken from the participant's perception lose 2.9 pp (Appendices A and C). Scored in chronological order, with the standardisation estimated from each participant's preceding imagery trials only, the transfer holds at $+1.90$ pp ($p = 0.0004$, 19 of 22) against $+1.88$ when the moments use all of the participant's imagery trials, so the readout can be computed causally.

Alignment and classifier capacity. With architecture and readout held fixed, learned recentring contributes $+1.6$ pp and a closed-form recentring without a network reaches $+4.1$, so learning the alignment does not improve on computing it. A search over convolutional networks, filter-bank front ends, Riemannian tangent-space classifiers, SPD networks with learned recentring and multi-source weighting, each paired against the readout on the same epochs with its own floor, produced no configuration that outperformed the ten-channel alpha-power readout on OVI or FVI (Appendix D, Table 10). A contrastive EEG-image encoder, the family shared by the cross-subject frameworks of Section 2 and by the encoder of Section 4.4, was run as a further baseline under the same protocol: the official encoder of Liu et al. (2026) decodes the held-out participant's perception ($+19.8$ pp over its null on OVI, 22 of 22) and does not transfer to imagery (-3.2 pp), where it is 8.0 pp below the alpha readout ($p = 0.0006$).

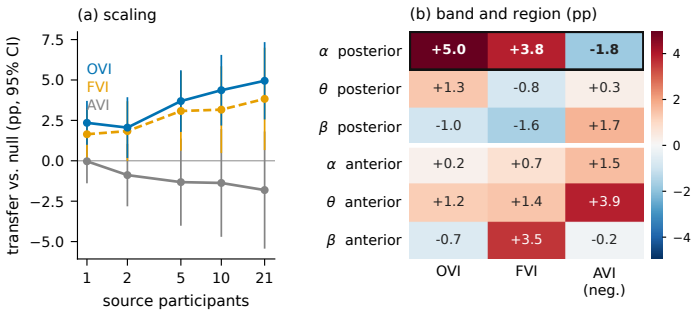

Figure 1: Controls on the transfer of Table 1. (a) Transfer against the number of source participants, mean and 95% CI. (b) Transfer by frequency band and scalp region, in points above each null; posterior alpha, fixed in advance, is boxed; anterior alpha is the ocular control.

4.2 Perception-to-Imagery Transfer in Single Neurons

The same question is asked at the single-neuron scale, in human ventral temporal cortex (Wadia et al., 2026), within each patient, because neurons differ between patients.

A perception population vector identifies the matching held-out imagery vector at **22.2%** against a measured permutation null of 12.9% ($p = 0.0315$), with no fitted transformation, within patient and leave-one-stimulus-out, over 4 patients (Figure 5). Fitting a per-neuron gain reaches **29.6%**, but its increment over identity, $+7.4$ pp, is not significant against its own permutation null ($p = 0.061$, carried by 3 of the 7 blocks and reversed in none). A general linear map is under-determined with at most eight stimuli and is not fitted; as in EEG (Section 4.1), the simplest map suffices. The transfer is asymmetric in one respect: in the reverse direction the identity map still transfers ($p = 0.049$) while a fitted per-neuron gain does not, consistent with over-fitting seven gains on seven stimuli.

Retrieval in the shared readout space. Read into the shared readout space instead, the same routing is null (imagery top-1 16.7% against a 13.2% null, detection floor 20.4%): a readout built from a block's 3 to 7 neurons has rank at most the number of neurons, and its own perception accuracy sits barely above that floor, whereas the population vector above carries no such constraint. Cross-patient transfer, evaluated along the shared stimulus axis, is not detected (detection floor 14.8%; Appendix F).

4.3 Temporal Dynamics, Geometry and Brain-Model Alignment

Time course. Reading the same decoder across the trial (Figure 2), the effect is present at image onset, falls to chance while the image is on screen, returns during the delay and reaches its maximum during imagery. Scored in the guarded windows on the final epoch set, the delay reads $+4.5$ ($p = 0.004$) and imagery $+5.9$ ($p = 0.00003$). The decodable window is therefore wider than the imagery epoch and precedes the cue, an observation Section 5 takes up. The delay and imagery peaks coincide with the post-stimulus alpha rebound (Pfurtscheller & Lopes da Silva, 1999).

Induced oscillations. Subtracting each participant's phase-locked response, the label-free grand mean over that participant's trials, leaves the OVI transfer at $+3.38$ pp ($p = 0.007$, 16 of 22), while the same subtraction removes 4.30 pp of within-modality perception transfer. The phase-locked component carries a substantial part of the within-modality readout and at most a small fraction of the transfer to imagery (Appendix C).

Geometry. The transport distance between the perception and imagery states falls along the same sequence of maps as in feature space, a translation and then a rescaling, each step significant after Holm correction on all three tasks and no further step supported by the recordings (Appendix H, Figure 7).

Brain-model alignment. An encoding model fitted on perception responses only (each of the 43 reactivated neurons on its own screening responses to the other 499 stimuli, least squares, stimulus-held-out) predicts their imagery responses with mean $r = +0.271$ ($p = 0.0010$, AlexNet fc7; Figure 6). The layer-depth gradient of the code is monotone in perception for both architectures



Figure 2: The same decoder read across the trial, in points above the permutation null: 0.60 s windows stepped by 0.30 s, mean and 95% CI, $n = 22$, source the pooled early perception window. Windows overlap by half and are unguarded, so every per-window value is descriptive.

Table 2: Test panel on THINGS-EEG2, 10 subjects, frozen decoder at depth 16 (420 M parameters) and depth 30 (2.1 B), on the seven identification metrics published by ENIGMA. Best method in bold, second underlined; the three rows below the rule are a floor, a degenerate baseline and a ceiling. D^2-FOSA reports only PCC and SSIM, hence its empty cells. The distribution metrics of D^2-FOSA, which do not rank reconstructions, are in Appendix E, Table 11.

	PCC↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	SwAV↓
ENIGMA, published	0.1718	0.4233	0.8364	0.8949	0.7765	**0.8148**	0.5403
D^2-FOSA, published	0.193	0.350					
VAR-d16, $K = 0$ (flat)	0.1798	0.4140	0.8707	0.9200	0.7550	0.7600	0.5167
VAR-d16, HINT	<u>0.2233</u>	**0.4285**	<u>0.9149</u>	<u>0.9358</u>	0.7701	0.7547	<u>0.4969</u>
VAR-d30, $K = 0$ (flat)	0.1935	0.4167	0.8805	0.9290	<u>0.7790</u>	0.7992	0.5015
VAR-d30, HINT	**0.2271**	<u>0.4277</u>	**0.9276**	**0.9411**	**0.7844**	<u>0.8013</u>	**0.4824**
permuted prefix, $K = 6$	0.0708	0.3688	0.5039	0.4830	0.5026	0.5028	0.6287
one flat colour ($d = 1$)	0.3079	0.5846	0.6602	0.6013	0.5142	0.5206	0.6662
oracle prefix (ceiling)	0.5099	0.4706	0.9999	0.9999	0.9959	0.9933	0.2214

(Spearman $\rho = +1.000$ over 8 AlexNet layers and 12 DINO blocks) and is preserved in imagery for AlexNet ($\rho = +0.833$, $p = 0.0102$) but not for DINO ($\rho = +0.315$, $p = 0.32$); AlexNet does not significantly outperform DINO on imagery ($p = 0.0599$), so no architecture dissociation is claimed (Appendix F). The same encoding convention gives the scalp counterpart of this single-neuron result: on THINGS-EEG2 (10 subjects, 200 test images, stimulus-held-out ridge encoding), all 21 visual feature spaces (AlexNet conv1 to fc8, DINO blocks 1 to 12, CLIP image) predict the EEG response better than all semantic spaces (GloVe, CLIP text, BERT), with no overlap between the two families (the weakest visual space exceeds GloVe by $+0.184$, $p = 0.0039$, 9 of 10 subjects), and earlier layers best predict earlier EEG time points. The split and its own-null permutation are implemented as a visual-imagery task family with a cross-state split in NeuralBench (Banville et al., 2026), released with the camera-ready version (Appendix H).

4.4 Image Reconstruction with a Frozen Autoregressive Decoder

Leaving half of the token positions of the brain-predicted prefix to the prior improves the semantic metrics for every one of the 10 subjects at $K = 10$ (CLIP-L/14 by $+0.078$, Inception by $+0.064$, SwAV by $+0.048$, each at the smallest Holm-corrected value attainable at $n = 10$) and at the configuration of Table 2; the pixel-level metrics are unchanged. The effect replicates across two samplers, two generation seeds and four bridge sizes spanning a factor of 52 in parameters; the mechanism controls, the effect of decoder depth and the cost accounting are in Appendix E.

Comparison with the reference methods. The configuration of Table 2 exceeds ENIGMA (Kneeland et al., 2026a) on five of seven identification metrics with VAR-d16 and six of seven with VAR-d30. D^2-FOSA (Yu et al., 2026) evaluates the same dataset with a distribution panel instead, and against it our reconstructions are ahead on pixel correlation and on within-subject 200-way retrieval (65.2% against 38.0%) and behind on SSIM, IS, FID and KID. The two panels disagree in a way

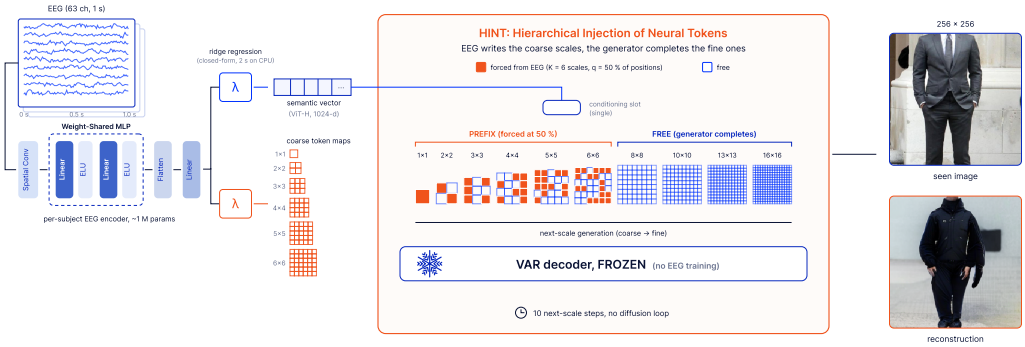

Figure 3: Reconstruction from perception EEG (THINGS-EEG2). Two ridge maps turn the EEG embedding into a semantic vector and coarse token maps; half of the positions of the first six scales are forced from EEG and the frozen VAR decoder generates the rest. Right: a seen image and its reconstruction.

that is informative about the metrics rather than about the methods. The three distribution metrics (Appendix E, Table 11) do not rank reconstructions on this dataset: a permuted prefix, which reconstructs a different image, matches the deployed configuration on all three. The uniform-colour image that scores above every published method on PixCorr and SSIM is in turn the worst row on all three, and flat conditioning, which forces no token at all, has a better FID and KID than the configuration that beats it on all seven identification metrics. Identification metrics reward correspondence to the particular target and distribution metrics reward typicality, and a reconstruction claim needs the first.

Reconstruction from imagery. An imagery decoder that chooses among a small candidate set can drive the generator only as retrieval-conditioned rendering, and the reconstruction is then no better than the selection. The cross-participant readout selects the imagined image among four in 40% of participant-class units (16 of 22 participants, $p = 0.004$), with each session standardised to its own trials, as the causal estimate of Section 4.1 does. The renders identify the true image just above the floor of a permuted selection, as a 40%-correct selection predicts; at single neurons the same procedure selects above chance and renders at its floor. Because each class is one fixed image, a correct render reflects the selection and carries no imagery content beyond class identity.

5 Discussion

We asked whether a decoder can read a person's mental imagery without ever having seen that person. On EEG it can, within the limits set out below. A decoder trained on other participants' perception classifies the imagery of a held-out participant, the same posterior alpha pattern recurring across people and across states, and it does so with a ten-channel readout that no learned alignment, no architecture search and no contrastive encoder improves upon. What makes the transfer possible is not a shared feature space learned from data but a per-participant standardisation of a fixed feature, estimated within the recording session, and a causal estimate of it matches the offline one ($+1.90$ against $+1.88$ pp), so the readout runs in real time as defined in the introduction. Single neurons in ventral temporal cortex show the same picture: the identity map transfers perception to imagery within a patient, a fitted map does not improve on it, and the transfer does not cross patients. Both modalities need only a low-complexity map: a translation and a rescaling in EEG, the identity within patient in single neurons.

The readout is weak at the single trial: it is real, causal and calibration-free, but a few percentage points above chance, which is why imagery does not yet drive the generator. The generator itself is not the limit. A frozen next-scale decoder driven by a closed-form bridge reconstructs seen images in 68 ms of inference per trial, and forcing half of the token positions outperforms forcing all of them. Driven by imagery, the participant-level selection is correct in 40% of cases among four images, and the rendered images are where that accuracy places them, just above the floor of a permuted selection: the bottleneck is selection, not rendering, in both modalities.

Limitations. Each imagery class corresponds to a single fixed image, so class and image identity are confounded and the transferred pattern is an image-anchored contrast rather than a category geometry. It is present before the cue, so it may reflect short-term memory, sustained attention or early-onset imagery; a design that reveals the item only at the cue would separate them. The single-neuron release is trial-averaged and axis-selected, so every neuron result is an upper bound with an effective sample of four patients. Reconstruction is offline on repetition averages, and two of the seven standard metrics are exceeded by a single flat-colour image for every method, so reconstruction scores are read against controls that test whether they require reconstruction at all. Prefix forcing is compared only with no forcing, so the panel measures the effect of the prefix; comparing conditioning mechanisms is left open. Future work will test the transfer on additional corpora with several exemplars per class, including a pre-registered replication on Alljoined-1.6M (Xu et al., 2025), recorded with a 32-channel consumer headset that carries all ten readout electrodes.

