A PATCHWISE AGGREGATION ENCODER FOR EEG-TO-IMAGE RETRIEVAL: AN ANALYSIS OF WHAT IMAGE FEATURES MATTER

Anonymous authors
Paper under double-blind review

ABSTRACT

What image features matter in EEG-to-image retrieval? Published approaches on the Things-EEG benchmark align EEG responses with features of a pre-trained visual encoder, but which image features drive this alignment remains unclear. We answer this question with the Patchwise Aggregation Encoder (PAE), a 1.57 M-parameter image encoder trained from scratch. The image is split into non-overlapping patches, each patch is processed independently by a shared stack of residual blocks, and the patch features are fused by a learned weighted average. On zero-shot intra-subject retrieval over 200 unseen images, PAE reaches a mean top-1 of 96.9% and top-5 of 99.7% across the 10 subjects, a new state of the art on this benchmark. A purely statistical variant, S-PAE, replaces the pixel input with 15 low-level statistics of each patch and is even stronger, at 97.6% top-1 and 99.9% top-5. Ablations then identify the features this alignment relies on. The feature space is dominated by luminance contrast and edge statistics, and the learned pooling weights concentrate on the image center, consistent with foveal dominance in early visual processing. Retraining with only three statistics retains 95.2% top-1 accuracy, and brightness and contrast alone reach 90.3%. EEG-to-image retrieval thus rests on the low-level statistics of local patches, computed independently and combined by learned spatial weights, which makes decoding compact and interpretable by construction. These findings offer a new perspective on EEG-image alignment and motivate further research on EEG-based visual decoding.

1 INTRODUCTION

Decoding visual content from non-invasive brain recordings is a central goal of computational neuroscience and of brain-computer interfaces (Khaligh-Razavi & Kriegeskorte, 2014; Horikawa & Kamitani, 2017; Spampinato et al., 2017; Zhu et al., 2010). Electroencephalography (EEG) offers millisecond temporal resolution at low cost, and the Things-EEG dataset (Gifford et al., 2022) established a benchmark for quantitative visual decoding on this signal. In the standard zero-shot protocol, a model is trained on EEG responses to 1,654 object concepts and must identify which of 200 held-out images the subject saw. Early contrastive approaches aligned EEG features with frozen CLIP embeddings (Radford et al., 2021) and remained below 20% top-1 accuracy (Song et al., 2024). Attention-based EEG encoders improved this to about 29% (Li et al., 2024). The most recent approaches keep the frozen features and instead modify the images before encoding. The Uncertainty-aware Blur Prior (UBP) (Wu et al., 2025) estimates the uncertainty within the paired data and dynamically blurs the high-frequency details of each image, reaching 50.9%, and the Visual EEG Decoding (VED) framework of Liu et al. (Liu et al., 2026) extends the blur to multiple levels with a feature selection module, reaching 80.0%.

These approaches share a design choice that has received little scrutiny. The alignment pairs EEG responses with image embeddings extracted by a frozen encoder pre-trained on millions of images, in every case a CLIP variant (Song et al., 2024; Wei et al., 2024; Li et al., 2024; Zhang et al., 2025; Wu et al., 2025; Liu et al., 2026; Xu et al., 2025a; Meng et al., 2026; Jo et al., 2026). Which image features support this alignment has mostly been studied by varying the image representation space. Replacing CLIP features with embeddings aligned to human perceptual similarity judgments improves retrieval by up to 21% top-1 (Rajabi et al., 2025), spreading the target over several layers of a vision transformer outperforms final layer alignment (Choi et al., 2026; Dosovitskiy et al., 2021),

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

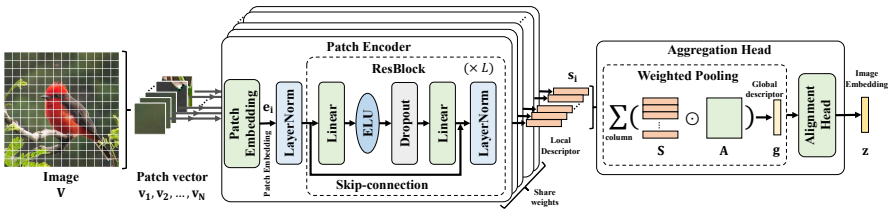

Figure 1: **The Patchwise Aggregation Encoder (PAE).** The image encoder f_θ consists of exactly two modules. (i) The *Patch Encoder* $\mathbf{G}$ (left) splits the image into N non-overlapping $w \times w$ patches and maps each patch to a d-dimensional embedding by the shared Patch Embedding (Eq. (1)), and L ResBlocks refine every token independently, with no token interaction (Eqs. (3)–(4)). (ii) The *Aggregation Head* $\mathbf{H}$ (right) fuses the local features into $\mathbf{g}$ by Weighted Pooling, each dimension by its own learned spatial weighting map (Eq. (5)), and a linear Alignment Head projects $\mathbf{g}$ to the image embedding $\mathbf{z}$ (Eq. (6)). The patch representations are computed independently and are combined only by the final aggregation.

and fusing several pre-trained encoders, two CLIP variants together with a variational autoencoder pixel latent, raises retrieval to 75.7% top-1, indicating that the signal aligns jointly with semantic and pixel-level content (Zheng et al., 2026). In all of these studies, however, the space is supplied by pre-trained visual models whose features pass through a complex extraction pipeline, so the image features that matter for EEG-to-image retrieval cannot be read out from the encoder, and their attribution remains indirect. To our knowledge, no published approach on this benchmark trains the image branch from scratch.

This paper removes the frozen branch and asks what image features matter in EEG-to-image retrieval. Two observations from visual neuroscience shape the design. First, a single feedforward pass of early visual processing supports scene identification within about 150 ms of stimulus onset (Thorpe et al., 1996). Second, the statistics of local contrast in natural images correlate with responses from early visual cortex and with evoked EEG potentials (Field, 1987; Scholte et al., 2009; Groen et al., 2013). Both suggest that the image information the EEG can carry is largely local and statistical in nature, and the encoder is designed to compute this representation. We introduce the Patchwise Aggregation Encoder (PAE), a 1.57 M-parameter image encoder trained from scratch. The encoder divides the image into patches, processes every patch independently through a shared stack of residual blocks, and fuses the patch features by learned spatial weights. This topology places PAE in the family of deep bag-of-local-features models (Brendel & Bethge, 2019; Csurka et al., 2004; Lazebnik et al., 2006), in which local features are computed independently and combined by simple aggregation. The design serves two purposes. First, it tests whether such an encoder can match the retrieval performance of approaches that use pre-trained image encoders on this benchmark. Second, its compact and interpretable design makes the features accessible to attribution, by reading out what the encoder computes, by retraining it under input constraints, and by testing which inputs are necessary. We also train a purely statistical variant, the Statistics Patchwise Aggregation Encoder (S-PAE), which replaces the pixel input with 15 low-level statistics of each patch (Field, 1987; Ruderman, 1994; Hyvärinen et al., 2009), to test whether the alignment-relevant content of a patch is preserved by these statistics.

Our main contributions can be summarized as follows. First, the PAE reaches a mean top-1 of 96.9% and top-5 of 99.7% across the 10 subjects, a new state of the art on this benchmark, and improves top-1 on every subject over approaches that use pre-trained image encoders. Second, the purely statistical S-PAE is even stronger, at 97.6% top-1 and 99.9% top-5, showing that the retrieval does not require the pixel input. Third, an attribution analysis shows that luminance contrast and edge statistics dominate the feature space and that the learned pooling weights concentrate on the image center, consistent with foveal dominance in early visual processing. Fourth, controlled suppression and architecture ablations show that the encoders rely primarily on the spatial layout of the patches and are nearly achromatic, and that adding interaction between the patch tokens substantially reduces accuracy.

Table 1: The 15 low-level patch statistics of the S-PAE. Notation: R, G, $B \in [0,1]^{H\times W}$ are the red, green, and blue color channels of the image, and $\mathrm{L} = 0.299\,R + 0.587\,G + 0.114\,B \in [0,1]^{H\times W}$ is the BT.601 gray image (International Telecommunication Union, 2011); $\nabla\mathrm{L} = (g_h, g_v)$ is its gradient computed with finite differences, with magnitude $\|\nabla\mathrm{L}\| = (g_h^2 + g_v^2)^{1/2}$, $\langle\cdot\rangle$ the spatial average over the w^2 pixels of a patch, $\mathbf{1}[\cdot]$ the indicator function, $\nabla^2\mathrm{L}$ the discrete Laplacian over the four neighboring pixels, $\mathrm{L}_{2\times2}$ the image after non-overlapping 2×2 block averaging, and $\bar{\mathrm{L}}$ the average gray level over the full image. Motivation and citations: Appendix D.

Family	Statistic	Symbol	Definition
Luminance	Brightness	μ_{L}	$\langle \mathrm{L} \rangle$
	Contrast	σ_{L}	$\left(\langle \mathrm{L}^2 \rangle - \langle \mathrm{L} \rangle^2\right)^{1/2}$
Edge	Gradient energy	E_∇	$\langle g_h^2 + g_v^2 \rangle$
	Edge fraction	f_{e}	$\langle \mathbf{1}[\|\nabla\mathrm{L}\| > 0.12] \rangle$
Orientation	Orientation energies	e_θ	$\frac{\langle (g_h \cos\theta + g_v \sin\theta)^2 \rangle}{\langle \|\nabla\mathrm{L}\|^2 \rangle}$, $\theta = 0°, 45°, 90°, 135°$
	Horizontal / vertical log-ratio	$\rho_{h/v}$	$\log_{10} \frac{\langle g_h^2 \rangle}{\langle g_v^2 \rangle}$
Frequency	Laplacian energy	E_Δ	$\left\langle (\nabla^2\mathrm{L})^2 \right\rangle$
	Fine-frequency fraction	f_{hf}	$\frac{\langle (\mathrm{L} - \mathrm{L}_{2\times2})^2 \rangle}{\langle (\mathrm{L} - \bar{\mathrm{L}})^2 \rangle}$
Color	Channel means	μ_R, μ_G, μ_B	$\langle R \rangle$, $\langle G \rangle$, $\langle B \rangle$
	Saturation	$\langle s \rangle$	mean of $s = \max(R,G,B) - \min(R,G,B)$

2 METHODS

In the intra-subject setting, each subject provides a training set of M paired samples $\{(\mathbf{V}_n, \mathbf{X}_n)\}_{n=1}^{M}$, where $\mathbf{V}_n \in \mathbb{R}^{3\times H\times W}$ is a natural image and $\mathbf{X}_n \in \mathbb{R}^{C\times T}$ is the subject's EEG response (C scalp electrodes, T time samples per channel). We learn an image encoder f_θ and an EEG encoder Φ_ϕ such that the **image embedding** $\mathbf{z} = f_\theta(\mathbf{V}) \in \mathbb{R}^{d_z}$ and the **EEG embedding** $\mathbf{y} = \Phi_\phi(\mathbf{X}) \in \mathbb{R}^{d_z}$ place each image next to the EEG it evoked.

Apart from the image encoder, all components are identical to Visual EEG Decoding (VED) (Liu et al., 2026). The EEG encoder is the brain branch of VED and applies channel-wise spatial filtering, temporal encoding, and a linear projection to d_z. It is trained jointly with the image encoder. The complete training objective, the optimization setup, and the model-selection rule are given in Appendix C. Each feature of the image encoder is computed from a single patch, each feature dimension carries an explicit spatial weighting map, and the embedding is a linear mixture of feature dimensions. Every contribution to the embedding is thus explicit, which makes the encoder *interpretable by construction* and enables the attribution of §3.3. Table 4 (Appendix C) lists the symbols with their default values.

2.1 IMAGE ENCODER: THE PATCHWISE AGGREGATION ENCODER (PAE)

The image encoder f_θ is designed around the hypothesis *that a weighted global average of independently computed local features suffices to carry the image information that supports the alignment*. We call the resulting architecture the **Patchwise Aggregation Encoder** (PAE) (Fig. 1). PAE has exactly two modules. The **Patch Encoder** $\mathbf{G}$ maps each patch to a local feature through the **Patch Embedding** followed by L nonlinear **ResBlocks**. The **Aggregation Head** $\mathbf{H}$ fuses the local features into the image embedding by **Weighted Pooling** followed by a linear **Alignment Head**. All nonlinear capacity resides in the per-patch stage $\mathbf{G}$, the global stage $\mathbf{H}$ is a linear weighted sum, and patch tokens never exchange information. The only cross-patch operation is the final pooling itself. This topology places PAE in the family of deep bag-of-local-features models (Brendel & Bethge, 2019; Csurka et al., 2004; Lazebnik et al., 2006). Two differences remain. The uniform spatial average is replaced by learned, per-dimension weighting maps over positions (Eq. (5)), and the linear readout is trained for contrastive alignment with EEG rather than for classification.

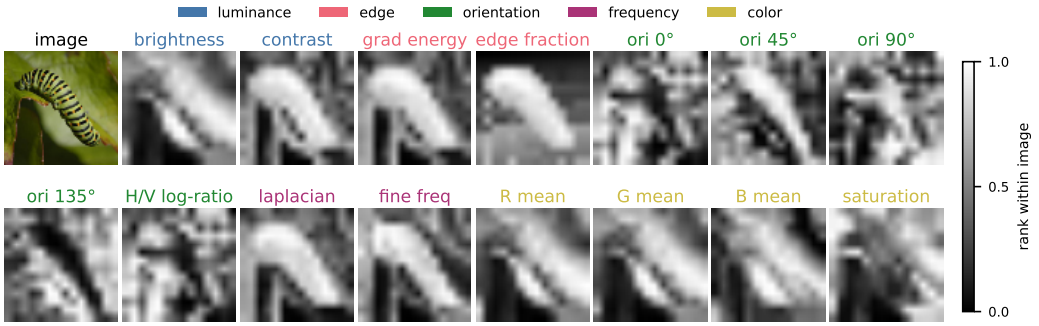

Figure 2: **The 15 low-level statistics of one test image.** The image is split into 196 non-overlapping 16×16 patches. Each panel shows one statistic of Table 1 as a 14×14 map over the patches, bilinearly upscaled to the image size for display, with label colors marking the five families. Gray level encodes the rank of each patch within the 196 values of the image (colorbar on the right). This rank scaling is for display only, whereas the encoder consumes the statistics standardized with the moments of the training set (§2.2).

Stage 1: Patch Embedding. The image $\mathbf{V} \in \mathbb{R}^{3\times H\times W}$ is partitioned into N non-overlapping patches of $w \times w$ pixels, and the i-th patch is vectorized to $\mathbf{v}_i \in \mathbb{R}^{3w^2}$. The Patch Embedding is a single convolutional layer with d kernels of size w^2 and stride w (equivalently, a linear projection $\mathbf{W}_e \in \mathbb{R}^{d\times 3w^2}$ shared across patches), mapping each patch to a **patch embedding** $\mathbf{e}_i$:

$$\mathbf{e}_i = \mathbf{W}_e\,\mathbf{v}_i \in \mathbb{R}^d, \qquad i = 1, \ldots, N. \tag{1}$$

An initial LayerNorm (Ba et al., 2016), applied to each patch token *independently*, produces the full token set $\mathbf{t}_i^{(0)} = \mathrm{LN}(\mathbf{e}_i)$, $\mathbf{T}^{(0)} = [\,\mathbf{t}_1^{(0)}; \ldots; \mathbf{t}_N^{(0)}\,] \in \mathbb{R}^{N\times d}$.

