BEAM: AUDITING THE SELECTION-SPECIFICITY GAP IN EEG AND VISION ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG–vision alignment connects brain recordings to visual concepts through learned representations. Interpreting these representations requires determining which coordinates support concept recognition and influence model predictions. The value of selecting concept-responsive coordinates depends on how their decoding and intervention outcomes compare with equally sized random subsets. We address this problem with the **Brain Encoder Audit with Matched Controls (BEAM)**. Our contributions are threefold. **i)** We introduce a matched-control audit that measures the benefit of coordinate selection for both linear decoding and downstream intervention, using shared activation sites, inputs, and subset sizes. **ii)** We identify a *selection-specificity gap* across the audited EEG and vision settings on THINGS-EEG2: selection benefits intervention more than decoding. A representative across-subject EEG comparison yields 10.74 times the random-control intervention effect and 1.25 times its probe accuracy. **iii)** We characterize how the value of selection depends on the operation through *removal–replacement asymmetry* in a CLIP case study and *steering across subsets* in EEG, and develop a *concentration-and-recoverability account* of the gap. BEAM helps EEG–vision researchers distinguish coordinates that carry decodable concept information from those whose selection disproportionately affects model predictions, and test whether targeted steering benefits from concept-based coordinate selection.

1 INTRODUCTION

EEG–vision alignment learns a mapping from brain recordings to visual representations, enabling recognition of the images a participant viewed (Song et al., 2024; Li et al., 2024). Understanding this mapping requires identifying how visual concepts are represented and how those representations affect predictions. Concept-responsive coordinates offer candidates for both interpretation and activation steering, building on concept-based analyses of vision models (Bau et al., 2017; Goh et al., 2021). Their usefulness depends on what selecting those coordinates adds to decoding and intervention.

A successful probe or intervention raises a shared question: how much does the result depend on the chosen coordinates? Measuring that contribution requires comparison with an equally large random subset. Prior work establishes that decodability and behavioral influence can differ (Hayne et al., 2022; Huang & Chang, 2025). We ask a more specific question: *how much does the same coordinate-selection rule benefit decoding and intervention?*

We introduce the **Brain Encoder Audit with Matched Controls (BEAM)** to answer this question. BEAM compares selected and random coordinate sets at the same activation site, on the same inputs, and at matched set sizes (Figure 1). For each procedure, *selection specificity* is the selected-set effect divided by the mean random-set effect. Readout specificity measures the advantage in probe accuracy; intervention specificity measures the advantage in downstream accuracy drop. Comparing the ratios measures how selection benefits the two procedures.

The results support three connected findings. The **selection-specificity gap** describes a larger selection advantage for intervention than for decoding. It recurs in the main THINGS-EEG2 comparisons of a CLIP vision encoder and within- and across-subject EEG settings. The largest gap occurs in across-subject EEG.

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.

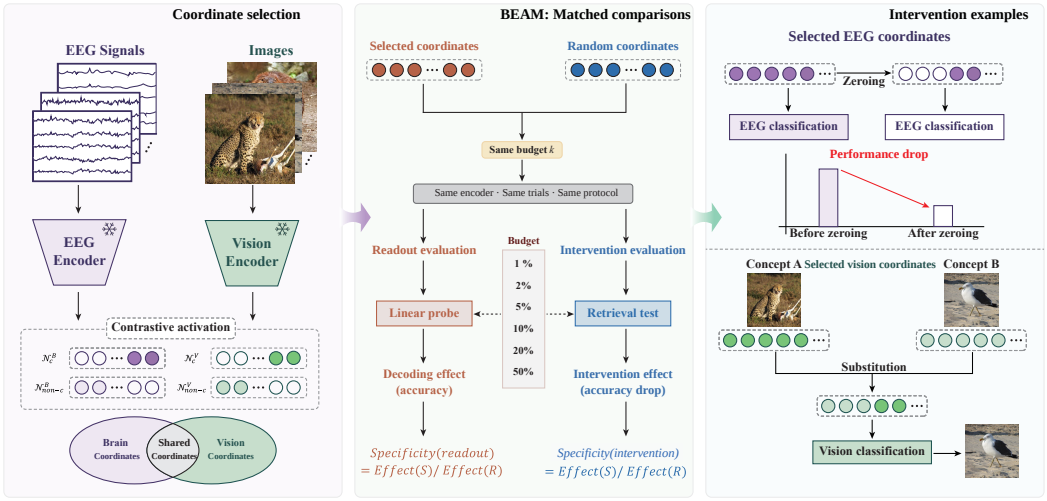

Figure 1: **BEAM overview.** Contrastive responses define a selected coordinate set. A size-matched random set provides a reference for both ridge-probe accuracy and downstream intervention loss. The same encoder site, evaluation inputs, and coordinate budget are used within each comparison. EEG interventions zero coordinates; vision interventions substitute a donor input's activations. Their selected/random effect ratios quantify selection specificity separately on the two axes.

Removal–replacement asymmetry shows why the operation matters. In the CLIP case study, many original predictions survive zeroing the selected coordinates, while replacing those coordinates with donor activations redirects predictions toward the donor. The same subset therefore has different effects when its activations are removed or replaced.

Steering across subsets describes improvements obtained by applying a learned transform through selected, random, and low-variance EEG coordinates. Random subsets achieve the highest reported accuracy in both subject settings. The transform improves retrieval through several subsets; the random control measures the added value of selecting coordinates.

We develop a **concentration-and-recoverability account** linking these observations. An encoder's predictions may depend strongly on particular coordinates, while a refitted probe can recover concept information from alternative subsets. This account predicts how the gap can change with subset size.

Contributions. We contribute **BEAM**, a matched-control audit of coordinate selection; evidence for the **selection-specificity gap**, with focused studies of **removal–replacement asymmetry** and **steering across subsets**; and a **concentration-and-recoverability account** of the gap. The main comparison spans one dataset; the operation study uses CLIP, steering uses EEG, and an additional EEG backbone supplies site-selectivity evidence.

2 Related Work

Interpreting EEG–vision alignment. EEG-to-image methods align brain recordings with visual representations (Song et al., 2024; Li et al., 2024; 2025; Wang et al., 2025). For motor-imagery EEG, PTSM studies cross-subject transfer (Jing et al., 2025). NICE examines temporal, spatial, spectral, and semantic structure, using Grad-CAM and attention visualizations to analyze spatial processing (Song et al., 2024). Rezvani et al. (2025) use semantic captions and gradient-based saliency to interpret EEG-to-image generation. These analyses relate EEG inputs and learned representations to visual content. BEAM asks how much selecting concept-responsive internal coordinates contributes to recovering concept identity and changing model predictions. It measures this contribution against equally sized random subsets in both procedures. This comparison helps determine whether a subset's decoding or intervention result benefits from the selection rule.