Conclusion. Imagery decoders have so far been fitted on labelled imagery, usually recorded from their user in a calibration session. For a closed set of known images, the readout reported here needs no such session: it is fitted on other people's perception, adapts to a new user through the mean and variance of that user's unlabelled imagery, and reads single trials at a fixed latency from ten posterior electrodes that a consumer headset already carries. It is a few percentage points above chance, and that accuracy sets what a neural interface built on it can display. Because it needs no data from its user, it can be evaluated on many people at once, the scale at which imagery decoding for real-world interfaces can be tested.

AI USE STATEMENT

We used large language models in three of the categories that require disclosure. For writing, they edited and shortened prose and proposed wording, which the authors then revised; the structure, the claims and the argument of the paper are the authors'. For retrieval and discovery, they located candidate related work, each item of which an author read before it was cited. For research execution, they wrote analysis and experiment scripts and ran them on the authors' compute under the authors' instructions. We did not use them to generate synthetic datasets, and the paper proves no mathematical claims. All AI-assisted work was reviewed. The EEG transfer analyses were pre-registered before they were run, with their endpoint, null and detection floor named in advance; every number in the paper is traced to the script and the output file that produced it; the analysis code was read by the authors; and results that could not be reproduced from the released assets are reported as not reproduced rather than adjusted. We take responsibility for the final content of this work, including its text, its claims and its artifacts.

REPRODUCIBILITY STATEMENT

Datasets are public: Gao et al. (2026) on NEMAR (nm000242), THINGS-EEG2 on Figshare (18470912) and Wadia et al. (2026) on OSF (g9xh6). Code, derived epoch sets and run ledgers will be released with the camera-ready version, and they trace every number in the paper to the script and output file that produced it. The EEG transfer estimator is closed-form; its eight seeds, permutation counts and windows are given in Appendix A. The reconstruction experiments ran on single NVIDIA L40S GPUs, except the one-time adaptation of the depth-30 decoder (one H100); the exact software environment, seeds and decoder snapshots are listed in Appendix G.

ETHICS STATEMENT

A decoder that reads a participant's imagery with no labelled data from that participant lowers the calibration barrier, which is both the assistive point and the risk. The threat model is covert decoding of an internal state without consent. The results bound it: the transfer requires the participant's active engagement in a cued imagery task, reads a closed set of ten known images at five points above chance, and does not transfer between patients at single neurons. The single-neuron data are a public release of recordings obtained under the original study's clinical consent; our analysis is secondary, on de-identified trial-averaged responses, and adds no procedure to what those patients consented to. Per-subject reconstruction models are fitted on thousands of paired trials that only a consenting participant can provide.

REFERENCES

Hubert Banville, Stéphane d'Ascoli, Simon Dahan, Jérémy Rapin, Marlène Careil, Yohann Benchetrit, Jarod Lévy, Saarang Panchavati, Antoine Ratouchniak, Mingfang Zhang, Elisa Cascardi, Katelyn Begany, Teon Brooks, and Jean-Rémi King. NeuralBench: A unifying framework to benchmark NeuroAI models, 2026.

Paolo Bartolomeo, Dounia Hajhajate, Jianghao Liu, and Alfredo Spagna. Assessing the causal role of early visual areas in visual mental imagery. *Nature Reviews Neuroscience*, 21(9):517, 2020. doi: 10.1038/s41583-020-0348-5.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2023.

Jesse L. Breedlove, Ghislain St-Yves, Cheryl A. Olman, and Thomas Naselaris. Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, 30(12):2211–2224.e6, 2020. doi: 10.1016/j.cub.2020.04.014.

Hugo Caselles-Dupré, Charles Mellerio, Paul Hérent, Alizée Lopez-Persem, Benoit Béranger, Mathieu Soularue, Pierre Fautrel, Gauthier Vernier, and Matthieu Cord. Mind-to-image: Projecting visual mental imagination of the brain from fMRI, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer, 2022.

Yan-Han Chang, Hsi-An Chen, Min-Jiun Tsai, Chun-Lung Tseng, Ching-Huei Lo, Kuan-Chih Huang, and Chun-Shu Wei. A human EEG dataset for multisensory perception and mental imagery, 2025. OpenNeuro ds005815.

Chi-Sheng Chen and Chun-Shu Wei. Mind's eye: Image recognition by EEG via multimodal similarity-keeping contrastive learning, 2024.

Zexuan Chen, Sichao Liu, Runhao Lu, Huichao Qi, Alexandra Woolgar, Xi Vincent Wang, and Lihui Wang. MindAlign: Bridging EEG, vision, and language for zero-shot visual decoding, 2026.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014. doi: 10.1038/nn.3635.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017. doi: 10.1080/2326263X.2017.1297192.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from EEG signals, 2026.

Daniela Dentico, Bing Leung Cheung, Jui-Yang Chang, Jeffrey Guokas, Melanie Boly, Giulio Tononi, and Barry Van Veen. Reversal of cortical information flow during visual imagery as compared to visual perception. *NeuroImage*, 100:237–243, 2014. doi: 10.1016/j.neuroimage.2014.05.081.

Nadine Dijkstra, Pim Mostert, Floris P. de Lange, Sander Bosch, and Marcel A. J. van Gerven. Differential temporal dynamics during visual imagery and perception. *eLife*, 7:e33904, 2018. doi: 10.7554/eLife.33904.

Nadine Dijkstra, Luca Ambrogioni, Diego Vidaurre, and Marcel van Gerven. Neural dynamics of perceptual inference and its reversal during imagery. *eLife*, 9:e53588, 2020. doi: 10.7554/eLife.53588.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG, adapting to any setup with large-scale pretraining on 25,000 subjects, 2025.

Jing'ao Gao, Yao Liu, Zhengshuang Li, Kaixin Huang, Fan Wang, Jiaping Xu, Lei Zhao, Tianwen Li, and Yunfa Fu. An EEG dataset for visual imagery-based brain–computer interface. *Scientific Data*, 13:194, 2026. doi: 10.1038/s41597-025-06512-5.

Yu Gao and José Miguel Diniz. From perception to imagination: a robust deep learning architecture for real-time EEG decoding and neurophysiological validation. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 14(1):2684109, 2026. doi: 10.1080/21681163.2026.2684109.

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.

Google Research. TabFM: A scikit-learn compatible tabular foundation model. `https://github.com/google-research/tabfm`, 2026. v1.0.0 pretrained weights.

Benjamin J. Griffiths, Stephen D. Mayhew, Karen J. Mullinger, João Jorge, Ian Charest, Maria Wimber, and Simon Hanslmayr. Alpha/beta power decreases track the fidelity of stimulus-specific information. *eLife*, 8:e49562, 2019.

Sven Guenther, Nataliya Kosmyna, and Pattie Maes. Image classification and reconstruction from low-density EEG. *Scientific Reports*, 14(1):16436, 2024. doi: 10.1038/s41598-024-66228-1.

J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6):2919–2934, 2016. doi: 10.1093/cercor/bhw068.

Hanna-Leena Halme and Lauri Parkkonen. Across-subject offline decoding of motor imagery from MEG and EEG. *Scientific Reports*, 8(1):10087, 2018. doi: 10.1038/s41598-018-28295-z.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Jun Kai Ho, Tomoyasu Horikawa, Kei Majima, Fan Cheng, and Yukiyasu Kamitani. Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, 271:120007, 2023. doi: 10.1016/j.neuroimage.2023.120007.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8:15037, 2017. doi: 10.1038/ncomms15037.

Tomoyasu Horikawa, Masako Tamaki, Yoichi Miyawaki, and Yukiyasu Kamitani. Neural decoding of visual imagery during sleep. *Science*, 340:639–642, 2013. doi: 10.1126/science.1234330.

Yiheng Hu and Qing Yu. Spatiotemporal dynamics of self-generated imagery reveal a reverse cortical hierarchy from cue-induced imagery. *Cell Reports*, 42(10):113242, 2023. doi: 10.1016/j.celrep.2023.113242.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, Siyi Sun, and Yongxiong Wang. NEED: Cross-subject and cross-task generalization for video and image reconstruction from EEG signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Baimam Boukar Jean Jacques, Brandone Fonya, Nchofon Tagha Ghogomu, Pauline Nyaboe, and Kipngeno Koech. Zero-shot neural priors for generalizable cross-subject and cross-task EEG decoding, 2026.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024.

Yukiyasu Kamitani, Misato Tanaka, and Ken Shirakawa. Visual image reconstruction from brain activity via latent representation, 2025.

Justin Kilmarx, Ivan Tashev, José del R. Millán, James Sulzer, and Jarrod Lewis-Peacock. Evaluating the feasibility of visual imagery for an EEG-based brain–computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:2209–2219, 2024. doi: 10.1109/TNSRE.2024.3410870.

Reese Kneeland, Wangshu Jiang, Ugo Bruzadin Nunes, Paul Steven Scotti, Arnaud Delorme, and Jonathan Xu. ENIGMA: EEG-to-image in 15 minutes using less than 1% of the parameters, 2026a.

Reese Kneeland, Cesar Kadir Torrico Villanueva, Tong Chen, Jordyn Ojeda, Shubh Khanna, Jonathan Xu, Paul S. Scotti, and Thomas Naselaris. MIRAGE: Robust multi-modal architectures translate fMRI-to-image models from vision to mental imagery. *PLOS Computational Biology*, 2026b. doi: 10.1371/journal.pcbi.1014263.

Reinmar J. Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Roger Koenig-Robert and Joel Pearson. Why do imagery and perception look and feel so different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1817):20190703, 2021. doi: 10.1098/rstb.2019.0703.

Nataliya Kosmyna, Jussi T. Lindgren, and Anatole Lécuyer. Attending to visual stimuli versus performing visual imagery as a control strategy for EEG-based brain-computer interfaces. *Scientific Reports*, 8(1):13222, 2018. doi: 10.1038/s41598-018-31472-9.

Sunghan Lee, Sehyeon Jang, and Sung Chan Jun. Exploring the ability to classify visual perception and visual imagery EEG data: Toward an intuitive BCI system. *Electronics*, 11(17):2706, 2022. doi: 10.3390/electronics11172706.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion, 2024.

Jianghao Liu, Minye Zhan, Dounia Hajhajate, Alfredo Spagna, Stanislas Dehaene, Laurent Cohen, and Paolo Bartolomeo. Visual mental imagery in typical imagers and in aphantasia: A millimeter-scale 7-T fMRI study. *Cortex*, 185:113–132, 2025a. doi: 10.1016/j.cortex.2025.01.013.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for EEG decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17580–17588, 2026. doi: 10.1609/aaai.v40i21.38813.

Yulong Liu, Yongqiang Ma, Guibo Zhu, Haodong Jing, and Nanning Zheng. See through their minds: Learning transferable brain decoding models from cross-subject fMRI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5730–5738, 2025b. doi: 10.1609/aaai.v39i6.32611.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations, 2022.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013, 2021. doi: 10.1016/j.isci.2021.103013.

Furkan Ozcelik and Rufin VanRullen. Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion, 2023.

Victoria Peterson, Nicolas Nieto, Dominik Wyser, Olivier Lambercy, Roger Gassert, Diego H. Milone, and Ruben D. Spies. Transfer learning based on optimal transport for motor imagery brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(2):807–817, 2022. doi: 10.1109/TBME.2021.3105912.

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. doi: 10.1016/S1388-2457(99)00141-8.

Vinay Prasad and Anupam B. Jena. Prespecified falsification end points: Can they validate true observational associations? *JAMA*, 309(3):241–242, 2013. doi: 10.1001/jama.2012.96867.

Daria Proklova, Daniel Kaiser, and Marius V. Peelen. MEG sensor patterns reflect perceptual but not categorical similarity of animate and inanimate objects. *NeuroImage*, 193:167–177, 2019. doi: 10.1016/j.neuroimage.2019.03.028.

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, 2021.

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?, 2018. bioRxiv.

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, 2023.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21, 2019. doi: 10.3389/fncom.2019.00021.

Hirokatsu Shimizu and Ramesh Srinivasan. Improving classification and reconstruction of imagined images from EEG signals. *PLOS ONE*, 17(9):e0274847, 2022. doi: 10.1371/journal.pone.0274847.

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C. Aoki, Kei Majima, Yusuke Muraki, and Yukiyasu Kamitani. Spurious reconstruction from brain activity: The thin line between reconstruction, classification, and hallucination. *Journal of Vision*, 24(10):321, 2024. doi: 10.1167/jov.24.10.321.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition, 2024. ICLR 2024.

Alfredo Spagna, Dounia Hajhajate, Jianghao Liu, and Paolo Bartolomeo. Visual mental imagery engages the left fusiform gyrus, but not the early visual cortex: A meta-analysis of neuroimaging evidence. *Neuroscience & Biobehavioral Reviews*, 122:201–217, 2021. doi: 10.1016/j.neubiorev.2020.12.029.

Alfredo Spagna, Zoe Heidenry, Michelle Miselevich, Chloe Lambert, Benjamin E. Eisenstadt, Laura Tremblay, Zixin Liu, Jianghao Liu, and Paolo Bartolomeo. Visual mental imagery: Evidence for a heterarchical neural architecture. *Physics of Life Reviews*, 48:113–131, 2024. doi: 10.1016/j.plrev.2023.12.012.

Rico Stecher and Daniel Kaiser. Representations of imaginary scenes and their properties in cortical alpha activity. *Scientific Reports*, 14:12796, 2024. doi: 10.1038/s41598-024-63320-4.

Rico Stecher, Radoslaw Martin Cichy, and Daniel Kaiser. Decoding the rhythmic representation and communication of visual contents. *Trends in Neurosciences*, 48(3):178–188, 2025. doi: 10.1016/j.tins.2024.12.005.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity, 2023. CVPR 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024.

