ENCODING WITHOUT ALIGNMENT: ANALYZING THE CORRESPONDENCE BETWEEN LLMS AND THE VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM) embeddings learned only from text can predict human visual-cortical responses to natural scenes, but whether this predictive correspondence implies shared representational geometry remains unclear. Using fMRI responses from all eight Natural Scenes Dataset subjects and Llama-3-8B representations of the corresponding COCO captions, we compare two notions of brain–model correspondence. First, voxel-wise ridge encoding yields positive held-out prediction accuracy across cortical regions, with higher accuracy in body-selective cortex than in early visual cortex and little variation across Llama layers. Second, we characterize local representational geometry within semantic neighborhoods. Principal-component variation patterns over the shared neighboring stimuli define subspaces in a common neighbor-index space, whose orientations we compare using chordal distance. Within each system, tangent-space distance increases with separation along the cortical ROI ordering and with Llama layer separation. Across systems, however, alignment remains close to an isotropic random-subspace baseline, while being modestly stronger in EBA and FBA than in early visual cortex. Local variance spectra also dissociate: cortical effective dimensionality varies comparatively little across regions, while Llama effective dimensionality increases with layer depth. An exploratory residual-stream intervention using encoding-derived directions produces mixed and nonspecific effects across four subjects. In contrast to Llama, CLIP text-encoder representations show increasing cortical encoding accuracy and stronger cross-system alignment with depth. These results demonstrate that accurate linear encoding can coexist with weak alignment of dominant local variation subspaces.

1 INTRODUCTION

Understanding the relationship between biological and artificial intelligence remains a key challenge in computational neuroscience and machine learning. One approach to investigating this relationship is by comparing representations of the same information in artificial models and the human brain. Because these representations occupy spaces with different dimensionality and coordinate systems, correspondence is typically established by comparing activation patterns with linear encoding models or representational similarity analysis (RSA) (Kriegeskorte et al., 2008). Recent studies of brain-model correspondence show that representations extracted from large language models (LLMs) capture information that predicts human cortical responses to natural language (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Tang et al., 2023) and, more recently, to visual scenes described linguistically (Doerig et al., 2025). A high encoding score indicates that model features are linearly informative about brain activity, but does not reveal whether the two representation spaces share similar geometric structure. RSA and related measures compare global similarity structure across stimuli (Kornblith et al., 2019), but not the orientation of local variation within semantic neighborhoods.

In this paper, we consider two geometric notions of model-brain alignment. First, *linear encoding* asks whether brain activations can be predicted by linear combinations of LLM features. Second, *tangent-space orientation alignment* asks whether local representational variation is similarly organized across brain and LLM representations. Because the two systems have different native coordinates, their

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.

principal directions cannot be compared directly. Instead, within each local semantic neighborhood, we express the corresponding patterns of variation over the same stimuli, providing a common stimulus space in which local subspace orientations can be compared. We evaluate these notions using 7T fMRI responses to natural scenes and text-only Llama-3-8B embeddings of captions describing the same scenes, comparing individual brain regions of interest (ROIs) with LLM layers. Our results show that linear predictability and tangent-space orientation provide distinct characterizations of brain-model correspondence, which we establish through four principal contributions:

1. We characterize local representational variation within semantic neighborhoods through its PCA spectrum in each system's native space (voxel space for cortical ROIs, activation space for LLM layers). Local spectra differ only modestly across the six cortical ROIs (V1, V2, V3, hV4, EBA, FBA), but in Llama-3-8B, local variance is dominated by the leading principal component in early layers and becomes distributed across more components with depth.
2. We find that local geometry reflects hierarchical organization within both systems. Within each semantic neighborhood, we define tangent spaces from the dominant variation patterns over the neighboring stimuli. Chordal distance between these local tangent spaces increases with separation along the cortical hierarchy (mean Spearman $\rho = 0.78$ across subjects), with the largest step between hV4 and EBA, and increases with layer separation in Llama-3-8B ($\rho = 0.84$).
3. We show that ridge regression from Llama-3-8B representations yields positive held-out prediction accuracy across all ROIs, with higher accuracy in body-selective regions than in early visual cortex in all eight subjects, but little variation across Llama layers. In contrast, brain–LLM tangent-space alignment is weak: chordal distances remain close to the isotropic random-subspace expectation, although alignment is modestly stronger in EBA and FBA than in early visual cortex. Neither measure provides evidence of systematic layer–region correspondence. A comparison using the CLIP ViT-L/14 text encoder on the same captions instead shows increasing encoding accuracy and stronger tangent-space alignment with depth (Appendix S5), indicating that Llama's layer-invariant correspondence is not a general property of caption representations.
4. We derive LLM steering vectors from ridge-encoding weights and brain PCA directions, characterize their organization across layers and ROIs, and evaluate them in an exploratory residual-stream intervention across four subjects. Injecting each subject's FBA-derived direction increases the probability of sports-related caption tokens in three of four subjects. However, we find that the effects are not specific to the predicted categories, and the corresponding V1-derived direction does not provide a clean negative control.

2 Related Work

Prior work in brain-LLM alignment has established predictive correspondence between model activations and brain activity, but provides limited comparison of the structure of their underlying representation spaces. Contextual embeddings from LLMs predict neural responses during language comprehension (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Toneva & Wehbe, 2019), with alignment improving with scale and training (Aw et al., 2024), and continuous language studies additionally report hierarchical alignment between LLM depth and both cortical organization and neural response timing (Raugel et al., 2025; Caucheteux & King, 2022; Caucheteux et al., 2023; Goldstein et al., 2022; Défossez et al., 2023). Linear encoding similarly predicts fMRI voxel responses to viewing natural images in high-level visual cortex from large language text and image embeddings (Doerig et al., 2025). Remarkably, many different vision models achieve similar encoding performance on NSD: Conwell et al. (2024) test 224 vision models and find that 126 of them reach occipitotemporal predictivity scores within $r = 0.1$ of one another despite wide variation in architecture, training objective, and image diet, which indicates that predictive alignment alone does not distinguish a model's representational organization.

Representational geometry provides a complementary framework for studying neural and model representations. Neural populations exhibit high-dimensional structure, including power-law eigenvalue spectra (Stringer et al., 2019) and effective dimensionality that increases unboundedly with

neuron count (Manley et al., 2024). Work on object-manifold geometry further characterizes representations in terms of distinct properties such as dimension, radius, and orientation (DiCarlo & Cox, 2007; Chung et al., 2018; Cohen et al., 2020; Chung & Abbott, 2021; Sorscher et al., 2022), allowing us to compare the organization of model and neural representations even when encoding performance is similar. Cross-model alignment has also been measured locally, through mutual nearest neighbors (Huh et al., 2024); our diagnostic instead compares the orientation of local variation subspaces.

Finally, the cortical visual hierarchy provides a known organizational structure against which geometric model-brain correspondence can be tested. The transition from retinotopic early visual cortex to object-selective lateral and ventral cortex is one of the most established divisions in visual neuroscience (Grill-Spector et al., 2001; Larsson & Heeger, 2006; Bao et al., 2020), with representational-similarity and model-based analyses repeatedly recovering an early-visual cluster and a late-ventral cluster (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). We therefore test whether geometric relationships between model and neural representations recover this established cortical organization.

3 METHODS

3.1 REPRESENTATIONS AND SEMANTIC NEIGHBORHOODS

For brain region u with $d_B^{(u)}$ voxels, we write $\mathbf{r}_B^{(u)}(I) \in \mathbb{R}^{d_B^{(u)}}$ for the trial-averaged voxel response to stimulus I. Language representations are derived from the COCO captions associated with each image (up to five) using the pretrained base checkpoint `meta-llama/Meta-Llama-3-8B`. Each caption is processed independently. For LLM layer ℓ with hidden dimension d_L, we mean-pool the contextual token representations and then average caption-level vectors with equal weight to obtain one image-level representation $\mathbf{r}_L^{(\ell)}(I) \in \mathbb{R}^{d_L}$.

To compare local geometry across brain and LLM representations, we define a common semantic neighborhood $\mathcal{N}_k(I)$ for each stimulus I from Euclidean distances between L2-normalized CLIP ViT-B/32 image embeddings (Radford et al., 2021); we use $k = 100$ neighbors, excluding I itself. CLIP provides a suitable common semantic reference because (i) its joint vision-language embedding space is trained to align corresponding text and image representations, (ii) it is pretrained independently of the NSD responses and of Llama (though not of the CLIP text encoder analyzed in Appendix S5), and (iii) it is used as a reproducible semantic-similarity reference (Doerig et al., 2025).

3.2 LOCAL SPECTRUM

For each stimulus I, we estimate the dimensionality of its response variation within its neighborhood index set $\mathcal{N}_k(I)$. For brain region u, we form the centered neighborhood response matrix

$$X_u(I) \;=\; \big[\,\mathbf{r}_B^{(u)}(J) - \bar{\mathbf{r}}_B^{(u)}(I)\,\big]_{J\in\mathcal{N}_k(I)}^{\top} \;\in\; \mathbb{R}^{k\times d_B^{(u)}}, \qquad \bar{\mathbf{r}}_B^{(u)}(I) \;=\; \tfrac{1}{k}\sum_{J\in\mathcal{N}_k(I)} \mathbf{r}_B^{(u)}(J). \quad (1)$$

Let γ_i denote the singular values of $X_u(I)$, so the local covariance eigenvalues are $\mu_i = \gamma_i^2/(k-1)$. We compute the cumulative fraction of variance explained by the first m principal components, $C_m(I) = \sum_{i\le m}\mu_i / \sum_i \mu_i$, and average it over all stimuli for each ROI. We repeat this on the LLM hidden states $\mathbf{r}_L^{(\ell)}$ (not standardized) over the same neighborhoods.