Decodability and causal use. Linear probes measure recoverable information (Alain & Bengio, 2016; Belinkov, 2022); control tasks help assess what a probe's performance reveals about its repre-

sentation (Hewitt & Liang, 2019). Predictive weights also differ from signal-generating mechanisms in neuroimaging (Haufe et al., 2014). Erasure methods test dependence on recoverable attributes (Elazar et al., 2021; Ravfogel et al., 2020).

Empirical comparisons show that decoding and causal relevance can diverge. Hayne et al. (2022) find a weak relationship between probe-identified and ablation-identified units in convolutional networks. Counting transformers exhibit different depth profiles for decodability and causal influence (Huang & Chang, 2025); bracket-sequence transformers encode hierarchical variables whose causal roles differ (Sharma et al., 2026). BEAM examines the contribution of coordinate selection to this divergence. Holding the activation site and subset size fixed, it compares how much the same selection rule changes each result relative to random coordinates. The selection-specificity gap summarizes this difference in relative effects.

Concept localization and functional importance. Network dissection associates units with visual concepts (Bau et al., 2017), while single-direction analyses relate selective responses to task importance (Morcos et al., 2018). Multimodal neurons respond across input forms (Goh et al., 2021), and sparse probing studies concept recovery from small coordinate sets (Gurnee et al., 2023). The intervention and metric determine which aspect of a representation's function is tested (Zhang & Nanda, 2024). BEAM compares coordinate selection under zeroing, donor substitution, and learned steering, building on activation interventions (Li et al., 2023). Its matched controls quantify selection effects separately for each operation.

Appendix I discusses feature dictionaries, probe evaluation, and circuit-level interventions. Table 8 compares the measurements used by the most directly related studies.

3 Measuring the Value of Coordinate Selection

BEAM starts with two sets of activation coordinates: the k coordinates selected by their responses to concepts, and k coordinates sampled at random. We test each set in two ways. A linear probe predicts concept identity from its activations. An intervention changes those activations and measures the resulting drop in the encoder's task accuracy. Comparing selected and random sets shows how much the selection rule contributes to each measurement.

Setup. Let a frozen encoder produce activations $a(x) \in \mathbb{R}^d$ at a specified site, and let $a_U(x)$ restrict them to coordinates $U \subseteq [d]$. The concept vocabulary is partitioned into disjoint selection and evaluation pools, $\mathcal{C}_{\text{sel}}$ and $\mathcal{C}_{\text{eval}}$. Write D_c for inputs with concept label c and T for the encoder's downstream task. An activation coordinate is one dimension at the audited site, including a flattened feature-map dimension.

One selected set per encoder. We first compute the contrastive response

$$s(u, c) = \mathbb{E}_{x \sim D_c}[a_u(x)] - \mathbb{E}_{x \sim D_{\neg c}}[a_u(x)]. \tag{1}$$

We call variation in activation responses across concepts *concept selectivity*. The selection statistic measures this variation:

$$v(u) = \text{Var}_{c \in \mathcal{C}_{\text{sel}}}[s(u, c)], \qquad S_k = \arg\text{top-}k_{u \in [d]} v(u), \qquad R_k \sim \text{Unif}\left(\binom{[d]}{k}\right). \tag{2}$$

High variance identifies coordinates whose responses differ across concepts. This produces one shared set S_k, evaluated on concepts excluded from selection. R_k is sampled uniformly without replacement and has the same cardinality. For cross-site comparisons, selectivity scores are standardized against each site's random controls to account for differences in activation scale.

Readout and intervention effects. The readout axis fits a ridge classifier to the restricted activations. For a training activation matrix A_U and one-hot labels Y,

$$\begin{aligned} W_U &= \arg\min_W \|A_U W - Y\|_F^2 + \lambda \|W\|_F^2, \\ \Phi_{\text{read}}(U) &= \mathbb{E}_{(x,c) \in D_{\text{test}}} \mathbf{1}\left[c = \arg\max_{c' \in \mathcal{C}_{\text{eval}}} \langle w_{c'}, a_U(x) \rangle\right]. \end{aligned} \tag{3}$$

Each cross-validation fold fits its probe on training trials and scores held-out trials. The intervention axis measures the drop in downstream accuracy after intervening on the same coordinates:

$$\Phi_{\text{int}}(U) = \text{Acc}(T \mid a) - \text{Acc}\big(T \mid \text{do}(a_U = \zeta)\big), \qquad \zeta \in \{\mathbf{0}, a_U(x')\}. \tag{4}$$

Zeroing is used in the EEG grid; substitution with a donor input x' is used in the vision grid. The intervention effect is specific to the chosen operation and downstream task. Readout uses probe accuracy; intervention uses the downstream accuracy drop.

Selection specificity. For either axis, define

$$\sigma_\bullet(k) = \frac{\Phi_\bullet(S_k)}{\mathbb{E}_{R_k}[\Phi_\bullet(R_k)]}, \qquad \bullet \in \{\mathrm{read}, \mathrm{int}\}. \tag{5}$$

Concept selectivity determines which coordinates are selected; selection specificity measures the benefit of that selection. A readout ratio above one indicates higher probe accuracy; an intervention ratio above one indicates a larger downstream accuracy drop than the mean random control. We call a disparity between the two ratios a *selection-specificity gap*. A particularly informative case is

$$\sigma_{\mathrm{read}}(k) \approx 1 \qquad \text{and} \qquad \sigma_{\mathrm{int}}(k) \gg 1. \tag{6}$$

In this case the selection criterion identifies coordinates that matter disproportionately to the intervention, although random coordinates support similar probe accuracy.

We include an intervention ratio in the cross-setting summary when the mean random-control accuracy drop is at least 0.05 percentage points. Table 1 reports all conditions, marks those below the threshold, and gives the selected and random accuracy drops alongside their ratios. Ratios with a denominator near zero are sensitive to small changes in that denominator; the absolute drops show the size of the measured intervention effect.

4 Concentration and Recoverability

An encoder may depend strongly on a small coordinate set, while a newly trained probe can recover the same concept from several alternative sets. We explain this possible source of the selection-specificity gap through two properties: the concentration of task-relevant contributions and the recoverability of concept information from a subset.