Stage 2: ResBlocks (per-token refinement). A sequence of L residual MLP blocks refines the tokens. The block MLP is a two-layer transform with ELU activation σ and hidden width h,

$$\mathrm{MLP}_\ell\big(\mathbf{t}_i^{(\ell-1)}\big) = \mathbf{W}_2^{(\ell)}\,\sigma\big(\mathbf{W}_1^{(\ell)}\mathbf{t}_i^{(\ell-1)}\big) \in \mathbb{R}^d, \qquad \mathbf{W}_1^{(\ell)} \in \mathbb{R}^{h\times d},\ \mathbf{W}_2^{(\ell)} \in \mathbb{R}^{d\times h}, \tag{2}$$

and the ℓ-th ResBlock updates each token independently:

$$\mathbf{t}_i^{(\ell)} = \mathrm{LN}_\ell\Big(\mathbf{t}_i^{(\ell-1)} + \mathrm{MLP}_\ell\big(\mathbf{t}_i^{(\ell-1)}\big)\Big), \qquad i = 1, \ldots, N,\ \ell = 1, \ldots, L, \tag{3}$$

where LN_ℓ is a per-token LayerNorm. Dropout at rate p is applied to the hidden activations during training. By induction over Eqs. (2)–(3), the output of the **Patch Encoder G** is the **local feature matrix**

$$\mathbf{S} = \mathbf{T}^{(L)} \in \mathbb{R}^{N\times d}, \qquad \mathbf{s}_i = \mathbf{t}_i^{(L)} = \mathbf{G}\big(\mathbf{v}_i\big) \in \mathbb{R}^d, \tag{4}$$

where $\mathbf{G}(\mathbf{v}_i)$ is computed independently for each patch by the same function, so that every entry of $\mathbf{S}$ is a fixed nonlinear function of patch $\mathbf{v}_i$ *alone*.

Stage 3: Weighted Pooling. The first stage of the Aggregation Head fuses the N local features into a **global feature** $\mathbf{g} \in \mathbb{R}^d$ by a weighted sum whose weights are learned parameters. Concretely, a learned parameter matrix $\mathbf{P} \in \mathbb{R}^{N\times d}$ is normalized by a softmax over the N positions inside the weighted sum,

$$g_j = \sum_{i=1}^{N} \frac{\exp(\mathbf{P}_{ij})}{\sum_{i'=1}^{N}\exp(\mathbf{P}_{i'j})}\,\mathbf{S}_{ij} = \sum_{i=1}^{N} \mathbf{A}_{ij}\,\mathbf{S}_{ij}, \qquad j = 1, \ldots, d. \tag{5}$$

The matrix $\mathbf{P}$ is initialized to zero, so pooling starts from the uniform average, and the spatial preferences are acquired during training. By construction, each column $\mathbf{A}_{:j} \in \mathbb{R}^N$ is nonnegative and sums to one, so that *every feature dimension has its own learned spatial weighting map*.

Stage 4: Alignment Head. The second stage of the Aggregation Head is a single linear map that produces the image embedding,

$$\mathbf{z} = \mathbf{W}_h\,\mathbf{g} \in \mathbb{R}^{d_z}, \qquad \mathbf{W}_h \in \mathbb{R}^{d_z\times d}. \tag{6}$$

Table 2: Intra-subject 200-way retrieval on Things-EEG (10 subjects, top-k accuracy in %). Rows: mean over three training seeds; the per-subject mean $\pm$ std are given in Appendix Table 11. UBP and VED rows: numbers as reported in the original papers (Wu et al., 2025; Liu et al., 2026). †The finer patch size ps=8. Highest value per column in **bold**.

		S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Mean
UBP	top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
VED	top-1	81.9	81.7	78.3	76.9	71.4	84.3	78.2	84.5	78.7	84.1	80.0
	top-5	96.8	96.8	96.0	97.2	93.7	98.7	96.7	98.1	96.8	98.1	96.9
PAE	top-1	96.0	97.5	97.0	86.8	91.5	97.7	96.5	98.3	97.0	98.3	95.7
	top-5	99.3	**100.0**	**100.0**	97.5	98.3	**100.0**	**99.5**	**100.0**	99.3	99.8	99.4
PAE†	top-1	**97.2**	**98.0**	98.0	91.5	92.7	98.2	97.2	**99.3**	**98.7**	98.3	96.9
	top-5	99.7	**100.0**	**100.0**	98.7	**100.0**	**100.0**	99.2	**100.0**	99.5	**100.0**	99.7
S-PAE	top-1	97.0	97.0	**98.3**	**94.8**	**95.2**	**98.7**	**98.2**	**99.3**	97.7	**99.3**	**97.6**
	top-5	**100.0**	**100.0**	**100.0**	**100.0**	99.7	**100.0**	**99.5**	**100.0**	**99.8**	**100.0**	**99.9**

Stages 3–4 together form the **Aggregation Head H**, and the full architecture is the composition

$$\mathbf{z} = f_\theta(\mathbf{V}) = \mathbf{H}\big(\mathbf{G}(\mathbf{v}_1), \ldots, \mathbf{G}(\mathbf{v}_N)\big). \tag{7}$$

2.2 Image encoder without pixels: the Statistics Patchwise Aggregation Encoder (S-PAE)

Building on the hypothesis of §2.1, we examine a stronger hypothesis, *that the alignment-relevant content of a single patch is largely preserved by its low-level image statistics*. Each patch is replaced by the vector of its 15 low-level statistics, which fall into five families chosen for their role in early visual processing (Field, 1987; Olshausen & Field, 1996), and the PAE pipeline operates on these vectors directly. We call the resulting encoder the **Statistics Patchwise Aggregation Encoder** (**S-PAE**). Fig. 2 illustrates the 15 statistics on one test image. Table 1 defines every statistic, and Appendix D details the motivation and the citations. Every statistic is standardized with the moments of the training set.

Architecture. Let $\mathbf{u}_i \in \mathbb{R}^{15}$ collect the 15 statistics of patch i (in the order of Table 1). S-PAE replaces the Patch Embedding (Eq. (1)), whose input is the vectorized RGB patch $\mathbf{v}_i \in \mathbb{R}^{3w^2}$, by a learned linear map on the statistics,

$$\mathbf{e}_i = \mathbf{W}_{\text{stat}}\,\mathbf{u}_i \in \mathbb{R}^d, \qquad \mathbf{W}_{\text{stat}} \in \mathbb{R}^{d\times 15}, \tag{8}$$

followed by the same per-token LayerNorm as in the PAE. Everything downstream is unchanged. The ResBlocks refine each token independently (Eq. (3)), and the Aggregation Head fuses the features into the image embedding (Eqs. (5)–(6)). The image branch is thus an explicit function of the $15 \times N$ low-level statistics of the image alone.

3 Experiments

3.1 Dataset, protocol, and implementation details

Dataset. We evaluate on **Things-EEG** (Gifford et al., 2022), a 63-channel EEG dataset recorded from 10 subjects viewing natural images in a rapid serial visual presentation paradigm. The training pool contains 16,540 images (1,654 concepts $\times$ 10 exemplars, 4 repetitions each). The test pool contains 200 disjoint concepts (1 image each, 80 repetitions). Training repetitions are averaged, a 95/5 train/validation split is held out for model selection, and the 80 test repetitions are averaged into a single test sample per image. Appendix B records the full recording and preprocessing details.

Training and evaluation. The image encoders are the **PAE** (§2.1) and the **S-PAE** (§2.2). For the PAE, input images of 224×224 pixels are split into N=196 patches of w=16 pixels and processed by L=3 ResBlocks of width d=192 into a d_z=1024 embedding, with 1.57 M parameters in total.

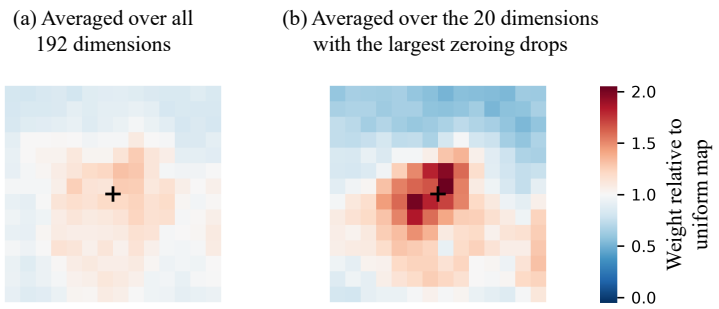

Figure 3: **Pooling maps averaged over all dimensions and over the twenty dimensions with the largest zeroing drops.** Each panel shows a weighted average of the learned per-dimension maps, divided by the uniform map ($1/196$ per position). Left: the map averaged over all dimensions, with eccentricity 0.17 and spread 3.92. Right: the mean map of the twenty dimensions with the largest zeroing drops (Appendix F.1), with eccentricity 0.66 and spread 3.76.

The S-PAE uses the same configuration, with the Patch Embedding replaced by the learned map on the 15 statistics. We also report the PAE at the finer patch size $w{=}8$ ($N{=}784$), which changes only the patch size. The image and EEG branches of both encoders are trained end-to-end from scratch. We use AdamW with learning rate 10^{-3}, batch size $B{=}1024$, and 200 epochs. The checkpoint with the best validation top-1 is selected per subject. Unless stated otherwise, every training configuration in this paper is run with three random seeds, and results are reported as mean $\pm$ std over seeds. Evaluation is *intra-subject* 200-way zero-shot retrieval, identical to the protocol of VED and UBP (Liu et al., 2026; Wu et al., 2025). Appendix C details the complete experimental setup.

3.2 Full-scale main results on 10 subjects

Table 2 reports intra-subject retrieval accuracy on all 10 subjects against the two published baselines VED (Liu et al., 2026) and UBP (Wu et al., 2025), both of which use frozen CLIP RN50 (Radford et al., 2021) image features. The PAE reaches a mean top-1 of 95.7 and a mean top-5 of 99.4 at the default patch size $ps{=}16$. At the finer patch size $ps{=}8$ the same recipe reaches a mean top-1 of 96.9 and a mean top-5 of 99.7. Relative to the baselines, the PAE improves top-1 on every subject by a mean of $+15.1\%$ over VED and $+44.8\%$ over UBP. The S-PAE (§2.2) replaces the pixel input entirely by the 15 per-patch statistics and reaches a mean top-1 of 97.6 and a mean top-5 of 99.9. It matches or exceeds the pixel PAE on nine of the ten subjects (mean $+1.9\%$) and is the strongest row of the table.

3.3 Which image features drive the alignment?

Every contribution to the embedding is explicit (§2), so the image content behind the alignment can be analyzed directly on the trained encoders. Three analyses follow. The pooling map analysis localizes where in the image the learned spatial weights concentrate (§3.3.1). The sufficiency analysis retrains the S-PAE on subsets of the 15 statistics to identify which of them support retrieval (§3.3.2). The controlled suppression analysis measures how retrieval responds when one image feature type is destroyed at test time (§3.3.3). Together they show that the feature space is dominated by luminance contrast and edge statistics and that the learned pooling weights concentrate on the image center.

3.3.1 Pooling map analysis

By Eq. (5), every feature dimension j has an explicit weighting map $\mathbf{A}_{:j}$ over the 14×14 grid. Following the population receptive field analysis of Dumoulin & Wandell (2008), each map is summarized by its centroid eccentricity and spatial spread, in grid units (Appendix F.2). The learned maps are broad but not uniform in position. The mean spread is 3.86 against 4.03 for the uniform map, and $154/192$ maps lie above the 95th percentile of their own permutation null. The map averaged over dimensions has centroid eccentricity 0.17 and spatial spread 3.92 (Fig. 3). The mean map of the twenty dimensions with the largest zeroing drops (Appendix F.1) has spatial spread 3.76

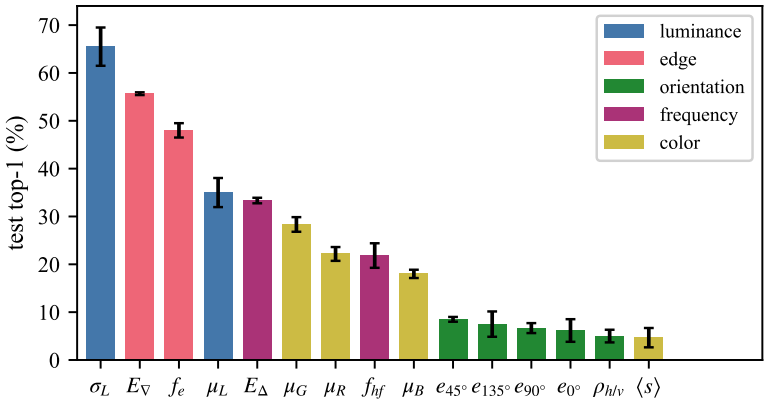

Figure 4: **S-PAE trained on a single statistic** (subject 1, test top-1, mean ± std). Each bar corresponds to the full S-PAE encoder trained with one of the 15 patch statistics as its only input, colored by statistic family. Chance is 0.5%. These runs also constitute the final round of the greedy backward elimination. Exact values are given in Table 6.

and eccentricity 0.66. The pooling maps are thus individually structured and collectively weighted toward the image center, consistent with foveal dominance in early visual processing (Cowey & Rolls, 1974), and the dimensions that matter most for retrieval have the smaller spatial spread. See Appendix F.2 for more details.

3.3.2 Sufficiency of the statistics

We retrain the S-PAE with subsets of the 15 statistics as its only input, changing the input dimension of the first linear map and leaving everything downstream unchanged (subject 1). Training on a single statistic yields between 4.7% ($\langle s \rangle$) and 65.5% (σ_{L}) top-1 accuracy, against 97.0% for the full set (Fig. 4). The edge statistics follow at 55.7% (E_∇) and 48.0% (f_{e}), and the four orientation energies stay below 9%. The feature space is thus dominated by luminance contrast and edge statistics. Retraining with the three statistics $\{\mu_{\mathrm{L}}, \sigma_{\mathrm{L}}, E_\nabla\}$ retains 95.2% top-1 accuracy, within 1.8% of the full set, and brightness and contrast alone ($\{\mu_{\mathrm{L}}, \sigma_{\mathrm{L}}\}$) reach 90.3%. The ablation removes in each round the statistic whose removal costs least accuracy and ends with this pair. See Appendix F.3 for more details.