3.3 TANGENT-BASIS ORIENTATION IN NEIGHBOR-INDEX SPACE

The SVD of the centered neighborhood matrix, $X_u(I) = U_u(I)\,\Gamma_u(I)\,V_u(I)^{\top}$, isolates the dominant directions of local response variation in both neighbor and voxel space. The right singular vectors (columns of $V_u(I)$) span the local tangent basis in native voxel space, while the leading $d = 8$ left singular vectors, $U_u(I) \in \mathbb{R}^{k\times d}$, span the corresponding variation patterns over the k neighbors. Because the same k neighbors are used for every region and layer, $U_u(I)$ lives in a common neighbor-index space and can be compared across regions and across systems.

For regions or layers u_1 and u_2, let $M = U_{u_1}(I)^{\top} U_{u_2}(I) \in \mathbb{R}^{d\times d}$. Its singular values $\sigma_i(M) = \cos\theta_i$ are the cosines of the principal angles between the two subspaces, and we use the (squared)

chordal distance

$$\mathrm{Chord}(u_1, u_2; I) \;=\; \sum_{i=1}^{d} \sin^2\theta_i \;=\; d - \|M\|_F^2 \;\in\; [0, d], \tag{2}$$

which equals 0 when the subspaces coincide and d when they are orthogonal. Equivalently, $\|M\|_F^2 = d - \mathrm{Chord}$ measures subspace overlap, with larger values indicating greater alignment. For every pair of ROI and layer representation spaces, we average the chordal distance across semantic neighborhoods. The resulting pairwise dissimilarities form a joint brain $\times$ LLM matrix, with within-brain comparisons in the top-left block, within-LLM comparisons in the bottom-right block, and cross-system comparisons in the off-diagonal blocks.

The maximum distance d is not an informative reference, because two *random* d-dimensional subspaces are not orthogonal. The columns of $U_u(I)$ are orthogonal to the all-ones vector (the data are centered), so we draw random frames uniformly (Haar measure) from the $(k-1)$-dimensional centered subspace, for which $E\big[\|M\|_F^2\big] = d^2/(k-1)$ and

$$E[\mathrm{Chord}] \;=\; d - \frac{d^2}{k-1}. \tag{3}$$

For $d = 8$ and $k = 100$ the expectation is 7.354 (chance overlap 0.65), and a single random pair has standard deviation 0.104 (Monte Carlo, 2×10^4 draws). Because each reported distance averages about 10^4 neighborhoods, the spread of a single random pair is not the relevant reference; we therefore test at the group level, with one-sample t-tests and bootstrap confidence intervals across subjects of the per-subject means against Equation (3).

3.4 Voxel-wise ridge encoding

For each voxel v in region u, we fit a linear map from layer-ℓ features to that voxel's response with ℓ_2 regularization,

$$\hat{\mathbf{w}}_v^{(u,\ell)} \;=\; \arg\min_{\mathbf{w}\in\mathbb{R}^{d_L}} \sum_{I\in\mathcal{I}_{\text{train}}} \big(r_v(I) - b_v - \mathbf{w}^\top \tilde{\mathbf{r}}_L^{(\ell)}(I)\big)^2 \;+\; \lambda_v \|\mathbf{w}\|_2^2, \tag{4}$$

where $\tilde{\mathbf{r}}_L^{(\ell)}(I) \in \mathbb{R}^{d_L}$ is the layer-ℓ representation of image I, z-scored with the training-set mean and standard deviation of each feature, and b_v is an intercept. We collect the fitted weights as rows of $W^{(u,\ell)} \in \mathbb{R}^{d_B^{(u)} \times d_L}$. For each subject, we use all images with available responses (10,000 for S1, S2, S5, and S7; fewer for S3, S4, S6 and S8, who completed 30-32 of the 40 NSD sessions) and a fixed 80/20 train-test split shared by all layers. The penalty $\lambda_v \in \{10^0, 10^1, \ldots, 10^6\}$ is selected independently for each voxel by efficient leave-one-out cross-validation on the training set (scikit-learn `RidgeCV`, mean squared error). Encoding performance is the Pearson correlation between predicted and observed responses on the held-out images, computed per voxel and averaged within each ROI.

3.5 Per-PC steering vectors

The ridge encoding of Equation (4) produces, for each (ROI, layer) pair, a weight matrix $W^{(u,\ell)}$. We decompose it into one LLM-space direction per global principal component of the ROI's response. Let $E_u \in \mathbb{R}^{d_B^{(u)} \times n_{\mathrm{pc}}}$ ($n_{\mathrm{pc}} = 8$) collect, as orthonormal columns, the top principal components of the centered $|\mathcal{I}| \times d_B^{(u)}$ response matrix over all images; these are the global analog of the local voxel-space tangent basis of Section 3.3. We define the steering matrix

$$S^{(u,\ell)} \;:=\; W^{(u,\ell)\top} E_u \;\in\; \mathbb{R}^{d_L \times n_{\mathrm{pc}}}, \tag{5}$$

whose i-th column $\mathbf{s}_i^{(u,\ell)} := W^{(u,\ell)\top}\mathbf{e}_i$ is the steering vector for the i-th brain PC. For any image, $(\mathbf{s}_i^{(u,\ell)})^\top \tilde{\mathbf{r}}_L^{(\ell)}(I) = \langle \mathbf{e}_i,\, \hat{\mathbf{r}}_B^{(u,\ell)}(I) - \mathbf{b}\rangle$: the steering vector reads out the predicted response along the i-th brain PC. The vectors live in the z-scored feature space; for interventions on the residual stream (Section 4.5) we map them to raw hidden-state coordinates by dividing each coordinate by the training-set feature standard deviation. We analyze two quantities: the *layer-layer cosine* $|\cos(\mathbf{s}_i^{(u,\ell)}, \mathbf{s}_i^{(u,\ell')})|$, which tracks how the direction read out for a brain PC changes with depth, and the *ROI-ROI cosine* $|\cos(\mathbf{s}_i^{(u,\ell)}, \mathbf{s}_i^{(u',\ell)})|$, which tests whether different regions read the same LLM direction for their i-th component. Absolute values are used because PC signs are arbitrary.

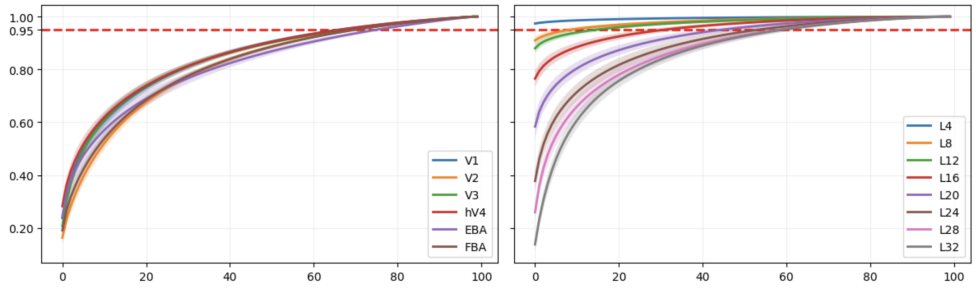

Figure 1: **Cumulative explained variance of local (neighborhood, $k = 100$) PCA for Subject 1**, averaged over neighborhoods. Left: the six cortical ROIs have broadly similar spectra. Right: Llama-3-8B spectra flatten with layer depth. The red dashed line marks 95% explained variance.

4 Results

We use the Natural Scenes Dataset (NSD) (Allen et al., 2022): 7T fMRI from eight subjects viewing COCO images (Lin et al., 2014), with single-trial GLMdenoise betas (`betas_fithrf_GLMdenoise_RR`, 1.8 mm) averaged over the (up to three) repetitions of each image. We analyze six ROIs, $\mathcal{U}_B = \{\mathrm{V1}, \mathrm{V2}, \mathrm{V3}, \mathrm{hV4}, \mathrm{EBA}, \mathrm{FBA}\}$, each intersected with the `nsdgeneral` mask: V1-hV4 from the NSD population-receptive-field ROIs, and the body-selective regions from the NSD `floc-bodies` localizer—EBA, in lateral occipitotemporal cortex (Downing et al., 2001), and FBA, the union of `FBA-1` and `FBA-2` in ventral occipitotemporal cortex (Peelen & Downing, 2005; Schwarzlose et al., 2005). We order them from early visual to body-selective cortex; the EBA→FBA order is a convention rather than an anatomical hierarchy. (In some appendix panels FBA carries its legacy label "vOTC".) Voxel counts in S1 are 1350, 1433, 1187, 687, 2238 and 696.

Model representations are derived from the COCO captions associated with each image (up to five). We process each caption independently, mean-pool its contextual token representations at each selected depth, and average caption-level vectors to obtain one representation per image. For the pretrained base model Llama-3-8B, we analyze representations after $\{4, 8, 12, 16, 20, 24, 28, 32\}$ decoder blocks; L32, unlike the intermediate representations, is returned after the model's final RMSNorm. For the CLIP ViT-L/14 text encoder, we analyze blocks $\{1, 3, 4, 6, 7, 9, 10, 12\}$. The tangent-space analyses use all eight subjects; the CLIP comparison and residual-stream intervention use the four subjects with complete data (S1, S2, S5, and S7); the steering-vector organization is shown for S1, S5, and S7.

4.1 Linear encoding predicts cortical responses with little layer specificity

Voxel-wise ridge regression (Equation (4)) from Llama-3-8B layer features predicts held-out responses in every ROI (Appendix Table S2). Across subjects, prediction accuracy ranges from $r = 0.14$-0.26 in early visual ROIs {V1, V2, V3, hV4} and from $r = 0.21$-0.40 in body-selective ROIs {EBA, FBA} (S1: 0.16-0.19 vs. 0.33-0.40). Both body-selective ROIs are predicted better than every early visual ROI in all eight subjects, whereas the ordering within the early visual group varies across subjects. In contrast, encoding accuracy shows little dependence on Llama layer: within each ROI, the layer-to-layer range in mean accuracy is at most 0.015 in every subject (Table S2). Cortical predictability varies systematically across ROIs but remains largely invariant across Llama layers.