Concentrated contributions under removal. Suppose each coordinate contributes nonnegative task-relevant mass m_u, with $\sum_u m_u = 1$, and let $m_{(1)} \geq \cdots \geq m_{(d)}$. Define $M(k) = \sum_{i=1}^{k} m_{(i)}$. If the selection score approximates this ordering and zeroing damage is locally proportional to removed mass, then

$$\sigma_{\mathrm{int}}(k) \approx \frac{M(k)}{k/d} = \frac{dM(k)}{k}. \tag{7}$$

Uniform random selection captures k/d of the mass in expectation. The ratio in Eq. 7 is nonincreasing with k, because it is proportional to the mean of the k largest masses. It equals one at $k = d$. Concentration can therefore produce high specificity at small budgets while the absolute selected-minus-random damage, proportional to $M(k) - k/d$, grows over part of the same range.

This additive approximation predicts equal expected removed mass under uniform contributions. Correlation, nonlinear downstream responses, and donor substitution can change the curve. We test its monotone prediction in the across-subject EEG zeroing sweep.

Recoverability after restricting coordinates. Consider a signal representation $a(x) = Bz(x) + \eta(x)$, where $B \in \mathbb{R}^{d\times r}$ maps an r-dimensional concept signal into the activation space and η is noise. A linear probe on U sees $B_U z(x) + \eta_U(x)$. In the noiseless case, every linear decision on z is expressible from a_U whenever

$$\mathrm{rank}(B_U) = r. \tag{8}$$

A left inverse of B_U then recovers z, preserving linear decodability. With finite data, performance also depends on the conditioning of B_U, noise, regularization, and class margins.

The orientation of the signal subspace matters. If a rank-one signal is confined to one coordinate, a random k-coordinate subset retains it with probability k/d. If the same noiseless signal has nonzero coefficients in every coordinate, any nonempty subset permits recovery. Both representations have rank one, yet their random-subset behavior differs. Recoverability therefore depends jointly on signal rank and the distribution of that signal across coordinates.

Implications for the audit. The account identifies three comparisons. First, selected and random probe accuracies may converge before intervention effects do, because a refitted probe can use alternative coordinates. Second, absolute damage and specificity can move in opposite directions as the budget widens. Third, a learned steering map may work through several coordinate subsets if each supports the relevant transform. The steering comparison tests an empirical extension of the recoverability account.

Signal-geometry and downstream-sensitivity measurements can test the assumptions underlying this account.

5 BEAM

BEAM returns two budget-dependent specificity curves and their underlying effects. It combines closed-form ridge fitting with intervention forward passes through a frozen backbone.

Matched comparisons. Selected and random sets share the activation site, evaluation trials, probe solver, regularization, and intervention rule. Set size is matched at each budget.

A budget sweep with absolute effects. We evaluate $k/d \in \{1, 2, 5, 10, 20, 50\}\%$ to track selection advantage as more coordinates become available. Each condition reports selected and mean random intervention effects in percentage points. The reporting threshold is applied to the mean random effect. Readout specificity uses raw accuracy ratios.

The two ratios answer different operational questions. Readout compares how well newly fitted classifiers recover concept labels, whereas intervention compares how strongly the frozen encoder responds to an activation edit. Their denominators retain those respective meanings. Reporting each selected and random effect makes the comparison interpretable when a probe approaches its accuracy ceiling or an intervention causes only a small accuracy drop.

Algorithm 1 BEAM with matched random controls

Require: Frozen encoder and site; disjoint selection and evaluation concept pools; budgets $\mathcal{K}$; random draws R; task T; threshold τ

Compute $s(u, c)$ and $v(u)$ on the selection pool (Eqs. 1–2)
for $k \in \mathcal{K}$ **do**
  Select top-k coordinates S_k and draw $R_k^{(1)}, \ldots, R_k^{(R)}$
  for each set $U \in \{S_k, R_k^{(1)}, \ldots, R_k^{(R)}\}$ **do**
    Fit and evaluate the ridge probe using shared cross-validation folds
    Apply the specified intervention and measure the downstream accuracy drop
  end for
  Compute $\sigma_\bullet(k) = \Phi_\bullet(S_k)/\bar{\Phi}_\bullet(R_k)$
  Mark the intervention ratio if $\bar{\Phi}_{\text{int}}(R_k) < \tau$
end for
return Ratios, absolute effects, budgets, and threshold flags

Comparing encoder settings. Within- and across-subject EEG share zeroing and the concept task; vision uses substitution. We compare the gap's direction across settings and interpret its magnitude jointly with modality, task, training split, and intervention.

6 RESULTS

Concept-based coordinate selection benefits intervention more than decoding across the main EEG and vision comparisons. The CLIP case study shows that replacing selected activations changes predictions more strongly than zeroing them. In EEG steering, learned transforms improve retrieval through several subsets, with random coordinates achieving the highest accuracy. Together, these findings connect the value of selection to the operation applied.

Table 1: The full audit grid on THINGS-EEG2 (Gifford et al., 2022). Readout specificity σ_{read} is the selected/random ratio of fitted-probe accuracy; intervention specificity σ_{int} is the ratio of downstream accuracy drops (Eq. 5). Φ_S and Φ_R give the selected and mean random intervention effects in percentage points. Readout specificity ranges from 0.75 to 1.61 across eighteen conditions; intervention specificity reaches 10.74 among conditions above the reporting threshold. Gray cells have $\Phi_R < 0.05$ points and are excluded from the cross-setting intervention-ratio count.

	EEG across subject				**EEG within subject**				**Vision**			
k/d (%)	σ_{read}	σ_{int}	Φ_S	Φ_R	σ_{read}	σ_{int}	Φ_S	Φ_R	σ_{read}	σ_{int}	Φ_S	Φ_R
1	1.61	8.99	0.34	0.04	1.47	1.12	0.05	0.04	0.75	1.43	0.05	0.04
2	1.53	16.81	0.50	0.03	1.26	1.68	0.25	0.15	0.91	0.93	0.25	0.27
5	1.25	**10.74**	0.69	0.06	1.19	1.96	0.89	0.46	1.01	1.88	0.70	0.37
10	1.12	**9.29**	0.98	0.11	1.14	1.96	1.72	0.88	1.01	1.62	1.80	1.11
20	1.15	**7.15**	1.56	0.22	1.12	2.20	4.37	1.98	1.00	1.81	6.35	3.52
50	1.10	**3.45**	2.79	0.81	1.12	2.04	13.84	6.78	1.00	1.29	67.50	52.31

6.1 Design and evidence scope

All experiments use THINGS-EEG2 (Gifford et al., 2022), which contains scalp EEG recordings from ten participants viewing natural images. The shared 200-concept evaluation vocabulary connects image-to-text, EEG-to-image, and EEG-to-text tasks. CLIP ViT-L/14 supplies frozen vision and text towers (Radford et al., 2021). The primary EEG encoder is NICE-EEG (Song et al., 2024). The full specificity grid comprises this EEG encoder under within- and across-subject evaluation and the vision tower. The additional DWT encoder (Zhang et al., 2025) contributes site-selectivity measurements, which are reported separately from the eighteen-cell grid.

