VISUAL REPRESENTATIONS IN THE HUMAN MEDIAL TEMPORAL LOBE DURING NATURALISTIC MOVIE VIEWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neurons in the human medial temporal lobe (MTL) respond to specific concepts, such as a particular person, across very different images. So far, this selectivity has mostly been measured with isolated images and predefined labels. Representations from vision models provide continuous descriptions of complete visual scenes, enabling MTL activity to be studied without such labels. Using the SUMMER dataset, comprising 2,286 units recorded from 29 intracranially implanted participants while they viewed a full-length commercial movie, we decode framewise representations of a wide range of vision models from MTL population activity. We show that MTL activity identifies held-out frames above a temporally matched baseline and locates them coarsely in the film, with performance still increasing at the full population size, whereas low-level image statistics cannot be reliably decoded. Importantly, size and diversity of training data affect decodability. For example, for the same ResNet-50 architecture, decoding performance markedly improves when trained with language supervision on larger datasets. Across all visual feature spaces, a parahippocampal cortex population outperforms size- and firing-rate-matched control populations from the remaining MTL, as well as control populations with twice as many neurons. Our results show that human MTL populations, in particular in parahippocampal cortex, carry information about the visual content of movies that can be used to compare representations from vision models.

1 INTRODUCTION

Neural activity in the medial temporal lobe (MTL) is thought to integrate distributed perceptual inputs into abstract, conceptual representations including objects, landmarks or persons such as the well-known "Jennifer Aniston neuron" (Squire et al., 2004; Quiroga et al., 2005). These representations have been predominantly studied using isolated, static images and it is unclear what visual information the MTL carries during continuous, naturalistic perception. Recent work decoded predefined labels, such as characters, visual transitions and locations from MTL activity while participants watched a naturalistic movie (Zhang et al., 2023; Gerken et al., 2025). Such labels, however, capture only selected aspects of the visual scene and do not capture the full range of visual information carried by MTL population activity.

Over the past decade, vision models have moved from task-specific supervised networks to large foundation models, including self-supervised (DINOv2 (Oquab et al., 2024)), multimodal (CLIP (Radford et al., 2021)) and agglomerative models (RADIO (Ranzinger et al., 2026)). Their feature embeddings describe complete scenes without predefined labels, so decoding them from neural activity measures how much of a scene's content a population carries. Because different models describe the same scene differently, decoding also reveals which representations are most closely aligned with the neural data.

Here, we decode the representations of a wide range of vision models from the SUMMER dataset (Darcher et al., 2026): A large collection of 2,286 human single units recorded while participants watched a continuous, full-length movie, containing a broad range of people, objects, scenes, and events. We ask three questions: what visual information can be decoded from MTL populations

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.

during naturalistic viewing; how this information is distributed across MTL subregions; and which model properties (architecture, training data, objective) affect decodability of representations.

Naturalistic movie stimuli pose two challenges for comparing vision models with neural activity. First, slowly evolving and recurring scenes create strong temporal dependencies in both neural responses and model features. To protect against this, we train and test on temporally separated segments and compare against temporally matched null distributions. Second, feature spaces of different models differ in geometry, which can make frame identification easier in some spaces for the same prediction quality. To mitigate this, we whiten feature spaces and reduce them to a common dimensionality.

We find that MTL population activity supports decoding of visual representations above a temporally matched baseline during continuous movie presentation, and that performance increases with population size. Low-level image statistics, in contrast, cannot be reliably decoded from MTL. Visual representations are most decodable from the parahippocampal cortex: 293 parahippocampal units outperform matched populations, and even populations twice as large, drawn from the remaining MTL subregions. Finally, decodability depends on training-data and objective. While transformer-based models in our set outperform CNN-based ones, this largely co-varies with differences in training data. We investigate the dependence on training-size using various ResNet-50 models, and find that, for the same architecture, decoding performance markedly improves when trained with language supervision on larger datasets. Beyond this, decodability does not increase consistently with model generation or complexity, and the best models cannot be distinguished reliably.

2 Related Work

Semantic representation in the human medial temporal lobe Using static image paradigms, neurons in the human medial temporal lobe (MTL) encode semantic content, such as people, objects or landmarks that are invariant to low-level visual features (Quiroga et al., 2005; Karkowski et al., 2025; Reber et al., 2019). Using naturalistic stimuli, MTL neurons have been shown to encode characters (Zhang et al., 2023; Gerken et al., 2025), scenes (Gerken et al., 2025), event boundaries (Zheng et al., 2022) and spatial information (Stangl et al., 2021). This modality-invariant representation makes the MTL a natural reference for modern vision models, many of which are trained to align visual and semantic information. Yet it remains unclear whether and how models' representations relate to MTL single-unit activity during naturalistic viewing.

Vision model –brain alignment Comparisons between deep networks and the primate visual system have focused on visual feature and object-class specific representations. Successive layers of supervised convolutional networks correspond to successive stages of increasing feature selectivity (Yamins et al., 2014; Cichy et al., 2016), and this correspondence is strong enough that images synthesized from a learned model-to-brain mapping can drive individual neurons in macaques (Bashivan et al., 2019; Ponce et al., 2019).

Modern vision models are instead trained with self-supervised, multimodal or agglomerative objectives on far larger and more diverse datasets (Oquab et al., 2024; Radford et al., 2021; Ranzinger et al., 2026), and their representations increasingly capture semantic structure beyond visual appearance (Wang et al., 2023). In parallel, encoding models have been extended to naturalistic stimuli such as movies (Wen et al., 2018; Khosla et al., 2021; Zhang et al., 2021). In humans, however, these comparisons have relied mostly on non-invasive measurements of visual cortex (Wang et al., 2023; Conwell et al., 2024; Baghel et al., 2026) with few studies using intracranial electrocorticography (ECoG) recordings focused on language (Schrimpf et al., 2021; Goldstein et al., 2022). Only one study has examined the MTL at the level of hippocampal voxel patterns (Choksi et al., 2022; Kell et al., 2018; Caucheteux et al., 2023; Tuckute et al., 2024).

Non-invasive methods cover large brain areas but trade off temporal or spatial resolution, whereas microwire recordings in epilepsy patients measure single-neuron spiking with high temporal and spatial precision. Whether modern vision models capture the content encoded by single MTL neurons during naturalistic viewing is currently unknown. Such comparisons are sensitive to analysis choices, as model rankings can be affected by temporal autocorrelation (Feghhi et al., 2024) and choice of similarity measure (Soni et al., 2024). Similarly, differences in architecture can be confounded with differences in training data (Conwell et al., 2024; Han et al., 2023).

We address both points by comparing representations from diverse architectures and training objectives with MTL single-unit activity during a full-length movie, using temporally separated evaluation, whitened feature spaces, and models with matched architectures but different training data.