4.2 Local variance spectra and tangent-space organization within systems

Spectrum analysis reveals distinct dimensionality profiles across the two systems (Figure 1). In S1, cortical local spectra differ only modestly across the six ROIs: the first local component explains about 17-28% of the variance and 95% is reached only after about 65-80 of the 99 available components. Differences between ROIs are larger in some subjects, but are substantially reduced after subsampling to match voxel counts and remain small compared with the change across Llama layers. In Llama-3-8B, the first local component explains about 97% of the variance at L4 but only about 26% at L28, decreasing monotonically with depth. Because the analysis uses unstandardized hidden states,

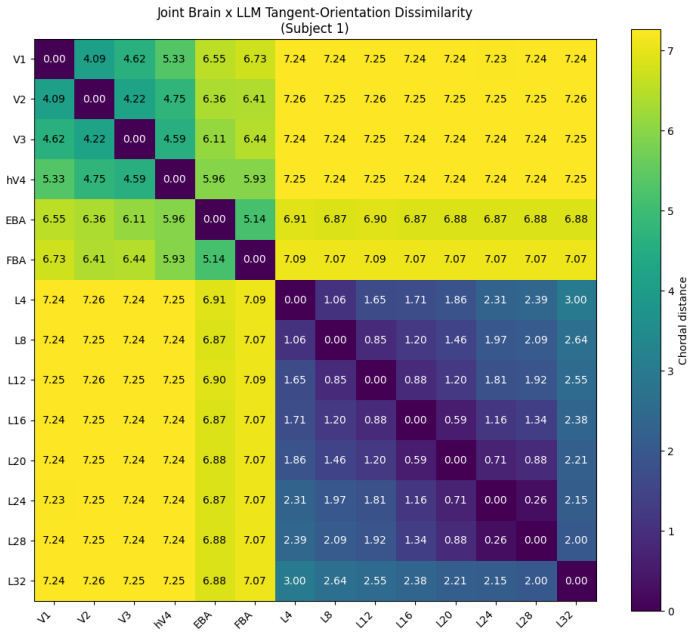

Figure 2: **Joint brain $\times$ LLM tangent-orientation dissimilarity (S1).** Chordal distance (Equation (2), $d = 8$, $k = 100$), averaged across semantic neighborhoods, between tangent spaces for six cortical ROIs and eight Llama-3-8B layers. The top-left block shows within-cortex distances, which increase across the ROI ordering and are largest between early visual and body-selective regions. The bottom-right block shows within-LLM distances, which increase with layer separation but are substantially smaller than within-cortex distances. Off-diagonal blocks show brain-LLM distances, which remain close to the random-subspace expectation of 7.354 but are lower for EBA and FBA than for early visual cortex.

the early-layer dominance may partly reflect a few high-variance dimensions. Likewise, PCA of trial-averaged fMRI responses includes measurement noise; the cortical spectra therefore characterize observed local variation rather than the latent signal spectrum.

We next examined the organization of local tangent spaces, which capture the dominant patterns of representational variation among stimuli within each semantic neighborhood. The joint brain-LLM orientation dissimilarity matrix (Figure 2) contains chordal distances averaged across semantic neighborhoods, with within-cortex distances in the top-left block and within-LLM distances in the bottom-right block. In cortex, distance increases with separation along the ROI ordering (S1: Spearman $\rho = 0.82$; mean $\rho = 0.78$ across eight subjects, range 0.70-0.82). The largest adjacent change occurs between hV4 and EBA in seven of eight subjects, indicating that the most pronounced reorientation generally occurs between retinotopic and body-selective cortex.

Llama shows a similar dependence on representational separation, although the changes in orientation are substantially smaller. Tangent-space distance increases with the number of intervening layers (Spearman $\rho = 0.84$ across the 28 layer pairs; Mantel $p = 5 \times 10^{-5}$). The association strengthens to $\rho = 0.96$ when comparisons involving L32 are excluded; L32 is the only representation returned after the model's final RMS normalization, so this difference may reflect both final-layer processing and the extraction convention. Distances between adjacent sampled Llama layers are approximately 4–$5\times$ smaller than those between adjacent cortical ROIs. Local tangent-space orientation changes systematically with separation in both systems, but the sampled Llama layers show much smaller orientation differences than the cortical ROIs.

4.3 CROSS-SYSTEM TANGENT-SPACE ALIGNMENT IS WEAK AND VARIES LITTLE ACROSS SAMPLED LAYERS

Despite systematic tangent-space organization within each system, cross-system alignment in the off-diagonal blocks of the joint matrix remains close to the random-subspace baseline (Figure 2). Averaged across layers, every ROI group has a chordal distance below the random-subspace expectation of Equation (3) (7.354) in all eight subjects (Appendix S1): V1-hV4 7.251 (8-subject bootstrap 95% CI $[7.236, 7.265]$; $t_7 = -13.0$, $p = 4 \times 10^{-6}$), FBA 7.133 ($[7.098, 7.172]$; $t_7 = -10.8$, $p = 1 \times 10^{-5}$), and EBA 6.994 ($[6.933, 7.062]$; $t_7 = -10.0$, $p = 2 \times 10^{-5}$). Thus, cross-system alignment is weak in magnitude but consistently greater than expected for random subspaces and exhibits a systematic regional gradient: alignment is strongest in EBA, followed by FBA and early visual cortex. In contrast, alignment varies little across Llama layers, providing no evidence of systematic layer-region correspondence.

Together with the encoding results, this frames a central dissociation: the leading directions of local variation in visual cortex and in Llama-3-8B hidden states are only weakly aligned, even though a linear map between them predicts cortical activity at $r = 0.14$-0.40. Linear prediction does not require preservation of the leading local variation subspaces. Two caveats bound this interpretation: a linear map need not preserve principal directions, and fMRI noise lowers the attainable alignment; we did not estimate this ceiling.

4.4 A VISION-ALIGNED TEXT ENCODER SHOWS DEPTH-DEPENDENT BRAIN CORRESPONDENCE

The layer-invariant pattern observed for Llama is not shared by CLIP caption embeddings. Using representations of the same captions from the CLIP ViT-L/14 text encoder (Appendix S5; S1, S2, S5, S7), encoding accuracy increases with depth in every ROI (EBA L1→L12: $0.354 \rightarrow 0.406$ in S1, $0.399 \rightarrow 0.433$ in S5; gains of 0.014-0.052 across ROIs and subjects). Because every ROI, including V1, is best predicted by the deepest layers, this is a global depth effect rather than a layer-region correspondence.

Tangent-space alignment shows a corresponding depth dependence: cross-system distance decreases with CLIP depth. At L12, the EBA gap below the random-subspace expectation is about $1.5\times$ that observed for Llama (6.74 vs. 6.94 averaged over S1, S2, S5, S7; per-subject ratios 1.2-1.7), though CLIP's first layer is less aligned with EBA than Llama is in all four subjects (Appendix S5). Unlike Llama's largely layer-invariant correspondence with cortex, CLIP shows increasing brain correspondence with depth in both linear encoding and tangent-space orientation. Because semantic neighborhoods are defined using CLIP image embeddings, however, part of this pattern may reflect the shared CLIP training objective.

4.5 PER-PC STEERING VECTORS

The dissociation raises a complementary question: in which LLM-space directions does the brain-relevant signal live? The steering matrix $S^{(u,\ell)} = W^{(u,\ell)\top} E_u$ (Equation (5)) gives $n_{\mathrm{pc}} = 8$ directions per (ROI, layer) cell without refitting. Across S1, S5, and S7 (Appendix S4), the cosine between adjacent sampled layers for PC0 rises from 0.32-0.43 (L4-L8) to 0.83-0.87 (L24-L28) and drops to 0.63-0.70 at L28-L32, though comparisons with L32 may partly reflect its final RMSNorm, while the L4-L28 cosine is only 0.12-0.19: the direction read out for a brain component rotates with depth. The ROI-ROI cosine shows the same two blocks at every layer (at L24: 0.80-0.96 within V1-hV4, 0.72-0.82 between EBA and FBA, 0.00-0.27 across blocks).

Residual-stream steering: adding $\varepsilon\,\mathbf{s}$ to a hidden state shifts the frozen encoding model's prediction along the corresponding brain PC by construction; whether it also shifts Llama's own next-token distribution is a separate question. For each subject, we took PC0 of FBA at L24, assigned it the four COCO super-categories with the largest $|d'|$ between images with and without the category (Appendix S4), added $\varepsilon\,\mathbf{s}$ ($\varepsilon \in \{0.5, 1, 2, 5\}$, raw hidden-state coordinates) to the L24 hidden state (the output of decoder block 23) while scoring held-out captions, and compared per-category changes in token log-probability with norm-matched random directions (Figure 3; Table S3). The results are mixed. Relative to random directions, sports tokens gain probability in S1, S5, and S7 ($z = 7.9$-8.2, 5.9-6.3 and 2.7-3.1; $\Delta \log p > 0$ in all three) but not in S2 ($z = -1.9$), and person tokens do so only