3.3.3 Controlled suppression

We apply the controlled suppression framework of Burgert et al. (2025) to the PAE and the S-PAE, each trained on subject 1 with the default configuration of §3.1. The nine transformations modify only the 200 test images, and each suppresses one of the three feature types of the framework. Patch shuffle on 3×3, 6×6, and 14×14 grids and per-patch 90° rotation on 6×6 and 14×14 grids suppress shape, bilateral and Gaussian filtering suppress fine texture by smoothing the high-frequency details of the image, and grayscale and channel shuffle suppress color. The encoders are evaluated on the transformed test sets, and reliance is measured by relative accuracy, the top-1 accuracy under a transformation divided by the accuracy on the unsuppressed images (Fig. 5). Both encoders rely primarily on the spatial layout of the patches and far less on edge continuity. Patch shuffle, which randomizes the spatial positions of the patches, collapses retrieval to between 16.2% and 35.2% relative accuracy, whereas rotation, which preserves the locality of the patches but breaks edge continuity by turning each patch in place, retains at least 82.7%. Both are robust to blur and nearly achromatic. Under grayscale, relative accuracy drops to 92.7% for the PAE and 95.4% for the S-PAE. Under channel shuffle, it drops to 89.6% for the PAE and 63.5% for the S-PAE. The gap between the encoders under channel shuffle is not color reliance but integrity of the luminance map. A per-pixel channel permutation scrambles the BT.601 luminance map from which twelve of the S-PAE's fifteen input statistics are computed, whereas grayscale deletes chromaticity and leaves the

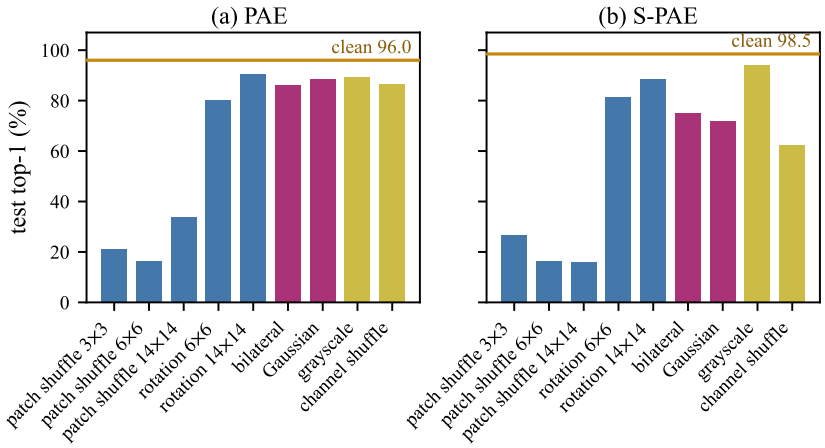

Figure 5: **Suppression profiles of the PAE and the S-PAE (subject 1).** Bars: test top-1 when one feature type is suppressed in the 200 test images, following the controlled suppression framework of Burgert et al. (2025) (blue = shape, magenta = texture, yellow = color). The orange line marks the unsuppressed clean accuracy, so relative accuracy, the framework's primary metric, is bar height divided by the line. Training and the EEG side are untouched. See Table 15 for more details.

map intact. The full protocol, the transformation examples (Fig. 11), and the extension of the suite to further image encoders are given in Appendix I.

3.4 Architecture ablations

We retrain the PAE with one component changed at a time on the four architecture axes of Table 3 (subject 1). The two patch sizes show the same trends, so we report the default ps=16. The token mixing variant inserts two learned cross-patch sub-layers between the residual blocks, in the style of ResMLP (Touvron et al., 2021), so that every patch token becomes a learned mixture of all 196 patch tokens, and the attention variant processes all patch tokens jointly through the full self-attention of a ViT of comparable size (Dosovitskiy et al., 2021). Adding interaction between the patch tokens substantially reduces retrieval accuracy, to 48.17% for token mixing and 44.67% for attention, compared with 96.00% for the PAE. Replacing the Weighted Pooling with a fixed uniform average reduces accuracy to 59.00%, and concat pooling, which concatenates all 196 patch features for a global linear head, reduces it to 52.00%. Applying the alignment head per patch feature before the weighted pooling (post-head pooling) yields 94.67%, and inserting one residual block between the pooling and the head (post-pooling ResBlock) yields 96.17%. Averaging with learned per-position weights is therefore substantially more accurate than direct averaging or feature concatenation, whereas changing the pooling position or inserting nonlinearity after the pooling yields no further gain over the default. Removing the ResBlocks entirely at L=0 makes the model linear and reduces accuracy to 36.00%. Accuracy reaches 94.17% at L=1 and stays between 93.67% and 96.00% for all larger depths, and it varies by less than 2% across widths from 48 to 384. The remaining ablations are reported in Appendix G.

4 Discussion

In EEG-to-image decoding, UBP and VED blur each image before encoding, motivated by the mismatch between sharp digital images and the blurred human perception (Wu et al., 2025; Liu et al., 2026), whereas the PAE uses no blur and its learned pooling maps still concentrate on the image center (§3.3.1). The engineered blur and the learned maps are thus two implementations of the same foveal prior. The statistics that carry retrieval are led by luminance contrast, gradient energy, and edge fraction, the statistics to which cortical cells in early visual cortex respond and that natural images supply in abundance (Field, 1987; Olshausen & Field, 1996; Hyvärinen et al., 2009), and brain responses correlate strongly with exactly these contrast and gradient statistics of natural images (Scholte et al., 2009). The reliance on the spatial layout of image structure agrees with the

Table 3: Architecture ablations on the four design axes (subject 1, top-1 %, mean ± std, one component changed at a time). Bold rows are the default PAE, the reference of every group. The columns are the default patch size (ps=16; 96.00 ± 0.87) and the finer patch size (ps=8; 97.17 ± 0.29, the best of Table 8). The complete table including the flat recipe axes is given in Appendix G.

Axis	Variant	ps=16	ps=8
token interaction	**none (PAE)**	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	token mixing	48.17 ± 1.15	48.17 ± 5.06
	attention (ViT)	44.67 ± 6.11	61.33 ± 15.22
spatial pooling	**Weighted Pooling (PAE)**	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	post-head pooling	94.67 ± 2.31	95.50 ± 0.87
	post-pooling ResBlock	96.17 ± 0.76	96.50 ± 0.50
	concat pooling	52.00 ± 3.91	41.50 ± 6.87
	fixed uniform average	59.00 ± 4.58	64.17 ± 1.26
depth L	0	36.00 ± 0.87	67.50 ± 1.00
	1	94.17 ± 1.04	97.17 ± 0.29
	2	95.33 ± 1.76	97.67 ± 0.76
	3 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	4	93.67 ± 1.04	97.00 ± 0.87
	6	93.67 ± 1.76	96.17 ± 1.04
width d	48	95.17 ± 1.89	94.83 ± 0.29
	64	94.50 ± 1.50	97.00 ± 0.87
	96	94.33 ± 2.08	97.50 ± 1.32
	128	95.67 ± 0.29	97.17 ± 1.04
	192 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	256	94.83 ± 1.44	96.50 ± 1.00
	384	95.17 ± 0.58	96.83 ± 0.58

spatial envelope representation of rapid scene recognition (Oliva & Torralba, 2001) and with the transition from image structure to scene category in evoked activity (Groen et al., 2013). The foveal weighting of the pooling maps mirrors retinal photoreceptor density, which drops sharply outside the fovea (Curcio et al., 1990), and cortical magnification, which assigns central vision a proportionally larger representation in early visual cortex (Cowey & Rolls, 1974; Dumoulin & Wandell, 2008). Fine texture contributes little against the spatial layout of image structure. The two features differ in where the visual system can resolve them. Fine detail is resolved only at the fovea, whereas coarse spatial structure is preserved in the periphery as summary statistics (Balas et al., 2009; Rosenholtz, 2016), and in this retrieval task the coarse structure dominates, since the encoders are robust to the smoothing that suppresses fine texture (§3.3.3). Color contributes little against luminance, consistent with fast evoked responses dominated by luminance (Derrington et al., 1984).

5 Conclusion

We asked what image features matter in EEG-to-image retrieval, and answered with the Patchwise Aggregation Encoder (PAE), a patchwise image encoder trained from scratch. On the Things-EEG benchmark, the PAE reaches 96.9% top-1 and 99.7% top-5 accuracy across the 10 subjects (patch size 8), a new state of the art. Replacing the pixel input of the PAE with 15 low-level statistics of each patch yields the purely statistical S-PAE, which is even stronger, at 97.6% top-1 and 99.9% top-5. The attribution analyses show that luminance contrast and edge statistics dominate the feature space and that the learned pooling weights concentrate on the image center, consistent with foveal dominance in early visual processing. Under controlled suppression, the PAE and the S-PAE rely primarily on the spatial layout of the patches and are nearly achromatic, and the architecture ablations show that adding interaction between the patch tokens substantially reduces accuracy. The retrieval is thus carried by the low-level statistics of local patches, computed independently and combined by learned spatial weights, which makes decoding compact and interpretable by construction. The answer to the opening question is that luminance contrast, edge statistics, and the spatial layout of an image matter more than its fine texture and color.

Reproducibility statement

The full training and evaluation code, the pretrained checkpoints behind the sub-01 results, and step-by-step reproduction instructions are available as an anonymized archive at https://figshare.com/s/c92d8907996cd2e0a719.

Ethics statement

This study uses only the publicly released Things-EEG dataset, whose data collection protocol was approved by its original authors. We performed no new recordings and analyze no identifiable personal data. The work targets basic research on visual decoding and has no foreseen clinical or surveillance application. The encoder operates on short, averaged responses to experimental stimuli, and its reliance on low-level image statistics shows how limited the decodable signal is. All results are reported at the subject level on a standard benchmark, and we adhere to the ICLR Code of Ethics.

AI use statement

Generative AI tools were used for text polishing only. All other aspects of this work, including the research questions, experimental design, implementation, data analysis, and scientific conclusions, were carried out by the authors. We have reviewed and edited all AI-polished text and take full responsibility for the content of this work.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 9(12):13, 2009. doi: 10.1167/9.12.13.

Yohann Benchetrit, Hubert Banville, and Jean-Remi King. Brain decoding: Toward real-time reconstruction of visual perception. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.19812.

Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations (ICLR)*, 2019.

Tom Burgert, Oliver Stoll, Paolo Rota, and Begüm Demir. Imagenet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.276769.

Chi-Sheng Chen and Chun-Shu Wei. Mind's eye: Image recognition by eeg via multimodal similarity-keeping contrastive learning. *arXiv preprint arXiv:2406.16910*, 2024.

Hongzhou Chen, Lianghua He, Yihang Liu, Longzhen Yang, Shaohua Shang, and MengChu Zhou. Visual neural decoding via improved visual-EEG semantic consistency. *arXiv preprint arXiv:2408.06788*, 2026.

Yebin Choi, Woohyeok Choi, Jun-Mo Kim, Songha Kim, and Tae-Eui Kam. Learning from perceptual to semantic visual representations of the brain via multi-level EEG-to-image alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Alan Cowey and Edmund T. Rolls. Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, 21(5):447–454, 1974. doi: 10.1007/BF00237164.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision (ECCV)*, 2004.

Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson. Human photoreceptor topography. *Journal of Comparative Neurology*, 292(4):497–523, 1990.

Andrew M. Derrington, John Krauskopf, and Peter Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):241–265, 1984.

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

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain–visual–linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Serge O. Dumoulin and Brian A. Wandell. Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660, 2008. doi: 10.1016/j.neuroimage.2007.09.034.

Teng Fei, Abhinav Uppal, Ian Jackson, Srinivas Ravishankar, David Wang, and Virginia R de Sa. Perceptogram: Reconstructing visual percepts from eeg. *arXiv e-prints*, pp. arXiv–2404, 2024.

Matteo Ferrante, Tommaso Boccato, Stefano Bargione, and Nicola Toschi. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. *Computers in Biology and Medicine*, 2024. doi: 10.1016/j.compbiomed.2024.108701.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987. doi: 10.1364/JOSAA.4.002379.

Sennay Ghebreab, H. Steven Scholte, Victor A. F. Lamme, and Arnold W. M. Smeulders. A biologically plausible model for rapid natural scene identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 22, pp. 629–637, 2009.

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.

Iris I. A. Groen, Sennay Ghebreab, Hielke Prins, Victor A. F. Lamme, and H. Steven Scholte. From image statistics to scene gist: Evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, 33(48):18814–18824, 2013.

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

Vijay Jayawant Harkare and Lakshmi Subramanian. BrainAlign: Leveraging EEG foundation models for symmetric, interpretable alignment with visual representations. In *Workshop on Data on the Brain and Mind (Findings Track), Neural Information Processing Systems (NeurIPS)*, 2025.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

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

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Chendong Qin, Xiao Lin, 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 (NeurIPS)*, 2025.

Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, 2009.

International Telecommunication Union. Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. Technical Report BT.601-7, ITU-R, 2011.

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. Hyfi: Hyperbolic feature interpolation for brain-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5575–5583, 2026.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 2169–2178, 2006.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Boang Li, Hui Cao, Badong Chen, Tao Wang, and Jie Zhang. EveryBrain: Generate EEG responses from images for specified individuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1893–1906, 2026a.

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

Dongyang Li, Kunpeng Xie, Mingyang Wu, Yiwei Kong, Jiahua Tang, Haoyang Qin, Chen Wei, and Quanying Liu. MindPilot: Closed-loop visual stimulation optimization for brain modulation with EEG-guided diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-MCRL: Neural multimodal contrastive representation learning for EEG-based visual decoding. *arXiv preprint arXiv:2412.17337*, 2025.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. ViEEG: Hierarchical visual neural representation for EEG brain decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for EEG decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 207(1167):187–217, 1980. doi: 10.1098/rspb.1980.0020.

Jiahe Meng, Weiming Zeng, Yueyang Li, Bo Chai, Hongjie Yan, Zhiguo Zhang, Wai Ting Siok, and Nizhuan Wang. STAMBRIDGE: Spectral-temporal amplitude-aware mid-feature bridge for EEG visual decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. doi: 10.1023/A:1011139631724.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. doi: 10.1038/381607a0.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nona Rajabi, Antonio H. Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-aligned image models improve visual decoding from the brain. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ruth Rosenholtz. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, 2:437–457, 2016. doi: 10.1146/annurev-vision-082114-035737.

Daniel L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5 (4):517–548, 1994. doi: 10.1088/0954-898X_5_4_006.

H. Steven Scholte, Sennay Ghebreab, Lourens Waldorp, Arnold W. M. Smeulders, and Victor A. F. Lamme. Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, 9(4):29, 2009. doi: 10.1167/9.4.29.

Enze Shi, Huawen Hu, Yincheng Yao, Sigang Yu, and Shu Zhang. EVA: Achieving discriminative and semantically faithful multi-scale EEG-vision alignment, 2025a. OpenReview manuscript.

Enze Shi, Huawen Hu, Qilong Yuan, Kui Zhao, Sigang Yu, and Shu Zhang. BrainAlign: EEG-vision alignment via frequency-aware temporal encoder and differentiable cluster assigner. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025b.