Jinze Tong and Wanzhong Chen. CNDNet: A cross-state neural decoding network for unified semantic decoding from perceptual and imagery EEG. *IEEE Transactions on Human-Machine Systems*, 56(4):887–896, 2026. doi: 10.1109/THMS.2026.3696743.

Joram van Driel, Christian N. L. Olivers, and Johannes J. Fahrenfort. High-pass filtering artifacts in multivariate classification of neural time series data. *Journal of Neuroscience Methods*, 352: 109080, 2021. doi: 10.1016/j.jneumeth.2021.109080.

V. S. Wadia, C. M. Reed, J. M. Chung, L. M. Bateman, A. N. Mamelak, U. Rutishauser, and D. Y. Tsao. A shared code for perceiving and imagining objects in human ventral temporal cortex. *Science*, 392:207–215, 2026. doi: 10.1126/science.adt8343.

Holly Wilson, Mohammad Golbabaee, Michael J. Proulx, Stephen Charles, and Eamonn O'Neill. EEG-based BCI dataset of semantic concepts for imagination and perception tasks. *Scientific Data*, 10(1):386, 2023. doi: 10.1038/s41597-023-02287-9. ; OpenNeuro ds004306.

Siying Xie, Daniel Kaiser, and Radoslaw M. Cichy. Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology*, 30(13):2621–2627.e5, 2020. doi: 10.1016/j.cub.2020.04.074.

Jonathan Xu, Ugo Bruzadin Nunes, Wangshu Jiang, Samuel Ryther, Jordan Pringle, Paul S. Scotti, Arnaud Delorme, and Reese Kneeland. Alljoined-1.6m: A million-trial EEG-image dataset for evaluating affordable brain-computer interfaces, 2025.

Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang Li. D^2-FOSA: Dual-diffusion guided EEG-to-image reconstruction with frequency-oriented semantic alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26698–26710, 2026.

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A Riemannian geometry framework with applications to brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018. doi: 10.1109/TBME.2017.2742541.

Tianyu Zhang, Fan Wan, Kaili Sun, Xingyu Miao, Yueming Sun, Minye Shao, and Yang Long. Dynamic convolution and graph-coupled attention for cross-subject EEG-vision decoding. In *British Machine Vision Conference (BMVC)*, 2025. ECHO-Net; paper 509.

A Specification of the EEG Transfer Estimator

This appendix specifies the estimator of Section 4.1 in the notation of that section, with every constant stated, since the two recording montages it connects differ.

Epoch set. The release yields 42 session files over 22 participants, two of whom contributed a single session. Four defects were identified and corrected before any transfer number was computed: two pairs of session files were byte-identical, one participant's OVI session carried a collapsed trigger code that merged two classes, and one session file's label table disagreed with the class table declared for its task. The duplicates are dropped, the merged trials are removed, and the disagreeing session is excluded. Table 1, the classifier comparison of Appendix D and the controls of Figure 1 are computed on the corrected set.

Three readouts are computed on earlier cuts and are reported as descriptive. The sliding time course of Figure 2 and the pair decomposition use a full-trial re-epoching that still contained two of the duplicate files; the shared-order control of Appendix C uses the original per-task cut, which also retained the collapsed trigger and the excluded session. Dropping the three affected participants as targets leaves both conclusions unchanged or slightly stronger: the order control moves from $+4.5$ to $+5.1$ pp and the pair profile from $r = 0.88$ to $r = 0.98$. The time course is not recomputed, since its inferential content lies in the guarded delay and imagery readouts of Section 4.1, which are computed on the corrected set. The $k = 1$ reading of Section 4.1 ($+1.9$ pp on OVI) is a single-seed value on the pre-exclusion cut and is quoted as a bound rather than as an estimate on the scale of Table 1.

Code and data. The dataset is public (Gao et al., 2026), NEMAR nm000242. Analysis code, the derived epoch set and the run record are released at an anonymised link with the submission. Software: MNE-Python 1.12, scikit-learn 1.9, NumPy 2.x.

Preprocessing. Notch at 50 Hz, zero-phase FIR band-pass 1 to 40 Hz (MNE defaults, kernel 3.30 s), resample to 250 Hz. No baseline correction, no re-referencing, no artefact rejection, no channel interpolation.

Electrodes. Three electrode sets are used (Table 3): the ten posterior channels are the decoder's features, the eight frontal channels enter only the posteriority index, and the ten anterior channels, the eight frontal plus FCz and FC3, enter only the band and region comparison of Figure 1b and the anterior window control. All ten posterior channels are present in all 42 sessions, so the feature vector is ten-dimensional throughout.

Table 3: The three electrode sets. All ten posterior channels are present in all 42 sessions.

Set	Used in	Channels
Posterior (10)	decoder features, Eq. (1)	Pz, P3, P4, P7, P8, PO3, PO4, Oz, O1, O2
Frontal (8)	posteriority index only	Fpz, Fp1, Fp2, Fz, F3, F4, F7, F8
Anterior (10)	band and region comparison and anterior window control only	the eight frontal plus FCz and FC3

Time-frequency decomposition. Morlet wavelets, $n_c = 7$ cycles at every bin, on a fixed grid of twenty centre frequencies evenly spaced from 5 to 31 Hz inclusive (step $26/19 \approx 1.368$ Hz), decimated to a 50 Hz output grid. Three bins fall inside the alpha band and enter the average: 9.11, 10.47 and 11.84 Hz. The band and region comparison uses the same selector, giving three theta bins and fourteen beta bins, with no bin shared between bands. The logarithm of Eq. (1) is taken after the frequency average and before the time average.

Trial timeline and windows. With $t = 0$ at the imagery cue, fixation onset is at $-g - 7.0$ s, image onset at $-g - 4.0$, image offset and mask onset at $-g$, and imagery ends at $+4.0$, where g is measured per run (mean 2.1247 s, sd 0.0031, range 2.112 to 2.131 over the 66 participant by task rows). Table 4 gives the readout windows: the source perception window is measured from image onset, the published imagery readout from the cue, and the three guarded readouts all have length $g - 1.050$ s, which is 1.075 s at the mean gap and lies between 1.062 and 1.081 s across runs. The sliding time course of Section 4.1 uses 0.60 s windows stepped by 0.30 s, 41 starts from -8.60 to $+3.40$ s. Figure 2 shows the resulting profile. The pattern appears at image onset, falls to chance while the image is on screen, returns through the delay and reaches its guarded maximum during imagery, so the readable window extends well beyond the epoch of the cue. The delay and imagery peaks coincide with the post-stimulus alpha rebound (Pfurtscheller & Lopes da Silva, 1999), where the account of Griffiths et al. (2019), on which fidelity is highest where power falls, predicts the opposite sign. Windows overlap by half and are unguarded, so each per-window value is descriptive and the inference is carried by Table 1.

Table 4: Readout windows, in seconds. Guarded windows are given with $t = 0$ at the imagery cue and g the per-run mask duration (mean 2.1247 s); all three share the length $g - 1.050$ s.

Window	Reference event	Interval
Source perception	image onset	$[0.200,\ 0.660]$
Published imagery readout	imagery cue	$[0.600,\ 2.280]$
Guarded imagery	imagery cue	$[0.600,\ g - 0.450]$
Guarded delay	imagery cue	$[-g + 0.600,\ -0.450]$
Guarded fixation	imagery cue	$[-g - 6.400,\ -7.450]$

Pseudo-trials and resampling. $k = 40$ on the source side and $k = 10$ on the target side, cut as disjoint blocks from a within-class permutation, the remainder of each class discarded on every draw. $R = 40$ observed and 40 null draws, except the time course (12 observed and 12 null draws per window) and the classifier comparison (10 and 10). Every pooled-source condition runs at

$k = 40/10$; the own-perception row of Table 1 is the one exception, at $k = 10/10$. Each condition's null is built at its own k, so conditions are compared only at matched k.

Readout. One linear SVM at $C = 1$ per unordered class pair, averaged with equal weight over pairs. The source pseudo-trial matrix is standardised a second time inside the fit and the same moments are applied to the target; training and test sets are truncated to equal class counts; and a pair is skipped when a class runs short of pseudo-trials, the accuracy being the mean over surviving pairs.

Standardisation. Column-wise, per participant and per modality (the standardisation of Eq. (1)), with the two sessions concatenated before the moments are taken, population standard deviation. Source participants are standardised each to their own moments before pooling. The moments are estimated on the target participant's unlabelled epochs, including those later scored.

Null. One fresh uniform permutation of the pooled source labels per null draw; the stored value is the mean over the 40 draws, so no permutation distribution is retained and the reported p values are parametric across participants rather than permutation p values.

Guard band. With $n_c = 7$ cycles and a lowest alpha bin at $f = 9.11$ Hz, the Morlet envelope of Eq. (1) has $\sigma_t = n_c/(2\pi f) = 0.122$ s and a conservative half-width of $n_c/2f = 0.384$ s. A mask window ending 26 ms before the imagery cue therefore integrates activity up to $+0.36$ s after it, and the bias runs towards the hypothesis under test. The 450 ms guard band stops the window's support 66 ms short of the cue, and the imagery and fixation readouts are length-matched to the guarded mask window (1.075 s at the mean measured gap) at the same latency from their own events. The wavelet is not the only non-causal element: preprocessing applies a zero-phase FIR band-pass whose kernel is 3.30 s long, so we measured the combined leakage rather than deriving it. Passing a simulated trial with doubled post-cue alpha through the identical filter and estimator, 6.31% of the post-cue change reaches an unguarded window ending 26 ms before the cue, against 0.053% for the guarded window: the guard band removes the leak by a factor of 119, and the residual is three orders below the effects read in that window.

B Choice of the Negative Control

Grounds for the designation. The designation of AVI as the negative control follows from published work that predates this study, and not from our own data. In M/EEG, decodability of natural objects is carried by perceptual rather than categorical dissimilarity: with perceptual dissimilarity controlled, MEG carries no category information at all (Proklova et al., 2019); same-category exemplar pairs are systematically less discriminable than cross-category pairs (Cichy et al., 2014); and imagery representations are late and coarse rather than finely resolved (Dijkstra et al., 2018; Spagna et al., 2021), with what survives in EEG being posterior alpha-band content (Xie et al., 2020).

Composition of the animal set. Three animal photographs, a bird, a dog and a fish, combine the second and third of these factors: they are the only one of the three sets drawn from a single superordinate category. They are not perceptually homogeneous, and the first factor therefore does not favour the designation, which is one reason we treat the resulting null as fixed in advance rather than mechanistically explained.

Falsification endpoint. Pre-specifying the condition predicted weakest as a falsification endpoint, a practice imported from observational epidemiology (Prasad & Jena, 2013), turns that prediction into a test (Section 3.2). This is a convergent extrapolation, since no published study tests the three-way comparison directly. The prediction was made before any transfer computation, and the test of it is what happened across participants: every transfer condition this study ran leaves AVI at its null (Section 4.1, Table 1).

Interpretation of the AVI readout. We do not count our own cross-modal readout of AVI as a second confirmation of the designation. That readout is one of the quantities the designation was made alongside, so treating it as evidence for the designation would be circular; and it is not a statement about within-participant imagery decodability in this dataset, which the paper accompanying the release (Gao et al., 2026) reports as high for animals.

C CONTROLS: CORRESPONDENCE, ORDER, ARTEFACTS AND LEAKAGE

Class correspondence. A cyclic derangement of the class correspondence abolishes the transfer: the observed condition is $+5.85$ pp above the deranged one on OVI (95% CI $[+2.28, +9.43]$, $p = 0.0027$).

Shared presentation order. The transfer does not depend on the order in which classes were presented. This control is needed because the release reuses class sequences across runs: the 41 or 42 runs of a task (two sessions of 22 participants, minus the absent recordings) follow 16 to 18 distinct sequences, so trial position alone predicts class at 67% pairwise (chance 50%) under the same leave-one-subject-out scheme. Label permutation cannot detect a decoder that exploits this, since permuting labels also breaks the agreement between sequences. For each target, source participants are ranked by position-wise class agreement with the target's sequence and split at the median, and the transfer estimator of Section 4.1 is run on each half. The least-agreeing half, at chance agreement (0.243 against 0.250), gives $+4.5$ pp ($p = 0.0011$); the two halves differ by $+0.1$ pp $[-2.0, +2.1]$, and random pools of the same size read $+4.2$ pp. The same partition, run on the pre-stimulus window of FVI where the $+1.7$ pp fixation residual of the guarded readouts lives, gives a paired difference of $+0.3$ pp $[-0.8, +1.4]$: the FVI fixation readout is not carried by the shared order either. This control was run on the original per-task cut of the epoch set, which still carried the collapsed trigger code and the excluded session repaired in Appendix A; dropping the three affected participants as targets moves the least-agreeing-half effect from $+4.5$ to $+5.1$ pp.

Frequency band and scalp region. Table 5 gives the transfer for six combinations of frequency band and scalp region on all three tasks, in percentage points above each comparison's own null. Posterior alpha was fixed in advance, and anterior alpha serves as the ocular control. Posterior alpha is the only combination positive on both OVI and FVI with AVI at the null. These eighteen values enter the thirty-seven-comparison family of Section 3.2.

Table 5: Transfer by frequency band and scalp region, in percentage points above each comparison's own null. Posterior alpha was fixed in advance; anterior alpha serves as the ocular control.