3 METHODOLOGY

3.1 NEURAL DATASET AND PROBLEM FORMULATION

We used the SUMMER dataset (Darcher et al., 2026), which provides single-neuron spiking activity from the human MTL recorded during the complete presentation of the movie *500 Days of Summer* (83 min, 25 fps, 125,743 frames). The dataset comprises recordings from 29 intracranially implanted patients, with single- and multi-units recorded from the amygdala (A), the hippocampus (H), the entorhinal cortex (EC), the parahippocampal cortex (PHC), the piriform cortex (PiC), and other neighboring structures. Because each patient watched the same film, spike trains can be aligned to a common timeline to yield a single pseudo-population of 2,286 units, following common practice (Kamiński et al., 2017; Jamali et al., 2021; Gerken et al., 2025).

Spikes were grouped into 80 ms bins, corresponding to two movie frames per bin, resulting in $59{,}080$ bins across the $4{,}726$ s of movie that remain after the title sequence and end credits were excluded. For each vision model, we decoded the visual feature embedding from the lagged window of neural activity. Since the dataset used a single continuous presentation of the movie, both the stimulus and the neural signal were exposed to temporal autocorrelation and recurring visual content across the presentation. We addressed this using buffered interleaved splits (3.4) and circular-shift nulls (3.6), and also performed frame-similarity controls to quantify recurring visual content (Section 3.4).

3.2 VISUAL FEATURE SPACES

As baseline, we extracted a 134-dimensional handcrafted feature space comprising luminance and color statistics, edge information, spatial-frequency content, and coarse spatial layout (Appendix A.1). We then compared learned representations spanning architectures, scales and training paradigms to this baseline space. Supervised models included ImageNet-trained (Russakovsky et al., 2015) AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan & Zisserman, 2015), ResNet-50 (He et al., 2015), a Vision Transformer (Dosovitskiy et al., 2021), and a Places365-trained ResNet-50 (Zhou et al., 2017) to contrast object- and scene-centric supervision. We further evaluated the self-supervised models DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025), the vision-language models CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025), and RADIOv4 (Ranzinger et al., 2026), an agglomerative foundation model. To isolate the effect of training-set size, we additionally evaluated three CLIP ResNet-50 variants trained with the same contrastive image-text objective on training data spanning two orders of magnitude: CC-12M (Ilharco et al., 2021; Changpinyo et al., 2021), a 15M-pair subset of YFCC100M (Ilharco et al., 2021; Thomee et al., 2016), and the 400M-pair WIT dataset (Radford et al., 2021). All pretrained vision models were frozen, and no model was trained or fine-tuned on the neural recordings (Appendix Table 1). Each frame was resized to the model's specific input resolution without cropping (input resolution from 224 to 512 pixels). Feature-space geometry is summarized in Appendix Table 2.

3.3 NEURAL AND VISUAL PREPROCESSING

All data-dependent preprocessing parameters were estimated independently within each cross-validation fold using the training partition and applied to the validation and test partitions. The binned spike rates of each neuron were normalized using the mean and standard deviation from the training partition. To accommodate temporal offsets between visual input and neural activity, each input sample concatenated 11 consecutive 80 ms bins spanning -400 ms to $+400$ ms relative to the target bin, following previous work (Gerken et al., 2025).

The learned embedding of each frame was $L2$-normalized and then centered by subtracting the training-set mean. Because the low-level features had heterogeneous scales, they were first standardized using training-set statistics and then $L2$-normalized and centered in the same manner. To place learned representations on a common dimensional and covariance scale, we fitted a separate PCA-whitening transformation for each feature space and cross-validation fold using only training

data. For each feature space m, let $\mathbf{U}_{m,k}$ contain the leading k eigenvectors of the training covariance matrix and let $\mathbf{\Lambda}_{m,k}$ be the diagonal matrix of their corresponding eigenvalues. Each centered feature vector $z^c_{m,t}$ was transformed as $\widehat{\mathbf{z}}_{m,t} = \mathbf{z}^c_{m,t}\mathbf{U}_{m,k}\mathbf{\Lambda}^{-1/2}_{m,k}$. We used $k = 128$ for the primary analyses. We additionally evaluated $k \in \{64, 256, 512\}$ and unwhitened representations as sensitivity analysis. Low-level features were evaluated without whitening because numerical rank of the space did not support larger values of k.

3.4 Temporal separation of data partitions

The continuous and naturalistic paradigm required careful control for temporal autocorrelation. We adopted the buffered five-fold cross validation scheme developed for the SUMMER dataset (Gerken et al., 2025). The movie was divided into 11 blocks, each comprising five contiguous segments. Within each block, one segment was assigned to the test set, the following segment to the validation set, and the remaining three to the training set. These assignments were rotated across folds and boundaries between partitions were separated by 32 s buffers. Under this setup, each fold contained approximately 30,932 training, 7,304 validation, and 7,229 test samples.

These temporal buffers excluded adjacent frames but, given the film's non-chronological narrative, could not fully eliminate recurring visual content. To quantify this recurrence, we computed, for each feature space and fold, the maximum cosine similarities between every validation or test frame and the training frames separated from it by at least 32 s. We compared these similarities with a matched within-training control and repeated the analysis across increasing temporal separations. This procedure assessed whether visual similarity was concentrated near partition boundaries or distributed throughout the movie and the resulting analysis is shown in Appendix Figure 4.

3.5 Decoding visual representations from MTL activity

As core analysis, we fitted a separate multi-output ridge decoder for each feature space and cross-validation fold. The decoder predicted the $k = 128$ whitened representations of a given frame at time t from the standardized activity of the 2,286-unit population across the 11 bins spanning $t-400$ ms to $t + 400$ ms. For each feature space and fold, the regularization strength α was selected by maximizing validation frame identification over 19 logarithmically spaced values from 10^0 to 10^9. The selected decoder was then evaluated on the test set.

3.6 Evaluation metrics and statistical significance

Decoders were evaluated on held-out bins from the five temporally buffered folds. For each test sample, we ranked the true target by cosine similarity against candidates from the same test split that were at least 32 s away and defined identification accuracy as the fraction of eligible candidates ranked below the true frame (chance: 0.5). This graded metric accommodates variable candidate set-sizes (median: approx. 6,700) and is less sensitive than top-k retrieval to recurring or nearly identical frames. For primary cross-space comparisons, we used the $k = 128$ PCA-whitened feature embeddings, matching their dimensionality and training data covariance. We calibrated identification using synthetic predictions $\tilde{z}_i(\lambda) = \lambda z_i + (1 - \lambda)z_{\pi(i)}$, where z_i is the correct target and $z_{\pi(i)}$ is the target of a temporally separated frame. The calibrated signal fraction λ was the correct-target weight whose identification matched the observed decoder, obtained by inverting a space- and fold-specific calibration curve averaged over five corruption draws. At fixed k, λ is comparable across spaces but is not explained by variance or the fraction of the native embedding recovered. We additionally measured identification within bounded temporal candidate windows and the temporal error of the highest-ranked frame. These analyses used all training, validation, and test bins as candidates, while retaining only test bins as queries.

Within-fold confidence intervals used 1,000 bootstrap resamples of 11 contiguous test blocks, preserving local temporal dependence. Cross-space comparisons applied the same bootstrap to paired per-bin identification differences. Significance was assessed separately for each feature space and fold using 100 circular shifts of targets relative to neural activity (minimum shift: 60 s). The regularization parameter α, selected on aligned validation data, was fixed across shifts. As leakage control, we also evaluated the 20% of test bins with lowest maximum similarity to any training frame.