Alvy Ray Smith. Color gamut transform pairs. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '78)*, pp. 12–19, 1978. doi: 10.1145/800249.807359.

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

Yonghao Song, Chunfeng Song, Yijun Wang, and Xiaorong Gao. BrainAE: Alignment-driven autoencoder for bidirectional visual encoding and decoding, 2025a. OpenReview manuscript.

Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from EEG with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

Cataldo Spampinato, Simone Palazzo, Ioannis Kavasidis, Daniele Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6809–6817, 2017.

Yueming Sun and Long Yang. Spatial-functional awareness transformer-based graph archetype contrastive learning for decoding visual neural representations from eeg. *arXiv preprint arXiv:2509.24761*, 2025.

YuYan Tan, Xuesong Yin, Xu Chen, Guodao Zhang, and Chenhao Xu. ReVIS: Reconstructing the individual visual perception via EEG signal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

J. Hans van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 265(1394):359–366, 1998. doi: 10.1098/rspb.1998.0303.

Jiyuan Wang, Li Zhang, Haipeng Lin, Qile Liu, Gan Huang, Ziyu Li, Zhen Liang, and Xia Wu. NeuroCLIP: Brain-inspired prompt tuning for EEG-to-image multimodal contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. MB2C: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, pp. 1–9, 2024. doi: 10.1145/3664647.3681292.

Fengjie Wu, Zhaolong Lin, Bangzheng Yin, Lihan Zhang, and Bohua Chen. EEG-HMLearner: Hybrid learning for visual stimulus decoding by extracting multimodal information from EEG signals. *Biomedical Signal Processing and Control*, 122, 2026.

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

Xin Xiang, Wenhui Zhou, Haonan Zhu, Yunrui Li, Guojun Dai, and Lili Lin. EEG-driven natural image reconstruction with regional semantic awareness. *Pattern Recognition*, 172:112589, 2026.

Chengjian Xu, Yonghao Song, Zelin Liao, Haochuan Zhang, Qiong Wang, and Qingqing Zheng. UMind: A unified multitask network for zero-shot M/EEG visual decoding. *arXiv preprint arXiv:2509.14772*, 2025a.

Ganxi Xu, Zhao-Rong Lai, Yuting Tang, Yonghao Song, Guoxu Zhou, Jian Zhu, Jinyi Long, et al. Image-to-brain signal generation for visual prosthesis with clip guided multimodal diffusion models. *arXiv preprint arXiv:2509.00787*, 2025b.

Haopei Xu and Ni Lin. Neurovista: A bidirectional masked cross-modal fusion network for robust eeg-to-image decoding. *Neural Networks*, pp. 108297, 2025.

Junkai Yang and Arno Onken. On the role of low-level visual features in EEG-based image reconstruction. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2025.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Kaifan Zhang, Lihuo He, Junjie Ke, Yuqi Ji, Lukun Wu, Lizi Wang, and Xinbo Gao. CognitionCapturerPro: Towards high-fidelity visual decoding from EEG/MEG via multi-modal information and asymmetric alignment. *arXiv preprint arXiv:2603.12722*, 2026.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. *arXiv preprint arXiv:2602.07495*, 2026.

Ding Zhu, Jan Bieger, Gary Garcia Molina, and Ronald M. Aarts. A survey of stimulation methods used in SSVEP-based BCIs. *Computational Intelligence and Neuroscience*, 2010:702357, 2010.

Shuqi Zhu, Ziyi Ye, Yi Zhong, Qingyao Ai, Yujia Zhou, and Yiqun Liu. Brain image reconstruction with retrieval-augmented diffusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, 2025. doi: 10.1145/3726302.3729909.

Yanan Zhu, Ziwei Xiang, Jiamin Wu, Jinyang Guo, Hongyuan Zhang, Chunfeng Song, Hongjian Fang, Qihao Zheng, Yufei Guo, Xuelong Li, and Xianglong Liu. BRENA: Brain-inspired hierarchical neural alignment framework for visual decoding from EEG signals. In *International Conference on Learning Representations (ICLR)*, 2026.

A Limitations

The scope of these conclusions has boundaries. The results are obtained per subject on one dataset and one protocol, with test repetitions averaged, and the attribution covers the image branch. Extending these analyses to other datasets and further modalities is the natural next step.

B Dataset details

All experiments in this paper use public datasets of evoked responses to natural object images from the THINGS database under a shared zero-shot protocol, in which the test concepts are disjoint from the training concepts.

Things-EEG. Things-EEG (Gifford et al., 2022) records EEG from 10 subjects who view natural object images under rapid serial visual presentation. The stimuli are split into 1,654 training concepts with 10 images each and 200 held-out test concepts with 1 image each. Each training image is shown 4 times and each test image is shown 80 times, so that one subject provides 16,540 training image conditions and 200 test image conditions. Images are presented centrally at 7° visual angle for 100 ms each with a stimulus onset asynchrony of 200 ms, and subjects perform an orthogonal target detection task while keeping central fixation. EEG is recorded with a 64-channel EASYCAP cap arranged according to the international 10–10 system at a sampling rate of 1,000 Hz, with online filtering between 0.1 and 100 Hz and online referencing to the Fz electrode. We use the preprocessed data of the VED framework (Liu et al., 2026), which averages the repetitions to improve the signal-to-noise ratio, resamples to 250 Hz, and keeps the 1 s post-stimulus window, so that each sample is a 63×250 array (C=63 channels, T=250 time samples in §2), giving 16,540 training samples and 200 test samples per subject. A random 5% of the training samples serve as a validation set for model selection, following the protocol of VED, and retrieval uses each of the 200 test EEG samples as a query against the 200 candidate images.

Things-MEG. Things-MEG (Hebart et al., 2023) records magnetoencephalography from 4 subjects with a 271-channel whole-head system while each image is presented for 500 ms. The zero-shot split mirrors Things-EEG, with 1,654 training concepts at 12 images each shown once and 200 held-out test concepts at 1 image each shown 12 times. Preprocessing filters the signals between 0.1 and 100 Hz, resamples to 200 Hz, and segments baseline-corrected epochs around the stimulus onset; the test repetitions are averaged. The dataset is used in the companion study of Appendix H.

C Experimental setup

Contrastive objective. Each optimization step draws a mini-batch of B image–EEG pairs $\{(\mathbf{V}_n, \mathbf{X}_n)\}_{n=1}^{B}$ and computes both embeddings, $\mathbf{z}_n = f_\theta(\mathbf{V}_n)$ and $\mathbf{y}_n = \Phi_\phi(\mathbf{X}_n)$. We use the symmetric CLIP-style contrastive loss (Radford et al., 2021) with the normalization protocol of VED (Liu et al., 2026): only the image side is ℓ_2-normalized at training time, the EEG side is left unnormalized, and the temperature is fixed at $\tau = 1$ (no learnable logit scale):

$$\tilde{\mathbf{z}}_n = \frac{\mathbf{z}_n}{\|\mathbf{z}_n\|_2}, \qquad \hat{\mathbf{y}}_n = \mathbf{y}_n, \qquad \mathbf{M}_{nm} = \hat{\mathbf{y}}_n^\top \tilde{\mathbf{z}}_m, \tag{9}$$

where $\mathbf{M} \in \mathbb{R}^{B\times B}$ collects all pairwise matching logits in the batch, rows indexed by the EEG side and columns by the image side. The loss averages the cross-entropy over the two retrieval directions,

$$\mathcal{L}_{\text{eeg}} = -\frac{1}{B}\sum_{n=1}^{B} \log \frac{\exp(\mathbf{M}_{nn})}{\sum_{m=1}^{B}\exp(\mathbf{M}_{nm})}, \qquad \mathcal{L}_{\text{img}} = -\frac{1}{B}\sum_{m=1}^{B} \log \frac{\exp(\mathbf{M}_{mm})}{\sum_{n=1}^{B}\exp(\mathbf{M}_{nm})}, \tag{10}$$

$$\mathcal{L} = \tfrac{1}{2}\big(\mathcal{L}_{\text{eeg}} + \mathcal{L}_{\text{img}}\big), \tag{11}$$

The first loss normalizes each row of $\mathbf{M}$ by a softmax over the images and the second each column by a softmax over the EEG samples, so that EEG→image and image→EEG matching are both enforced; both branches are trained end-to-end from scratch with this single objective.

Optimization and model selection. We use AdamW with learning rate 10^{-3} on all parameters of both branches, batch size B=1024, for 200 epochs. Per subject, 5% of the training images are held out as a validation set; the checkpoint with the best validation top-1 retrieval is selected (§3.1). Every training configuration is run with three random seeds, and results are reported as mean ± std over seeds.

Table 4 collects every constant of the default architecture and training protocol. The PAE is reported at two patch sizes: the default w=16 (N=196) and the finer w=8 (N=784); the patch size changes only w and N, and every other constant is unchanged. The S-PAE is trained at w=16.

Table 4: Architecture and optimization constants (defaults). The architecture is the PAE at the default w=16; 1.57 M parameters in total.

Symbol	Meaning	Value
$H \times W$	input image size (pixels)	224×224
w	patch size (pixels)	16 (PAE also 8)
N	patches per image	196 (14×14); 784 at w=8
C, T	EEG input	63 channels, 250 time samples
d	token / feature width	192
h	MLP hidden width	1024
L	number of ResBlocks	3
d_z	embedding (alignment) width	1024
p	dropout rate (MLP hidden, training only)	0.85
τ	contrastive temperature (fixed)	1
B	batch size	1024
	optimizer	AdamW, learning rate 10^{-3}
	training epochs / random seeds	200 / 3

Parameter budget. At the default w=16, the image encoder has 1.57 M parameters. The patch embedding contributes $3w^2d \approx 0.15$ M, the L ResBlocks $2Ldh \approx 1.18$ M, the alignment head $dd_z \approx 0.20$ M, and the pooling weights $Nd \approx 0.04$ M, with the LayerNorms accounting for the remainder.

D THE 15 LOW-LEVEL PATCH STATISTICS

Why these statistics. The choice of statistic families follows what is known about early visual processing. Luminance and contrast are not only canonical image statistics (Field, 1987; Ruderman, 1994). The contrast values of natural images follow Weibull distributions whose two parameters summarize local contrast and texture, and these parameters correlate strongly with brain responses, from fMRI BOLD in early visual cortex (Scholte et al., 2009) to evoked EEG potentials (Groen et al., 2013), and suffice to drive biologically plausible models of rapid natural scene identification (Ghebreab et al., 2009). Gradient, frequency, and orientation statistics describe the local spatial structure to which cortical simple-cell receptive fields are selective (Field, 1987; Olshausen & Field, 1996). Such statistics are extracted by the first, largely feedforward pass of visual processing, which supports rapid scene categorization within about 150 ms of stimulus onset (Thorpe et al., 1996); the S-PAE deliberately restricts itself to these families.

The low-level patch statistics. With $\mathrm{L}(u,v) = 0.299\,R(u,v) + 0.587\,G(u,v) + 0.114\,B(u,v)$ the BT.601 gray image (International Telecommunication Union, 2011), whose intensities lie in

$[0, 1]$, $\nabla \mathrm{L} = (g_h, g_v)$ its gradient, with g_h and g_v the horizontal and vertical derivatives, computed by central finite differences at unit pixel spacing and by one-sided finite differences at the image border, $\|\nabla \mathrm{L}\| = (g_h^2 + g_v^2)^{1/2}$ its magnitude, $\mathbf{1}[\cdot]$ the indicator function, and $\langle \cdot \rangle$ the spatial average over the w^2 pixels of a patch, the 15 statistics are the following.

- **Luminance.** *Brightness* $\mu_{\mathrm{L}} = \langle \mathrm{L} \rangle$, the mean gray level of the patch (Hyvärinen et al., 2009). *Contrast* $\sigma_{\mathrm{L}} = \left(\langle \mathrm{L}^2 \rangle - \langle \mathrm{L} \rangle^2 \right)^{1/2}$, the standard deviation of the gray levels, i.e. the local reversal contrast, a canonical statistic of natural images (Ruderman, 1994).
- **Edge.** *Gradient energy* $E_\nabla = \langle g_h^2 + g_v^2 \rangle$, the total energy of the edge map $\|\nabla \mathrm{L}\|$, i.e. the total local contrast carried by intensity gradients. Derivative statistics of this form are the classical description of local spatial structure (Field, 1987; van Hateren & van der Schaaf, 1998). *Edge fraction* $f_{\mathrm{e}} = \langle \mathbf{1}[\|\nabla \mathrm{L}\| > 0.12] \rangle$, the fraction of pixels whose gradient magnitude exceeds a fixed threshold, i.e. the area of strong edges, as in edge detection with a gradient threshold (Canny, 1986). The two statistics are two summaries of the same edge map, its total energy and the area of its strong edges.
- **Orientation.** *Orientation energies* $e_\theta = \langle (g_h \cos\theta + g_v \sin\theta)^2 \rangle \,/\, \langle \|\nabla \mathrm{L}\|^2 \rangle$ for $\theta = 0°$, $45°$, $90°$, and $135°$, the gradient energy projected onto each of the four orientations, normalized so that the four values sum to two, forming an orientation histogram of the patch that matches the orientation statistics of natural images and of oriented cortical receptive fields (Field, 1987; Olshausen & Field, 1996). *Horizontal / vertical log-ratio* $\rho_{h/v} = \log_{10} \left(\langle g_h^2 \rangle / \langle g_v^2 \rangle \right)$, positive for horizontally dominated patches, negative for vertically dominated ones, and zero for isotropic ones (Field, 1987).
- **Frequency.** With $\nabla^2 \mathrm{L} = 4\mathrm{L} - \sum_{u'v' \sim uv} \mathrm{L}(u', v')$ the discrete Laplacian over the four neighboring pixels and $\mathrm{L}_{2\times2}$ the image after non-overlapping 2×2 block averaging, we define *Laplacian energy* $E_\Delta = \langle (\nabla^2 \mathrm{L})^2 \rangle$, squared curvature, which is extremely sensitive to fine texture; the Laplacian is a classical operator for edge detection (Marr & Hildreth, 1980). *Fine-frequency fraction* $f_{\mathrm{hf}} = \langle (\mathrm{L} - \mathrm{L}_{2\times2})^2 \rangle \,/\, \langle (\mathrm{L} - \bar{\mathrm{L}})^2 \rangle$, with $\bar{\mathrm{L}}$ the average gray level over the full image rather than the patch, the fraction of the patch's intensity energy about the image mean that is removed by 2×2 block averaging, i.e. the energy in the finest spatial frequencies, the band that carries most of the contrast in natural images (Field, 1987; Ruderman, 1994).
- **Color.** *Channel means* $\mu_R = \langle R \rangle$, $\mu_G = \langle G \rangle$, $\mu_B = \langle B \rangle$, the average color of the patch (Hyvärinen et al., 2009). *Saturation* $\langle s \rangle$, with $s = \max(R, G, B) - \min(R, G, B)$ per pixel, i.e. the standard max–min saturation, zero for gray pixels and near one for saturated colors (Smith, 1978).