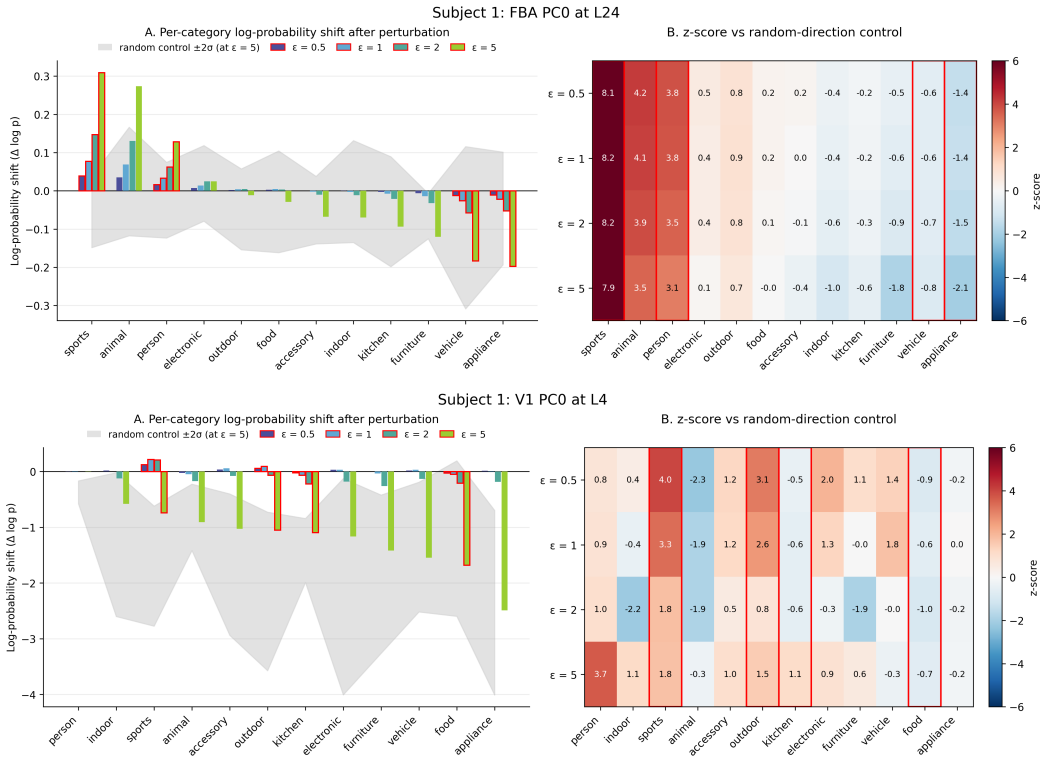

Figure 3: **Steering test in Subject 1, the subject with the largest effect.** Top: FBA PC0 at L24. Bottom: V1 PC0 at L4. Left panels: per-category $\Delta \log p$ for $\varepsilon \in \{0.5, 1, 2, 5\}$ (red outline: predicted categories; gray band: $\pm 2\sigma$ of random directions at $\varepsilon = 5$). Right panels: z-scores against random directions. Subjects S2, S5, and S7 are shown in Appendix Figure S12.

in S1 and S5. In every subject, at least one non-predicted category shifts as strongly as a predicted one (e.g., furniture, $z \approx 6$ in S5). The same protocol applied to V1 PC0 at L4 produces maximal $|z|$ of 4.0, 4.5, 6.6 and 11.0 in S1, S2, S5, and S7; at L4, random directions of the same norm also cause large, non-specific losses (down to $\Delta \log p \approx -4$ at $\varepsilon = 5$), so this cell is not a clean negative control. We therefore interpret the residual-stream intervention as exploratory rather than as causal validation of the brain-derived direction.

5 Discussion

Llama caption embeddings predict cortical responses to the corresponding images despite weak brain-model alignment in the dominant patterns of variation within semantic neighborhoods. Encoding accuracy is higher in body-selective than early visual regions but varies little across Llama layers. Similarly, cross-system tangent-space orientation alignment remains close to the random-subspace baseline and is largely layer-invariant, yet exhibits consistently stronger alignment in EBA and FBA than early visual cortex. Both measures reveal ROI-specific differences in brain-LLM correspondence, but show that successful linear prediction can occur without strong alignment of local representational variation.

The weak cross-system alignment occurs despite systematic organization of local representational variation within both cortex and Llama. Llama tangent spaces are substantially more similar across layers than cortical tangent spaces are across ROIs, suggesting greater preservation of local variation across model depth than across cortical regions, consistent with the sequential transformation of representations across layers. Within cortex, tangent spaces are more similar among the early visual ROIs (V1-hV4) and between the two body-selective ROIs (EBA and FBA), but are less similar when retinotopic and body-selective regions are compared. The largest change between adjacent ROIs

generally occurs from hV4 to EBA, consistent with a transition from retinotopic to category-selective organization of visual representations (Grill-Spector et al., 2001; Larsson & Heeger, 2006). Within Llama, tangent space similarity decreases systematically with layer separation, indicating gradual reorientation across depth. The variance spectra further distinguish the systems, with comparatively modest variation in cortical effective dimensionality across ROIs but a strong increase in Llama effective dimensionality with depth. These results show that cortex and Llama each exhibit systematic local representational structure, yet this within-system organization does not translate into strong alignment of local variation between the two systems.

The CLIP ViT-L/14 text encoder comparison shows that vision-aligned text representations exhibit a different pattern of correspondence with the brain. For the same caption inputs, representations extracted from progressively deeper layers of CLIP's text encoder show increasing cortical encoding accuracy and stronger cross-system tangent-space alignment, while both metrics are largely layer-invariant in Llama-3-8B. These increases with depth occur across all ROIs, indicating a common depth dependence rather than region-specific layer correspondence. Although the shared caption input controls for input content, differences in training objective, architecture, and training data leave the source of this contrasting depth dependence unresolved. This contrast demonstrates that models receiving the same linguistic input can differ in how their internal transformations correspond to cortical representations, which echoes Conwell et al. (2024), who found similar encoding scores across very different vision models: similar encoding accuracy can hide different local geometry, and the orientation diagnostic offers a complementary criterion.

The steering construction derives its targets from cortical responses to natural images rather than from prompt contrasts, and so complements prompt-based activation engineering (Turner et al., 2023; Rimsky et al., 2023). In our data, however, the causal effects of these directions are not consistent across subjects. Stronger tests would use matched layers and perturbation sizes, many more random and permuted directions, and a baseline direction learned directly from category labels. Such a baseline matters because ridge weights are not activation patterns (Haufe et al., 2014).

Limitations. We examine one LLM and one CLIP model, with LLM representations derived from captions rather than visual input. We do not estimate a noise ceiling for tangent-space alignment, so fMRI noise may contribute to the weak cross-system alignment. Semantic neighborhoods are defined using CLIP embeddings, which may preferentially capture structure represented by CLIP and should therefore be considered when interpreting the CLIP comparison. Finally, the CLIP and steering analyses are limited to four subjects, with steering limited to one ROI-layer pairing per subject.

AI USE STATEMENT

We used generative AI tools as a coding and writing assistant: to help implement the analysis and figure code, and to draft and edit prose. All experiments were run and their outputs inspected by the authors, all numerical claims were checked against the saved outputs, and the authors take full responsibility for the content, claims, and artifacts of this work.

ETHICS STATEMENT

This work analyzes the publicly available Natural Scenes Dataset (Allen et al., 2022), collected with informed consent under the approval of the original study's institutional review board and used under the NSD terms of use. No new human data were collected. The models used are publicly released research models.

REPRODUCIBILITY STATEMENT

Section 3 specifies the neighborhood construction, local PCA, chordal distance and its null distribution, the ridge encoding protocol, and the steering-vector construction; Section 4 and Appendix S2 specify the data, ROIs, layers and preprocessing, and Appendix S4 the residual-stream steering protocol.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Alexander Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. In *Conference on Language Modeling*, 2024.

Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023.

SueYeon Chung and LF Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, 2021.

SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746, 2020.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1):9383, 2024. doi: 10.1038/s41467-024-53147-y.

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

James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

Adrien Doerig, Tim C Kietzmann, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7:123–135, 2025.

Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022.

Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11):1409–1422, 2001.

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.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Jonas Larsson and David J Heeger. Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, 26(51):13128–13142, 2006.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Jason Manley, Sihao Lu, Kevin Barber, Jeffrey Demas, Hyewon Kim, Daniel Meshulam, and Alipasha Vaziri. Simultaneous, cortex-wide dynamics of up to 1 million neurons reveal unbounded scaling of dimensionality with neuron number. *Neuron*, 112(10):1694–1709, 2024.

Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of Neurophysiology*, 93(1):603–608, 2005.

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*, 2021.

Joséphine Raugel, Stéphane d'Ascoli, Jérémy Rapin, Valentin Wyart, and Jean-Rémi King. Scaling and context steer LLMs along the same computational path as the human brain. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schubert, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. arXiv preprint arXiv:2312.06681, 2023.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105641118, 2021.

Rebecca F Schwarzlose, Chris I Baker, and Nancy Kanwisher. Separate face and body selectivity on the fusiform gyrus. *Journal of Neuroscience*, 25(47):11055–11059, 2005.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, 2019.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. arXiv preprint arXiv:2308.10248, 2023.

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

SUPPLEMENTARY MATERIAL

Appendix map.

- **S1.** Steering-vector and encoding panels for S1 and per-subject steering replication. *Supports §4.1, §4.5.*
- **S2.** Local spectrum and orientation: controls, per-subject replication, and group statistics for cross-system distances. *Supports §4.2, §4.3.*
- **S3.** Per-subject ridge encoding accuracy. *Supports §4.1.*
- **S4.** Residual-stream steering: protocol and per-subject results. *Supports §4.5.*
- **S5.** CLIP ViT-L/14 text-encoder comparison (S1, S2, S5, S7). *Supports §4.1, §4.4, §5.*

S1 STEERING-VECTOR AND ENCODING PANELS

Figures S1-S4 show the S1 steering vectors and encoding accuracy for all ROIs and layers; Figure S5 replicates the steering panels in S5 and S7. In the Llama analysis the $6 \times 8 \times 8 = 384$ steering vectors live in $\mathbb{R}^{d_L}$ with $d_L = 4096$.