	OVI	FVI	AVI (neg.)
α posterior	$+5.0$	$+3.8$	-1.8
θ posterior	$+1.3$	-0.8	$+0.3$
β posterior	-1.0	-1.6	$+1.7$
α anterior	$+0.2$	$+0.7$	$+1.5$
θ anterior	$+1.2$	$+1.4$	$+3.9$
β anterior	-0.7	$+3.5$	-0.2

Artefact removal. The topographic ocular control is that the transfer is posterior and anterior alpha sits at chance (Table 5). The complementary control removes ocular components explicitly. An independent component analysis was fitted per recording on its perception and imagery epochs jointly, with surrogate ocular references built from the montage (vertical, the mean of Fp1, Fp2 and Fpz; horizontal, F7 minus F8). Components whose mean absolute correlation with either reference exceeded a threshold were zeroed and the epochs rebuilt with rank preserved. Two analyses used thresholds of 0.60 (A, 1.9 of 20 components removed per recording on average) and 0.50 (B, 2.2 of 20, Table 6); the frontal collapse and the posterior stability hold at both. The procedure uses no labels but is fitted on epochs later scored, as is the standardisation of Eq. (1).

Table 6: Analysis B, raw against ocular-corrected epochs: perception-to-imagery transfer of the log-alpha descriptor, in percentage points above each comparison's own null (Section 3.2).

	OVI	FVI	AVI (neg.)
log-alpha descriptor, raw	$+5.53$	$+3.66$	-1.95
log-alpha descriptor, ocular components removed	$+5.09$	$+3.67$	-1.54

Localisation of the ocular contribution. On the eight frontal electrodes, perception decoding of the negative control falls from $+30.4$ to -1.2 pp under the removal (paired -31.6 $[-36.1, -27.0]$, below zero in 22 of 22), and FVI frontal perception from $+9.9$ to -1.1. On the ten posterior electrodes the decoder uses, the same removal changes the negative control by -0.1 $[-1.4, +1.1]$

($p = 0.85$), OVI by -0.1 ($p = 0.83$) and FVI by $+0.4$ ($p = 0.19$), and analysis A agrees on the negative control ($+13.5$ to $+12.8$, $p = 0.38$). Reconstruction alone accounts for none of the frontal collapse: rebuilding from the same decomposition with no component removed leaves frontal perception where it was ($+0.3$, $p = 0.20$; -0.3, $p = 0.43$ on FVI). The condition the negative-control null is read against therefore survives the correction, and the frontal appearance of group activation-pattern maps refit on all 32 channels reflects channels the ten-electrode decoder never sees. This bounds the ocular account by surrogate reference; the release records no eye movements. The corrected transfer stands on its own, at $+5.09$ pp on OVI ($p = 0.0003$, 20 of 22) and $+3.67$ on FVI ($p = 0.023$, 15 of 22), with the negative control negative throughout. The change the correction makes is bounded: -0.43 $[-1.09, +0.22]$ on OVI and $+0.02$ $[-0.62, +0.66]$ on FVI. The interval bounds any surrogate-detectable ocular contribution at about a fifth of the effect (-1.09 pp of $+5.5$) while the point estimate leaves it at full size; the topographic control and the component removal agree.

Phase-locked subtraction. Analysis B subtracted each participant's phase-locked response, the label-free grand mean over that participant's trials in a modality, and re-ran the transfer on the residual. On OVI the residual still transfers at $+3.38$ pp ($p = 0.007$, 16 of 22), and the two conditions correlate at $r = +0.95$ across participants. The subtraction removes -4.30 pp of within-modality perception transfer ($p < 0.001$) while the loss to the cross-modality transfer is bounded at -0.06 $[-0.92, +0.81]$. It rests on a single seed, one task and one band.

Comparison family and choice of test. The multiplicity correction of Section 3.2 counts every primary comparison run: the eighteen transfer comparisons (six readouts by three tasks), three of which are in Table 1, the eighteen band and region values (Table 5), and the shared-order primary contrast above, thirty-seven in all. The time course, the topography and the scaling with the number of source participants are descriptive readouts of comparisons already counted, and carry no separate claim. The scaling in particular is predicted by any component shared across source participants, including a stimulus-driven one under this fixed-image design, so it is consistent with a shared pattern and is not evidence against that confound. At the Bonferroni bar of $0.05/37 = 0.0014$, the OVI transfer passes with the parametric test ($p = 0.0002$) and the conclusion is unchanged under a Wilcoxon signed-rank test ($p = 0.0003$) and a participant-level sign-flip permutation ($p = 0.0001$); a sign test (18 of 22 positive, two-sided $p = 0.004$) does not pass the corrected threshold, as expected when effect sizes are discarded at $n = 22$.

Leakage control. The standardisation of Eq. (1) uses the mean and variance of the target participant's imagery epochs, including the epochs that are then classified. No label is used; the question is whether computing these moments on the scored epochs inflates the transfer. Each participant's imagery trials were split into two class-balanced halves, one scored and one never scored, and the transfer was measured under four standardisations, each against its own permutation null (Table 7). Moments from the never-scored half (C) give the same transfer as moments from the scored half itself (B): the difference, $+0.4$ pp ($p = 0.19$), is below the smallest difference detectable at this n (0.7 pp). Moments from the participant's perception (D) lose 2.9 pp against C ($p = 0.0098$). The standardisation must therefore use imagery epochs, but not the scored ones. AVI stays at its null in every condition.

Table 7: Leakage control on OVI: transfer under four standardisations, in percentage points above each condition's own permutation null.

	moments estimated on	OVI	p
A	all imagery trials (Table 1)	$+5.2$	0.0009
B	the scored half	$+5.5$	0.0012
C	the never-scored half	$+5.1$	0.0004
D	the participant's perception	$+2.2$	0.010

Causal evaluation. In real time, the moments can only come from trials already recorded. Single trials were scored in chronological order, with the moments estimated from the participant's preceding imagery trials only, after a 20-trial burn-in that is never scored. On OVI the transfer is $+1.90$ pp ($p = 0.0004$, 19 of 22), against $+1.88$ pp when the moments use all imagery trials: the difference, 0.02 pp, is below the 0.26 pp detectable. Moments from the participant's preceding perception epochs alone, with no imagery epoch, keep $+1.57$ pp ($p = 0.0002$), 0.33 pp below imagery moments

Table 8: Alternative readouts on the frozen features. Percentage points above each condition's own source-label-permutation null, mean ± between-participant SE, $n = 22$, at 10 observed and 10 null draws. All rows are pairwise (chance 0.5; six class pairs on OVI, three on FVI and AVI) except the K-way logistic row, which is a single K-way classifier read against its own $1/K$-chance null. Lots A and B share their draws and their reference; lot C is a separate stream with its own reference, and conditions are only ever compared within a lot.

	OVI	FVI	AVI (neg.)
Lots A and B, one stream of draws			
linear SVM, inline reference	$+5.6 \pm 1.3$	$+3.9 \pm 1.9$	-1.8 ± 1.7
TabPFN, in-context	$+5.6 \pm 1.4$	$+3.5 \pm 1.8$	-1.0 ± 1.8
TabFM, in-context	$+5.5 \pm 1.4$	$+3.3 \pm 2.1$	-2.4 ± 1.7
K-way logistic	$+5.2 \pm 1.4$	$+2.5 \pm 2.1$	-1.5 ± 1.9
Lot C, a separate stream			
linear SVM, inline reference	$+4.8 \pm 1.3$	$+3.9 \pm 1.8$	-2.1 ± 1.6
Gaussian process	$+5.2 \pm 1.3$	$+4.6 \pm 2.0$	-1.6 ± 1.5
nearest centroid	$+4.4 \pm 1.2$	$+4.5 \pm 2.1$	-1.2 ± 1.6

($p = 0.036$). FVI shows the same ordering without reaching significance ($+1.27$ pp, $p = 0.050$), and AVI stays at or below its null.

D Alternative Classifiers and Alignment Methods

D.1 Classifier Capacity

The readout of Section 4.1 is a linear SVM, and both that choice and the pairwise decomposition were tested. A comparison of alternative classifiers was frozen before any real-data run, on the same frozen features and windows as Table 1 but with its own pseudo-trial draws and permutation nulls (10 observed and 10 null resamples, against 40 and 40 for the main estimator). It was collected in three lots (Table 8); lots A and B share one stream of draws and one inline linear SVM reference, and lot C carries its own. Every comparison is paired within a lot, because the same SVM re-run on fresh draws moves by 0.8 pp. The reading was fixed in advance: a condition counts as an improvement only above $+2$ pp paired against its matched SVM with the negative control still at chance. The two multi-condition lists were read under Bonferroni correction ($0.05/6$ and $0.05/3$), and the K-way row is descriptive.

No alternative differs from its matched SVM by more than 1.5 pp on any task, and the largest difference favouring an alternative is $+0.8$ pp, below the $+2$ pp threshold fixed in advance. Two tabular foundation models used as in-context classifiers (Hollmann et al., 2025; Google Research, 2026) reproduce the linear readout to within a fraction of a point (paired $+0.02$ and -0.04 on OVI, -0.45 and -0.61 on FVI), and a Gaussian process (Rasmussen and Williams, 2006) and a nearest-centroid prototype sit within 0.7 pp of their group's SVM on both imagery tasks. A single multinomial regression over all K classes transfers on OVI as well as the pairwise readout (paired -0.4, $p = 0.32$), so nothing hinges on the pairwise decomposition there; on FVI the two available draw sets disagree in sign (-1.4 and $+0.6$) against a $1/K$ chance level, so we do not read that cell. A second frozen lot consuming more structure than the frozen features (time-resolved alpha trajectories, learned spatial filters, Riemannian tangent-space classifiers; Barachant et al., 2012) tied at best and lost by up to 6.9 pp, with the negative control at chance throughout; one losing condition, regularised CSP with shrinkage LDA, reads nominally above chance on the fixation gate ($+1.5$ and $+1.8$), which we record for completeness; no claim rests on it.

Pooling and a contrastive encoder. Pooling twenty-one source participants adds no resolvable accuracy over the participant's own perception ($+0.4$ $[-1.7, +2.5]$ on OVI); with a single source participant the transfer is halved (-2.4 $[-4.6, -0.1]$, $p = 0.04$), and pooling restores it. Pooling therefore replaces the participant's own perception data at equal accuracy. A contrastive EEG-image encoder, the family shared by the cross-subject frameworks of Section 2 and by the encoder of Section 4.4, was run as a further baseline under the same protocol : the official encoder of Liu et al. (2026), unmodified, trained on the pooled perception EEG of the source participants against

Table 9: Recentring with one architecture and one readout held fixed, on all 32 channels. OVI transfer in percentage points above each comparison's own null, mean over four seeds, each a full retrain, with the between-seed SD and the parameter count.

	OVI transfer	seed SD	parameters
one network, no recentring	$+1.28$	0.42	73,348
same network, learned recentring (Kobler et al., 2022)	$+2.88$	0.84	73,348
closed-form recentring, no network	$+4.08$	0.41	0

the vision-model embeddings of the stimulus images, with leave-one-participant-out splitting, per-participant standardisation of the raw EEG, epoch selection on two source participants' perception, its own source-label-permutation null, and the held-out participant's imagery scored without any imagery data in the fit (one seed, recipe fixed before the run). The encoder decodes the held-out participant's perception ($+19.8$ pp over its null on OVI, 22 of 22; $+11.6$ on FVI; $+31.7$ on AVI) and does not transfer to imagery (-3.2 pp on OVI, $p = 0.18$; $+2.8$ on FVI, $p = 0.14$; -0.1 on AVI); per participant it is 8.0 pp below the alpha readout on OVI at $k = 10$ ($p = 0.0006$, 4 of 22) and 3.0 pp below at $k = 1$ ($p = 0.0017$).

D.2 Alignment

The standardisation of Eq. (1) is a Euclidean approximation of Riemannian recentring (Zanini et al., 2018); this comparison measures what the exact operation adds. It varies only the recentring, with one architecture and one readout fixed, on all 32 channels, so its values are not comparable with the ten-channel Table 1 (Table 9).

The first two rows differ only in whether the representations are recentred: with the participant as unit and each participant's value averaged over seeds, recentring contributes $+1.61$ pp (paired, 95% CI $[+0.15, +3.07]$, $p = 0.033$, better in 14 of 22); the per-seed gains are $+2.95$, $+0.33$, $+1.11$ and $+2.04$, so one seed in four resolves it alone. The closed-form row has no network and is not paired with the others. It is ahead of the learned recentring on every seed but not significantly ($+1.19$ pp $[-0.78, +3.17]$, $p = 0.22$), so learning the operation does no better than computing it in closed form. The null of the learned conditions retrains the network under permuted labels, whereas the closed-form null permutes labels on fixed features, so the two nulls differ. A single-seed run of the same contrast in a separate pipeline gave $+3.63$ pp; the four-seed value above replaces it.

Affine invariance. Affine invariance, usually cited to justify the affine-invariant metric (Congedo et al., 2017) because it makes distances immune to the invertible linear mixing that volume conduction imposes, does not explain the gain: the log-Euclidean metric, which lacks it, ties the affine-invariant metric ($+0.40$ pp, 95% CI $[-0.71, +1.51]$, $p = 0.46$, $r = +0.86$ across participants). With recentring, the 32-channel tangent-space conditions tie the linear readout (-0.42 pp $[-3.80, +2.96]$, $p = 0.80$), while the ten-feature tangent-space conditions of the classifier comparison lost by up to 7 pp; on both montages, a deeper classifier does not improve the transfer.

D.3 Architecture Search: Space, Selection and Results

The search scored various configurations, each on all 22 participants . It re-epoched the raw recordings of Gao et al. (2026) (32 channels, 250 Hz) and rebuilt the validation script around the reference readout. For each held-out participant, the 21 source participants' perception epochs (0.200 to 0.660 s after image onset) are pooled; features are averaged over pseudo-trials of 40 source and 10 target trials; a pairwise linear SVM ($C = 1$) is fitted per class pair and scored on the held-out participant's imagery epochs (0.600 to 2.280 s after the cue); the condition's own null repeats the identical pipeline with the source labels permuted. A condition's score is the mean over the 22 participants of accuracy minus its own null, in percentage points, over 20 observed and 20 null pseudo-trial draws (the original reference script used 40 and 40; the record's reference rows were re-run in this analysis's validation script at 20 and 20), read with a one-sided t test across participants. Learned encoders replace the feature block only: the network is trained on the pooled source perception, its time-averaged convolutional output is standardised per participant, and the same pseudo-trial SVM and the same null are applied to the embedding. The reference in every comparison is the ten-channel alpha-power descriptor on the same ICA-cleaned epochs, $+5.09$ pp on OVI, $+3.67$ on

FVI and -1.54 on AVI, and each condition is compared with its task's reference by a paired test across the 22 participants; the search record flags a configuration as beating the reference at paired $p < 0.05$.