3.7 REGIONAL ANALYSIS OF THE PARAHIPPOCAMPAL CORTEX

To assess regional contributions, we grouped neurons based on anatomical labels provided into amygdala (A, 559 neurons), hippocampus (H, 846 neurons), entorhinal cortex (EC, 493 neurons), and parahippocampal cortex (PHC, 293 neurons). The remaining 95 neurons were located in various other areas and were excluded from the regional comparison.

To localize visually predictable activity independently of regional population size, we fitted a separate ridge encoding model for each neuron, predicting square-root-transformed spike counts from the $k = 128$ whitened visual features, using the same lagged design and temporally buffered folds. Per-neuron performance was quantified as held-out Pearson r minus the mean correlation under 100 circular shifts. PHC was compared with pooled A, H, and EC using 20,000 permutations of the anatomical labels. This grouping was based on findings which suggest that the PHC is functionally a higher-order visual area (Doeller & Kaplan, 2011; Rajimehr et al., 2011), as opposed to a site of multimodal integration. Common firing-rate ranges controlled for differences in neuronal excitability, and robustness was evaluated across all learned visual feature spaces (see Appendix A.3).

We applied the same $k = 128$ pipeline for each feature space to the 293-unit PHC population. PHC was compared with 50 random and 50 firing-rate-matched subpopulations of 293 units from the pooled A, H, and EC. We additionally evaluated 20 random and 20 firing-rate matched populations of 586 pooled control units. Firing rates and matching ranges were estimated independently from the training data of each fold. Every population was evaluated using the same movie bins and target representations, with α selected independently on the corresponding validation data and significance assessed using its own circular-shift null. Regional contrasts were computed within each fold by comparing PHC with the mean of the corresponding control samples.

4 RESULTS

4.1 DECODING VISUAL REPRESENTATIONS FROM THE HUMAN MTL

We asked whether visual representations can be decoded from MTL activity, and in a first experiment, decoded the CLIP embeddings and low-level visual features. For CLIP, the MTL activity reliably predicted the visual representations on the held-out data. Across five folds, the CLIP feature space achieved a mean identification accuracy of 0.785 among eligible frames spanning the full movie (fold range 0.74 - 0.821), compared with 0.575 under circularly shifted spike-movie alignments. No null shift reached the observed score in any fold. Decoding was substantially weaker for the low-level visual features: their mean identification accuracy was 0.622 against a null of 0.547 and identification accuracy did not reliably exceed the null in two out of five folds (Fig. 1a). The difference in accuracy between the two feature spaces was greater when discrimination was restricted to temporally local moments: within a 60-s-candidate window, CLIP remained above its null in all five folds (identification 0.605; gap to null: $+0.106$, $5/5$ folds), whereas low-level features were at chance (0.5, gap to null -0.004, $0/5$ folds; Fig. 1b).

For an additional retrieval analysis, the top-ranked CLIP frame was a median of 233 s from the true frame, with 7.6% of frames retrieved within 60 s compared to 1.8% by chance. Thus, MTL activity localized frames at a resolution of minutes, not at the level of individual shots. Furthermore, the result was not driven by recurring or near-duplicate frames: mean identification remained high when the analysis was restricted to the least-duplicated 20% of held-out bins (0.766 for CLIP, compared with 0.592 for low-level features, see Appendix Table 3). We also found that performance increased with population size, and showed no signs of saturation. CLIP identification accuracy increased from 0.566 with 36 units to 0.785 with 2,286 units, and the final doubling in population size contributed a further $+0.030$ (see Fig. 1c).

Overall, these findings demonstrate that MTL activity carries visual information during continuous, naturalistic viewing—particularly high-level CLIP representations—whereas low-level visual features cannot be reliably decoded. Full population performance provides a lower bound that can be linearly decoded from the sampled MTL population.

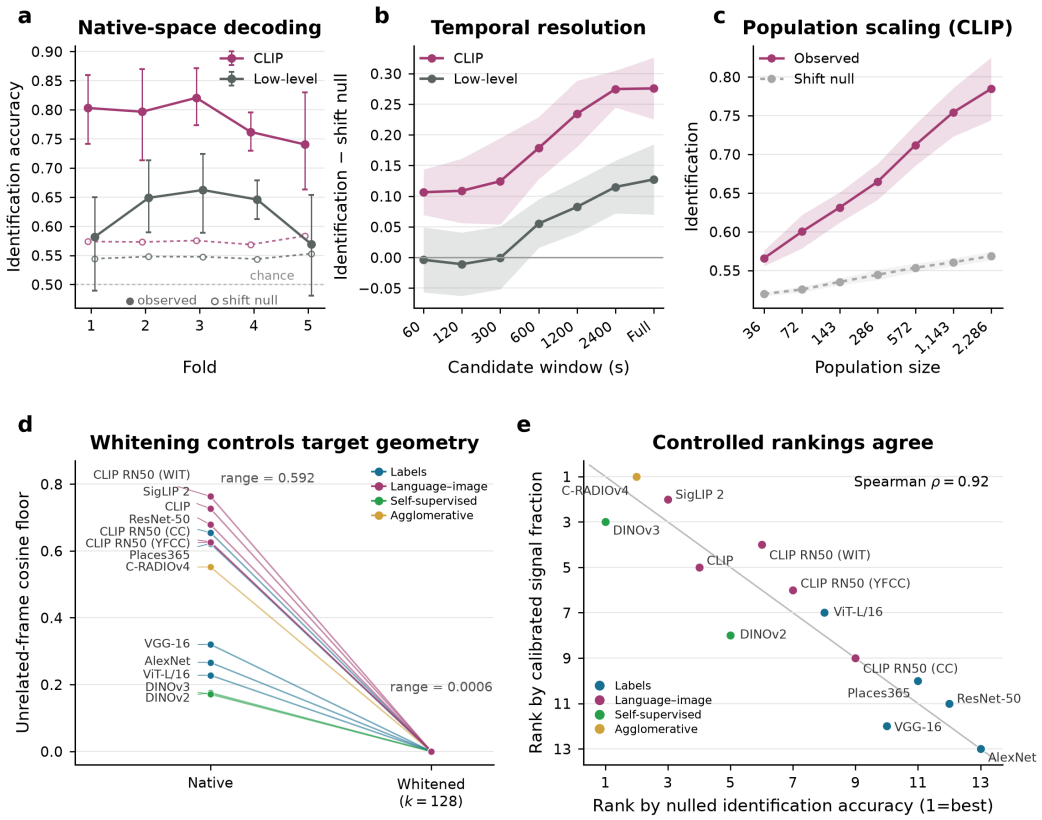

Figure 1: **Continuous visual-feature decoding from the MTL population and control of target-space geometry. (a)** Identification accuracy for CLIP and low-level features across the five cross-validation folds. Performance (filled markers) and corresponding circular shift null-means (open markers). Error bars: 95% confidence intervals obtained by resampling held-out movie segments. Chance identification is 0.5. **(b)** Identification above the circular-shift null as the candidate set is restricted to a symmetric temporal window around the target frame, guaranteeing a minimum separation of 32 s. Lines: means across five folds. Shaded regions: 95% t-intervals across folds. **(c)** CLIP identification as a function of population size. For each subsample size, five random draws were averaged within each fold, lines and shaded regions show the mean and 95% t-interval across five folds. Full population ($N = 2{,}286$) evaluated once per fold. Dashed gray line: circular-shift null. **(d)** Whitening to $k = 128$ removes large differences in unrelated-frame similarity across all visual spaces. **(e)** After whitening, rankings by nulled identification accuracy and calibrated signal fraction remain consistent (Spearman $\rho = 0.92$).