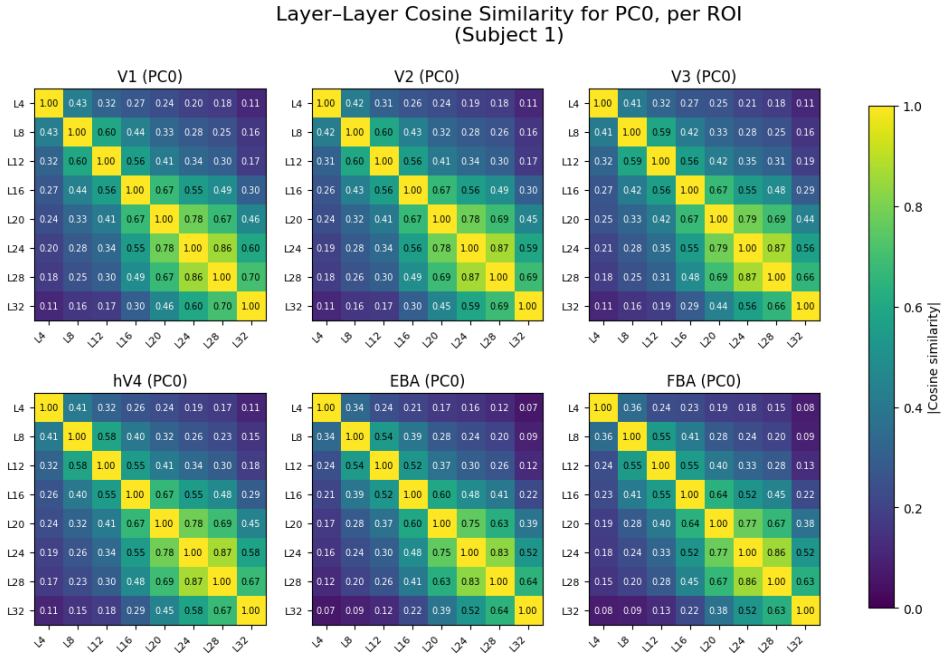

Figure S1: **Layer-layer steering-vector cosine for PC0, per ROI (S1).** Each panel shows $|\cos(\mathbf{s}_0^{(u,L_a)}, \mathbf{s}_0^{(u,L_b)})|$ across the eight Llama-3-8B layers. Cosines between adjacent sampled layers rise from 0.34-0.43 (L4-L8) to 0.83-0.87 (L24-L28) and drop to 0.63-0.70 (L28-L32); L4-L32 cosines are 0.07-0.11.

S2 LOCAL SPECTRUM AND ORIENTATION: CONTROLS AND PER-SUBJECT REPLICATION

Local variance spectra and effective dimensionality. We characterize local effective dimensionality using cumulative PCA variance spectra, computed within each semantic neighborhood in the native representation space (voxel space for cortex; activation space for Llama) and averaged across neighborhoods. Effective dimensionality is defined as the number of principal components required to explain 95% of the variance. Figure S6 shows cortical spectra across all eight subjects, which are broadly similar across ROIs, with larger differences in some subjects. In the two subjects shown in

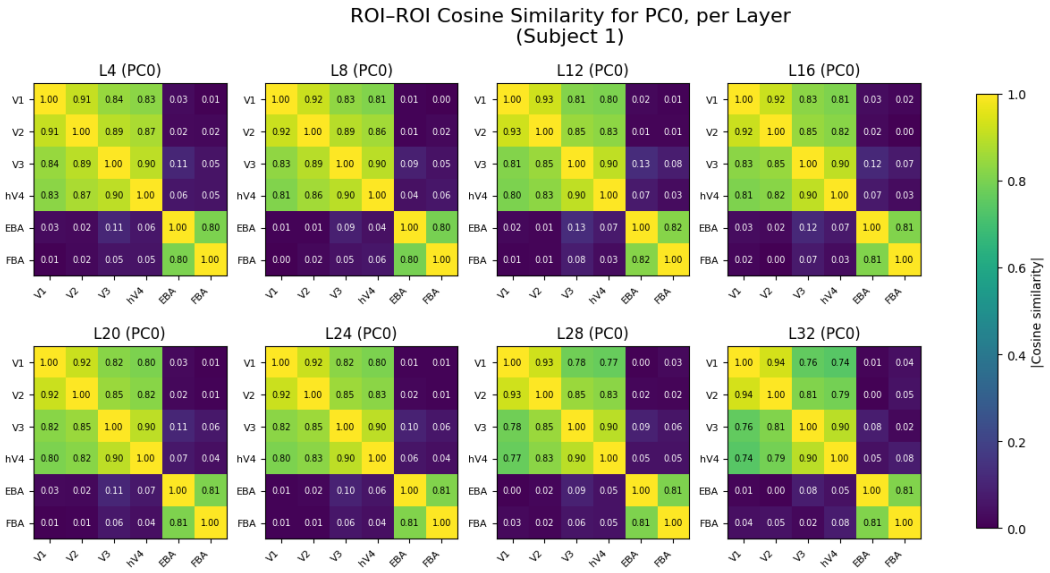

Figure S2: **ROI-ROI steering-vector cosine for PC0, per Llama-3-8B layer (S1).** A two-block structure, {V1, V2, V3, hV4} vs. {EBA, FBA}, is present at every layer from L4 to L32 (within V1-hV4: 0.74-0.94; EBA-FBA: 0.80-0.82; across blocks: 0.00-0.13).

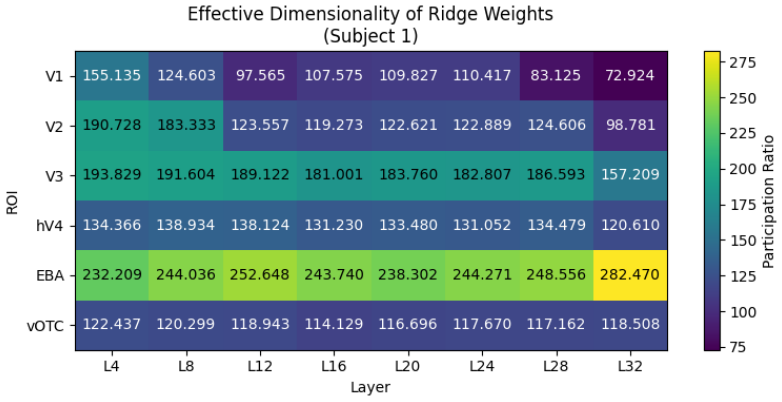

Figure S3: **Effective Dimensionality measured by participation ratio of the ridge weight matrices $W^{(u,\ell)}$ (S1)**, computed from the singular values γ_j of $W^{(u,\ell)}$ (standardized feature coordinates) as $(\sum_j \gamma_j)^2 / \sum_j \gamma_j^2$. EBA has the highest value at every layer (232-282); V1 falls, non-monotonically, from 155 at L4 to 73 at L32.

Figure S8, these differences are substantially reduced after matching ROI voxel counts, indicating a contribution of ROI size. In contrast, Llama spectra change strongly with depth (Figure S7), with deeper layers requiring progressively more components to explain 95% of the variance.

Brain orientation gradient per subject. Per-subject Spearman ρ between chordal distance and ROI separation is 0.82 (S1), 0.79 (S2), 0.79 (S3), 0.82 (S4), 0.82 (S5), 0.70 (S6), 0.74 (S7) and 0.77 (S8). Two-cluster separability (every within-cluster distance smaller than every cross-cluster distance) holds in S1, S2, S4, S5, S6 and S8 and is partial in S3 and S7, where the EBA-FBA distance exceeds the smallest cross-cluster distance (Figure S9).

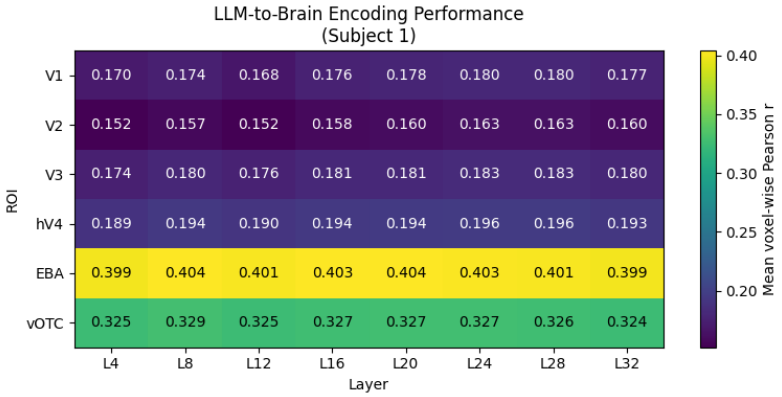

Figure S4: **Held-out ridge encoding accuracy for all ROIs and Llama-3-8B layers (S1).** Mean voxel-wise Pearson r; the layer-to-layer range within each ROI is at most 0.012. Heatmap form of the S1 column of Table S2. (The row labelled vOTC is FBA.)

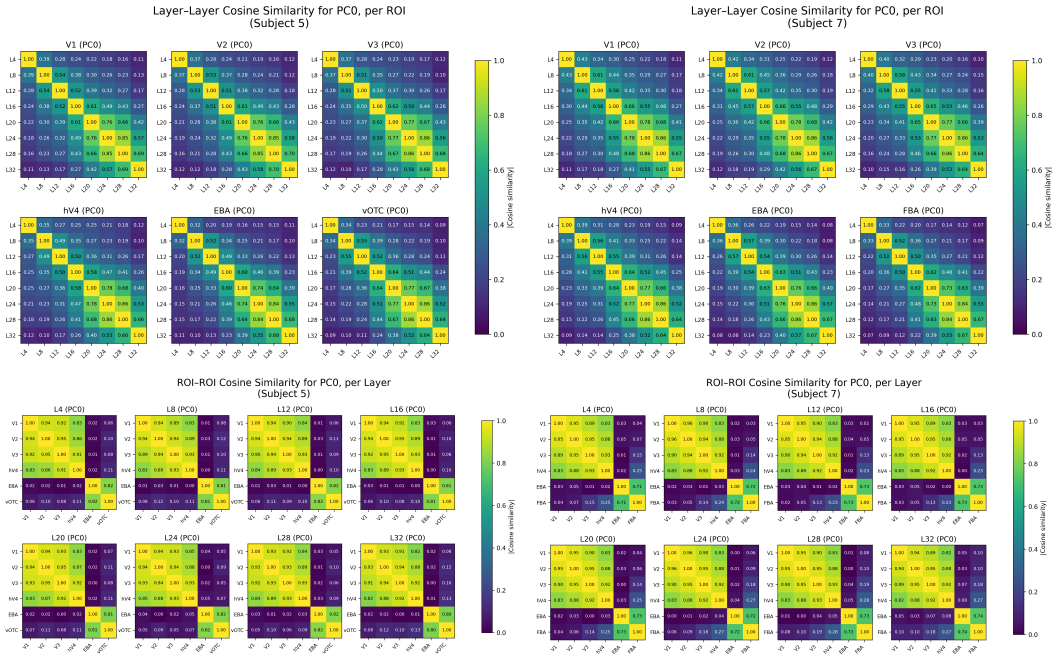

Figure S5: **Steering-vector panels in S5 (left) and S7 (right).** Top: layer-layer cosine for PC0 per ROI (adjacent layers 0.32-0.86; L4-L32 0.07-0.12). Bottom: ROI-ROI cosine for PC0 per layer (two blocks at every layer; at L24 EBA-FBA 0.82 in S5 and 0.72 in S7).