Configuration selection differs by family. The convolutional conditions were tuned by a Bayesian search with hyperband early termination, eight trials per agent, on four development targets (participants 01, 05, 11 and 17) maximising the same bridge score; the top one to three configurations were then run on the full 22-participant leave-one-out, where the 18 remaining participants had never been used for tuning. Inside every training run, early stopping reads four held-out source participants and never the target. Every other family is one configuration per condition at a single seed, with the closed-form tangent-space condition replicated at three further seeds. Table 10 lists the families. Twenty-five rows are exact re-runs of one configuration launched inside different arrays as built-in validity checks (ten groups; the alpha descriptor on the clean epochs, for instance, appears four times at $+5.53$ pp), so 190 distinct configurations were scored. Sixteen rows use labelled trials of the target participant (a perception calibration block, or 5 to 20 labelled imagery trials per class) and are not held-out-participant conditions. Eleven rows carry the flag: one few-shot condition on OVI ($+2.05$ pp over the reference, $p = 0.014$, with 20 labelled imagery trials per class) and ten conditions on AVI, where the reference sits at its null. No held-out-participant condition beat the reference on OVI or FVI. Paired against the readout, the largest held-out-participant gains on OVI and FVI ($+1.1$ and $+1.6$ pp, over 156 and 14 comparisons) lie below their detectable floors of 6.1 and 5.6 pp; on the negative control, nine of 29 conditions, eight of them on one tangent-space feature pipeline, exceed the readout at uncorrected $p < 0.05$, none after Holm correction over the 199 comparisons, and each difference lies below its own floor.

Table 10: The scored configurations of the architecture search, by family (190 distinct; 25 are validity re-runs). Every configuration is scored by the protocol of Section 3.2 and paired against the alpha-power readout on the same epochs.

family	n	what varies
Riemannian tangent space, closed-form recentring	65	filter bank, channel set, metric, PCA, seed
Convolutional networks (EEGNet, ShallowConvNet, analytic)	30	sweep-selected configurations, channels, band, window
Training-set augmentation (imagery, calibration, few-shot)	24	source of added trials, weight, shots
Fixed filter-bank front ends (Morlet, FBCNet)	23	spatial width, trained or frozen, channels
Classifier, imagery window, pseudo-trial size	21	LDA, logistic, SVM C; window; k
Cross-modal contrastive encoder	9	alignment weight, subject-adversarial weight
Feature-level fusion	8	feature blocks combined
Riemannian Procrustes across modalities	7	rotation, stretch, bank
Multi-source weighting	7	aggregation rule, vividness weighting
Reference re-runs	6	clean or ocular-corrected epochs
SPD networks with learned recentring	5	domain normalisation, dimensions, embedding

E Reconstruction: Controls, Reference Comparison and Real-Time Cost

This appendix reports the controls, the rank of the brain-to-image map, the comparison with the reference method and the real-time cost that support Section 4.4. All decoder generation and panel evaluation ran on single NVIDIA L40S GPUs (one job per subject per configuration, a few minutes each); the ridge fits, rank sweeps, statistics and figures are CPU-only. The exploratory compute exceeds what is reported here, as is usual for an empirical study with many pre-registered branches that returned null results.

Floor. Giving the decoder a prefix computed from a different image, keeping the correct start token, collapses performance at every K; the gap to the true prefix grows from $+0.094$ to $+0.201$

AlexNet(2) as K increases, on 10 of 10 subjects. The prefix carries image-specific information, and more of it at larger K.

Ceiling. Giving the decoder the true token prefix reaches 7 of 7 columns against the target (Table 2). The frozen decoder is therefore not the limiting component; the bridge is.

Trivial nulls. Downsampling the target to $d \times d$ and re-enlarging it traces exactly the shape of the K ladder: high PixCorr, SSIM and AlexNet, collapsing Inception and CLIP-L/14, the signature that Ozcelik & VanRullen (2023) report for a low-resolution VDVAE stage that beats their full pipeline on precisely these columns. At matched AlexNet(2), our configurations carry $+0.02$ to $+0.18$ more CLIP-L/14 than this null, which knows the image and we do not, though that margin collapses to $+0.023$ at $K = 10$ (against $+0.152$ at $K = 6$, $q = 0.5$), so the configuration most like a low-pass filter is exactly the one that partial forcing dominates. At its degenerate limit $d = 1$, a single uniform colour, the same family scores PixCorr 0.3079 and SSIM 0.5846, above the published 0.1718 and 0.4233, and above our own configuration (Table 2). PixCorr and SSIM cannot support reconstruction claims on this benchmark; this applies to our numbers first. This is the metric-level counterpart of a known failure mode: reconstruction scores that a degenerate output can earn measure the evaluation, not the decoding (Shirakawa et al., 2024).

Which positions are forced. The number of forced positions is fixed per scale and not drawn per image. At scale k with $N_k = p_k^2$ positions, the $\lfloor q\, N_k \rceil$ positions of highest fit-set correlation are forced, at least one per scale, and the same positions are used for every image of a participant. At the setting of Table 2 ($K = 6$, $q = 0.5$) that is 1, 2, 4, 8, 12 and 18 positions, 45 of 91 in total. The rounding is to nearest rather than down, which coincides with $\lfloor q\, N_k \rfloor$ at $q = 0.5$ but not in general: at $q = 0.75$ the two rules give 1, 3, 7, 12, 19, 27 and 1, 3, 6, 12, 18, 27. The random control forces the same number of positions at each scale, drawn without replacement from a seed that depends on the scale and on q but not on the participant, so its draw is shared across the ten participants rather than repeated independently.

Position-selection controls. A static per-scale confidence rule improves over random position selection by less than the declared threshold, and a per-image token classifier performs worse: within the rules tested, the amount of forcing is more consequential than the particular positions selected. Choosing the forced positions by ridge confidence rather than at random changed AlexNet(2) by $+0.007$ and PixCorr by $+0.01$ (9 to 10 of 10 subjects) and nothing else; the rate q, not the position rule, carries the effect.

Control for the mechanism. An unforced position is generated by the autoregressive prior, whereas a forced position replaces the prior with a low-SNR EEG estimate. To test whether the gain of partial forcing comes from leaving positions to the prior rather than from transmitting less information, the predicted continuous band map was shrunk toward the training-set mean before quantisation: the map is centred on the mean target map of the training images, decomposed into its ten telescoping pyramid bands, each band multiplied by a single scalar, the training-set correlation between prediction and truth at that scale (0.54 at the coarsest scale, 0.10 to 0.19 at the next five), the mean re-added and the shrunk map passed through the frozen residual quantiser. With all scalars at one the bank reproduces the original prefix to 6×10^{-7}. Under full forcing at $K = 3$, 6 and 10 (greedy, guidance 1.0, validation split), the shrunk prefix degrades AlexNet(2) by 0.071, 0.106 and 0.192, with 0 of 10 subjects improved: sending a smoothed, lower-information prefix hurts, whereas leaving half of the positions to the prior helps.

Encoder retraining at matched retrieval. The ceiling above leaves the encoder as the remaining candidate limit, and the natural objection to a ridge readout is that an encoder trained against the token target would deliver more. We trained one: a $1024 \rightarrow 448$ linear head on the VQ band target (scales 0 to 2), fitted jointly with the existing contrastive loss so that retrieval is unchanged by construction; the loss weight is swept and selected on the same 400-image validation split. At exactly matched retrieval (Δ 200-way top-1 $= 0.0000$) it raises the 2×2 band readout by $+0.0069$ (10 of 10 subjects, $p = 0.0020$) and the 1×1 scale the forcing consumes by $+0.0016$ (7 of 10, $p = 0.084$); both remain below a closed-form ridge fit on the untouched frozen encoder at every loss weight tried. Within the retrieval constraint, retraining the encoder yields no headroom over the ridge.

Snapshot selection for the frozen decoder. The slot fine-tune that teaches a VAR decoder (Tian et al., 2024) to accept a ViT-H embedding (Radford et al., 2021) in place of a class label sees

only images, never EEG, and its training loss decreases monotonically for the full 15,120 steps (to 0.69 nat at depth 30). Validation cross-entropy has no interior minimum either: no training scalar tracks downstream quality. We therefore select the snapshot on the validation panel in the deployed configuration, and it selects early, at 5,040 steps for depth 16 and 3,024 for depth 30, out of 15,120. On one subject this is a plateau rather than a peak; the intervention is resolved only on the paired 10-subject contrast, where running depth 30 to convergence instead of stopping at 3,024 costs -0.013 ± 0.004 CLIP-L/14 in the deployed configuration (1 of 10 subjects improve). The sign reverses when the same decoder is given the true image embedding, which gains $+0.012 \pm 0.004$ Inception. A decoder trained to convergence reads its conditioning vector more sharply, which helps a clean vector and hurts a noisy one; stopping early leaves it smooth enough to tolerate the EEG estimate. This is also what makes classifier-free guidance usable here: at 15,120 steps guidance 2.5 costs -0.026 CLIP-L/14, at 5,040 it costs -0.004, and on the early-stopped depth-30 decoder the penalty is gone on 10 subjects (-0.0001) while Inception gains $+0.014 \pm 0.004$ (8 of 10). Early stopping, not depth, is what removes the guidance penalty. Replacing the depth-16 decoder with the depth-30 decoder at the same (K, q) and the same bridges raises Inception from 0.7701 to 0.7844 and CLIP-L/14 from 0.7547 to 0.8013; the gain is confined to the semantic columns, and the (K, q) optimum is unchanged between depth 12 and depth 30. Snapshot selection and guidance scale were both chosen on validation, and the test panel was evaluated once per reported condition.

Rank of the brain-to-image map. Re-sweeping the ridge penalty at every rank gives held-out band-0 correlation 0.3946 at rank 4, 0.4880 at rank 16, 0.4996 at rank 32, 0.5026 at rank 64, and 0.5060 at full rank (2.98 M); naive truncation of the solution fitted at the full-rank-optimal penalty understates achievable accuracy by roughly an order of magnitude, because a lower-rank map needs more regularisation, not less. Semantic cosine follows the same shape, 0.6163 at rank 8 to 0.6313 at full rank. The deployed band-matching step sums only the six coarsest pyramid bands before the prefix reaches the quantiser (`bandmatch.patch_nums = [1..6]`), so the structural map's effective output space is $32 \times 80 = 2560$ dimensions, not 8192, a second, independent $3.2\times$ reduction with no accuracy cost, verified by an exact-reproduction gate against the deployed bank (maximum deviation 6×10^{-6}, structure; 2.9×10^{-8}, semantic; 10 of 10 subjects). At the regenerated rank-32 operating point, Inception degrades monotonically with rank (-0.022 at rank 32) while PixCorr, SSIM and CLIP-L/14 do not, the reverse of the pre-registered prediction that degradation would fall on the low-level columns. Per image, the full 8192-wide bridge takes 56.7 μs and the truncated 2560-wide bridge 17.0 μs; the frozen decoder is identical across all four bridge sizes and dominates by three orders of magnitude at both depths, so shrinking the bridge is a storage claim, not a speed claim.

Re-run of the reference method. ENIGMA (Kneeland et al., 2026a) appears in the literature with two different identification rows, and a re-run of its released pipeline reproduces both, from the same weights, under two training protocols. Training on the four averaged repetitions, which is the protocol its paper states, gives 0.170, 0.424, 0.834, 0.893, 0.775, 0.817 and 0.545 against a published 0.172, 0.423, 0.836, 0.895, 0.777, 0.815 and 0.540. Training on single trials, which is what its released code does, gives 0.156, 0.429, 0.819, 0.882, 0.750, 0.791 and 0.547, the row quoted by Xu et al. (2025). Each arm is within 0.01 of the row it reproduces on every metric. The row of Table 2 is the first, so that the identification and distribution columns of the reference come from one configuration; the second arm scores IS 6.30, FID 179.1 and KID 0.022, so the conclusion of Table 11 does not depend on the choice.

Comparison with the reference method. The deployed configuration exceeds ENIGMA (Kneeland et al., 2026a) on five of seven metrics with VAR-d16 and six of seven with VAR-d30 (Table 2); only CLIP-L/14 remains below the reference. Flat conditioning already exceeds the reference on four and five metrics respectively, so the evidence for partial forcing is the paired within-method contrast reported above, which does not depend on an external reference. PixCorr and SSIM are exceeded by a single flat colour on this dataset. A reproducible $1/199$ off-by-one error in the released two-way identification code is corrected wherever a gap to the reference is quantified. D^2-FOSA (Yu et al., 2026) reports 200-way retrieval (38.0% within subject, 13.1% leave-one-subject-out) and an IS, FID, KID, SSIM and PCC panel on the same dataset rather than the identification panel of Table 2; computed on our reconstructions with torchvision Inception features (10 subjects, 200 test images each), those metrics read PCC 0.212 ± 0.005 against their 0.193, SSIM 0.250 ± 0.003 against 0.350, IS 6.13 ± 0.04 against 11.81, FID 195.7 ± 1.2 per subject and 172.1 pooled against 146.33, and KID 0.040 ± 0.001 against 0.025. The pooled values of the other generated rows are 165.1 for

Table 11: Distribution metrics on the same renders as Table 2, per subject: 200 renders against the 200 targets at 256 px, IS with 10 splits, KID over 100 subsets of 100, averaged over the 10 subjects. Best in bold, second underlined. ENIGMA publishes none of the three; its row is recomputed here from a re-run of its released pipeline under this protocol, on the arm that reproduces the line of Table 2. †D^2-FOSA does not state its protocol for these three metrics, and IS and FID depend strongly on the number of images, so its values are not ranked against ours.