4.2 Geometry-aware evaluation of neural decoding

The geometries of the models' native feature spaces differed substantially: mean cosine similarity between unrelated frames ranged from 0.171 to 0.763, which could affect identification accuracies. Indeed, rankings based on identification accuracy diverged from those based on calibrated signal fractions: for example, DINOv2 exceeded Places365 in raw identification (0.792 vs 0.730), but their calibrated signal fractions reversed this ordering ($\lambda = 0.114$ vs. 0.168). Subtracting each space's circular-shift null was insufficient, as the resulting ranking remained strongly correlated with raw identification (Spearman $\rho = 0.91$).

Whitening each learned representation to $k = 128$ dimensions reduced the range of unrelated-frame similarities to below 0.001 (Fig. 1d) while preserving significant decoding in 64 of 65 space-fold combinations. Null corrected identification and calibrated signal fraction under whitening produced similar rankings (Spearman $\rho = 0.92$, Fig. 1e). Circular shift nulls, however, remained space-dependent, likely reflecting residual temporal structure not removed by whitening (Appendix Ta-

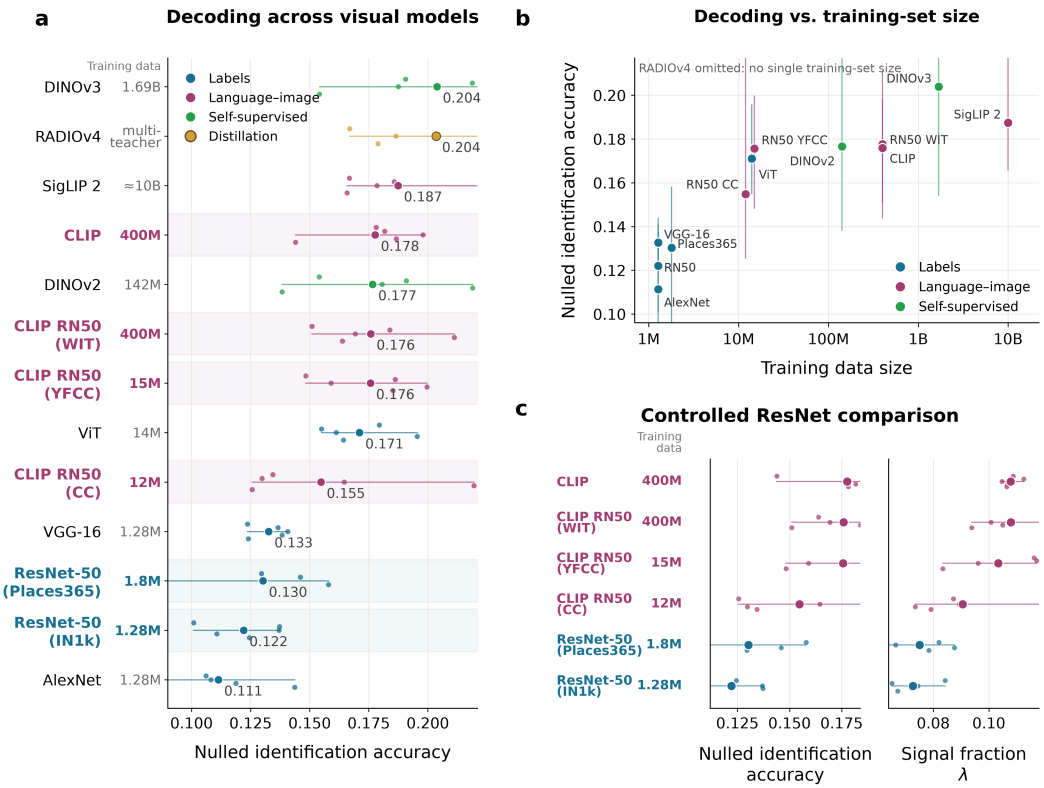

Figure 2: **Geometry-controlled decoding across visual feature spaces. (a)** Nulled identification accuracy for 13 models, ordered by the five-fold mean; small points show individual folds and horizontal lines their range. **(b)** Identification accuracy as a function of training-set size; color denotes the supervision type. Identification was evaluated against all eligible frames from the complete movie. **(c)** Controlled comparison of ResNet-50 models trained on different datasets and objectives (with CLIP ViT-L/14 as reference), shown for nulled identification accuracy and calibrated signal fraction (λ). All results use $k = 128$ whitening and the pooled 2,286-unit population; identification is evaluated against the full eligible frame pool.

ble 4). We report both identification accuracy and calibrated signal fraction to better describe the overall prediction performance. Rankings were stable across $k \in \{64, 128, 256, 512\}$, with $k = 128$ providing the best balance between geometric comparability and fold-level sensitivity (Appendix Fig. 5). Feature spaces whitened to $k = 128$ explained between 81.7% and 95.3% of variability.

4.3 Decodability varies with training dataset size

We evaluated decodability across all visual feature spaces using nulled identification accuracy. Modern, large-scale representations achieved the highest decoding performance, led by DINOv3 (0.204), the multi-teacher distillation model RADIOv4 (0.204) and SigLIP2 (0.187). In contrast, traditional label-supervised convolutional networks such as AlexNet and ResNet-50 yielded the lowest nulled identification accuracies (0.111 and 0.122, respectively). When analyzing performance across the complete set of vision models, nulled identification accuracy exhibited a monotonic increase with the size of the training dataset. Models trained on massive datasets exceeding one billion samples (i.e., DINOv3, SigLIP) consistently achieved the highest scores (Fig. 2a,b).