Within-subject evaluation trains and evaluates the encoder on the same participant; across-subject evaluation holds out that participant. This tests transfer across people. CLIP supplies the visual comparison, and the additional EEG architecture tests site selection. All comparisons share one dataset.

The selection pool and the 200 evaluation concepts are disjoint. On the readout axis, a closed-form ridge classifier predicts concept identity using five-fold cross-validation and ten trials per concept, with the same regularization in both arms. On the intervention axis, the grid measures the loss in the encoder's native retrieval or classification task after EEG zeroing or vision substitution. The 50-concept substitution study below additionally measures how often the donor concept becomes top-1. Steering and sparse anchoring target EEG-to-text accuracy. Table 3 lists sites and widths; Appendix A describes the reported implementation.

6.2 The selection-specificity gap

The selection-specificity gap is largest in the across-subject EEG setting and recurs, at smaller magnitude, in the within-subject and vision settings. Table 1 supplies both ratios and the absolute intervention effects. Figure 2 visualizes the budget dependence and the separation between axes.

At a 5% budget, zeroing 72 selected EEG coordinates costs 0.69 percentage points of retrieval accuracy, versus 0.06 for random coordinates. The ratio before display rounding is 10.74; the readout ratio is 1.25. This combines a large relative intervention effect with a sub-percentage-point accuracy drop.

Across budgets of 5%–50%, the across-subject intervention ratio decreases from 10.74 to 3.45, while the readout ratio ranges from 1.10 to 1.25. Over the same range, selected-minus-random damage increases from approximately 0.63 to 1.98 percentage points. Figure 2a should therefore be read together with the absolute-effect columns: the reduction in relative specificity accompanies an increase in absolute excess damage. This ordering is consistent with the concentration model of Section 4.

Selection advantage across EEG and vision. In the within-subject EEG setting, intervention specificity ranges from 1.68 to 2.20 across retained cells, compared with readout specificity of 1.12–1.26. In CLIP, the readout ratios at 5%, 10%, 20%, and 50% are 1.0138, 1.0083, 1.0039, and 1.0011, re-

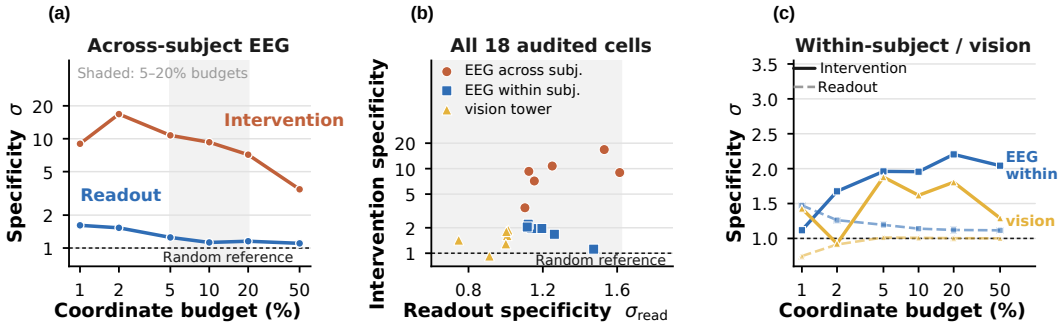

Figure 2: **Selection specificity differs between decoding and intervention.** **(a)** Across-subject EEG ratios over six budgets. The shading highlights 5%–20%; the 50% cell also passes the reporting threshold. **(b)** All eighteen cells, including four below-threshold intervention ratios retained for completeness. **(c)** Within-subject EEG and vision curves; solid lines denote intervention and dashed lines decoding. Table 1 identifies excluded cells and provides absolute effects. Vision readout is near unity at budgets of 5%–50%.

spectively. The corresponding intervention ratios are 1.88, 1.62, 1.81, and 1.29. These are small decoding advantages alongside larger substitution effects. The 2% vision cell is a boundary case: both ratios fall below one (0.91 readout, 0.93 intervention), so selection offers no advantage on either measurement there.

All fourteen cells above the intervention reporting threshold satisfy $\sigma_{\text{int}} > \sigma_{\text{read}}$. The four below-threshold conditions are marked in Table 1: across-subject EEG at 1% and 2%, and within-subject EEG and vision at 1%. Because cells share encoders, inputs, and nested budgets, this count measures consistency across conditions. Monotone decay is clearest in the across-subject EEG sweep.

Readout sensitivity at small budgets. At the smallest EEG budgets, the readout advantage reaches 1.61 across subjects and 1.47 within subjects. Selection thus matters more for decoding when few coordinates are available, while the across-subject intervention gap persists at wider budgets. A ratio near one describes similar selected and random accuracies; its interpretation depends on their absolute level.

6.3 Structured concept responses coexist with the specificity gap

To interpret the gap, we examine the concept structure from which coordinates were selected. Both EEG and vision representations show measurable concept structure. Figure 3 connects three observations: EEG-to-text errors share structure with image-to-text errors, selectivity varies by activation site, and concept profiles align across the two modalities.

The EEG-to-text pathway transfers the alignment learned through EEG–image training and CLIP pretraining, with no direct EEG–text training objective (Appendix C). We compare its errors with image-to-text errors using off-diagonal entries of the category-confusion matrices. The pooled Spearman correlation is $\rho = 0.227$, and correlations are positive for each of ten participants. Thus, the two pathways share a pattern in which visual categories they confuse. Appendix D gives the per-participant results and the label-permutation procedure.