Every statistic is standardized with the moments of the training set before use; together these forms are the classical vocabulary of natural image statistics (Hyvärinen et al., 2009).

E Sanity check

Every retrieval result in this work is evaluated on the 200-way task, whose ideal chance level is $1/200 = 0.50\%$ at top-1, 1.50% at top-3, and 2.50% at top-5. Following Song et al. (Song et al., 2024), we additionally provide an experimental chance level. The EEG–image pairing of the 200 test pairs, with all 80 repetitions averaged, is randomly shuffled, and the standard 200-way retrieval is evaluated against the shuffled pairing instead of the true one. The shuffle is repeated 100 times and averaged over the three training seeds of the default PAE of Table 2 on subject 1, without retraining. The default PAE is also retrained from scratch under a consistently mismatched pairing, in which the image lists of the train, validation, and test splits are each permuted with a fixed random permutation when the data are loaded while the EEG rows keep their order, so that training and testing operate on the same wrong pairing, under the otherwise identical recipe over three seeds. Table 5 reports the ideal chance level, the experimental chance, and the PAE retrained under the mismatched pairing.

F Supplementary analyses

Table 5: Experimental chance level of the 200-way retrieval on subject 1, mean ± std over three seeds. Following the sanity check of Song et al. (Song et al., 2024), the EEG–image pairing of the 200 test pairs, with all 80 repetitions averaged, is randomly shuffled 100 times per seed, and the standard 200-way retrieval is evaluated against the shuffled pairing. The last row retrains the default PAE from scratch with the EEG–image pairing consistently permuted in the train, validation, and test splits when the data are loaded. The ideal chance is $1/200$ at top-1, $3/200$ at top-3, and $5/200$ at top-5.

	top-1 (%)	top-3 (%)	top-5 (%)
Ideal chance	0.50	1.50	2.50
Experimental chance	0.51 ± 0.01	1.45 ± 0.03	2.52 ± 0.10
Mismatched training	0.83 ± 1.04	2.50 ± 0.50	3.83 ± 1.04

F.1 Zeroing drop analysis

The zeroing analysis measures the contribution of each dimension of the learned embedding. Each dimension g_d of the global feature $\mathbf{g}$ of Eq. (5), the output of Weighted Pooling, is set to zero at test time, and the 200-way retrieval is recomputed with the EEG as query. The EEG embeddings are fixed throughout, no retraining takes place, and the intervention is applied once per dimension, giving 192 recomputes. The analysis uses the PAE trained on subject 1. Of the 192 dimensions, 82 show a non-zero drop, 55 of them exceed 0.5%, and the largest drop is 23.5%. Fig. 6 shows the 192 drops sorted from the largest to the smallest.

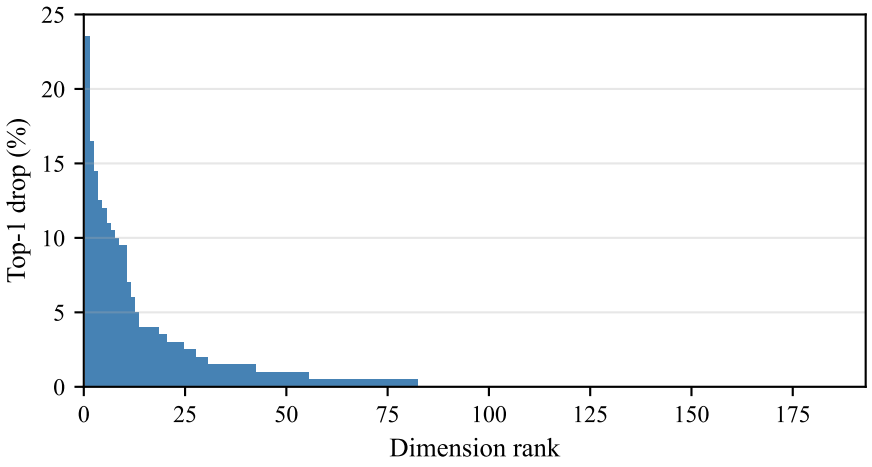

Figure 6: **Zeroing drop of each dimension of the global feature.** Each bar is one of the 192 dimensions, sorted from the largest drop to the smallest. The drop is the decrease in top-1 accuracy when the dimension is set to zero at test time and the 200-way retrieval is recomputed with the EEG as query.

F.2 Pooling maps

This appendix documents the pooling maps behind §3.3.1 and defines the centroid eccentricity and the spatial spread used to summarize a map, together with the two null distributions referenced there. Fig. 7 shows all 192 individual maps of the PAE trained on subject 1 in the leaf order of the clustering dendrogram, Fig. 8 the mean map of each cluster, and Fig. 9 the centroid eccentricity and spatial spread of every map against the two null distributions.

Map moments and nulls. Each pooling map $\mathbf{A}_{:j}$ of Eq. (5) assigns a nonnegative weight to each of the 196 grid positions $\mathbf{p} = (r, c)$ with $r, c \in \{0, \ldots, 13\}$ and sums to one over positions, and the grid center is $\mathbf{c} = (6.5, 6.5)$ in grid units. The centroid $\boldsymbol{\mu}_j$ is the first moment $\sum_{\mathbf{p}} \mathbf{A}_{\mathbf{p}j}\, \mathbf{p}$, and the

centroid eccentricity $\mathrm{ecc}_j = \|\boldsymbol{\mu}_j - \mathbf{c}\|$ is the distance of the center of mass from the grid center, so a centered map has ecc 0. The spatial spread summarizes the dispersion of the map around its centroid and is the analogue of the Gaussian size parameter of the population receptive field model (Dumoulin & Wandell, 2008),

$$\mathrm{sd}_j = \left(\tfrac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}_j\right)^{1/2}, \qquad \boldsymbol{\Sigma}_j = \sum_{\mathbf{p}} \mathbf{A}_{\mathbf{p}j}\,(\mathbf{p}-\boldsymbol{\mu}_j)(\mathbf{p}-\boldsymbol{\mu}_j)^\top. \tag{12}$$

The uniform map assigns $1/196$ to every position and has ecc 0 and sd 4.03. It is the pooling map at initialization, because the pooling parameters $\mathbf{P}$ of Eq. (5) are initialized to zero and the softmax over them is exactly uniform. Two null distributions calibrate these measures. The permutation null shuffles the 196 weights of a map over the grid positions, destroying the spatial arrangement while keeping the value distribution, with 400 shuffles per map. The Dirichlet(1) null draws 2×10^5 random maps, each a set of 196 nonnegative weights summing to one, with all such maps equally likely, and serves as the spread reference in Fig. 9.

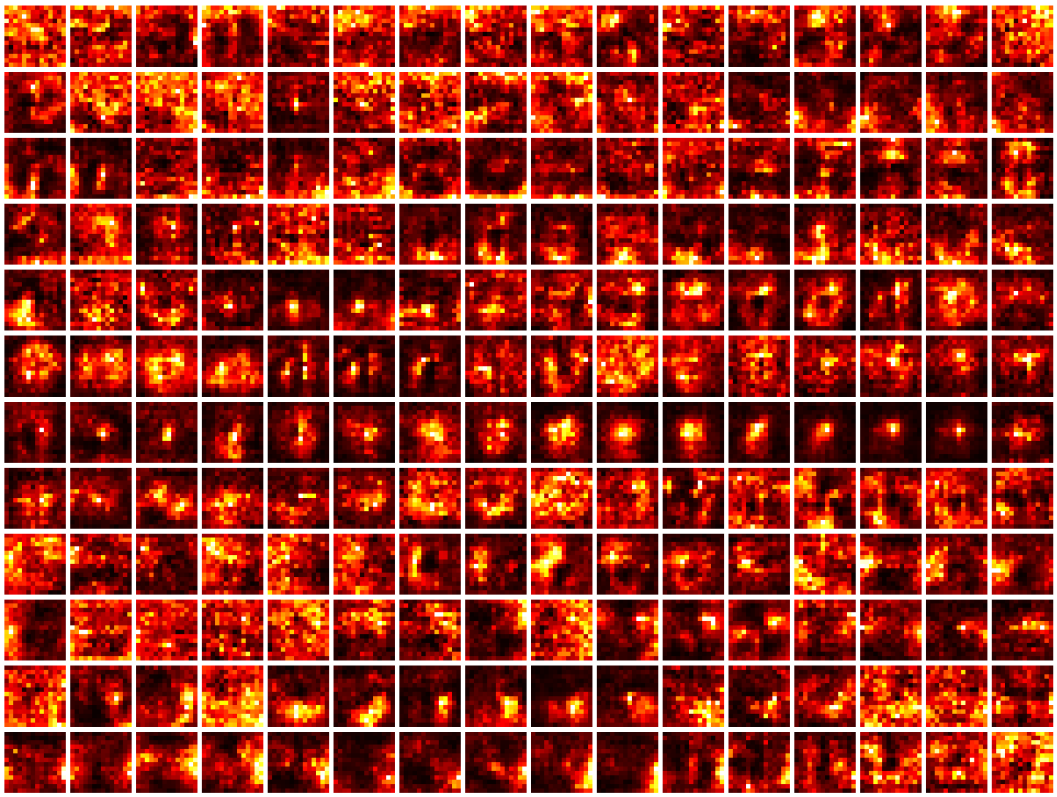

Figure 7: **All 192 learned pooling maps (subject 1).** Every per-dimension weighting map $\mathbf{A}_{:j}$ of Eq. (5).

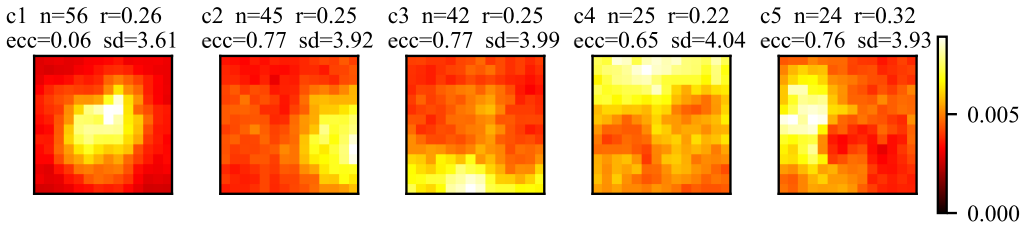

Figure 8: **The learned pooling maps form five clusters.** The 192 per-dimension weighting maps (subject 1), clustered by map profile. Shown is the mean map of each cluster, annotated with the cluster size n, the mean correlation within the cluster r, and the centroid eccentricity and spatial spread of the cluster.

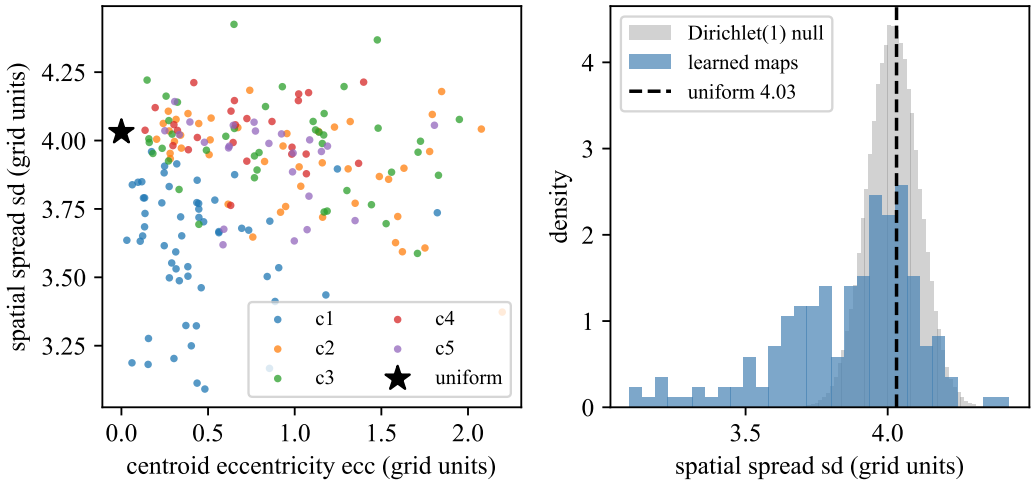

Figure 9: **Pooling-map center and dispersion.** Left: every one of the 192 per-dimension maps in the eccentricity–spread plane, colored by cluster; the star marks the exactly uniform map (ecc 0, sd 4.03). Right: the spread distribution of the learned maps against a Dirichlet(1) null of random normalized maps.

F.3 S-PAE INPUT SUBSETS

This appendix documents the trained input subsets behind §3.3.2. Each subset serves as the only input of the S-PAE, the input dimension of the first linear map is adjusted, and everything downstream is unchanged. Every configuration is trained with the recipe of §3.1 on subject 1 over three seeds. Table 6 lists all trained subsets in five groups. The first group is the full set, the default S-PAE with all 15 statistics as the input. The second group is the chosen path of the greedy backward elimination, which starts from the full set and removes in each round the statistic whose removal costs least accuracy, retraining the S-PAE after every removal. The third group is the forward accumulation, which adds the statistics one at a time from the strongest to the weakest according to Fig. 4 and retrains the S-PAE at every prefix. The fourth group retrains the S-PAE with all members of one statistic family of Table 1 as the input. The last group trains the S-PAE with a single statistic as the only input, the same runs as those plotted in Fig. 4.