To isolate the impact of training data and objectives from architectural differences, we conducted a controlled comparison using only models sharing an identical ResNet-50 backbone. Across these structurally equivalent models, both nulled identification accuracy and the calibrated signal fraction (λ) scaled monotonically with dataset size. Performance increased sequentially from the baseline ResNet-50 trained on ImageNet-1k (1.28M samples) through intermediate CLIP variants (CC at

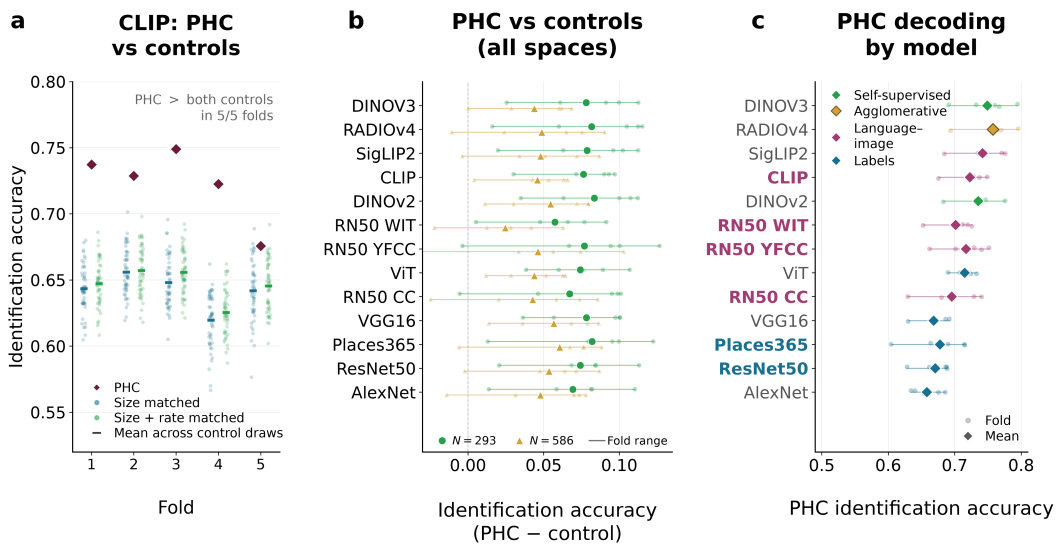

Figure 3: **Visual information is concentrated in PHC. (a)** Identification accuracy for the 293-unit PHC population and 50 size-matched or size-and-rate-matched draws from pooled amygdala, hippocampus, and entorhinal cortex. Diamonds denote PHC; translucent points denote draws; horizontal bars denote their means. **(b)** Fold-wise PHC advantages over firing-rate-matched populations of equal size ($N = 293$) and doubled size ($N = 586$) across seven visual feature spaces. Large symbols denote across-fold means and horizontal lines the fold range. **(c)** Identification accuracy of the PHC population across all visual feature spaces. Small circles show individual folds, diamonds five-fold means and horizontal lines indicate fold ranges. Feature spaces in panels **(b)** and **(c)** are ordered by descending nulled identification accuracy in the MTL population, as in Fig. 2. The CLIP-ResNet-50 models are abbreviated as RN50, appended with the dataset identifier.

12M (Changpinyo et al., 2021), and plateaued for the CLIP ResNet-50 variants trained on YFCC-15M(Thomee et al., 2016; Radford et al., 2021) and WIT-400M (Radford et al., 2021) (Fig. 2c). With the architecture held fixed, training thus changed decodability by about as much as the difference between the original CNN and transformer groups.

For very large training-data sizes, we did not see clear evidence for further improvement: The 27-fold larger WIT-400M did not exceed YFCC-15M, a supervised ViT trained on ImageNet-21k (Deng et al., 2009) (14M images, 0.171) came close to CLIP (400M) and DINOv2 (142M), and SigLIP2 (∼10B) scored below DINOv3 (1.7B). We note that ImageNet and CLIP ResNet-50 differ in training objective as well as data, so the effect cannot be attributed to dataset size alone.

4.4 Parahippocampal cortex leads visual decoding within the MTL

Previous neuroimaging work suggests that PHC activity is driven by spatial and visual features, as opposed to higher-order category-level information (Rajimehr et al., 2011; Doeller & Kaplan, 2011). To test whether visually predictable activity was enriched in the PHC at the level of individual neurons, we compared the extent to which CLIP embeddings could predict the firing rate of neurons recorded from across MTL regions. Mean null-corrected CLIP encoding was $2.46\times$ greater in PHC than in pooled amygdala, hippocampus, and entorhinal cortex (Appendix Fig. 6). However, stronger per-neuron encoding does not necessarily imply a stronger population readout, particularly when neurons carry redundant stimulus information. We therefore decoded the visual features from the 293-unit PHC population and compared its performance to size- and firing-rate-matched populations sampled from the pooled amygdala, hippocampus, and entorhinal cortex. Tested on the CLIP feature space, the PHC population achieved 0.723 identification accuracy and a 0.091 calibrated signal fraction across held-out frames. This outperformed size-matched (0.642, $\lambda = 0.06$) and rate-matched (0.646, $\lambda = 0.062$) controls across all five folds (Appendix Table 5, Fig. 3a). Despite comparable null values (0.5509 - 0.5701), PHC maintained the highest identification accuracy.

This advantage generalized across all visual feature spaces. Relative to equally sized firing-rate-matched populations, identification margins ranged from 0.058 to 0.084 (Fig. 3b). The advantage remained positive in every space even when the matched control contained twice as many neurons ($N = 586$), with margins of $0.025 - 0.061$. PHC also exceeded the size-matched control on average for the 20% least-similar test frames in every space, indicating that the regional advantage was not explained by recurrent or near-duplicate movie frames. Decoding from PHC reproduced the ordering of the full population, with ImageNet- and Places-trained CNNs outperformed by models trained on larger datasets, while null values in the PHC only varied slightly (Fig. 3c, Appendix Table 6). Finally, Places365 and ImageNet-trained ResNet-50 produced nearly identical PHC signal fractions ($\lambda = 0.0694, 0.0693$), with no evidence for a PHC-specific preference for scene-supervised representations. Overall, the regional analysis indicated a concentration of visual information in the PHC compared to other MTL subregions, with the same ordering of vision models.

5 Discussion and conclusion

We compared model-derived visual feature spaces against population activity of a large set of single units recorded from the human medial temporal lobe (MTL) during continuous, naturalistic movie viewing. Visual features could be reliably decoded from MTL activity, with three core findings: First, MTL activity contained substantially more information about learned visual representations than low-level, manually derived features. Second, this information was unevenly distributed within the MTL: the parahippocampal cortex (PHC) carried a disproportionately large amount of visual information. Third, after controlling for temporal autocorrelation and feature-space geometry, performance depended on how the model was trained: at a fixed architecture, CLIP training on web-scale data outperformed ImageNet classification, whereas leading foundation models could not be reliably distinguished by generation, input modality or architecture.

Our reported advantage of learned over low-level feature decoding is consistent with the MTL's role in integrating perceptual input into more abstract representations. However, low-level features are also lower-dimensional and change rapidly between frames, which may contribute to their poor decodability overall. Within the MTL, PHC yielded the strongest visual decoding, outperforming size-matched and larger control populations across all visual feature spaces. This fits with its position as the MTL region closest to visual cortex and its established role in scene processing, both relevant for a scene-rich movie stimulus. Because models were ranked similarly in PHC and in the remaining MTL, PHC may also carry a stronger version of a signal shared across the MTL.

Our controlled ResNet-50 comparison indicates that decodability depends strongly on how a model was trained: the same architecture decoded markedly better after CLIP training on web data than after ImageNet classification, consistent with results obtained with fMRI (Conwell et al., 2024). Among the foundation models trained on large datasets differences were small and not reliable. However, decoding performance scaled with the number of units up to the full population of $2{,}286$ neurons without saturating, so this may reflect limited neural data rather than model equivalence.