For site selectivity, the primary EEG flatten block has standardized selected-minus-random scores of 2.77 within subjects and 2.75 across subjects. In the vision tower, 49 of 50 concepts reach their selectivity maximum in block 22 or 23. Across 200 concepts, the EEG and vision selectivity profiles correlate at $\rho = 0.250$ within subjects and 0.218 across subjects. These measurements motivate the audited sites and establish structured responses; the matched specificity grid then asks what is gained by choosing their most selective coordinates.

In the additional EEG backbone's one-participant, across-subject comparison, the flatten block ranks first among fifteen sites (1.61), followed by spatial (1.44) and temporal (1.43) sites (Appendix E; Figure 5). Peak selectivity thus occurs at a similar site.

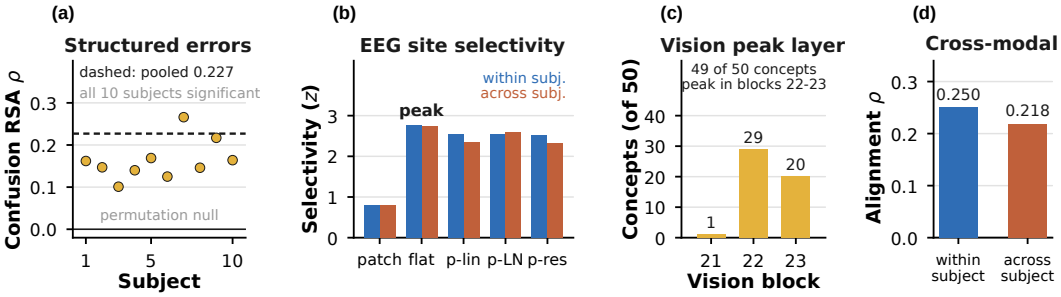

Figure 3: **Concept structure across EEG and vision representations.** **(a)** EEG-to-text versus image-to-text confusion correlation by participant, with the pooled value dashed. **(b)** Standardized site selectivity in the primary EEG encoder under two subject splits. **(c)** Peak-selectivity vision block for 50 concepts. **(d)** Across-concept alignment of vision and EEG selectivity profiles. These measurements establish structured responses and motivate the audited sites; the selected/random grid tests their selection specificity.

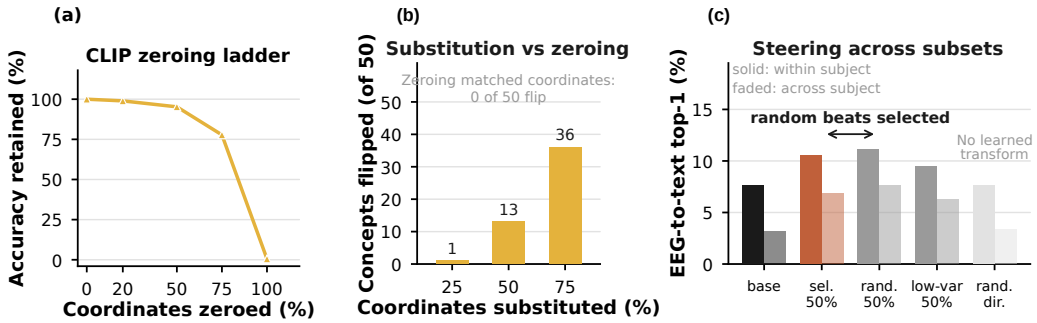

Figure 4: **Removal, replacement, and learned steering.** **(a)** CLIP accuracy retained relative to its unmodified baseline under coordinate zeroing. **(b)** Donor-identity flips under substitution in 50 concepts; zeroing the matched 75% coordinates gives zero donor flips. **(c)** EEG-to-text steering with selected, random, and low-variance subsets; the last condition uses a random direction without the learned transform. The no-transform values are close to baseline. Table 2 additionally reports sparse anchoring.

6.4 Removal–Replacement Asymmetry in the Vision Encoder

The CLIP case study shows *removal–replacement asymmetry*: a representation can tolerate zeroing while responding strongly to donor substitution. At a 75% budget, donor substitution makes the donor concept top-1 for 36 of 50 concepts; the matched zeroing control produces zero donor-identity flips. Figure 4b compares these operations at the same coordinates.

The substitution effect grows with budget: donor-identity flips rise from $1/50$ at 25% to $13/50$ at 50% and $36/50$ at 75%. At the largest budget, the original concept remains top-1 in only $2/50$ cases after substitution, compared with $41/50$ after zeroing. Substitution both suppresses the original prediction and promotes donor identity.

Figure 4a provides the CLIP removal ladder. Accuracy is 94.5% before removal, 90.0% after removing half the selected coordinates, and 73.5% after removing three quarters. Zeroing therefore leaves substantial recognition performance at these budgets. The 50-concept comparison in panel b tests a different outcome: whether substitution directs the prediction toward a specific donor. Read together, the panels show both preservation of recognition after removal and directional changes after replacement. The matched substitution-versus-zeroing comparison establishes the operation effect at the same selected coordinates.

Table 2: Selected versus random coordinates for EEG-to-text interventions at a 50% budget. Accuracies are percentages; chance is 0.5%. Random subsets outperform selected subsets in both steering rows. Anchoring instead gives positive selected-minus-random differences of 0.05, 0.05, and 0.40 percentage points. Trainable fraction refers to updated EEG-encoder parameters; the steering map is fitted separately with the backbone frozen.

Intervention family	Trainable	None	Selected	Random	Lowest-var	No transform
Learned linear steering with frozen backbone, 50% budget						
within subject	0%	7.65	10.50	**11.15**	9.45	7.60
across subject	0%	3.20	6.85	**7.60**	6.25	3.35
Sparsity-constrained anchoring during training, across subject, 50% budget						
projection tuning	94.25%	3.20	3.45	3.40	n/a	n/a
1×1 convolution	0.091%	3.20	2.85	2.80	n/a	n/a
learnable anchor vector	0.029%	3.20	3.30	2.90	n/a	n/a

6.5 Steering across subsets

Steering across subsets extends the comparison from disrupting predictions to improving retrieval. A learned transform raises EEG-to-text accuracy through selected, random, and low-variance subsets; random subsets perform best in both subject settings. Table 2 separates these results from sparse anchoring during training, and Figure 4c compares the selected, random, low-variance, and no-transform controls.

A linear transform maps EEG embeddings toward concept directions in the frozen text tower and is applied without the test concept label. With selected coordinates, accuracy increases from 3.20% to 6.85% across subjects and from 7.65% to 10.50% within subjects. Random half-coordinate sets reach 7.60% and 11.15%, exceeding the selected conditions by 0.75 and 0.65 percentage points. Even the lowest-variance half reaches 6.25% and 9.45%. The no-transform control yields 3.35% and 7.60%, close to the corresponding baselines. The learned transformation improves accuracy across all three subset-selection rules.