Cross-system distance per subject. All 24 values lie below the isotropic random-subspace expectation 7.354 (Equation (3)), and the ordering EBA $<$ FBA $<$ V1-hV4 holds in every subject (Table S1).

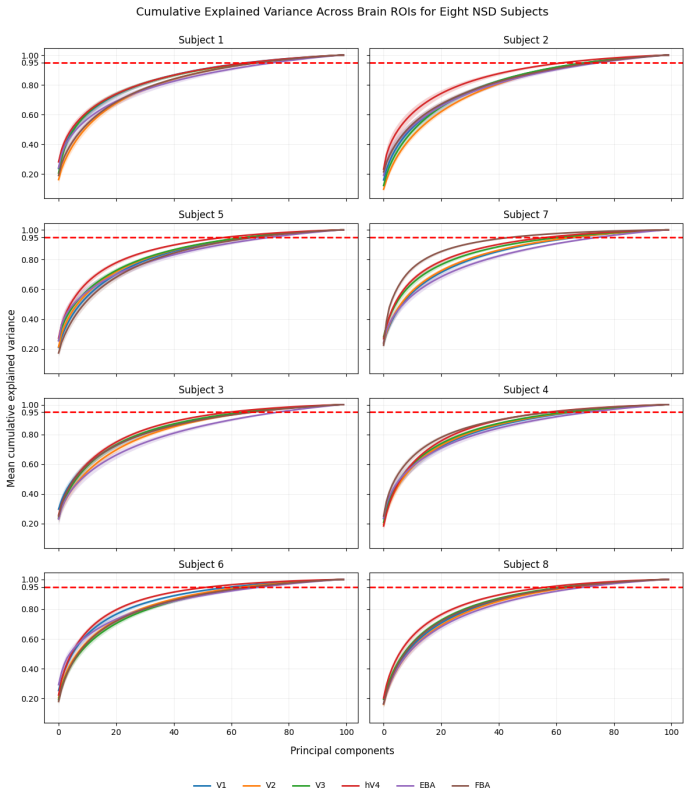

Figure S6: **Mean local cumulative variance spectra across cortical ROIs.** Curves show cumulative explained variance, averaged across semantic neighborhoods, for each ROI in all eight subjects. The dashed line marks the 95% threshold used to define local effective dimensionality.

Table S1: Mean cross-system chordal distance over the eight Llama layers for each subject and ROI group. V1–hV4 is additionally averaged across the four early visual ROIs. Smaller values indicate greater alignment.

ROI group	S1	S2	S3	S4	S5	S6	S7	S8
V1-hV4	7.245	7.243	7.263	7.251	7.208	7.251	7.257	7.288
FBA	7.075	7.085	7.121	7.159	7.067	7.135	7.210	7.214
EBA	6.883	6.993	7.002	7.020	6.855	6.986	7.023	7.191

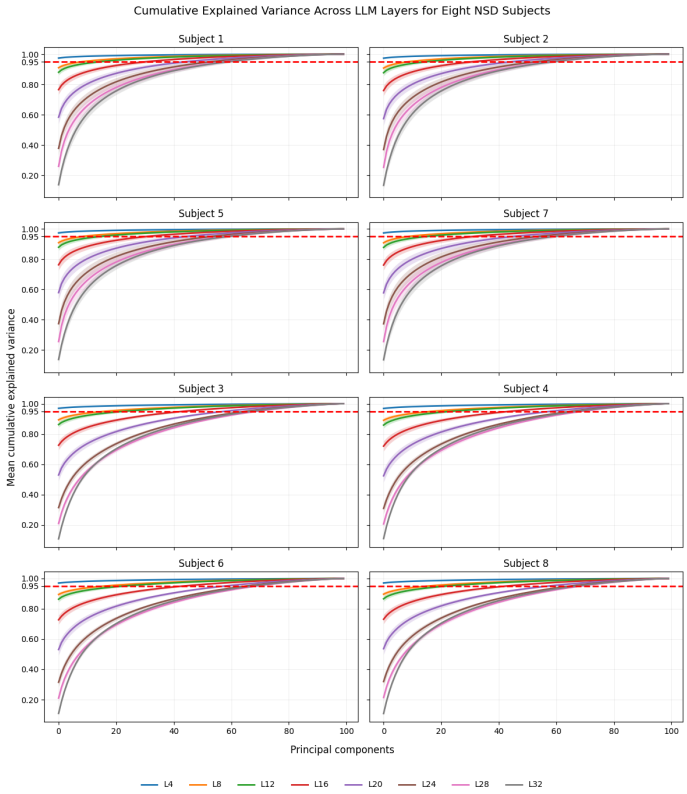

Figure S7: **Mean local cumulative variance spectra across Llama layers.** Curves show cumulative explained variance, averaged across semantic neighborhoods, for each Llama-3-8B layer. The dashed line marks the 95% threshold used to define local effective dimensionality.

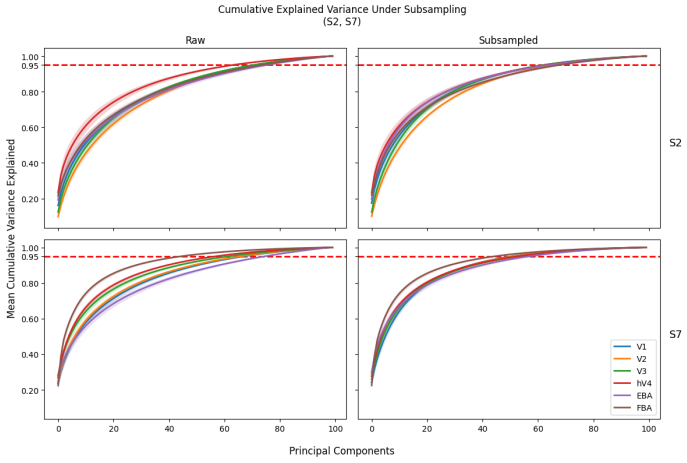

Figure S8: **Effect of ROI voxel count on cortical local variance spectra.** Mean cumulative variance spectra are shown before and after subsampling each ROI to match the voxel count of the smallest ROI for two subjects. ROI differences are reduced after voxel-count matching.

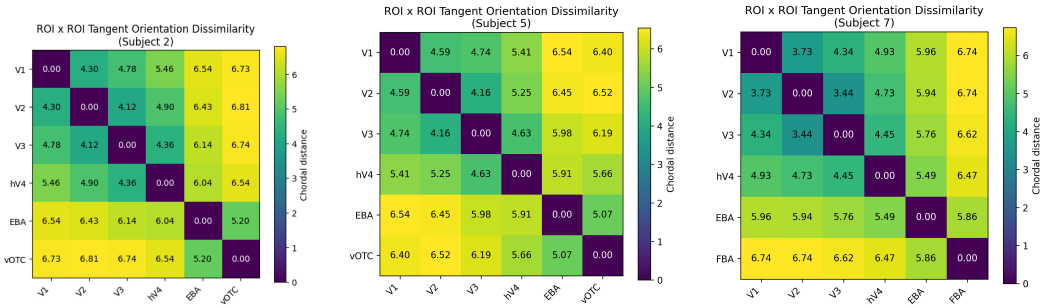

Figure S9: **Brain orientation dissimilarity matrices (raw), S2, S5, and S7.** Spearman ρ with ROI separation: 0.79, 0.82 and 0.74.

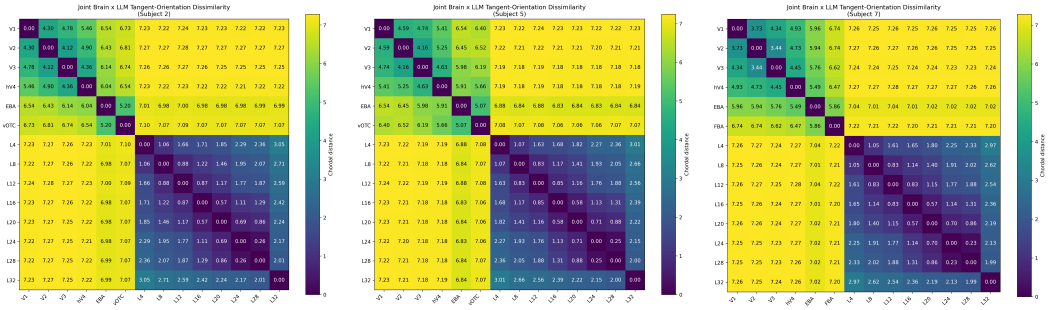

Figure S10: **Joint brain $\times$ Llama-3-8B tangent-orientation dissimilarity (raw), S2, S5, and S7.** Same convention as Figure 2 (S1).

S3 PER-SUBJECT RIDGE ENCODING ACCURACY

Table S2: Held-out ridge encoding accuracy for all eight NSD subjects: mean voxel-wise Pearson r averaged over the eight Llama-3-8B layers $\pm$ half of the layer-to-layer range.