A central methodological point is that feature spaces differ in geometry. Decoding performance therefore reflects not only how accurately neural activity predicts a representation, but also how that representation arranges candidate frames. Whitening each space to an equal number of variance-normalized directions reduced this difference, after which identification accuracy and calibrated signal fraction yielded consistent rankings. Because such differences change the baseline difficulty of the task, model–brain comparisons should control for feature-space geometry as part of the evaluation design (Soni et al., 2024).

Naturalistic viewing also introduces challenges. First, temporal autocorrelations in both input and neural data need to be controlled for. Second, movie frames co-vary with speech, sound, narrative structure, memory and attention, whereas the extracted feature spaces describe visual content only. As in other model–brain comparisons, rankings might depend on evaluation choices (Soni et al., 2024; Han et al., 2023), and decodability does not imply that a model shares the computations of the brain. Future work could test whether feature spaces transfer across recording modalities using fMRI on the same movie (Aliko et al., 2020), and whether the decoded representations are relevant for behavior, using tasks that link visual content to perceptual decisions.

AI USE STATEMENT

In this work, we used generative AI tools for research idea generation, literature search and identification, writing manuscript text, method implementation, and modifying scientific figures. We have not used generative AI tools for primary experimental data collection, and experimental dataset collection tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming, copyediting, and rephrasing prose. We have reviewed all AI-assisted work. LLM-generated research ideas and literature search results were verified through manual literature surveys; code for implementing methods was tested and debugged by the authors; AI-assisted scientific figure modifications and manuscript text were thoroughly reviewed and validated for accuracy and clarity. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sarah Aliko, Jiawen Huang, Florin Gheorghiu, Stefanie Meliss, and Jeremy I. Skipper. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, 7(1): 347, October 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00680-2.

Shashank Baghel, Kshitij Dwivedi, Dinesh Singh, and Sanjeev Nara. Do Vision Model See Like the Brain? A Comparison Across EEG Encoding Model, September 2026.

Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aav9436.

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, March 2023. ISSN 2397-3374. doi: 10.1038/s41562-022-01516-2.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, 2021.

Bhavin Choksi, Milad Mozafari, Rufin VanRullen, and Leila Reddy. Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks*, 154:538–542, 2022. doi: 10.1016/j.neunet.2022.07.033.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, June 2016. ISSN 2045-2322. doi: 10.1038/srep27755.

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*, 15:9383, 2024. doi: 10.1038/s41467-024-53147-y.

Alana Darcher, Franziska Gerken, Johannes Niediek, Marcel S. Kehl, Thomas P. Reber, Stefanie Liebe, Laura Nett, Attila Racz, Lukas Kunz, Bernhard Staresina, Rachel Rapp, Pedro J. Gonçalves, Ismail Elezi, Valeri Borger, Rainer Surges, Laura Leal-Taixé, Jakob H. Macke, and Florian Mormann. Human neuron activity during an 83-minute movie from 2,286 neurons and 29 patients. *Scientific Data*, 13(1):1105, July 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07955-0. URL https://doi.org/10.1038/s41597-026-07955-0.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Christian F. Doeller and Raphael Kaplan. Parahippocampal Cortex: Translating Vision into Space. *Current Biology*, 21(15):R589–R591, August 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.06.023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A. Blank, and Jonathan C. Kao. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*, 2024.

Franziska Gerken, Alana Darcher, Pedro J Gonçalves, Rachel Rapp, Ismail Elezi, Johannes Niediek, Marcel S Kehl, Thomas P Reber, Stefanie Liebe, Jakob H Macke, Florian Mormann, and Laura Leal-Taixé. Decoding movie content from neuronal population activity in the human medial temporal lobe. August 2025. doi: 10.7554/elife.106758.1. URL http://dx.doi.org/10.7554/eLife.106758.1.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01026-4.

Yena Han, Tomaso Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Mohsen Jamali, Benjamin L. Grannan, Evelina Fedorenko, Rebecca Saxe, Raymundo Báez-Mendoza, and Ziv M. Williams. Single-neuronal predictions of others' beliefs in humans. *Nature*, 591(7851):610–614, March 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03184-0. URL https://doi.org/10.1038/s41586-021-03184-0.

Jan Kamiński, Shannon Sullivan, Jeffrey M Chung, Ian B Ross, Adam N Mamelak, and Ueli Rutishauser. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature Neuroscience*, 20(4):590–601, April 2017. ISSN 1546-1726. doi: 10.1038/nn.4509. URL https://doi.org/10.1038/nn.4509.

Katharina Karkowski, Marcel S. Kehl, Yingjie Qin, Alana Darcher, Philipp Karkowski, Valeri Borger, Rainer Surges, Martin N. Hebart, Thomas P. Reber, Stanislas Dehaene, Yair Lakretz, and Florian Mormann. Semantic Tuning of Single Neurons in the Human Medial Temporal Lobe, October 2025. ISSN 2692-8205.

Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018. ISSN 08966273. doi: 10.1016/j.neuron.2018.03.044.

Meenakshi Khosla, Gia H. Ngo, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Science Advances*, 7(22):eabe7547, May 2021. doi: 10.1126/sciadv.abe7547.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals

Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10, May 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.04.005.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Reza Rajimehr, Kathryn J. Devaney, Natalia Y. Bilenko, Jeremy C. Young, and Roger B. H. Tootell. The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLOS Biology*, 9(4):e1000608, April 2011. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000608.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-RADIOv4 (tech report), 2026. URL https://arxiv.org/abs/2601.17237.

Thomas P. Reber, Marcel Bausch, Sina Mackay, Jan Boström, Christian E. Elger, and Florian Mormann. Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. *PLOS Biology*, 17(6):e3000290, June 2019. ISSN 1545-7885. doi: 10.1371/journal.pbio.3000290.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Martin Schrimpf, Idan Asher 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):e2105646118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2105646118.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Ansh Soni, Sudhanshu Srivastava, Konrad Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024. doi: 10.1101/2024.08.07.607035.

Larry R Squire, Craig EL Stark, and Robert E Clark. The medial temporal lobe. *Annu. Rev. Neurosci.*, 27:279–306, 2004.

Matthias Stangl, Uros Topalovic, Cory S. Inman, Sonja Hiller, Diane Villaroman, Zahra M. Aghajan, Leonardo Christov-Moore, Nicholas R. Hasulak, Vikram R. Rao, Casey H. Halpern, Dawn Eliashiv, Itzhak Fried, and Nanthia Suthana. Boundary-anchored neural mechanisms of location-encoding for self and others. *Nature*, 589(7842):420–425, January 2021. ISSN 1476-4687. doi: 10.1038/s41586-020-03073-y.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, March 2024. ISSN 2397-3374. doi: 10.1038/s41562-023-01783-7.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5:1415–1426, 2023. doi: 10.1038/s42256-023-00753-y.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cerebral Cortex*, 28 (12):4136–4160, December 2018. ISSN 1047-3211. doi: 10.1093/cercor/bhx268.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1403112111.

Yipeng Zhang, Zahra M Aghajan, Matias Ison, Qiujing Lu, Hanlin Tang, Guldamla Kalender, Tonmoy Monsoor, Jie Zheng, Gabriel Kreiman, Vwani Roychowdhury, et al. Decoding of human identity by computer vision and neuronal vision. *Scientific Reports*, 13(1):1–16, 2023.