Sparse anchoring adds a training objective on the selected coordinates. Across projection tuning, a 1×1 convolution, and a learned anchor vector, selected accuracies are 3.45%, 2.85%, and 3.30%; random-control accuracies are 3.40%, 2.80%, and 2.90%. The selected-minus-random differences are therefore $+0.05$, $+0.05$, and $+0.40$ percentage points. Compared with the unchanged 3.20% baseline, the selected arms change by $+0.25$, -0.35, and $+0.10$ points. Across these parameterizations, the selection advantage accompanies either a small gain or a loss relative to the untouched model. Appendix G gives the additional reported regularization setting.

Consequences for method choice. Comparing each intervention with the untouched baseline measures task improvement; comparing selected and random subsets measures the benefit of selection. For these EEG evaluations, random subsets are a useful reference when choosing coordinates for learned steering: they improve retrieval more than the concept-selected subsets. Sparse anchoring illustrates why its selected–random difference should also be compared with the untouched model.

7 Conclusion

BEAM measures what choosing concept-responsive coordinates contributes to interpreting a frozen encoder. The selection-specificity gap shows that this contribution is larger for intervention than for decoding across the audited EEG and vision settings. Removal–replacement asymmetry in CLIP and steering across subsets in EEG connect the value of selection to the operation applied.

The concentration-and-recoverability account provides a possible explanation: strong dependence on particular coordinates can coexist with concept information recoverable from alternative subsets. Signal-geometry and downstream-sensitivity measurements could test this account further. Evidence spans one dataset; the additional backbone contributes site-selectivity comparisons.

Ethics Statement

This study uses the public THINGS-EEG2 dataset of EEG recordings collected while participants viewed natural images (Gifford et al., 2022). It involves no new participant recruitment or data collection. The analysis concerns recognition among a fixed set of visual concepts under controlled stimulus presentation. The methodological aim is to improve the evaluation of claims about concept-responsive coordinates by comparing selected and random coordinate sets under matched procedures.

Reproducibility Statement

Equations 1–5 define the audit and Algorithm 1 specifies the matched comparisons. Table 1 reports the complete main grid, including absolute intervention effects and threshold flags. The appendix provides the available site summaries, retrieval baselines, intervention ladders, and steering controls. The accompanying source contains the figure assets, plotting scripts, and reported derived-number summaries. The source supports recompiling the manuscript and inspecting the reported summaries; independent experimental reproduction additionally requires the underlying measurements and checkpoints.

AI Use Statement

Generative AI tools assisted with manuscript drafting and language editing, and with developing and debugging experimental code. The authors are responsible for the final manuscript, code, and experimental results.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1704.05796.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.14997. arXiv:2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Paper-Conference.pdf.

Jianrong Ding, Muxi Chen, Chenchen Zhao, and Qiang Xu. Concept-SAE: A controllable and invertible concept interface for sparse autoencoders, 2026. URL https://arxiv.org/abs/2509.22015. arXiv:2509.22015v2; accepted at ECML PKDD 2026.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://arxiv.org/abs/2209.10652.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in CLIP. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ff73af253974e6144c4eddf896d3095a-Abstract-Conference.html.

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. URL https://doi.org/10.1016/j.neuroimage.2022.119754.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. URL https://distill.pub/2021/multimodal-neurons/.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://arxiv.org/abs/2305.01610. arXiv:2305.01610.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. doi: 10.1016/j.neuroimage.2013.10.067. URL https://doi.org/10.1016/j.neuroimage.2013.10.067.

Lucas Hayne, Abhijit Suresh, Hunar Jain, Rahul Kumar, and R. McKell Carter. Much easier said than done: Falsifying the causal relevance of linear decoding methods. *arXiv preprint arXiv:2211.04367*, 2022. URL https://arxiv.org/abs/2211.04367.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025. URL https://arxiv.org/abs/2510.09794.

Changhong Jing, Yan Liu, Shuqiang Wang, Bruce X. B. Yu, Gong Chen, Zhejing Hu, Zhi Zhang, and Yanyan Shen. PTSM: Physiology-aware and task-invariant spatio-temporal modeling for cross-subject EEG decoding, 2025. URL https://arxiv.org/abs/2508.11357.

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 (AAAI)*, volume 40, pp. 5575–5583, 2026. doi: 10.1609/aaai.v40i7.37476. URL https://ojs.aaai.org/index.php/AAAI/article/view/37476.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 102822–102864, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba5f1233efa77787ff9ec015877dbd1f-Abstract-Conference.html.

Dongyang Li, Haoyang Qin, Mingyang Wu, Chen Wei, and Quanying Liu. BrainFLORA: Uncovering brain concept representation via multimodal neural embeddings. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5577–5586, 2025. doi: 10.1145/3746027.3754996. URL https://doi.org/10.1145/3746027.3754996.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03341.

Shaotian Li, Shangze Li, Chuancheng Shi, Wenhua Wu, Yanqiu Wu, Xiaohan Yu, Fei Shen, and Tat-Seng Chua. Latent anomaly knowledge excavation: Unveiling sparse sensitive neurons in vision-language models, 2026a. URL https://arxiv.org/abs/2604.07802.

Yanshi Li, Xueru Bai, Shuman Liu, Haibo Zhang, and Anxiang Zeng. Decomposing and steering functional metacognition in large language models. *arXiv preprint arXiv:2605.08942*, 2026b. URL https://arxiv.org/abs/2605.08942.

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*, volume 40, pp. 17580–17588, 2026. doi: 10.1609/aaai.v40i21.38813. URL https://ojs.aaai.org/index.php/AAAI/article/view/38813.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Ari S. Morcos, David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1803.06959.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/89e83382abeee53b932a6df62edbf9cc-Abstract-Conference.html.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4609–4622, 2020. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Arshak Rezvani, Ali Akbari, Kosar Sanjar Arani, Maryam Mirian, Emad Arasteh, and Martin J. McKeown. Interpretable EEG-to-image generation with semantic prompts, 2025. URL https://arxiv.org/abs/2507.07157. ICML 2025 Actionable Interpretability Workshop (non-archival).