ROI	S1	S2	S3	S4	S5	S6	S7	S8
V1	$0.175 \pm .006$	$0.202 \pm .005$	$0.194 \pm .007$	$0.197 \pm .004$	$0.256 \pm .007$	$0.189 \pm .006$	$0.170 \pm .004$	$0.142 \pm .005$
V2	$0.158 \pm .006$	$0.157 \pm .003$	$0.184 \pm .004$	$0.188 \pm .003$	$0.242 \pm .005$	$0.183 \pm .005$	$0.175 \pm .003$	$0.139 \pm .004$
V3	$0.180 \pm .005$	$0.173 \pm .003$	$0.181 \pm .004$	$0.158 \pm .003$	$0.227 \pm .004$	$0.165 \pm .004$	$0.164 \pm .002$	$0.143 \pm .002$
hV4	$0.193 \pm .004$	$0.233 \pm .003$	$0.192 \pm .002$	$0.175 \pm .002$	$0.254 \pm .004$	$0.167 \pm .005$	$0.165 \pm .002$	$0.162 \pm .002$
EBA	$0.402 \pm .003$	$0.363 \pm .003$	$0.311 \pm .003$	$0.325 \pm .003$	$0.404 \pm .005$	$0.337 \pm .003$	$0.323 \pm .004$	$0.212 \pm .002$
FBA	$0.326 \pm .003$	$0.344 \pm .003$	$0.270 \pm .003$	$0.266 \pm .003$	$0.364 \pm .004$	$0.286 \pm .004$	$0.306 \pm .004$	$0.218 \pm .003$

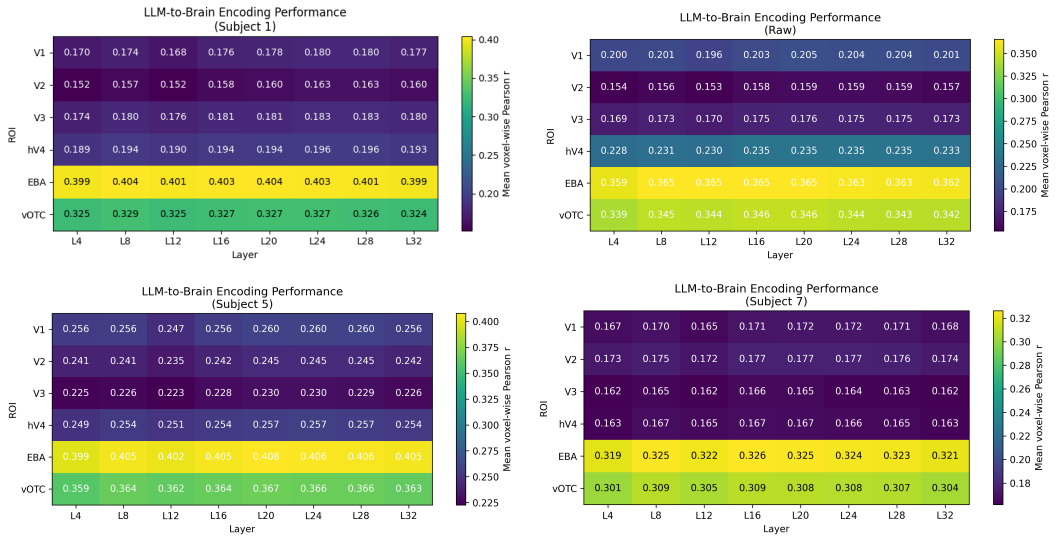

Figure S11: **Ridge encoding accuracy by ROI and Llama-3-8B layer for S1, S2, S5, and S7** (row vOTC = FBA). The body-selective advantage and the layer-flatness replicate in all four subjects.

S4 RESIDUAL-STREAM STEERING: PROTOCOL AND PER-SUBJECT RESULTS

Semantic assignment. For ROI u and PC i, we compute the score $y_i(I) = (\mathbf{r}_u(I) - \bar{\mathbf{r}}_u)^\top \mathbf{e}_i$ over the training images and, for each of the 12 COCO super-categories, the signed $d'_{i,c} = (\mu_{\text{in}} - \mu_{\text{out}})/\sigma_{\text{pooled}}$ between images that contain category c and images that do not. The predicted set is the four categories with the largest $|d'_{i,c}|$, and the sign of $d'_{i,c}$ gives the predicted direction of the shift under $+\varepsilon\,\mathbf{s}_i$; flipping the sign of $\mathbf{e}_i$ flips both $\mathbf{s}_i$ and $d'_{i,c}$, so the prediction does not depend on PC polarity.

Perturbation and controls. We add $\varepsilon\,\mathbf{s}$ to the output of Hugging Face decoder block 23 (the hidden state we call L24) at every token position, run the frozen model on held-out captions, and aggregate the per-token change in log-probability, $\Delta \log p$, by COCO super-category. Controls are $\varepsilon = 0$, which reproduces the baseline exactly, and norm-matched random Gaussian directions, which define the z-scores. Because $\Delta \log p$ for random directions is itself biased (mostly negative), a positive z means that $\mathbf{s}$ lowers a category's probability less than random directions do, not necessarily that it raises it; the sign of $\Delta \log p$ is shown in the left panels of Figures 3 and S12.

S5 CLIP VIT-L/14 TEXT-ENCODER COMPARISON (S1, S2, S5, S7)

To test whether layer-flat encoding and weak cross-system alignment are properties of caption embeddings in general, we ran the ridge encoding, tangent-orientation and steering analyses on the CLIP ViT-L/14 text encoder (Radford et al., 2021) (12 layers, width 768), applied to the same

Table S3: Residual-stream steering per subject (z against norm-matched random directions, range over $\varepsilon \in \{0.5, 1, 2, 5\}$). Predicted categories are the four with the largest $|d'|$ for FBA PC0.

Subj.	FBA PC0 at L24: predicted categories (z)	Largest non-predicted shift	V1 PC0 at L4: max $\lvert z \rvert$
S1	sports 7.9-8.2, person 3.1-3.8, vehicle -0.8–-0.6, appliance -2.1–-1.4	animal 3.5-4.2	4.0
S2	vehicle 1.9-2.2, outdoor 0.9-1.0, appliance -0.5-0.7, sports -1.9	food -3.2, animal -3.3–-2.6	4.5
S5	sports 5.9-6.3, person 4.9-5.9, appliance -0.4-0.1, vehicle -0.9–-0.6	furniture 6.1-6.5	6.6
S7	sports 2.7-3.1, outdoor 0.7-0.8, vehicle 0.4-0.6, appliance -4.8–-3.7	electronic -4.2–-3.0	11.0

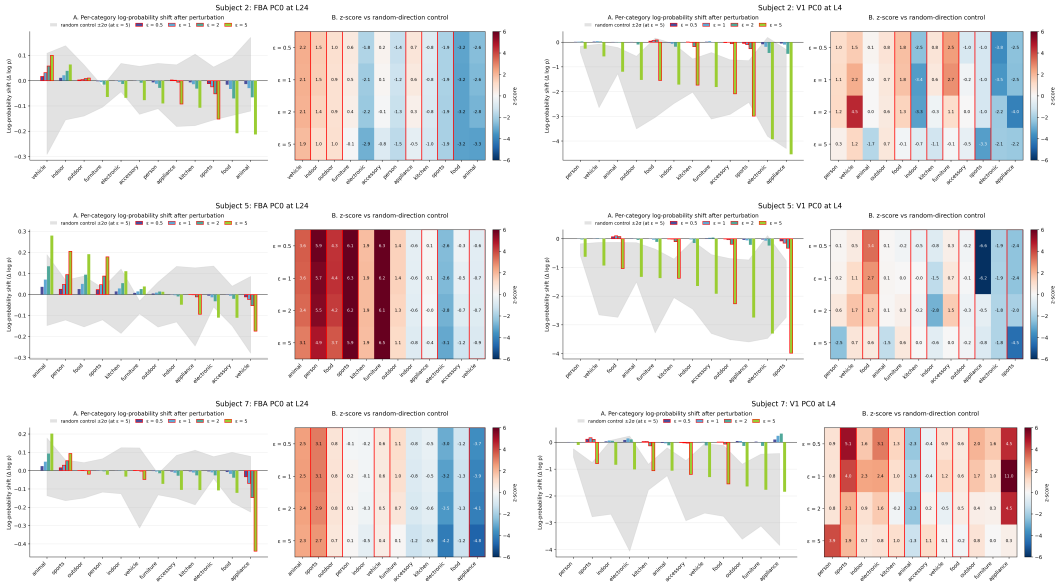

Figure S12: **Steering test in S2 (top), S5 (middle) and S7 (bottom).** Left: FBA PC0 at L24. Right: V1 PC0 at L4. Same conventions as Figure 3.

captions and pooled in the same way. We used layers 1, 3, 4, 6, 7, 9, 10 and 12, approximately evenly spaced. The neighborhoods are the same CLIP ViT-B/32 image-embedding neighborhoods as in the main analysis.

Encoding. Accuracy increases with depth in every ROI and subject (gains 0.014-0.052 from L1 to L12; Table S4, Figure S13). The increase is not strictly monotone: small dips occur (e.g., between L3 and L4), and in S7 every ROI peaks at L9-L10. Because every ROI, including V1, is best predicted by the deepest layers, this is a depth effect rather than a layer-to-region correspondence. At L12 the text encoder predicts every ROI at least as well as the layer-averaged Llama features.

Table S4: CLIP ViT-L/14 text encoder ridge encoding accuracy (mean voxel-wise r) per ROI and subject at L1 and L12.