Yizhen Zhang, Jung-Hoon Kim, David Brang, and Zhongming Liu. Naturalistic stimuli: A paradigm for multiscale functional characterization of the human brain. *Current Opinion in Biomedical Engineering*, 19:100298, September 2021. ISSN 2468-4511. doi: 10.1016/j.cobme.2021.100298.

Jie Zheng, Andrea G. P. Schjetnan, Mar Yebra, Bernard A. Gomes, Clayton P. Mosher, Suneil K. Kalia, Taufik A. Valiante, Adam N. Mamelak, Gabriel Kreiman, and Ueli Rutishauser. Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature Neuroscience*, 25 (3):358–368, March 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01020-w.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A APPENDIX

A.1 VISUAL FEATURE SPACES AND THEIR GEOMETRY

Each visual feature space was derived from a pre-trained vision model, with features extracted for every frame of the movie used in the SUMMER dataset (Darcher et al., 2026).

The only exception is the low-level feature space, which uses no pretrained model. From each squashed frame we computed 134 handcrafted image statistics: color and luminance moments and percentiles (15), per-channel intensity histograms (48), opponent-color means and standard deviations (4), Sobel edge magnitude and density (3), the spatial-frequency power spectrum pooled into log-spaced radial by orientation bins together with total power (48), and a 4×4 average-pooled luminance map giving coarse spatial layout (16). Because the columns differ in scale by up to three orders of magnitude, they were standardized per column on the training bins before decoding.

Learned visual feature spaces were read at their conventional readouts: the post-ReLU fc7 layer for AlexNet and VGG16, global average pooling of the final stage for all ResNet-50 spaces, the CLS token for the supervised ViT and the DINO models, the projected image embedding for CLIP ViT-L/14 and SigLIP2, and the backbone summary token for RADIOv4. We summarize the details about the visual feature spaces and the vision models used for extraction in Table 1.

Table 1: **Visual representation spaces.** Each movie frame was mapped to one global feature vector using either a frozen pretrained model or the handcrafted low-level control. The complete frame was resized without applying cropping to the input resolution (px). Output feature dimensionality is denoted by d. *Data* is the training dataset and *Size* its approximate number of images (label- and self-supervised models) or image–text pairs (language-aligned models). The three CLIP-RN50 spaces share an architecture, readout, resolution and objective and differ only in the size of their training set.

Space	Architecture	Training objective	Data	Size	px	d
Low-level control						
Low-level	Manual	None	—	—	224	134
Label-supervised						
AlexNet	AlexNet	Classification	ImageNet-1k	1.28M	224	4,096
VGG16	VGG-16	Classification	ImageNet-1k	1.28M	224	4,096
ResNet-50	ResNet-50	Classification	ImageNet-1k	1.28M	224	2,048
Places365	ResNet-50	Classification	Places365	1.8M	224	2,048
ViT	ViT-L/16	Classification	ImageNet-21k[a]	14M	224	1,024
Language-aligned						
CLIP	ViT-L/14	Contrastive img–txt	WIT	400M	224	768
SigLIP2	ViT-g/16	Sigmoid img–txt	WebLI	∼10B	384	1,536
CLIP-RN50(WIT)	ResNet-50[b]	Contrastive img–txt	WIT	400M	224	2,048
CLIP-RN50(YFCC-15M)	ResNet-50[b]	Contrastive img–txt	YFCC-15M	15M	224	2,048
CLIP-RN50(CC-12M)	ResNet-50[b]	Contrastive img–txt	CC-12M	12M	224	2,048
Self-supervised						
DINOv2	ViT-L/14	Self-distillation	LVD-142M	142M	224	1,024
DINOv3	ViT-7B/16	Self-distill. + reg.	LVD-1689M	1.69B	512	4,096
Agglomerative						
RADIOv4	ViT-H/16	Multi-teacher distill.	Teachers[c]	—	512	2,560

[a] Pretrained on ImageNet-21k, then finetuned on ImageNet-1k.
[b] CLIP's ResNet-50 variant, which differs from the supervised ResNet-50 in its deep stem, anti-aliased downsampling and attention-pooled head. The readout used here is the average pool of the final stage, matching the supervised ResNet-50 rows rather than the learned attention pool.
[c] Distilled from SigLIP2-g, DINOv3 and SAM3, so it has no training set of its own and is not comparable on the size axis.

By extracting visual features from each vision models for every frame of the movie, we obtain movie-specific representation for each feature space. Because movie contain complex, naturalistic content - including recurring scenes and many temporally correlated adjacent frames - we assess each feature space's geometry by analyzing frame-to-frame similarity across the movie and evaluating the test/train partitions from the five-fold cross-validation setup. Table 2 summarizes each space's geometry using three metrics: the mean cosine similarity between randomly sampled frame pairs (floor), the normalized median of test-to-train similarity, and the percentage of test frames with a raw maximum cosine similarity to training above 0.95.

Table 2: **Native geometry of the feature spaces on the movie stimulus.** Statistics were pooled over 36,515 test bins from the five disjoint cross-validation folds. *Floor* denotes the mean cosine similarity between randomly sampled, temporally separated frame pairs. *Normalized median* is the median maximum test-to-training similarity, normalized relative to the space-specific floor f. The final column reports the percentage of test bins whose maximum raw cosine similarity to a training frame exceeded 0.95.

Space	**Floor**	**Norm. median test-to-train**	**Raw** > 0.95
Low-level control			
Low-level	0.000	0.727	4.7%
Label-supervised			
AlexNet	0.266	0.592	3.4%
VGG16	0.320	0.562	3.5%
ResNet-50	0.655	0.621	9.2%
Places365	0.622	0.713	13.1%
ViT	0.228	0.543	2.8%
Language-aligned			
CLIP	0.679	0.637	8.4%
SigLIP2	0.727	0.683	22.7%
CLIP-RN50 (WIT)	0.763	0.687	17.6%
CLIP-RN50 (YFCC-15M)	0.626	0.660	10.0%
CLIP-RN50 (CC-12M)	0.627	0.688	11.4%
Self-supervised			
DINOv2	0.176	0.588	5.8%
DINOv3	0.171	0.576	8.3%
Agglomerative			
RADIOv4	0.552	0.679	12.4%

Additionally, we evaluated the frame similarity between test-and-training frames while progressively increasing the temporal separations from the implemented 32 s distance up to a maximum of 1,000 s separations (Fig. 4).