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers. *arXiv preprint arXiv:2604.22128*, 2026. URL https://arxiv.org/abs/2604.22128.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. VL-SAE: Interpreting and enhancing vision-language alignment with a unified concept set. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/407106f4b56040b2e8dcad75a6e461e5-Abstract-Conference.html.

Chuancheng Shi, Shangze Li, Shiming Guo, Simiao Xie, Wenhua Wu, Jingtong Dou, Chao Wu, Canran Xiao, Cong Wang, Zifeng Cheng, Fei Shen, and Tat-Seng Chua. Where culture fades: Revealing the cultural gap in text-to-image generation, 2025. URL https://arxiv.org/abs/2511.17282.

Chuancheng Shi, Shangze Li, Wenjun Lu, Wenhua Wu, Cong Wang, Zifeng Cheng, Fei Shen, and Tat-Seng Chua. TraceRouter: Robust safety for large foundation models via path-level intervention, 2026. URL https://arxiv.org/abs/2601.21900.

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. URL https://arxiv.org/abs/2308.13234. arXiv:2308.13234.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

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, 2025. URL https://arxiv.org/abs/2511.09250.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

Enshang Zhang, Zhicheng Zhang, and Takashi Hanakawa. Category-aware EEG image generation based on wavelet transform and contrast semantic loss. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7922–7930, 2025. doi: 10.24963/ijcai.2025/881. URL https://www.ijcai.org/proceedings/2025/881.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

Xianhui Zhang, Chengyu Xie, Linxia Zhu, Yonghui Yang, Weixiang Zhao, Zifeng Cheng, Cong Wang, Fei Shen, and Tat-Seng Chua. Who transfers safety? identifying and targeting cross-lingual shared safety neurons, 2026. URL https://arxiv.org/abs/2602.01283.

A Implementation Details

Corpus and preprocessing. THINGS-EEG2 contains recordings from ten participants acquired with a 64-channel system while they viewed natural images (Gifford et al., 2022). The reported evaluation uses 200 concepts held out from EEG encoder training. Within-subject evaluation uses the same participant for training and testing; across-subject evaluation holds out the evaluated participant. Each backbone follows its original preprocessing convention, including channel selection, temporal resampling, and normalization where applicable.

Table 3: Audit configuration. Backbone weights remain frozen during the audit; probes are fitted separately. d is the width of the audited site; post-LN is a tower's final layer normalization; T1 is top-1 accuracy. The frozen text tower of the same backbone (Radford et al., 2021) is the EEG-to-text intervention target of Table 2. All rows are evaluated on THINGS-EEG2 (Gifford et al., 2022).

Encoder	Role in the audit	Site	d	Downstream task T
CLIP ViT-L/14 (Radford et al., 2021)	vision comparison	post-LN	1024	200-way image-to-text T1
NICE-EEG (Song et al., 2024)	brain encoder, primary	flatten	1440	200-way EEG-to-image T1
DWT (Zhang et al., 2025)	site comparison	flatten	1479	200-way EEG-to-image T1

Selection and readout. The selection score is the variance across selection concepts of the contrastive activation difference in Eq. 1. A shared selected set is formed for each encoder and budget. The ridge classifier uses five-fold cross-validation with ten trials per evaluation concept. The regularizer is held fixed across budgets and selected/random arms. Random sets are drawn uniformly without replacement and their effects are averaged over independent draws.

Interventions. The EEG grid zeros audited coordinates and measures EEG-to-image retrieval loss. The vision grid substitutes donor activations and measures image-to-text classification loss. The separate vision ladder compares substitution and zeroing over 50 concepts. EEG-to-text steering is evaluated at a 50% coordinate budget. Its learned linear map leaves backbone weights frozen; the anchoring variants update the trainable components listed in Table 2.

Computation. The reported experiments used a consumer graphics card with 16 GB memory. The audit consists of activation extraction, closed-form probe fitting, and intervention forward passes. Sparse anchoring additionally updates encoder parameters.

B Retrieval Baselines

Table 4 retains the reported retrieval baselines. The rows compare retrieval across EEG architectures; the specificity grid evaluates coordinate selection in the primary EEG and vision encoders. An unmarked value is reported as a reproduction of a configuration evaluated in the cited work; † denotes an added cross-subject evaluation.

Table 4: 200-way zero-shot retrieval on THINGS-EEG2, reported means over ten subjects. Values are accuracy fractions; T1 and T5 denote top-1 and top-5. † marks an added cross-subject evaluation. n/a denotes an unavailable baseline. The audit grid and intervention ladders each use their own evaluation baselines.

Backbone	Within T1	Within T5	Across T1	Across T5
NICE-EEG (Song et al., 2024)	0.135	0.377	n/a	n/a
DWT (Zhang et al., 2025)	0.334	0.6115	$0.0855^{\dagger}$	$0.236^{\dagger}$
HyFI (Jo et al., 2026)	0.687	0.918	0.135	0.373
Blur-prior decoder (Liu et al., 2026)	0.805	0.969	0.203	0.485

C THREE-PATH ZERO-SHOT TRANSFER

The image-to-text and EEG-to-image paths use the alignment learned during CLIP pretraining and EEG-encoder training, respectively. EEG-to-text composes these learned alignments to retrieve text concepts from EEG, with no direct EEG–text training objective.

D CONFUSION-STRUCTURE ANALYSIS

Off-diagonal entries of the 27×27 confusion matrices are compared using Spearman correlation. The reported null distribution uses 5,000 label permutations applied symmetrically to rows and columns. EEG-to-text versus image-to-text correlations are positive in each participant. The pooled coefficients are computed from subject-aggregated confusion matrices and can therefore exceed the individual coefficients.

Table 5: Per-participant off-diagonal confusion correlations. Reported label-permutation tests give $p < 0.01$ for all entries. The pooled row uses subject-aggregated confusion matrices.

Subject	EEG-to-image vs image-to-text	EEG-to-text vs image-to-text	EEG-to-image vs EEG-to-text
S1	0.135	0.162	0.089
S2	0.239	0.147	0.116
S3	0.089	0.101	0.179
S4	0.217	0.140	0.139
S5	0.131	0.169	0.095
S6	0.130	0.125	0.156
S7	0.173	0.266	0.121
S8	0.237	0.146	0.105
S9	0.181	0.217	0.129
S10	0.225	0.164	0.127
pooled	**0.243**	**0.227**	**0.270**

E SITE-WISE SELECTIVITY

Standardization against site-specific random controls makes the selectivity summaries comparable across activation scales. In the primary EEG encoder, the projection layer-normalization site has the largest reported raw score (1.44), whereas the flatten block has the largest standardized score. Table 6 gives the full site comparison. Figure 5 visualizes the additional backbone's fifteen-site, one-participant result.