	IS↑	FID↓	KID↓
ENIGMA, re-run	**6.75**	175.6	<u>0.019</u>
D^2-FOSA, published	11.81†	146.3†	0.025†
VAR-d16, $K = 0$ (flat)	6.06	182.1	0.023
VAR-d16, HINT	5.88	183.5	0.033
VAR-d30, $K = 0$ (flat)	<u>6.41</u>	**170.5**	**0.017**
VAR-d30, HINT	6.04	<u>173.2</u>	0.024
permuted prefix, $K = 6$	6.08	173.1	0.024
one flat colour ($d = 1$)	1.05	350.3	0.282
oracle prefix (ceiling)	9.27	103.6	0.0006

flat conditioning, 178.4 for the permuted prefix and 351.2 for the flat colour. The comparison is indicative: FID on 200 images is biased upward and their sample per FID is not stated, their Inception weights are the TensorFlow ones, their generator renders through a diffusion pipeline at a higher resolution, and SSIM is exceeded by a single flat colour on this dataset.

Distribution metrics. Table 11 reports IS, FID and KID for the renders of Table 2; these metrics do not measure reconstruction on this dataset. The permuted prefix, which reconstructs a different image than the one viewed and which the identification metrics place at chance, matches the deployed configuration on all three (IS 6.08 against 6.04, FID 173.1 against 173.2, KID 0.024 against 0.024): a metric a wrong reconstruction passes cannot separate a right one. This is the same objection the flat colour raises against PixCorr and SSIM. Read as a measure of the realism of the renders, the three columns show that forcing lowers realism: flat conditioning scores better than the deployed configuration on all three metrics at both depths, and the same holds against the reference method: paired over the ten subjects (Wilcoxon exact, Holm over twelve comparisons), the deployed configuration is below ENIGMA on all three at depth 16 (0 of 10 subjects on each, $p = 0.023$) and on IS (1 of 10, $p = 0.035$) and KID (0 of 10, $p = 0.023$) at depth 30. Its FID at depth 30 is 2.4 points better than ENIGMA's, and that difference is not resolved (7 of 10, $p = 0.26$). Forcing a brain-predicted prefix therefore buys fidelity to the viewed image at the price of some realism, while a prefix that is correct rather than predicted improves realism sharply (FID 103.6 for the oracle).

Two controls rule out the obvious confounds in the ENIGMA comparison. Resolution does not carry it: downsampling ENIGMA's renders from 512 to 256 px, the size at which every row here is scored, moves its values to 6.68, 175.5 and 0.018. Nor does the choice of render: averaging over the ten renders ENIGMA produces per image gives 6.63, 175.7 and 0.019 against 6.75, 175.6 and 0.019 for the first render. The comparison against D^2-FOSA remains indicative only, for the protocol reasons given below.

Reproduction of the generated rows. The three distribution metrics are reported for no published configuration of this decoder, so every row that carries them in Table 11 was generated and scored here under one procedure: 200 test images for each of 10 subjects, torchvision Inception-v3 features, Frechet and kernel distances against that subject's 200 target images, generation seed 1000. The scoring path reproduces exactly, since the flat-colour row, which does not sample, matches its published identification panel to four decimals on all seven metrics; the sampled rows reproduce their published panels to within 0.022 on every metric, the residual being the sampler operating point, which was probed directly (guidance 2.5 moves every identification metric away from the published values, and greedy decoding moves the brain-driven rows towards them). Two rows carry no distribution metrics. The depth-30 rows would need that decoder, which is not part of the released weights. The ceiling row is not reproduced from them: the true token prefix at the forced scales gives PixCorr 0.73 at full forcing and 0.45 at the deployed fraction against a published 0.5099, and that value is unchanged at every sampler setting tested, so the published row uses a conditioning embedding this reproduction does not have. The target image encoded and decoded by the frozen tokenizer, the bound of this decoder family, scores PixCorr 0.946, IS 11.13, FID 26.5 and KID -0.003.



Figure 4: **Structure versus semantics and realism.** Twelve held-out stimuli, chosen quantitatively. For every (subject, stimulus) of the deployed cell (VAR-d30, $K{=}6$, $q{=}0.5$, greedy, guidance 2.5) we take the mean of our reconstruction's four identification metrics and break ties by PixCorr, keeping each stimulus once at its best subject. Those four metrics saturate here, so the labels report PixCorr. Rows: seen image; ours; no prefix ($K{=}0$); the reference method, re-run here (averaged-repetition arm, first render, $512 \to 256$ px); true prefix. The columns illustrate a split the tables measure. The prefix carries layout and colour, and we lead there on PixCorr, SSIM and AlexNet (Table 2); the reference leads on Inception, CLIP and all three distribution metrics (Table 11), rendering a plausible image of a different layout. The $K{=}0$ row is already semantically right: the category comes from the conditioning embedding, not the prefix. Alignment was checked against the reference's own saved target, which matches ours to 0.24 of 255 in mean absolute difference, its JPEG residual.

Real-time cost. Both brain-to-image maps are closed-form ridge solutions with 8.39 M and 1.05 M parameters per subject. The structural map is well approximated at rank 32, which, regenerated end to end, retains five of seven metrics and the (K, q) effect ($+0.0355$ CLIP-L/14, 10 of 10 subjects) at $1/52$ of the bridge parameters. The rank-32 map factors exactly through the encoder's final linear layer, $z = Ah + a$, with numerical error below 9.5×10^{-7}; storing $A^\top U_r$ removes 98.2% of the corresponding inference-path weights, leaving 0.53 M EEG-side parameters per subject against 1.39 M for ENIGMA (Kneeland et al., 2026a), whose encoder is counted in the same way. Parameters, FLOPs and wall-clock at batch 1 and batch 25 are measured back-to-back on the same L40S with each method's released inference path (bf16 autocast, median of 30 calls; Table 12); ENIGMA reports 294 G FLOPs for its EEG model alone. The generator is $10.6\times$ smaller than the reference, needs $9.1\times$ fewer FLOPs per image, and is $2.3\times$ faster at batch 1 ($6.2\times$ at batch 25). Measured end to end on one L40S at batch 1, the chain from preprocessed EEG to image takes 68 ms per trial, of which the decoder takes 66 ms and the two ridge maps 0.04 ms, a fixed latency per trial; the speed advantage is due to the decoder. Unlike the EEG transfer, reconstruction needs per-subject training: 37 to 49 s of encoder training and 2 s of ridge solves on the subject's full training set. The one-time VAR adaptation takes approximately 3.5 GPU-hours on an L40S, against 5.5 h published for ENIGMA on 30 subjects; because adaptation and one-time training are compared with published figures on different hardware, Table 12 reports absolute times for those rows rather than ratios. Scaling from depth 16 to depth 30 spends most of the margin to gain the sixth metric.

F Single-Neuron Analyses

This appendix reports the single-neuron analyses behind Section 4.2: a reproduction of the published axis-tuning analysis, a map-complexity ladder, an encoding benchmark under the shared feature convention, a cross-modal decoding analysis with its leakage audit, and four follow-up experiments. Every number is computed by our protocol on the public release of Wadia et al. (2026); none reproduces a published value, and the proximity of our diagonal-map accuracy (29.6%) to the published cross-decoding accuracy (31.35%) is a coincidence of two different analyses. Three conventions hold throughout. The effective sample size is the number of patients (6 in the ladder; 4 in cross-

Table 12: System cost against ENIGMA, per generated image. The last five rows are measured back-to-back on one NVIDIA L40S with ENIGMA's released pipeline; the first two rows compare our measurements with ENIGMA's published figures on different hardware, so no ratio is given. Ratios in parentheses are ENIGMA over ours. Our one-time training uses no EEG. Classifier-free guidance doubles the batch at every step in our sampler, so raising the guidance scale to 2.5 changes neither FLOPs nor latency.

	ENIGMA	Ours, VAR-d16	Ours, VAR-d30
Per-subject adaptation	15 min of EEG; 10 min GPU (RTX 3090)	66,160 trials; 37 to 49 s GPU + 2 s CPU (L40S)	
One-time training	5.5 h, 30 subjects	≈3.6 GPU-h (L40S)	≈5 GPU-h (H100)
Generator parameters	4,451 M	420 M (10.6×)	2,121 M (2.1×)
FLOPs per image	8,736 G	957 G (9.1×)	4,052 G (2.2×)
Latency, batch 1	203 ms	90 ms (2.3×)	163 ms (1.25×)
Latency, batch 25	120 ms	19 ms (6.2×)	50 ms (2.4×)
Raw count vs ENIGMA		5/7	6/7

decoding and in the cross-patient experiments), never the number of stimuli or of held-out trials. The release is trial-averaged, so no noise ceiling can be computed. All 367 screening neurons were selected for axis tuning, so every statement concerns tuned neurons of ventral temporal cortex rather than the area.

Design and relation to the EEG protocol. The within-patient analyses fit the perception-to-imagery map on the paired responses of the $S-1$ training stimuli of one patient and recording block, map the held-out stimulus's perception vector, and rank the block's imagery vectors against it; the held-out stimulus's imagery never enters a fit. This leave-one-stimulus-out design within patient is not the cross-subject protocol of Section 4.1; the protocols differ, but both find that the simplest map carries the transfer. The cross-patient experiment shares the shape of the EEG protocol (source perception from other subjects, target imagery of the held-out subject) and its architecture-search control, and nothing else: trial-averaged data have no within-condition covariance, patients have no neuron correspondence, and spike counts have no alpha band.

Reproduction of the published analysis. Before any new analysis, the axis-tuning result of Wadia et al. (2026) is reproduced from the release: the preferred-axis correlation is 0.540 against the published 0.54, and the imagery value 0.203 against the published 0.20. The stimulus-to-file mapping is verified by decoding the release's five-way category labels from image features at 84.0% against a 35% permuted-alignment null. We reimplemented the published axis-tuning routine and compared it with the published values on the released data (Table 13). The screening and imagery preferred-axis correlations reproduce exactly; the paired test of preferred against orthogonal axes agrees at the smaller released n (367 rather than 456 neurons); the imagery orthogonal-axis value differs in sign ($+0.084$ against -0.08), both being correlations near zero. Three published quantities cannot be reproduced from the release, which contains only the 367 axis-tuned screening neurons (the fraction 367/456 needs the 89 untuned neurons), only the 43 reactivated imagery neurons (the published cross-decoding accuracies need all 231), and no reconstruction or decoding code. The analyses below therefore use the 43 reactivated neurons with their screening responses, with axis-tuning parity as the entry check.

Table 13: Stage 0 parity between the published axis-tuning values and our reimplementation.

Quantity	Published	Ours
Screening r_{pref} (mean)	0.54	0.540
Screening r_{ortho}	≈ 0	0.000
Paired test, pref. vs ortho.	$p = 4.28 \times 10^{-270}$ ($n = 456$)	$p = 6.7 \times 10^{-62}$ ($n = 367$ released)
Imagery r_{pref}	0.20 ($p = 0.001$)	0.203
Imagery r_{ortho}	-0.08 ($p = 0.93$)	$+0.084$

Map complexity. Imagery neurons group by recording session, each session using its own subset of 6 to 8 stimuli, so the population unit is a (patient, stimulus set) block: 13 blocks over 6 patients, averaged within patient so that the paired statistic stays at $n = 6$. At this n, 80% power and $\alpha = 0.05$, the smallest detectable paired effect is $d_z = 1.43$, stated before any result. Rungs that require within-condition trial variance (Gaussian optimal transport, the 2-Wasserstein distance between trial

clouds, every covariance rung) cannot be computed on trial-averaged data, and decoding in a 50-dimensional feature space is not identifiable from 6 to 8 stimuli. The imagery gain is measured in two ways: a through-origin gain with median 0.759 (IQR $[0.597, 1.010]$, $n = 43$ neurons) and a regression slope with intercept of median 0.097 (IQR $[0.028, 0.212]$). Imagery firing therefore sits at about 76% of the perception level with weak stimulus-by-stimulus proportionality, consistent with the published imagery $r_{\text{pref}} = 0.20$. Table 14 gives the leave-one-stimulus-out relative error of four maps per patient. The complexity-versus-fit curve saturates early: the first rung, a common scalar gain, brings the largest gain, the per-neuron diagonal adds little, and a general linear map slightly hurts held-out prediction. No step clears the $d_z = 1.43$ floor, so this is a statement of shape with its interval. Because the 43 neurons were selected for axis-proportional imagery firing, the result is an upper bound on map simplicity.

Table 14: Stage 1 ladder. Left: leave-one-stimulus-out relative error per patient (lower is better), averaged over that patient's blocks. Right: paired steps across the 6 patients (Wilcoxon; every d_z is read against the 1.43 floor).