	S1		S2		S5		S7	
ROI	L1	L12	L1	L12	L1	L12	L1	L12
V1	0.140	0.179	0.198	0.223	0.257	0.285	0.167	0.185
V2	0.121	0.160	0.149	0.170	0.242	0.268	0.176	0.190
V3	0.142	0.180	0.164	0.189	0.226	0.251	0.165	0.179
hV4	0.156	0.195	0.216	0.253	0.249	0.281	0.163	0.178
EBA	0.354	0.406	0.358	0.389	0.399	0.433	0.319	0.345
FBA	0.281	0.331	0.338	0.369	0.355	0.389	0.304	0.331

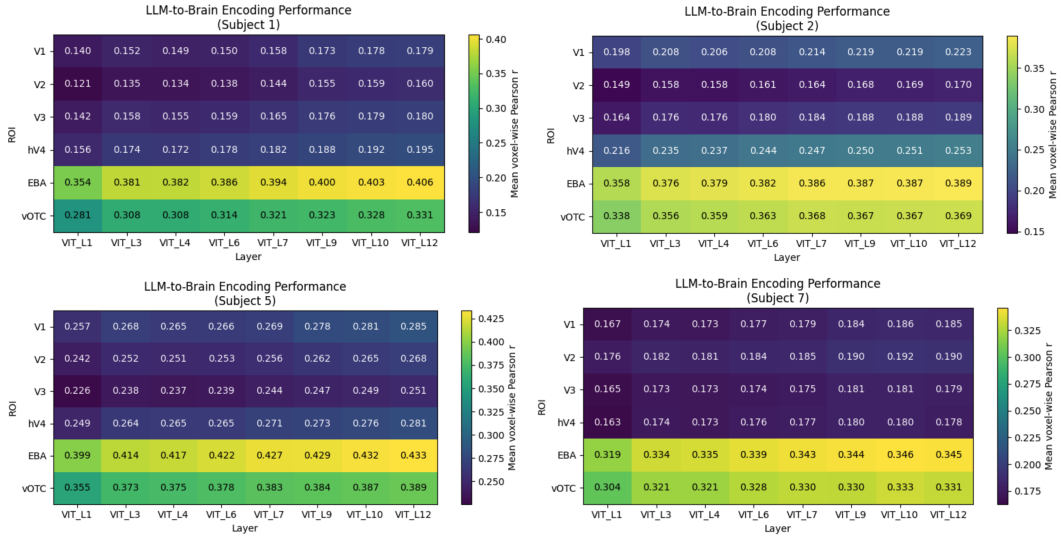

Figure S13: **CLIP ViT-L/14 text encoder ridge encoding accuracy across layers, S1, S2, S5, S7** (panel titles read "LLM"; row vOTC = FBA).

Cross-system alignment. Cross-system distances decrease with depth (Table S5, Figure S14). At L1 the text encoder is less aligned with EBA than Llama is in all four subjects; at L12 the EBA gap below chance is 1.18, 1.56, 1.65 and 1.52 times that of Llama in S1, S2, S5, and S7 (EBA 6.80, 6.79, 6.53, 6.85 vs. 6.88, 6.99, 6.86, 7.02; 1.47× on average). The minimum EBA distance is 6.51 (S5, L10). Because the neighborhoods come from CLIP image embeddings, this advantage may partly reflect the shared CLIP training objective.

Table S5: Cross-system chordal distance for the CLIP ViT-L/14 text encoder at L1 and L12, with the layer-averaged Llama-3-8B value for reference. V1-hV4 is the mean over the four early visual ROIs. Random-subspace expectation: 7.354.

ROI group		S1	S2	S5	S7
V1-hV4	CLIP L1	7.28	7.22	7.18	7.24
	CLIP L12	7.20	7.17	7.09	7.19
	Llama	7.25	7.24	7.21	7.26
FBA	CLIP L1	7.22	7.09	7.09	7.20
	CLIP L12	7.04	6.93	6.92	7.12
	Llama	7.08	7.09	7.07	7.21
EBA	CLIP L1	7.11	7.01	6.87	7.04
	CLIP L12	6.80	6.79	6.53	6.85
	Llama	6.88	6.99	6.86	7.02

Steering vectors. The ROI-ROI cosine of PC0 steering vectors shows the same two blocks as for Llama at every layer and in every subject (within V1-hV4 0.64-0.97; EBA-FBA 0.56-0.89; across blocks 0.00-0.32; Figure S15). The layer-layer cosine shows the same rotation with depth (adjacent layers 0.37-0.88, first versus last layer 0.13-0.31; Figure S16). Adjacent sampled layer pairs are one or two layers apart, so these values are not directly comparable with Llama's four-layer spacing.

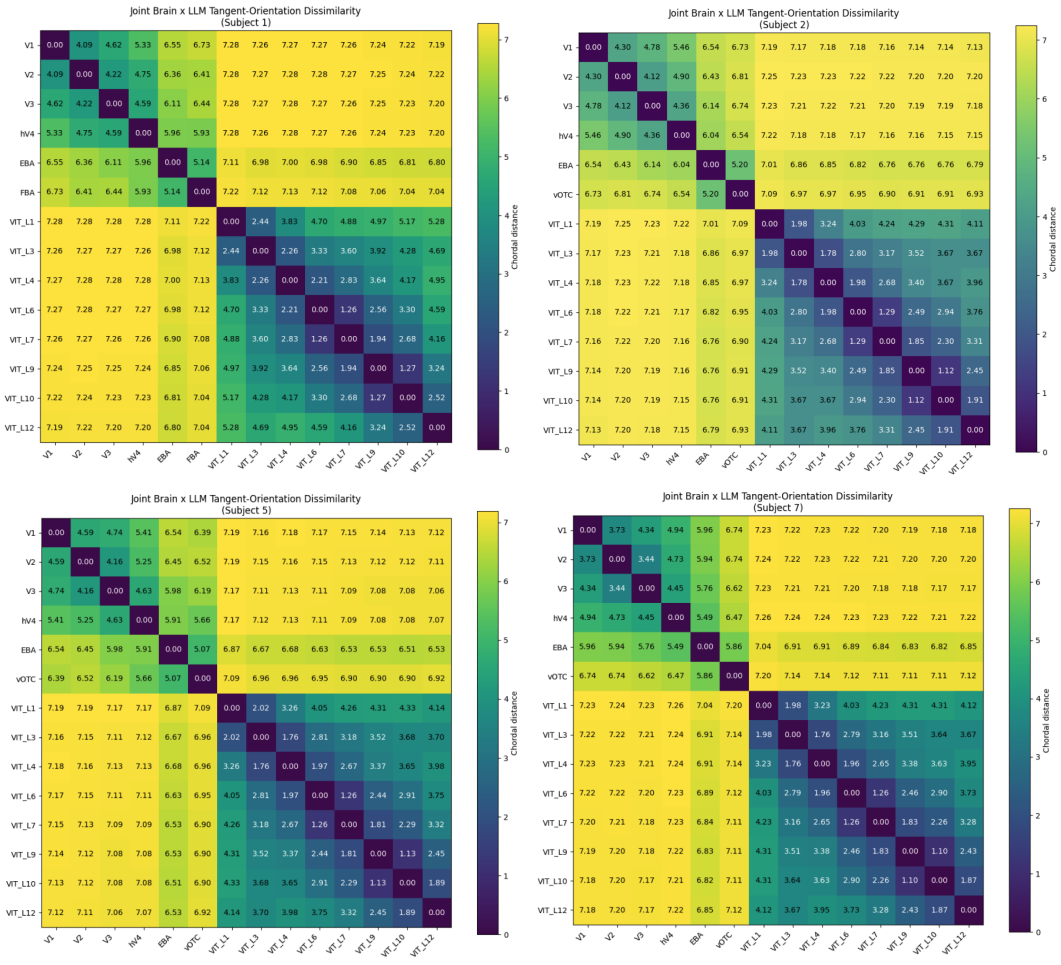

Figure S14: **Joint brain $\times$ CLIP ViT-L/14 text encoder tangent-orientation dissimilarity, S1, S2, S5, S7** (panel titles read "LLM"; vOTC = FBA).

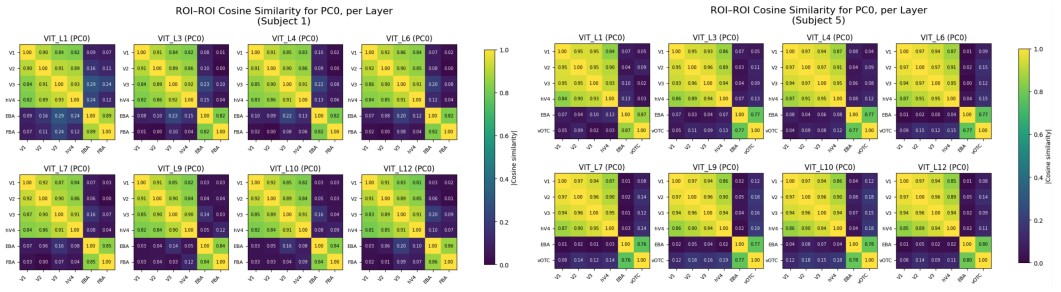

Figure S15: **ROI-ROI steering-vector cosine for PC0 at each CLIP text-encoder layer, S1 (left) and S5 (right).**

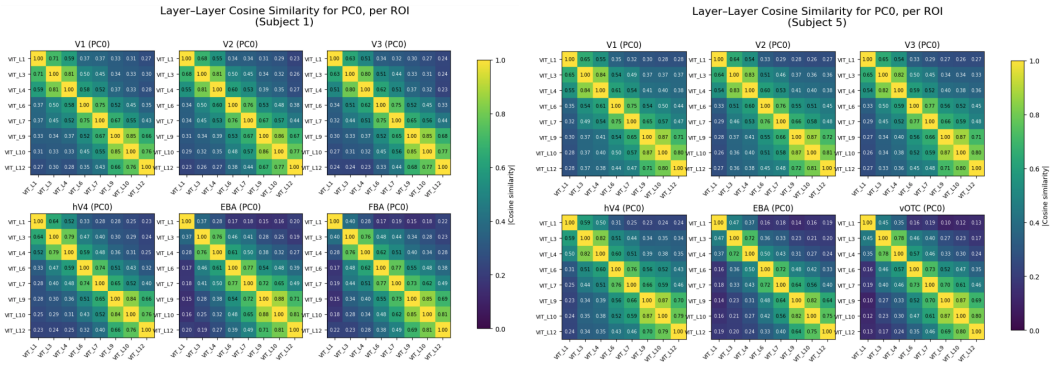

Figure S16: **Layer-layer steering-vector cosine for PC0 per ROI, CLIP text encoder, S1 (left) and S5 (right).**