G ABLATIONS

All cells in this section use subject 1, the same recipe as §3.1, and change exactly one component at a time relative to the default PAE or, for the last table, the default S-PAE. Every cell is trained

Table 6: S-PAE accuracy for the trained subsets of the 15 input statistics (subject 1, test top-1 %, mean ± std). Each row is one trained subset, and k is the number of statistics in the subset. The rows are grouped as the full set, the greedy backward elimination chosen path, the forward accumulation, the statistic families, and the single statistics. Each input statistic is named in its column for the rows that use it, and a blank cell means the statistic is not used.

		luminance		edge	orientation					frequency		color				
#	k	*input statistics (named when used)*														top-1 (%)
Full set																
1	15	μ_{L}	σ_{L}	E_{∇}	f_{e}	e_{0°	e_{45°	e_{90°	e_{135°	$\rho_{h/v}$	E_{Δ}	f_{hf}	$\langle s \rangle$	μ_R	μ_G	μ_B	97.00 ± 1.32
Greedy backward elimination, chosen path																
2	14	μ_{L}	σ_{L}	E_{∇}	f_{e}	e_{0°	e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}	$\langle s \rangle$	μ_R	μ_G	μ_B	98.33 ± 0.76
3	13	μ_{L}	σ_{L}	E_{∇}	f_{e}	e_{0°	e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}	$\langle s \rangle$		μ_G	μ_B	98.17 ± 0.58
4	12	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}	$\langle s \rangle$		μ_G	μ_B	98.17 ± 0.29
5	11	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}			μ_G	μ_B	98.33 ± 0.29
6	10	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°			E_{Δ}	f_{hf}			μ_G	μ_B	98.17 ± 0.76
7	9	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°				f_{hf}			μ_G	μ_B	97.17 ± 0.29
8	8	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°							μ_G	μ_B	98.50 ± 0.50
9	7	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°							μ_G		97.50 ± 1.00
10	6	μ_{L}	σ_{L}	E_{∇}	f_{e}			e_{90°							μ_G		96.50 ± 0.00
11	5	μ_{L}	σ_{L}	E_{∇}				e_{90°							μ_G		97.17 ± 1.04
12	4	μ_{L}	σ_{L}	E_{∇}				e_{90°									96.83 ± 1.04
13	3	μ_{L}	σ_{L}	E_{∇}													95.17 ± 0.29
14	2	μ_{L}	σ_{L}														90.33 ± 1.04
Forward accumulation																
15	1		σ_{L}														65.50 ± 4.00
16	2		σ_{L}	E_{∇}													79.17 ± 0.76
17	3		σ_{L}	E_{∇}	f_{e}												80.33 ± 1.53
18	4	μ_{L}	σ_{L}	E_{∇}	f_{e}												93.83 ± 1.04
19	5	μ_{L}	σ_{L}	E_{∇}	f_{e}						E_{Δ}						94.17 ± 1.61
20	6	μ_{L}	σ_{L}	E_{∇}	f_{e}						E_{Δ}				μ_G		96.00 ± 1.50
21	7	μ_{L}	σ_{L}	E_{∇}	f_{e}						E_{Δ}			μ_R	μ_G		95.50 ± 1.50
22	8	μ_{L}	σ_{L}	E_{∇}	f_{e}						E_{Δ}	f_{hf}		μ_R	μ_G		96.00 ± 0.00
23	9	μ_{L}	σ_{L}	E_{∇}	f_{e}						E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	95.67 ± 1.26
24	10	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°				E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	96.00 ± 0.50
25	11	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°		e_{135°		E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	95.17 ± 0.76
26	12	μ_{L}	σ_{L}	E_{∇}	f_{e}		e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	97.17 ± 1.89
27	13	μ_{L}	σ_{L}	E_{∇}	f_{e}	e_{0°	e_{45°	e_{90°	e_{135°		E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	97.33 ± 0.58
28	14	μ_{L}	σ_{L}	E_{∇}	f_{e}	e_{0°	e_{45°	e_{90°	e_{135°	$\rho_{h/v}$	E_{Δ}	f_{hf}		μ_R	μ_G	μ_B	97.83 ± 0.58
Statistic families																
29	2			E_{∇}	f_{e}												66.17 ± 2.75
30	5					e_{0°	e_{45°	e_{90°	e_{135°	$\rho_{h/v}$							26.67 ± 5.69
31	2										E_{Δ}	f_{hf}					44.67 ± 1.89
32	4												$\langle s \rangle$	μ_R	μ_G	μ_B	59.17 ± 3.75
Single statistic																
33	1	μ_{L}															35.00 ± 3.04
34	1			E_{∇}													55.67 ± 0.29
35	1				f_{e}												48.00 ± 1.50
36	1					e_{0°											6.17 ± 2.36
37	1						e_{45°										8.50 ± 0.50
38	1							e_{90°									6.67 ± 1.04
39	1								e_{135°								7.50 ± 2.65
40	1									$\rho_{h/v}$							5.00 ± 1.32
41	1										E_{Δ}						33.33 ± 0.58
42	1											f_{hf}					21.83 ± 2.57
43	1												$\langle s \rangle$				4.67 ± 2.02
44	1													μ_R			22.17 ± 1.44
45	1														μ_G		28.33 ± 1.53
46	1															μ_B	18.00 ± 0.87

at the default patch size ps=16 and the finer ps=8. Table 8 trains the PAE at every patch size and covers all divisors of 224. Table 9 trains the PAE at batch sizes 256, 512, 2048, and the default 1024.

Table 10 varies the input, the depth, and the pooling of the S-PAE. Table 7 summarizes the seven axes. The variants of these axes are the following.

- **Token interaction.** The token mixing variant inserts two learned cross-patch sub-layers between the residual blocks, in the style of ResMLP (Touvron et al., 2021), and the attention variant processes all patch tokens jointly through the full self-attention of a ViT of comparable size (Dosovitskiy et al., 2021).
- **Spatial pooling.** Post-head pooling applies the alignment head to each patch feature before the pooling. Post-pooling ResBlock inserts one residual block between the pooling and the head. Concat pooling concatenates all patch features for a global linear head. The fixed uniform average replaces the Weighted Pooling with the mean over the patch features.
- **Depth** L**.** The number of residual blocks, which are removed entirely at $L{=}0$.
- **Width** d**.** The token width shared by the Patch Embedding, the residual blocks, and the pooling.
- **Activation.** The nonlinearity inside the residual blocks.
- **Alignment width** d_z**.** The output dimension of the alignment head.
- **MLP dropout** p**.** The dropout rate on the hidden activations of the residual blocks.
- **Patch size.** The side length of the square patches, which determines the number of tokens.
- **Batch size.** The number of training pairs per optimization step.

H SWAPPING THE IMAGE ENCODER

This appendix reports the experiments in which image features are extracted with a frozen MAE ViT-L and fused with the EEG branch through two image-side heads, the CPE and the PWE. §H.1 specifies the two heads, and §H.2 reports the main result tables (intra-subject Things-EEG, cross-subject LOSO, Things-MEG), which include the PAE and the S-PAE rows for comparison, together with a repetition averaging analysis. All runs share the protocol of §3.1 of the main text.

Table 7: Architecture and training recipe ablations (subject 1, top-1 %, mean ± std, one component changed at a time). Bold rows are the default PAE, the reference of every group. The columns are the default patch size (ps=16; 96.00 ± 0.87) and the finer patch size (ps=8; 97.17 ± 0.29).

Axis	Variant	ps=16	ps=8
token interaction	**none (PAE)**	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	token mixing	48.17 ± 1.15	48.17 ± 5.06
	attention (ViT)	44.67 ± 6.11	61.33 ± 15.22
spatial pooling	**Weighted Pooling (PAE)**	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	post-head pooling	94.67 ± 2.31	95.50 ± 0.87
	post-pooling ResBlock	96.17 ± 0.76	96.50 ± 0.50
	concat pooling	52.00 ± 3.91	41.50 ± 6.87
	fixed uniform average	59.00 ± 4.58	64.17 ± 1.26
depth L	0	36.00 ± 0.87	67.50 ± 1.00
	1	94.17 ± 1.04	97.17 ± 0.29
	2	95.33 ± 1.76	97.67 ± 0.76
	3 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	4	93.67 ± 1.04	97.00 ± 0.87
	6	93.67 ± 1.76	96.17 ± 1.04
width d	48	95.17 ± 1.89	94.83 ± 0.29
	64	94.50 ± 1.50	97.00 ± 0.87
	96	94.33 ± 2.08	97.50 ± 1.32
	128	95.67 ± 0.29	97.17 ± 1.04
	192 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	256	94.83 ± 1.44	96.50 ± 1.00
	384	95.17 ± 0.58	96.83 ± 0.58
activation	**ELU (default)**	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	ReLU	95.83 ± 1.15	96.67 ± 0.76
	GELU	97.33 ± 0.76	96.83 ± 0.76
alignment width d_z	256	96.67 ± 1.04	97.33 ± 1.76
	512	95.67 ± 0.29	97.67 ± 0.58
	1024 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	2048	96.67 ± 0.29	97.00 ± 0.50
MLP dropout p	0.70	96.83 ± 1.53	97.33 ± 1.04
	0.75	97.00 ± 0.50	97.00 ± 0.50
	0.80	95.67 ± 2.02	97.67 ± 1.76
	0.85 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
	0.90	95.67 ± 1.04	97.00 ± 0.50
	0.95	94.33 ± 0.76	95.83 ± 1.15

Table 8: Patch size ablation (subject 1, top-1 %, mean ± std). Each row trains the PAE at the given patch size with every other condition unchanged, and the patch sizes cover all divisors of 224. The row with patch size 16 is the default PAE.

Patch size	Grid	Tokens	top-1
32	7×7	49	89.50 ± 2.29
28	8×8	64	92.17 ± 1.26
16 (default)	14×14	196	$\mathbf{96.00 \pm 0.87}$
14	16×16	256	96.33 ± 0.29
8	28×28	784	97.17 ± 0.29
7	32×32	1,024	96.67 ± 0.58
4	56×56	3,136	97.00 ± 0.87
2	112×112	12,544	88.67 ± 1.87

Table 9: Batch size ablation (subject 1, top-1 %, mean ± std). Each row trains the PAE at the given batch size with every other condition unchanged. The columns are the default patch size ps=16 and the finer patch size ps=8.

Batch size	ps=16	ps=8
256	96.17 ± 1.04	95.83 ± 0.29
512	97.17 ± 0.58	96.50 ± 1.80
1024 (default)	$\mathbf{96.00 \pm 0.87}$	$\mathbf{97.17 \pm 0.29}$
2048	92.33 ± 2.57	96.83 ± 0.29

Table 10: S-PAE ablations (subject 1, top-1 %, mean ± std, one component changed at a time). The default is the S-PAE of §2.2, in which the 15 low-level statistics of each patch replace the Patch Embedding, the depth is L=3, and the pooling is the Weighted Pooling. Each group varies one axis under the same protocol, and the default row is the same anchor cell repeated in each group. The columns are the default patch size (ps=16; 97.00 ± 1.32) and the finer patch size (ps=8; 95.67 ± 0.76).

Axis	Variant	ps=16	ps=8
input	**15 statistics (default)**	$\mathbf{97.00 \pm 1.32}$	$\mathbf{95.67 \pm 0.76}$
	+ luminance Patch Embedding (residual)	97.50 ± 1.00	95.67 ± 0.76
depth L	0	93.00 ± 3.04	90.17 ± 0.58
	1	96.83 ± 1.04	93.83 ± 1.15
	2	97.50 ± 1.00	95.50 ± 0.50
	3 (default)	$\mathbf{97.00 \pm 1.32}$	$\mathbf{95.67 \pm 0.76}$
	4	97.00 ± 0.50	96.00 ± 0.50
pooling	**Weighted Pooling (default)**	$\mathbf{97.00 \pm 1.32}$	$\mathbf{95.67 \pm 0.76}$
	fixed uniform average	65.50 ± 4.77	60.33 ± 2.75

H.1 IMAGE SIDE: THE CPE AND THE PWE ON A FROZEN MAE VIT-L

The image encoder uses the image foundation model `facebook/vit-mae-large` (He et al., 2022), a Vision Transformer (Dosovitskiy et al., 2021) with patch size 16 and image size 224 pre-trained via Masked Autoencoding, to extract image features. The foundation model produces 197 tokens, one CLS token and 196 patch tokens, each of dimension 1024, and is kept frozen. We design two heads that aggregate these tokens into the image embedding $\mathbf{z} \in \mathbb{R}^{d_z}$, the concatenation head (CPE) and the patchwise head (PWE).

- **Concatenation head (CPE).** The CLS token $\mathbf{c} \in \mathbb{R}^{1024}$ and the mean $\overline{\mathbf{p}}$ of the 196 patch tokens are concatenated, and the 2048-dimensional concatenation is mapped to $\mathbf{z}$ by an MLP with two linear layers and a hidden width of 768. The CPE has 3,148,544 learnable parameters.
- **Patchwise head (PWE).** An MLP with three linear layers of width 1024, ELU activations, layer normalisation, and dropout 0.85 is applied to each of the 197 tokens independently, giving one embedding $\mathbf{e}_i$ per token. The image embedding is their weighted sum $\mathbf{z} = \sum_{i=1}^{197} \mathbf{w}_i \odot \mathbf{e}_i$, where the weights are the softmax over the 197 positions of a learned 197×1024 parameter tensor, the same mechanism as the Weighted Pooling of the PAE. The PWE has 3,354,624 learnable parameters.

H.2 MAIN RESULTS

Table 11 reports the intra-subject Things-EEG results, Table 12 the cross-subject (leave-one-subject-out) Things-EEG results, and Table 13 the Things-MEG results (Hebart et al., 2023). The UBP and VED* rows quote the numbers reported in the original papers; every other row is a run of this study under the protocol of the corresponding VED experiment, with only the image encoder replaced, so that the comparison is fair. Table 14 and Fig. 10 report the repetition averaging curve: the first N of the 80 test repetitions of each image are averaged before retrieval, for N from 1 to the full 80, with the models behind Table 11, under the same conditions as the repetition averaging analysis of VED (Liu et al., 2026).

Table 11: Intra-subject 200-way retrieval on Things-EEG (10 subjects, top-k accuracy in %). PAE, PAE†, and S-PAE rows: mean ± std over three training seeds; the CPE, PWE, and VED rows are single seed. UBP and VED*: numbers as reported in the original papers (Wu et al., 2025; Liu et al., 2026); VED: our evaluation of the released VED code. †The finer patch size ps=8. Highest value per column in **bold**.

		S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Mean
UBP	top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
VED*	top-1	81.9	81.7	78.3	76.9	71.4	84.3	78.2	84.5	78.7	84.1	80.0
	top-5	96.8	96.8	96.0	97.2	93.7	98.7	96.7	98.1	96.8	98.1	96.9
VED	top-1	82.0	81.0	79.5	76.5	72.0	86.5	78.0	85.0	80.5	84.5	80.6
	top-5	98.0	96.5	95.5	96.5	93.5	99.0	97.0	98.0	96.5	99.0	97.0
CPE	top-1	90.0	88.0	90.0	88.5	80.0	90.5	89.0	93.0	88.0	93.0	89.0
	top-5	98.0	99.0	98.5	98.0	97.5	98.0	99.5	99.0	97.0	99.5	98.4
PWE	top-1	**99.5**	95.5	**99.0**	93.5	90.5	96.5	97.5	97.0	95.5	96.0	96.1
	top-5	**100.0**	99.5	**100.0**	99.5	99.5	99.5	**100.0**	**100.0**	99.5	99.5	99.7
PAE	top-1	96.0±0.9	97.5±0.9	97.0±0.5	86.8±1.4	91.5±0.5	97.7±0.6	96.5±0.5	98.3±0.6	97.0±2.3	98.3±0.8	95.7
	top-5	99.3±0.3	**100.0**±0.0	**100.0**±0.0	97.5±0.5	98.3±0.6	**100.0**±0.0	99.5±0.0	**100.0**±0.0	99.3±0.3	99.8±0.3	99.4
PAE†	top-1	97.2±0.3	**98.0**±0.9	98.0±0.5	91.5±4.0	92.7±1.4	98.2±0.6	97.2±0.6	**99.3**±0.3	**98.7**±0.8	98.3±1.3	96.9
	top-5	99.7±0.3	**100.0**±0.0	**100.0**±0.0	98.7±0.8	**100.0**±0.0	**100.0**±0.0	99.2±0.8	**100.0**±0.0	99.5±0.0	**100.0**±0.0	99.7
S-PAE	top-1	97.0±1.3	97.0±1.7	98.3±1.0	**94.8**±1.0	**95.2**±1.2	**98.7**±1.0	**98.2**±0.8	**99.3**±0.6	97.7±0.3	**99.3**±0.8	**97.6**
	top-5	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	**100.0**±0.0	99.7±0.3	**100.0**±0.0	99.5±0.0	**100.0**±0.0	**99.8**±0.3	**100.0**±0.0	**99.9**

I CONTROLLED SUPPRESSION

We apply the controlled suppression framework of Burgert et al. (2025) to the PAE, the S-PAE, the CPE, and VED. The framework quantifies the reliance of a model on a feature type by suppressing that feature type in the test images and reading the reliance off the drop in retrieval accuracy. Nine transformations in three feature types are applied to the 200 test images. Shape is suppressed by patch shuffle on 3×3, 6×6, and 14×14 grids and by per-patch $90°$ rotation on the 6×6 and

Table 12: Cross-subject 200-way retrieval on Things-EEG in the leave-one-subject-out protocol (10 subjects, top-k accuracy in %). All rows of this study use a single seed and the checkpoint selected on the pooled 5% validation split. UBP and VED*: numbers as reported in the original papers (Wu et al., 2025; Liu et al., 2026). Highest value per column in **bold**.