Patient	Blocks	Identity	Scalar	Diagonal	Linear (ridge)
P76CS	2	0.700	0.559	0.529	0.626
P79CS	3	0.536	0.424	0.365	0.358
P80CS	2	0.575	0.303	0.303	0.303
P84CS	2	0.934	0.819	0.580	0.529
P85CS	1	0.803	0.724	0.724	0.724
P92CS	3	0.500	0.570	0.586	0.652

Step	Δ error	d_z	p
identity $\rightarrow$ scalar gain	-0.108	0.99	0.062
scalar $\rightarrow$ per-neuron diagonal	-0.052	0.54	0.144
diagonal $\rightarrow$ general linear	$+0.017$	-0.32	1.000

Depth gradient and layer ordering. Table 15 gives the median leave-one-stimulus-out R^2 over the 367 neurons for every layer bank at 10, 20 and 50 components, with the Spearman monotonicity of the median against depth. The gradient is monotone in two architecturally unrelated networks at every dimensionality, with no interior peak. The published ordering of fc6 over fc8 is tested with a Wilcoxon test paired across the 367 neurons (Table 16): fc8 is favoured at every dimensionality, including at 20 components where the fc6 median is higher ($p = 1.0 \times 10^{-4}$, 57.5% of neurons), so there is no dimensionality at which fc6 beats fc8 on these data when tested within neurons. The authors' released fc6 features score $+0.0878$ in the same pipeline, so the reversal is not an artefact of our feature extraction. What the release cannot settle is the scoring procedure behind the published percentages (48.52% for fc6, 32.56% for fc8, 53.12% for DINO-ViT, 44.23% for GloVe): only the ordering, not the absolute values, is comparable.

Conditioning control. Deeper layers are lower-dimensional and more category-aligned, so an axis model on 500 trial-averaged stimuli might fit a well-conditioned low-dimensional space better regardless of what the neurons encode. Table 17 reports each layer's participation ratio and design condition number next to its median R^2 at 50 components. The sharpest case refutes the alternative: AlexNet fc8 has the lowest participation ratio of the eight AlexNet banks (10.59, below conv1 at 13.70) and the worst condition number (40.8), and scores the highest R^2. Over all 20 banks, depth predicts R^2 ($\rho = +0.985$, $p = 3.6 \times 10^{-15}$) while the participation ratio does not ($\rho = +0.269$, $p = 0.25$); the partial correlation of depth with R^2 given the participation ratio is $+0.984$ ($p = 4 \times 10^{-14}$), that of the participation ratio given depth -0.023 ($p = 0.93$), and the same holds within each architecture separately. The gradient tracks depth, not conditioning.

Cross-decoding from perception to imagery. Neurons of different patients are not recorded simultaneously, so no cross-patient population vector is formed; decoding is leave-one-stimulus-out identification within a block, with the identity, scalar-gain and per-neuron-gain maps of the ladder (the general linear map is under-determined at $S \leq 8$). The null permutes the imagery labels within block (2,000 draws), so chance is measured rather than assumed; blocks with fewer than 3 neurons are excluded, leaving 7 blocks, 36 neurons and 4 patients (Table 18, Figure 5). Two caveats are irreducible: the 43 neurons were selected on imagery-perception axis correlation over all their stimuli, which favours gain-like transfer, and the release is trial-averaged. What stands is the trans-

Table 15: Stage 1b under the shared convention. Top: median leave-one-stimulus-out R^2 over 367 neurons per layer bank and number of principal components. Bottom: Spearman monotonicity of the median R^2 against depth.

Layer	$n_{\mathrm{pc}} = 10$	$n_{\mathrm{pc}} = 20$	$n_{\mathrm{pc}} = 50$
AlexNet conv1	$+0.0411$	$+0.0357$	-0.0129
AlexNet conv2	$+0.0580$	$+0.0545$	-0.0015
AlexNet conv3	$+0.0809$	$+0.0770$	$+0.0268$
AlexNet conv4	$+0.0861$	$+0.0943$	$+0.0579$
AlexNet conv5	$+0.1023$	$+0.1055$	$+0.0664$
AlexNet fc6	$+0.1071$	$+0.1237$	$+0.0898$
AlexNet fc7	$+0.1121$	$+0.1272$	$+0.1031$
AlexNet fc8	$+0.1138$	$+0.1206$	$+0.1041$
DINO block 1	$+0.0338$	$+0.0251$	-0.0396
DINO block 2	$+0.0503$	$+0.0413$	-0.0165
DINO block 3	$+0.0519$	$+0.0551$	-0.0024
DINO block 4	$+0.0609$	$+0.0645$	$+0.0301$
DINO block 5	$+0.0830$	$+0.0850$	$+0.0505$
DINO block 6	$+0.0870$	$+0.0987$	$+0.0704$
DINO block 7	$+0.0933$	$+0.1130$	$+0.0816$
DINO block 8	$+0.1010$	$+0.1229$	$+0.0883$
DINO block 9	$+0.1049$	$+0.1213$	$+0.0938$
DINO block 10	$+0.1068$	$+0.1195$	$+0.0956$
DINO block 11	$+0.1102$	$+0.1175$	$+0.0976$
DINO block 12	$+0.1108$	$+0.1186$	$+0.1028$
AlexNet, 8 layers: ρ	$+1.000$	$+0.929$ ($p = 0.00086$)	$+1.000$
DINO, 12 blocks: ρ	$+1.000$	$+0.867$ ($p = 0.00026$)	$+1.000$

Table 16: fc6 against fc8, Wilcoxon paired across the 367 neurons, and the best DINO block against fc8 under the patch-token convention; at 20 components the DINO block is marginally higher.

n_{pc}	fc6	fc8	paired diff.	p	fc8 better
10	$+0.1071$	$+0.1138$	$+0.0076$	3.8×10^{-10}	65.1%
20	$+0.1237$	$+0.1206$	$+0.0037$	1.0×10^{-4}	57.5%
50	$+0.0898$	$+0.1041$	$+0.0077$	9.4×10^{-7}	61.6%
n_{pc}	best DINO block	fc8		p	
10	12: $+0.1108$	$+0.1138$		0.57	
20	8: $+0.1229$	$+0.1206$		0.039	
50	12: $+0.1028$	$+0.1041$		0.11	

Table 17: Conditioning control at 50 components: participation ratio of the raw feature bank, condition number of the design, and median R^2.

Layer	Depth	Participation ratio	Condition number	Median R^2
AlexNet conv1	1	13.70	9.8	-0.0129
AlexNet conv3	3	61.63	17.3	$+0.0268$
AlexNet conv5	5	77.07	7.3	$+0.0664$
AlexNet fc6	6	51.42	26.2	$+0.0898$
AlexNet fc7	7	46.80	17.8	$+0.1031$
AlexNet fc8	8	10.59	40.8	$+0.1041$
DINO block 1	1	6.78	131.5	-0.0396
DINO block 9	9	17.72	22.1	$+0.0938$
DINO block 12	12	12.44	32.3	$+0.1028$

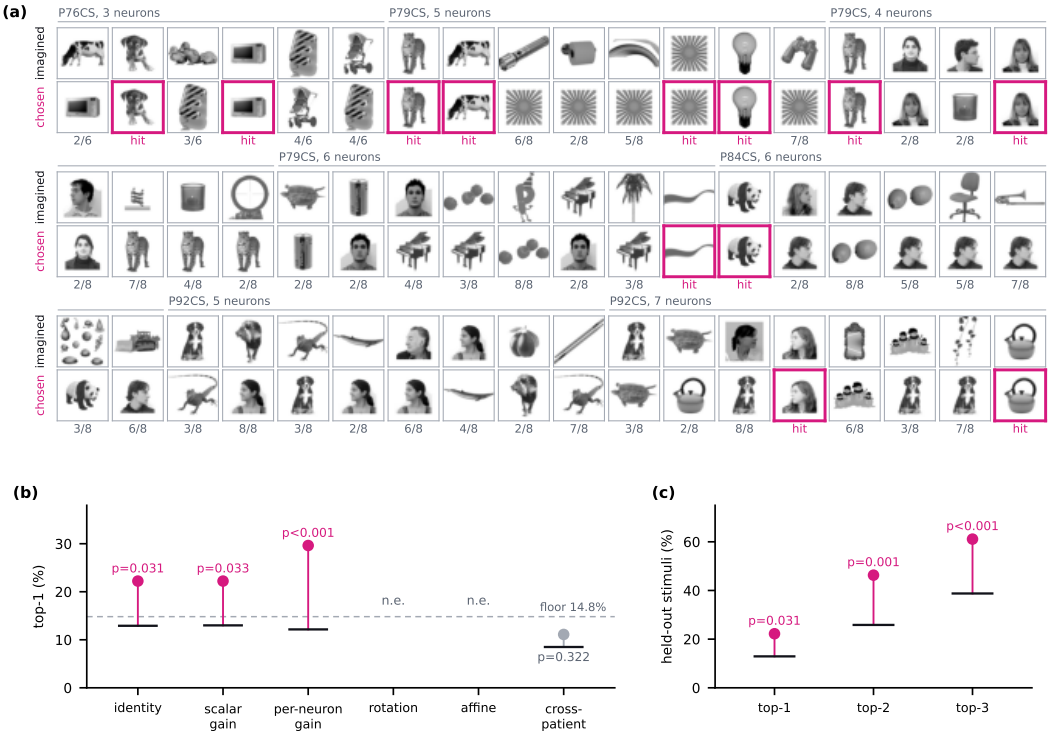

Figure 5: Perception-to-imagery identification in single neurons, identity map, within patient. (a) Each of the 54 held-out imagined stimuli of the 7 blocks (upper) and the perception population vector's first choice among the block's candidates (lower); hits framed, otherwise the rank of the true stimulus. (b) Top-1 for each map against its measured permutation null (black bars); n.e., not estimable at 6 to 8 stimuli. (c) Top-1 to top-3 against the same null. Effective n: 4 patients.

fer under the simplest map: identity identifies the imagined stimulus at 22.2% against a measured null of 12.9% ($p = 0.0315$), the scalar gain at the same 22.2% ($p = 0.034$). The per-neuron gain reaches 29.6% against its null of 12.2% ($p = 0.0005$), but its increment over identity, $+7.4$ pp, tested against its own label-permutation null, does not reach significance ($p = 0.061$; $d_z = 0.73$ at $n = 7$ blocks, carried by 3 of the 7 blocks). In the reverse direction identity and scalar gain give 20.4% ($p = 0.049$) while the fitted gain collapses to 11.1% ($p = 0.41$), the signature of seven gains fitted on seven training stimuli. The leakage audit found no fitted transform touching held-out data: no standardisation, centring or projection pools information across stimuli, the only fitted quantities are the gains computed from the training columns, and a corruption test (100 standard deviations of noise on the held-out imagery response) leaves every shipped prediction unchanged while a deliberately leaked fit is caught. Four of the 50 distinct stimuli appear in two blocks, so the effective n of the pooled p-value is closer to the 4 patients than to the 54 held-out stimuli.

Stimulus mapping. Every encoding number assumes that the 500 image files sorted by their code correspond to response indices 1 to 500. The release provides a trial-level category label per stimulus (five categories, consistent across the four repetitions of every stimulus), so under the correct mapping the category is decodable from the image features and under a wrong mapping it is at chance. A logistic regression on standardised features decodes the category at 79.8% from AlexNet fc6 (permuted-mapping null 34.9 ± 1.3%), 78.0% from fc8 and 84.0% from DINO block 12 (z of 34.3, 19.1 and 24.6). All 78 imagined stimuli index into the same screening set, so every imagined stimulus has perception responses from all 367 neurons in all 16 patients.

Perception-fitted models applied to imagery. For each of the 43 imagery neurons and each of the 20 layer banks at 50 components, the axis model is fitted on the neuron's perception responses to the 499 screening stimuli excluding the imagined stimulus being predicted, and its predictions are correlated with the neuron's imagery responses; no imagery datum enters any fit (Figure 6). The statistic is the mean Fisher-z correlation over neurons against a within-neuron permutation null

Table 18: Stage 2 cross-decoding, perception to imagery, 7 blocks, 36 neurons, 4 patients. Top: pooled top-1 identification against the within-block label-permutation null (2,000 draws), and the increment over identity against its own null. Bottom: per-block top-1.

Map	Top-1	Null	p	Δ identity	p (Δ)
Identity	22.2%	12.9%	0.0315	reference	reference
Scalar gain	22.2%	13.0%	0.034	+0.0 pp	n/a
Per-neuron gain	29.6%	12.2%	0.0005	+7.4 pp	0.061
Block	**Identity**	**Scalar**	**Diagonal**	**Δ identity**	
P76CS, 6 stimuli	33.3%	33.3%	33.3%	+0.0 pp	
P79CS, 8 stimuli	50.0%	50.0%	50.0%	+0.0 pp	
P79CS, 8 stimuli	25.0%	25.0%	25.0%	+0.0 pp	
P79CS, 8 stimuli	12.5%	12.5%	37.5%	+25.0 pp	
P84CS, 8 stimuli	12.5%	12.5%	25.0%	+12.5 pp	
P92CS, 8 stimuli	0.0%	0.0%	12.5%	+12.5 pp	
P92CS, 8 stimuli	25.0%	25.0%	25.0%	+0.0 pp	
Pooled	22.2%	22.2%	29.6%	+7.4 pp	

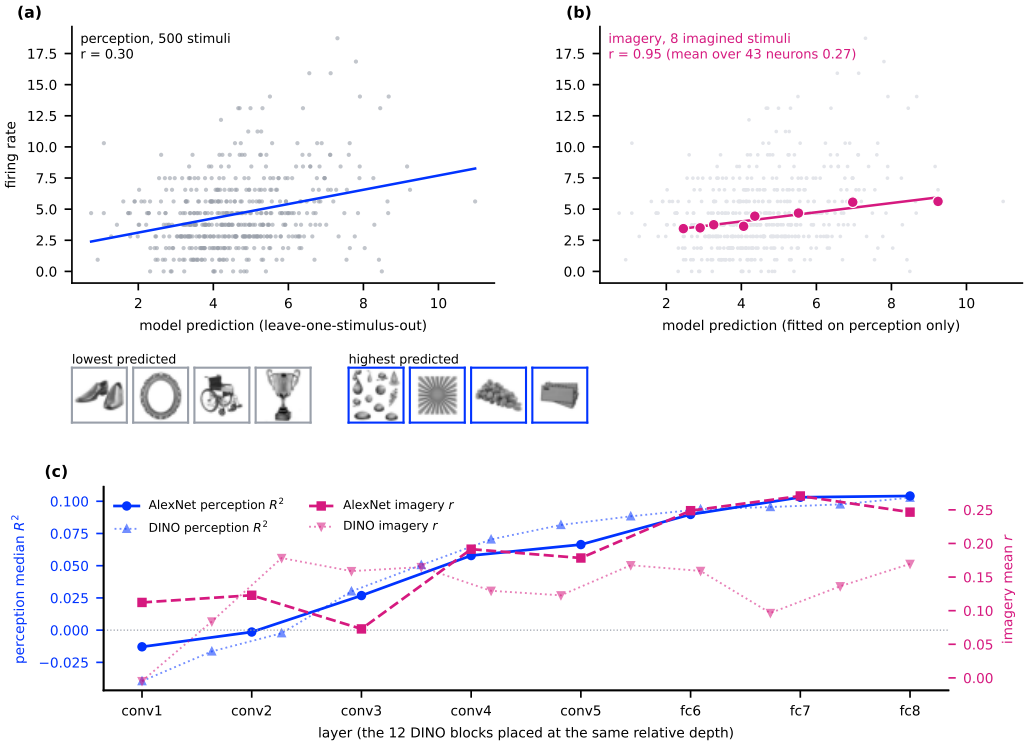

Figure 6: A perception-only encoding model applied to imagery. (a) Held-out perception prediction of one neuron from AlexNet fc7. (b) The same model's predictions for that neuron's imagery responses; this neuron is the best of 43, and the mean over neurons is $r = 0.27$. Below: the stimuli with the lowest and highest predicted response. (c) Perception R^2 and imagery r across layer depth.