Table 6: Standardized site-wise selectivity. Left: primary EEG encoder, within- and across-subject settings. Right: additional backbone, one participant under across-subject evaluation. The additional encoder has 1479 coordinates at its flatten site.

Site (primary)	Within	Across	Site (additional)	Width	Across
patch tokens	0.81	0.80	**eeg flatten**	1479	**1.61**
flatten	**2.77**	**2.75**	eeg spatial	40	1.44
projection linear	2.55	2.36	eeg temporal	40	1.43
projection layer norm	2.55	2.60	projection residual	1024	1.40
projection residual	2.51	2.32	projection layer norm	1024	1.40
			eeg fusion	40	1.40
			fusion convolution	63	1.39
			eeg projection	40	1.39
			wavelet	64	1.38
			local electrode	63	1.37
			projection linear	1024	1.30
			inverse wavelet	63	1.29
			local time	63	1.27
			embedding value	256	0.37
			transformer	256	0.37

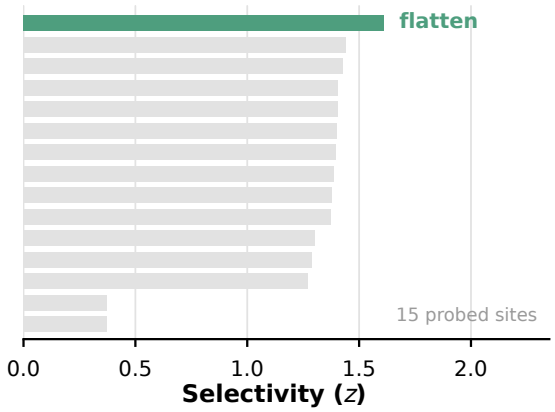

Figure 5: Site selectivity for the additional EEG backbone: fifteen sites, one participant, across-subject setting. The flatten block has the highest reported score. Table 6 lists all sites for the site comparison.

F Intervention Ladders

Table 7 reports CLIP accuracy under coordinate zeroing and the original- and donor-identity outcomes under substitution. The removal ladder measures classification accuracy; the 50-concept comparison counts which concept becomes top-1 after the intervention.

Table 7: CLIP removal and substitution outcomes. Left: top-1 accuracy after selected-coordinate zeroing. Right: top-1 identity in the 50-concept comparison; A is the original concept and B is the donor concept.

Zeroed	Accuracy (%)	Substituted	A still top-1	B becomes top-1
0%	94.5	0%	48/50	0/50
20%	93.5	25%	37/50	1/50
50%	90.0	50%	17/50	13/50
75%	73.5	75%	2/50	**36/50**
100%	0.5	75% (zeroed)	41/50	0/50

Amplification control. Scaling only the selected EEG coordinates gives EEG-to-text accuracy of 7.65% at baseline, followed by 7.75%, 7.50%, 6.65%, 6.40%, and 6.50% for $\alpha \in \{1.5, 2, 3, 5, 10\}$. Scaling a subset changes both its magnitude and potentially the direction of the full representation. Accuracy peaks at $\alpha = 1.5$ and is below baseline at larger tested values.

G Sparse Anchoring

Table 2 reports three parameterizations at a 50% coordinate budget. Projection tuning uses regularization strength 0.1 and a 50-epoch configuration; the convolution and learnable-vector entries also use strength 0.1. The additional learnable-vector entry at strength 1.0 gives 3.25% for selected coordinates and 2.85% for random coordinates, a 0.40-point difference. Its selected arm is 0.05 points above the 3.20% untouched baseline.

H Reporting Threshold

The intervention ratio is included in cross-setting summaries when the mean random-control drop is at least 0.05 percentage points. Applied to the full grid, this excludes the across-subject EEG 1% and 2% cells, the within-subject EEG 1% cell, and the vision 1% cell. The vision 2% cell is retained: its random effect is 0.27 points, although its intervention specificity is below one. Fourteen of eighteen cells therefore enter the count in the main text. The threshold selects ratios by the magnitude of their random-control denominator.

I Related Methods by Measurement Role

Probe evaluation. Description length (Voita & Titov, 2020) and information-theoretic analysis (Pimentel et al., 2020) examine the information recovered by probes. Probe-derived steering links recoverable states to behavioral effects (Li et al., 2026b). These approaches inform interpretation of the readout measurement; BEAM focuses on the relative benefit of choosing a coordinate subset.

Feature dictionaries and neuron analyses. Feature dictionaries address representations distributed across neuron directions (Elhage et al., 2022; Cunningham et al., 2024), including vision-language features (Pach et al., 2025; Shen et al., 2025) and user-defined concept interfaces (Ding et al., 2026). Second-order effects provide another view of neuron function in CLIP (Gandelsman et al., 2025). Selected neurons have also been used to study cross-lingual safety, anomalies, and cultural representations (Zhang et al., 2026; Li et al., 2026a; Shi et al., 2025). BEAM evaluates subsets of the encoder's activation coordinates; dictionary features define a different representation for selecting and intervening on concepts.

Interventions at different granularities. Causal mediation, model editing, and circuit discovery target behavioral mechanisms at different granularities (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Shi et al., 2026). BEAM holds the activation site and coordinate count fixed within each selected–random comparison. This isolates the choice of coordinates for the specified intervention and metric.

Table 8: Related work grouped by measurement role. The final column identifies the comparison relevant to BEAM.

Study	Measurement	Connection to the audit
Hayne et al. (2022)	Linear decoding and ablation	Relates selectivity to unit importance.
Huang & Chang (2025)	Probes and activation patching	Compares depth profiles of decoding and causal influence.
Sharma et al. (2026)	Hierarchical probes and interventions	Tests causal roles of decodable variables.
Gurnee et al. (2023)	Sparse probing	Examines information in small coordinate subsets.
Pach et al. (2025)	Vision-language feature dictionaries	Evaluates interpretable features and interventions.
Zhang & Nanda (2024)	Intervention metrics and methods	Shows why the intervention protocol matters.
Song et al. (2024)	EEG temporal, spatial, and semantic analyses	Relates EEG inputs and representations to visual concepts.
Rezvani et al. (2025)	Semantic prompts and EEG saliency	Interprets EEG-to-image generation.
BEAM	Selected/random effects on two axes	Measures the benefit of selection at matched budgets and sites.