		S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	Mean
UBP	top-1	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
	top-5	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
VED*	top-1	27.3	32.0	11.1	18.6	16.9	16.1	16.0	18.3	13.6	30.1	20.0
	top-5	54.9	63.0	34.5	46.0	40.7	43.2	45.4	49.0	41.2	62.1	48.0
PAE	top-1	33.0	23.5	15.0	18.5	24.0	13.5	**30.5**	**31.0**	24.0	**36.0**	24.9
	top-5	**71.0**	57.5	41.0	40.0	50.0	43.5	**57.5**	59.5	57.5	66.5	54.4
S-PAE	top-1	30.0	25.5	**21.5**	16.0	**25.0**	18.5	26.5	30.0	**32.0**	23.0	24.8
	top-5	70.5	57.0	**45.0**	40.0	**50.5**	47.0	52.0	**62.5**	**58.0**	58.0	54.1
CPE	top-1	**35.5**	**32.5**	17.5	**25.0**	18.5	**22.0**	24.0	25.5	15.0	34.0	**25.0**
	top-5	66.0	**71.5**	**45.0**	**52.5**	50.0	**50.0**	53.5	57.0	43.0	**70.0**	**55.9**
PWE	top-1	12.1	12.3	11.3	12.3	12.5	12.3	11.5	11.9	11.6	10.6	11.8
	top-5	28.7	29.9	28.2	29.3	30.4	30.1	28.0	29.2	29.0	26.3	28.9

Table 13: Intra-subject 200-way retrieval on Things-MEG (4 subjects, top-k accuracy in %). All rows of this study use a single seed. UBP and VED*: numbers as reported in the original papers (Wu et al., 2025; Liu et al., 2026). Highest value per column in **bold**.

		S01	S02	S03	S04	Mean
UBP	top-1	15.0	46.0	27.3	18.5	26.7
	top-5	–	–	–	–	55.2
VED*	top-1	25.4	78.1	37.3	35.3	44.0
	top-5	–	–	–	–	70.9
PAE	top-1	25.0	93.0	76.5	44.5	59.8
	top-5	54.0	99.5	93.0	78.5	81.3
S-PAE	top-1	**42.0**	**97.0**	**81.5**	**53.0**	**68.4**
	top-5	**76.0**	**100.0**	**95.5**	**83.0**	**88.6**
PWE	top-1	33.0	93.5	78.0	38.5	60.8
	top-5	65.5	98.5	94.5	69.0	81.9

Table 14: Repetition averaging curve for the S-PAE, PAE, and the PWE on Things-EEG intra-subject retrieval. Mean and between-subject std over the 10 subjects. Top-k accuracy in %.

	S-PAE			PAE			PWE		
N reps	top-1	top-3	top-5	top-1	top-3	top-5	top-1	top-3	top-5
1	26.50±7.67	44.90±9.12	54.75±9.10	22.70±7.16	40.80±8.56	49.80±8.52	26.10±6.20	44.60±8.37	54.50±8.93
2	45.20±11.46	64.55±10.00	74.25±8.99	39.30±10.89	59.90±11.22	68.50±10.67	46.20±11.86	66.70±10.74	75.10±9.60
3	57.05±11.16	77.10±9.57	83.60±6.42	49.95±10.63	70.70±11.63	78.20±9.78	57.80±10.35	77.90±8.41	84.95±7.28
5	70.80±11.67	87.45±7.03	91.80±4.43	65.45±11.10	82.00±8.53	89.05±6.79	71.80±10.01	88.65±8.09	93.50±5.37
8	81.30±8.10	94.00±3.77	96.35±2.21	76.30±10.20	89.85±6.37	93.80±4.87	81.35±8.54	93.80±4.26	96.45±2.62
10	85.40±6.53	95.95±2.73	97.95±1.66	79.90±9.77	92.75±5.32	95.75±3.71	84.15±7.91	95.50±3.09	97.60±2.02
15	89.55±6.11	97.70±2.07	98.95±1.21	85.90±8.93	95.25±4.35	97.25±3.07	86.55±7.40	97.05±2.47	98.65±1.33
20	91.60±5.53	98.30±1.60	99.20±1.14	87.90±8.51	96.10±4.32	98.20±1.99	89.05±5.87	97.25±2.37	98.80±1.65
80 (full)	97.70±1.81	99.75±0.35	99.85±0.24	95.75±4.12	98.85±1.63	99.35±0.91	96.05±2.62	99.25±0.72	99.70±0.26

14×14 grids, where the 14×14 grid is our addition and matches the patch grid of the S-PAE. Texture is suppressed by bilateral filtering with a diameter of 11 and color and spatial sigmas of 170 and 75, and by Gaussian filtering with a kernel size of 11 and a sigma of 2.0. Color is suppressed by grayscale conversion with the BT.601 luminance weights and by a per-pixel channel shuffle. All transformations use the validated parameter settings of the framework. Fig. 11 illustrates the nine transformations on five test images. Table 15 reports the results: the clean row gives the

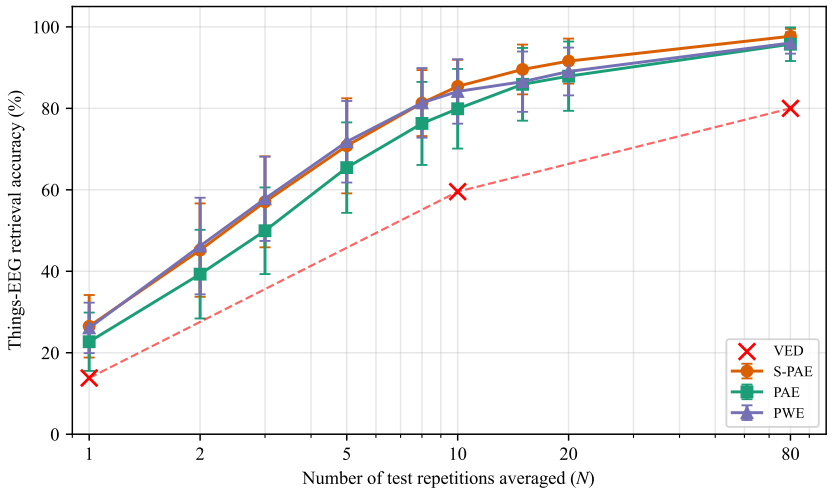

Figure 10: **Accuracy rises logarithmically with the number of test repetitions.** Repetition averaging curves (top-1) for the S-PAE, the PAE, and the PWE. Mean ± std over 10 subjects. VED (Liu et al., 2026) reference points are from the original paper.

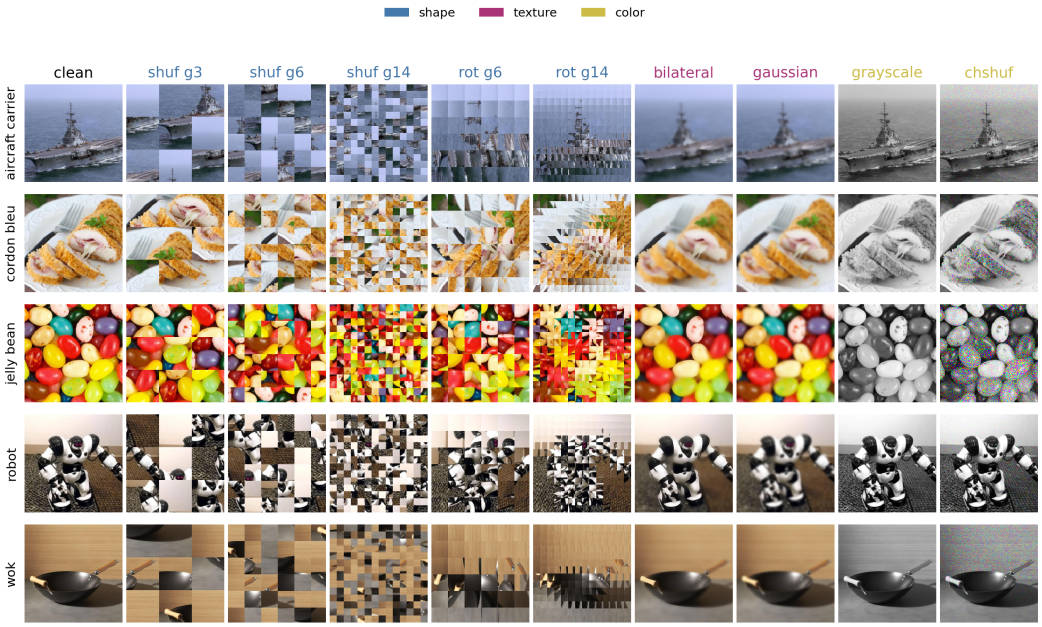

Figure 11: **The nine suppression transformations illustrated on five test images.** Each row shows one test image, with the unsuppressed image in the left column and the nine transformations of Fig. 5 in the remaining columns, colored by feature type (blue = shape, magenta = texture, yellow = color). All transformations use the validated parameter settings of Burgert et al. (2025) and are applied to the test images only.

top-1 accuracy on the unsuppressed images, and each remaining row gives the accuracy under one transformation, with the relative accuracy, the ratio of the suppressed accuracy to the clean accuracy.

Table 15: Suppression profiles of the PAE, the S-PAE, the CPE, and VED (subject 1). Cells give the absolute test top-1 (%) with relative accuracy (top-1_{supp} / top-1_{clean}). The clean row reports the top-1 accuracy on the unsuppressed images.

Condition	PAE	S-PAE	CPE	VED
clean (no suppression)	96.0	98.5	91.0	82.0
shape				
patch shuffle 3×3	21.0 (21.8)	26.5 (26.9)	59.5 (65.4)	50.5 (61.6)
patch shuffle 6×6	16.5 (17.1)	16.5 (16.8)	39.0 (42.9)	33.5 (40.9)
patch shuffle 14×14	34.0 (35.2)	16.0 (16.2)	34.0 (37.4)	9.5 (11.6)
rotation 6×6	80.0 (82.9)	81.5 (82.7)	44.0 (48.4)	31.5 (38.4)
rotation 14×14	90.5 (93.8)	88.5 (89.8)	61.0 (67.0)	39.0 (47.6)
texture				
bilateral	86.0 (89.1)	75.0 (76.1)	71.0 (78.0)	68.0 (82.9)
Gaussian	88.5 (91.7)	72.0 (73.1)	59.0 (64.8)	71.0 (86.6)
color				
grayscale	89.5 (92.7)	94.0 (95.4)	74.0 (81.3)	55.5 (67.7)
channel shuffle	86.5 (89.6)	62.5 (63.5)	63.0 (69.2)	55.0 (67.1)

J ARCHITECTURE BASELINES

Table 16 compares the PAE with four classic vision architectures. For a fair comparison, every model is trained and evaluated with the same recipe and EEG branch as the PAE (§3.1) on subject 1. LeNet (LeCun et al., 1998) is a modernised variant of the original network with four 3×3 convolutional layers of 32, 64, 128, and 128 channels, each followed by batch normalisation, a ReLU, and 2×2 max pooling, and ends with global average pooling and a linear head to the shared 1024-dimensional embedding (0.37 M parameters). AlexNet (Krizhevsky et al., 2012) keeps the layout of the original network with its five convolutional layers (48, 96, 192, 192, and 256 channels, the first convolution with stride 4) and, in place of the large fully connected layers, applies global average pooling and a linear head (1.25 M parameters). ResNet-10 (He et al., 2016) is a 7×7 convolutional stem with 48 channels followed by one residual block per stage at widths 48, 96, 192, and 384, and each transition between stages downsamples by a stride of 2, then global average pooling and a linear head (3.16 M parameters). The ViT (Dosovitskiy et al., 2021) embeds each 16×16 patch into a 192-dimensional token, prepends a CLS token and a learned positional embedding, applies four Transformer blocks with four heads and a feedforward width of 768 without dropout, and reads out through a final layer norm and a linear head on the CLS token (2.16 M parameters). LeNet reaches 84.7% top-1, AlexNet 92.0%, and ResNet-10 89.5%, i.e., 4.0–11.3% below the PAE, whereas the ViT collapses to 44.7%; the PAE reaches 96.0% with 1.57 M parameters.

Table 16: Architecture baselines (subject 1, test top-1 %, mean $\pm$ std over three seeds).

Architecture	top-1
LeNet	84.67 ± 2.57
AlexNet	92.00 ± 2.29
ResNet-10	89.50 ± 0.50
ViT	44.67 ± 6.11
PAE	$\mathbf{96.00 \pm 0.87}$

K THE DERIVATION OF THE PAE FROM A VIT

Table 17 derives the PAE from a ViT at depth 4. Each variant removes or replaces one component of the previous one and is trained and evaluated with the same recipe and EEG branch on subject 1 over three seeds. The four variants are:

- **E0, the ViT:** the reference model with patch size 16, width 192, depth 4, four attention heads, and a feedforward width of 768, reading out on the CLS token.
- **E1, attention only in the readout layer:** the attention of the first three blocks is removed, leaving three per-token residual MLP blocks; the fourth layer keeps its multi-head attention followed by an MLP, and the CLS token, the positional embedding, the final LayerNorm, and the linear head on the CLS token are kept.
- **E2, Weighted Pooling readout:** the readout attention is replaced by Weighted Pooling, a learned per-dimension weight over the 196 positions that is independent of the input; the CLS token, the positional embedding, the $K/Q/V$ projections, the final LayerNorm, and the linear head on the CLS token are removed, and the readout MLP is retained.
- **E3, the PAE:** the readout MLP, which is the fourth block, is removed, leaving three per-token MLP blocks, Weighted Pooling, and the single linear Alignment Head, giving the default PAE of Table 2.

Table 17: Derivation of the PAE from a ViT (subject 1, test top-1 %, mean ± std over three seeds).