(2,000 draws; Table 19). In AlexNet the best layer is fc7 ($r = +0.271$, $p = 0.0010$) with a depth gradient of $\rho = +0.833$ ($p = 0.0102$); in DINO the best block is block 3 ($r = +0.178$) and the gradient is not significant ($\rho = +0.315$, $p = 0.32$). Paired across the 43 neurons (Table 20), fc7 against DINO block 12 gives $+0.106$ in z with $p = 0.0599$, so no architecture dissociation is claimed; fc7 exceeds conv1 ($p = 0.045$) and fc6 and fc8 do not differ on imagery ($p = 0.94$).

Cross-patient transfer. The source is the perception of the other patients' screening neurons, the target one block's imagery, and the shared axis is the stimulus; no neuron correspondence is used. Three conditions are run on the same 7 blocks: representational similarity between the other patients'

Table 19: Experiment 2: mean Fisher-z correlation between perception-fitted predictions and actual imagery responses over 43 neurons, per layer bank, with its within-neuron permutation p.

AlexNet layer	r	p	DINO block	r	p
conv1	$+0.112$	0.049	block 1	-0.005	0.54
conv2	$+0.123$	0.042	block 2	$+0.084$	0.118
conv3	$+0.073$	0.145	block 3	$+0.178$	0.0085
conv4	$+0.191$	0.0055	block 4	$+0.159$	0.015
conv5	$+0.178$	0.0025	block 5	$+0.165$	0.012
fc6	$+0.249$	0.0005	block 6	$+0.130$	0.029
fc7	$+0.271$	0.0010	block 7	$+0.123$	0.043
fc8	$+0.247$	0.0010	block 8	$+0.168$	0.0075
			block 9	$+0.159$	0.015
			block 10	$+0.097$	0.092
			block 11	$+0.136$	0.026
			block 12	$+0.170$	0.0085
Depth ρ (8 layers)	$+0.833$	0.0102	Depth ρ (12 blocks)	$+0.315$	0.32

Table 20: Experiment 2b: paired comparisons across 43 neurons of the Fisher-z imagery correlation.

A	B	Mean z A	Mean z B	Diff.	p	Neurons favouring A
AlexNet fc7	DINO block 12	$+0.278$	$+0.171$	$+0.106$	0.0599	60.5%
AlexNet fc7	DINO block 3	$+0.278$	$+0.180$	$+0.097$	0.26	53.5%
AlexNet fc7	AlexNet conv1	$+0.278$	$+0.113$	$+0.165$	0.045	62.8%
AlexNet fc6	AlexNet fc8	$+0.254$	$+0.252$	$+0.002$	0.94	48.8%

perception and the block's imagery, leave-one-stimulus-out identification by similarity profile, and a fitted cross-patient ridge map as the architecture-search control. All three are null ($\rho = +0.001$, $p = 0.49$; 11.1% against $8.6 \pm 4.1\%$, $p = 0.33$; 11.1% against $6.1 \pm 3.4\%$, $p = 0.12$; Table 21). The detectable floor of the design, the 95th percentile of the cross-patient null, is 14.8%, so the null is informative for an effect of within-patient size (22.2%) and uninformative below 14.8%; the direct paired contrast of within against cross over the 7 blocks ($+11.9$ pp, $d_z = 0.51$, $p = 0.236$) is not significant, and no claim that within-patient transfer exceeds cross-patient transfer is made.

Table 21: Experiments 3 and 3b per block: within-patient identity top-1 (stage 2) and the three cross-patient conditions. Condition A is a Spearman ρ; conditions B and C are top-1 identification.

Block	Within, identity	A: RSA ρ	B: RSA identification	C: ridge identification
P76CS	33.3%	-0.411	0.0%	0.0%
P79CS	50.0%	-0.291	0.0%	0.0%
P79CS	25.0%	-0.008	12.5%	25.0%
P79CS	12.5%	-0.091	25.0%	12.5%
P84CS	12.5%	$+0.517$	0.0%	12.5%
P92CS	0.0%	-0.067	12.5%	0.0%
P92CS	25.0%	$+0.258$	25.0%	25.0%
Pooled (weighted)	22.2%	$+0.001$	11.1%	11.1%

Layer ordering. The published ordering of fc6 over fc8 does not reproduce: fc8 outperforms fc6 in a paired comparison over all 367 neurons ($p = 9.4 \times 10^{-7}$, 62% of neurons), the authors' own released fc6 features also score lower ($+0.0878$), the ordering corresponds to the top of a monotone depth gradient in both architectures, and a conditioning confound is refuted (partial correlation of depth given dimensionality $+0.984$, $p = 4 \times 10^{-14}$; of dimensionality given depth -0.023).

Limitations of the released data. The release does not support rotation or affine rungs of the map ladder (6 to 8 imagery stimuli per patient), covariance-based rungs (trial-averaged data), a reproduction of the published 31.35% cross-decoding accuracy, or an unbiased characterisation of the map (the 43 neurons were selected on the property under test).

Laterality, per-neuron gain and encoder readout. Laterality is null on this release. Over the six patients who contribute imagery neurons, the odds ratio is 2.03 at $p = 0.076$, and it is dominated by implant asymmetry, so the sample cannot separate a laterality effect from where the electrodes happen to sit. The per-neuron gain over the identity map is $+7.4$ pp and does not clear its own permutation null ($p = 0.061$), which is why Section 4.2 reports the identity map and not the fitted gain; read against chance instead of against its own null the same increment reaches $p = 0.0005$, and that is the comparison the null replaces. Encoder readouts use the patch-token mean rather than the CLS token, since the CLS readout is an artefact on this release; under the patch-token mean the ordering of fc6 against fc8 at 20 components is not established, because the ordering has to be tested within neurons rather than between their medians.

G Reconstruction Components: Architectures and Training

Software, seeds and decoder snapshots. The reconstruction pipeline of Section 4.4 runs under PyTorch 2.12.1 with CUDA 13.0 on NVIDIA L40S GPUs (Python 3.12, open_clip 3.3.0, torch_dct 0.1.6). Generation seeds are 1000 and 2000 for the test split and 1500 for validation; the permuted-prefix floor uses seed 7 and the random-position controls seeds 42 to 44; images are sampled in chunks of 25 with per-chunk seeds. The decoder adaptation used seed 0. The rows of Table 2 use the final 15,120-step snapshot. CLIP-L/14 scores use the QuickGELU variant of the OpenAI weights, which differs from the GELU variant by about 0.02.

H Alignment Metrics and Benchmark Task Family

Encoding-model alignment on THINGS-EEG2. On THINGS-EEG2 (10 subjects, 200 test images, split-half noise ceiling, stimulus-held-out ridge encoding), the hierarchical correspondence between model depth and response latency is observed without tuning: earlier model layers best predict earlier EEG time points and deeper layers later ones (layer-by-time diagonal $\rho = +0.42$, 10 of 10 subjects), occipital channels dominate, and the raw-r latency is 0.105 s. All 21 visual feature spaces (AlexNet conv1 to fc8, DINO-ViT-S/16 blocks 1 to 12, CLIP image) predict the EEG response to seen objects better than all semantic feature spaces (GloVe, CLIP text, BERT), with no overlap between the two families: the weakest visual space exceeds GloVe by $+0.184$ ($p = 0.0039$, floor 0.107, 9 of 10 subjects), although visual spaces differ among themselves by more than the visual-semantic gap. This is the EEG-scale counterpart of the single-neuron finding that the ventral temporal code is visual rather than semantic (Wadia et al., 2026), and it applies to perception only: on the Gao dataset, with ten unique images, encoding-model alignment is not identifiable (stimulus-held-out r numerically zero for unrelated feature sets), a property of the dataset that the task family below states explicitly.

Transport-distance metric. We measured the state change between perception and imagery as the 2-Wasserstein distance between a participant's perception and imagery clouds in the feature space of Eq. (1), computed exactly on class-stratified halves of matched size and read against a within-perception split-half floor, as a function of the alignment map applied. Before standardisation the distance is dominated by the change of alpha level between seeing and imagining ($+1.17$ above the floor on OVI, 22 of 22 participants); after the standardisation of Eq. (1) a residual of $+0.43$ remains ($p < 10^{-6}$, 22 of 22, floor 0.18), of the same size on the negative control ($+0.39$), so this residual measures the state, not the transferred content. Along the ladder of cross-fitted maps the residual falls from $+0.89$ at the identity to $+0.27$ after recentring (-0.62, $p < 10^{-5}$, 20 of 22 participants) and to $+0.10$ after per-axis rescaling (-0.17, within the 0.24 floor); a rotation between the covariance eigenbases, although detectable (0.40 rad above a 0.21 rad floor, $p < 10^{-4}$), brings the clouds no closer ($+0.005$), and a full affine map fitted on one half moves them apart on the other ($+0.05$, $p = 0.006$). Figures and animals follow the same ordering. The map from imagery to perception is therefore a translation and a diagonal rescaling, twenty parameters, which is the transport-space counterpart of the result of Section 4.1 that recentring is necessary and added capacity is not.

Benchmark release. The protocol of Section 3.2 is implemented as a task family in NeuralBench (Banville et al., 2026): a Study class for the imagery dataset, a cross-state split (fit on perception events, evaluate on imagery events, leave-one-participant-out and within-participant across sessions), the ten-electrode alpha readout as a baseline model restricted to the published analysis win-

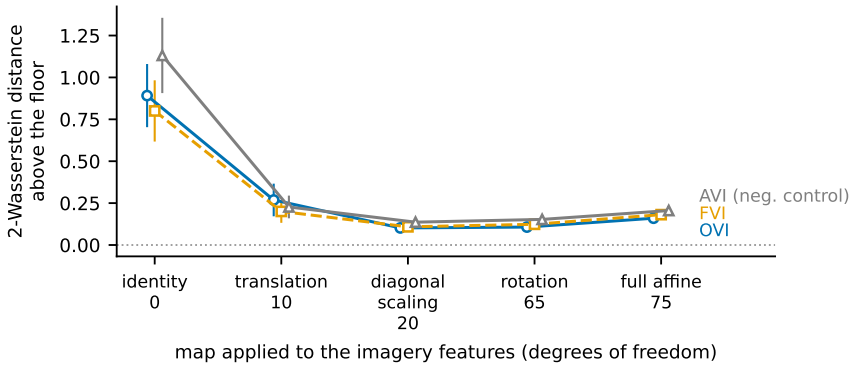

Figure 7: Transport distance between perception and imagery on Gao after each map, above the split-half floor; mean ± SE over 22 participants.

dows, DINOv2 and AlexNet-fc6 targets, and the own-null-subtracted transfer as the quantity every arm reports, a grid axis over the source-label permutation subtracted in the family's own metric module. The designated negative control and the published analysis windows are carried as dataset and model properties and scored in the driver. The transport distance ships as a function with its split-half floor and is not yet a task metric; the pre-cue window and the readout ceiling stay in the analysis code. The single-neuron recordings are a second task in the same family, one timeline per block of simultaneously recorded neurons, and the cross-decode of Section 4.2 reproduces through it to the reported precision, permutation nulls included. Both tasks supply the benchmark's study, events, split and extraction; the two readouts are driver scripts shipped with the family, because the framework's fit-once model interface receives neither participant identity nor the other state's population vector. The existing mental-imagery task in NeuralBench concerns motor and mental-task imagery; this family is its first visual-imagery task and its first cross-state split. The code accompanies the camera-ready version.

Category-level test. The two datasets share two categories, animals and objects, and a category-level direction learned from one modality's perception readouts could in principle be evaluated on the other's imagery readouts. On Gao this test is not identifiable: each category is a separate recording run, so a direction that separates animals from objects also separates runs, and no run-free contrast exists in the data. We state the design and do not claim it.

Constraints. The Gao dataset contains 10 unique images, so encoding-model alignment is not identifiable and retrieval is among a small fixed set; the ten images retain 17 to 25% of their fc6 energy in the shared 50-component basis; the Wadia imagery data contain 6 to 8 candidates per patient and admit no noise ceiling. On Gao, the imagery-period signal cannot be separated from the pre-cue period, so results are described as transfer to the post-stimulus period in which imagery is instructed.