Rung	Removed	Architecture	top-1
E0	—	ViT, depth 4 (CLS head)	44.67 ± 6.11
E1	attention of blocks 1–3	MLP body + readout attention	88.00 ± 2.18
E2	readout attention, CLS, pos. emb.	MLP body + Weighted Pooling	93.67 ± 1.04
E3	readout MLP	**PAE**	$\mathbf{96.00 \pm 0.87}$

E0 reaches 44.7% top-1, E1 88.0%, E2 93.7%, and E3, the PAE, 96.0%.

L Related work

L.1 Contrastive EEG-to-image alignment on Things-EEG

Two EEG resources support this literature. Things-EEG (Gifford et al., 2022) recorded ten subjects with 63 channels during rapid serial visual presentation of 16,740 images from 1,854 concepts, with 1,654 training concepts and 200 held out test concepts, and it is the benchmark used in the main text. An earlier release covered fifty subjects for one session (Grootswagers et al., 2022), and a MEG companion covers four subjects with 271 channels (Hebart et al., 2023).

BraVL introduced trimodal variational learning over brain, visual, and linguistic features and set the early baseline at 5.8% top-1 on the 200-way zero-shot task (Du et al., 2023). NICE established the standard recipe, a temporal and spatial convolutional EEG encoder trained with a symmetric InfoNCE objective (van den Oord et al., 2018) against frozen CLIP image embeddings (Radford et al., 2021; Song et al., 2024). Most later work refines the EEG branch or the loss while keeping the image branch frozen. MB2C reuses the NICE backbone and adds bidirectional cycle consistency with feature space adversarial translation and mixup augmentation (Wei et al., 2024). MUSE develops a spatial then temporal convolution variant of the same recipe with a similarity keeping loss inspired by the organization of inferotemporal cortex (Chen & Wei, 2024). NICE++ refines image features toward core object semantics with generated captions before alignment (Song et al., 2025b). Neural-MCRL adds spectral attention and text embeddings as semantic anchors (Li et al., 2025). SFTG learns over an electrode graph with spatial and functional connectivity and per identity archetypes (Sun & Yang, 2025). STAMBRIDGE conditions on amplitude spectra without hard frequency masking and distils a mid-feature bridge from a frozen vision language encoder (Meng et al., 2026). ViEEG decomposes each stimulus into contour, foreground, and scene views and aligns them hierarchically (Liu et al., 2025). UMind trains one encoder for retrieval, classification, and reconstruction with coarse and fine text (Xu et al., 2025a). EEG-HMLearner decodes image, class, and caption modalities jointly in one mixed expert encoder (Wu et al., 2026). VE-SDN decouples shared semantics from modality specific noise in a joint space on top of the ATM-S backbone (Chen et al., 2026), and NeuroVista adds bidirectional cross-modal attention with channel level masking (Xu & Lin, 2025). ATM itself contributes the channel and temporal attention EEG encoder used by several of these approaches and the first EEG guided diffusion reconstruction (Li et al., 2024). Foundation model encoders have been transplanted to this task as well, with CBraMod fine-tuned for symmetric

alignment (Harkare & Subramanian, 2025), although a broad benchmark finds that EEG foundation models do not reliably beat compact task specific decoders (Yang et al., 2025). A fusion strategy enriches the pre-trained branch itself, combining two CLIP encoders with a VAE pixel latent into a single hierarchical token, and reaches 75.7% intra-subject top-1 on Things-EEG (Zheng et al., 2026).

A second group changes the image side of the pairing. UBP degrades high-frequency detail with an uncertainty calibrated fovea blur and motivates the transform by the photoreceptor topography of the retina (Wu et al., 2025; Curcio et al., 1990), reaching 50.9% top-1. CognitionCapturerPro inherits this strategy as uncertainty weighted masking within a multimodal framework (Zhang et al., 2026). VED extends the blur to a multi-level bank with a feature selection module and reaches 80.0% (Liu et al., 2026). HyFI fuses semantic features of fovea blurred images with perceptual features of Gaussian blurred images by interpolation in hyperbolic space, with the UBP projection as default brain encoder, and reaches 68.2% (Jo et al., 2026; Wu et al., 2025). NeuroCLIP tunes visual prompts of a frozen CLIP and treats non-matching pairs as soft negatives because visually similar images elicit similar EEG, reaching 63.2% (Wang et al., 2025). BRENA aligns groups of EEG channels with selected local sub-features of the visual input and reaches 58.2% (Zhu et al., 2026). Frequency aware temporal encoding with differentiable channel clustering offers a further EEG side route (Shi et al., 2025b), and BrainAE shares one autoencoder between encoding, decoding, and masked signal reconstruction (Song et al., 2025a). A multi-scale alignment framework reports 30.6% (Shi et al., 2025a).

L.2 What visual information does the EEG carry

Closest to our question are studies that vary the image representation space. The largest controlled comparison replaces image embeddings with versions aligned to human perceptual similarity judgments and improves retrieval by up to 21% top-1 across six encoder families, several EEG encoders, and EEG, MEG, and fMRI datasets, which supports the view that rapid responses encode early impressions of shape, orientation, and color rather than category level similarity (Rajabi et al., 2025). Within one local alignment framework, self-supervised DINO and DINOv2 targets beat scale matched CLIP backbones (Zhu et al., 2026; Oquab et al., 2024). A comparison of ResNet-50, CORNet-S, and CLIP finds the brain inspired recurrent model numerically best but not significantly so, and reports that rare categories appear to be retrieved by color and shape rather than by semantics (Harkare & Subramanian, 2025). Aligning EEG with three layer groups of a vision transformer, selected by kernel alignment and validated against color, texture, and semantic attribute correlations, outperforms final layer alignment, and removing either the low or the high group degrades retrieval (Choi et al., 2026). A fusion framework that aligns EEG with two CLIP encoders and a VAE pixel latent reports gains of up to 21.4% top-1 from adding the pixel-level latent to a semantic encoder, and concludes that the signal aligns jointly with semantic and pixel-level content (Zheng et al., 2026). A linear decoder from EEG into CLIP and VAE latent spaces already matches previous reconstruction approaches, and the resulting feature maps place hue and texture near 100 ms and animacy near 180 ms, with an explicit warning that natural image statistics co-vary (Fei et al., 2024). On the MEG side, category level information dominates from about 250 ms after stimulus onset, while low-level image features are decodable mainly at response onsets and offsets (Benchetrit et al., 2024). On the reconstruction side, a dedicated pathway from EEG to VAE latents improves pixel correlation but recovers little beyond coarse color (Yang & Onken, 2025). Modality expert studies report that depth carries structure without color or texture, that captions are abstract, and that fusion prevents fine-grained information from being drowned out (Zhang et al., 2025; 2026). Hierarchical use of contour, foreground, and scene views shows foreground features ahead of pure contours (Liu et al., 2025). Blur studies find that removing high-frequency detail improves retrieval, with electrodes showing selectivity for particular blur levels (Wu et al., 2025; Liu et al., 2026), and EEG generation work reports that occipital channels are more sensitive to global image properties than to fine detail (Li et al., 2026a).

L.3 Reconstruction and generation from EEG

The aligned representations also drive image reconstruction. ATM maps EEG embeddings into CLIP space with a diffusion prior and generates with an image conditioned diffusion model, with a second branch into VAE latents for low-level detail (Li et al., 2024). Knowledge distillation from a

CLIP teacher targets semantic content with a class conditional latent diffusion model (Ferrante et al., 2024). BReAD retrieves a visually similar prior from 8.5 million images to initialize generation and shows that conditioning in CLIP space limits low-level fidelity (Zhu et al., 2025). Further approaches add multimodal experts (Zhang et al., 2025; 2026), regional semantic maps (Xiang et al., 2026), dual alignment to CLIP and VAE latents (Tan et al., 2026), unified multitask training (Xu et al., 2025a), cross subject and cross task transfer (Huang et al., 2025), or one shared autoencoder for both directions (Song et al., 2025a). In the inverse direction, diffusion models generate individualized EEG responses from images (Xu et al., 2025b; Li et al., 2026a), and closed-loop frameworks use an aligned encoder as a reward model for stimulus optimization (Li et al., 2026b).

L.4 Patchwise models and controlled suppression

The PAE descends from the bag-of-features tradition in vision, in which images are represented by unordered sets of local descriptors (Csurka et al., 2004; Lazebnik et al., 2006). BagNet showed that a network whose receptive field is confined to small regions, with local features combined only by a spatial average, reaches high ImageNet accuracy (Brendel & Bethge, 2019). The token mixing variant of our ablations follows the cross-patch sub-layer of ResMLP (Touvron et al., 2021), and our reading of the first layer continues the tradition of visualizing trained convolutional filters (Krizhevsky et al., 2012; Zeiler & Fergus, 2014). The suppression suite follows the controlled suppression framework of Burgert et al. (2025), which measures reliance on a feature type by degrading that feature type in the test images and reading the drop in accuracy.

L.5 Position

In summary, published approaches on Things-EEG either keep the image branch frozen and improve the EEG side (Song et al., 2024; Wei et al., 2024; Chen & Wei, 2024; Song et al., 2025b; Li et al., 2025; Sun & Yang, 2025; Meng et al., 2026; Liu et al., 2025; Xu et al., 2025a; Wu et al., 2026; Chen et al., 2026; Xu & Lin, 2025), or keep it frozen and transform the images before encoding (Wu et al., 2025; Liu et al., 2026; Zhang et al., 2026; Jo et al., 2026; Wang et al., 2025; Zhu et al., 2026), or fuse several pre-trained encoders (Zheng et al., 2026). Feature level studies probe the frozen branch by swapping it or slicing it (Rajabi et al., 2025; Zhu et al., 2026; Choi et al., 2026; Harkare & Subramanian, 2025; Fei et al., 2024). This paper instead trains the image branch from scratch, and its attribution shows that the feature space is dominated by luminance contrast and edge statistics and that the learned pooling weights concentrate on the image center, consistent with foveal dominance in early visual processing; the retrieval rests on the low-level statistics of local patches, computed independently and combined by learned spatial weights.

M Implementation Pseudocode

This appendix records the two algorithms as Python code, in the form in which they are implemented, as a complement to the equations of §2.2 and Appendix C. Listing 1 is the computation of the 15 low-level patch statistics of §2.2, and Listing 2 is the encoder forward pass and the training objective of Appendix C.

Listing 1: Low-level per-patch statistics of one image, the S-PAE input.

```
# Inputs:
#   V: (3, 224, 224) uint8 image
#   mu_s, sd_s: (15,) standardization moments per statistic, computed
#     once on the training split, over all training images and all
#     196 patches pooled together
# Output:
#   U_std: (196, 15) standardized statistics of the image; test
#     images reuse the same frozen mu_s, sd_s
# block(m): average of the map m over each of the 196 non-overlapping
#   16x16 patches, row-major; m has shape (C, 224, 224) -> out (196, C)

def block(m):
    c = m.reshape(-1, 14, 16, 14, 16).mean(dim=(2, 4)) # (C, 14, 14)
    return c.permute(1, 2, 0).reshape(196, -1)
```

```
V = V.float() / 255.0                                    # in [0, 1]
rgb = block(V)                                           # (196, 3)
L = 0.299 * V[0] + 0.587 * V[1] + 0.114 * V[2]           # (224, 224)
gy, gx = torch.gradient(L)  # central differences, one-sided at border

gx2 = block(gx**2)                                       # (196,)
gy2 = block(gy**2)                                       # (196,)
gm2 = gx2 + gy2

lum_mean = block(L)                                      # (196,)
lum_std  = (block(L**2) - lum_mean**2).clamp_min(0).sqrt()
grad_energy = gm2 / 256.0
edge_frac = block(((gx**2 + gy**2).sqrt() > 0.12).float())

ori = []
for th in (0, np.pi/4, np.pi/2, 3*np.pi/4):
    proj = gx * np.cos(th) + gy * np.sin(th)
    ori.append(block(proj**2))
ori_e = torch.stack(ori, dim=1) / (gm2[:, None] + 1e-8) # (196, 4)
hv = torch.log10((gx2 + 1e-6) / (gy2 + 1e-6))

lap = torch.zeros_like(L)                                 # zero border
lap[1:-1, 1:-1] = (4*L[1:-1, 1:-1] - L[:-2, 1:-1] - L[2:, 1:-1]
                   - L[1:-1, :-2] - L[1:-1, 2:])
lap_energy = block(lap**2) / 256.0

low = L.reshape(112, 2, 112, 2).mean(dim=(1, 3))         # 2x2 block
    average
low = low.repeat_interleave(2, 0).repeat_interleave(2, 1)
fine_freq = block((L - low)**2) / (block((L - L.mean())**2) + 1e-8)

sat = block(V.max(dim=0).values - V.min(dim=0).values)  # (196,)

U = torch.stack([lum_mean, lum_std, grad_energy, edge_frac,
                 ori_e[:, 0], ori_e[:, 1], ori_e[:, 2], ori_e[:, 3],
                 hv, lap_energy, fine_freq, sat,
                 rgb[:, 0], rgb[:, 1], rgb[:, 2]], dim=1) # (196, 15)
U_std = (U - mu_s) / (sd_s + 1e-8)
```

Listing 2: PAE and S-PAE forward pass and one training step.

```
# Inputs per mini-batch:
#   V:   (B, 3, 224, 224) float images, ImageNet normalization   [PAE]
#   U:   (B, 196, 15) standardized statistics                    [S-PAE]
#   X:   (B, 63, 250) EEG responses
# Output:
#   one AdamW update of all parameters on the batch loss

# image branch, PAE
t = Conv2d(3, 192, kernel_size=16, stride=16)(V)  # (B, 192, 14, 14)
t = t.flatten(2).transpose(1, 2)                  # (B, 196, 192)
# image branch, S-PAE: t = Linear(15, 192)(U)     # (B, 196, 192)

t = LayerNorm(t)
for _ in range(3):                  # residual blocks, shared over
    patches
    h = Linear(192, 1024)(t)
    h = Dropout(0.85)(ELU(h))       # dropout on the hidden activations
    t = LayerNorm(t + Linear(1024, 192)(h))

w = softmax(pool_scale, dim=1)      # pool_scale (1, 196, 192),
    initialized to zero
g = (t * w).sum(1)                  # (B, 192) learned weighted average
```

```
z = Linear(192, 1024)(g)           # image embedding (B, 1024)
y = EEG_encoder(X)                 # EEG branch inherited from VED

# one training step
# l2_norm(t) = t / t.norm(dim=-1, keepdim=True), row-wise l2 norm
img_f = l2_norm(z)                 # image side l2 norm at train
eeg_f = y                          # EEG side unnormalized at train
logits_eeg = eeg_f @ img_f.T       # (B, B)
logits_img = img_f @ eeg_f.T
labels = arange(B)
loss = 0.5 * (cross_entropy(logits_eeg, labels).mean()
              + cross_entropy(logits_img, labels).mean())
AdamW(loss).step()                 # lr 1e-3, batch 1024, 200 epochs

# evaluation, 200-way retrieval
sim = l2_norm(y) @ l2_norm(z).T    # both sides l2 norm, cosine
top1 = (sim.argmax(dim=1) == arange(N)).float().mean()
```