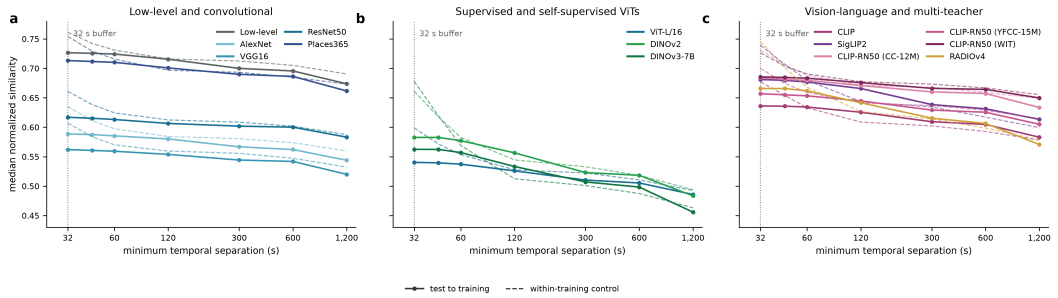

Figure 4: **Feature similarity across temporal separations.** For each feature space, we computed the maximum cosine similarity between each test frame and eligible training frames while progressively increasing their temporal separation. Similarities were normalized relative to the space-specific unrelated-frame floor as $(s - \text{floor})/(1 - \text{floor})$. Solid lines show the median test-to-training similarity, and dashed lines show the matched within-training control; values were computed within each fold and averaged across all five folds. The vertical dotted line marks the realized 32 s temporal buffer. Feature spaces are grouped for visualization into **(a)** low-level and convolutional representations, **(b)** supervised and self-supervised Vision Transformers, and **(c)** vision-language and multi-teacher representations. Curves represent individual feature spaces; no averaging was performed across models. All analyses use the original, pre-whitened representations.

A.2 Decoding visual representation from human MTL population activity

We decoded CLIP and low-level feature representations from human MTL population activity. The primary evaluation metric was identification accuracy across all held-out test bins. Complementary evaluation are summarized in Table 3, including identification accuracy restricted to the 20% least-duplicate test bins and the temporal error measuring the absolute temporal distance between true and highest-ranked frame. Additionally, we report the percentage of frames retrieved withing a 60s window. In addition to our baseline choice of $k = 128$ for whitening, we evaluated $k \in \{64, 256, 512\}$. The resulting model hierarchies, ranked by nulled identification accuracy, remained highly consistent across all values of k (Fig. 5). Decoding performances across all visual feature spaces are summarized in Tabel 4.

Table 3: **Decoding and temporal-retrieval metrics for CLIP and low-level features.** Identification accuracy is reported for all held-out frames and for the least-duplicated 20% of held-out bins. The temporal error denotes the absolute temporal distance between the true frame and the highest-ranked retrieved frame. Values are means across five folds, with fold ranges in parentheses. Chance retrieval was estimated by randomly sampling from the eligible candidate frames. A separate circular-shift null was not computed for the least-duplicated subset.

Measure	CLIP	Low-level	Reference
Identification, all frames	0.785 (0.740–0.821)	0.622 (0.569–0.662)	0.575 / 0.547 (null)
Identification, least-duplicated 20%	0.766 (0.710–0.815)	0.592 (0.551–0.635)	-
Median temporal error	233.2 s (202.2–313.2)	621.4 s (357.5–824.5)	1,403 s (chance)
Retrieved within 60 s	7.64% (5.61–10.35%)	3.59% (3.29–4.08%)	1.83% (chance)

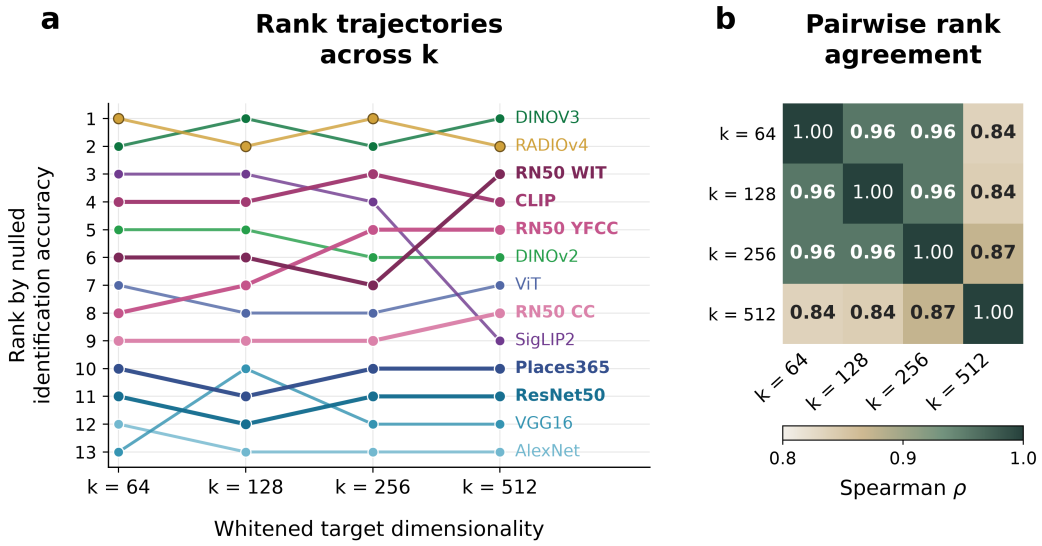

Figure 5: **Visual space decoding hierarchy is stable across whitening dimensionalities. (a)** Model ranks based on five-fold mean nulled identification accuracy at $k = 64, 128, 256, 512$. **(b)** Pairwise Spearman correlation between rankings. The braod hierarchy remains consistent across k ($\rho = 0.835 - 0.956$), with the greatest rearrangement at $k = 512$.

A.3 Decoding from the parahippocampal cortex

For each visual feature space and fold, one ridge model was fitted per neuron to predict square-root-transformed spike counts from the temporally lagged, $k = 128$ training-whitened features. A single regularization parameter was selected per space and fold by maximizing the mean validation correlation across neurons and was reselected within each circular-shift null. Encoding performance was defined for each neuron as its held-out Pearson correlation minus the mean correlation obtained from 100 circular shifts of neural activity relative to the visual features, with a minimum displacement of 60s; scores were subsequently averaged across the five folds. The PHC contrast used pooled amygdala, hippocampus, and entorhinal cortex neurons as the comparison group and 20,000 label permutations that preserved the observed group size. Using CLIP, mean null-corrected encoding performance was $2.46\times$ higher in PHC than in pooled amygdala, hippocampus, and entorhinal cortex ($p < 5 \times 10^{-5}$, label permutation; Fig. 6a). To examine firing-rate dependence, common ranges with boundaries 0.21, 0.58, 1.34 and 2.85 Hz were applied to every region. The PHC advantage persisted within all five firing-rate strata and generalized across all visual feature spaces, inidicating broad enrichment of visually predictable activity rather than an effect of firing rate or affinity for one particular representation family (Fig. 6b,c).

	Training set		Identification			Signal fraction	
Space	Data	Scale	Obs.	Null	Gap	λ	s.e.m.
RADIOv4	teachers	—	0.7988	0.5952	+0.2036	0.1328	0.0071
SigLIP2	WebLI	∼10B	0.7822	0.5947	+0.1874	0.1234	0.0064
DINOv3	LVD-1689M	1.69B	0.7923	0.5884	+0.2039	0.1231	0.0099
CLIP-RN50 (WIT)	WIT-400M	400M	0.7548	0.5790	+0.1759	0.1079	0.0053
CLIP	WIT-400M	400M	0.7643	0.5866	+0.1777	0.1078	0.0014
CLIP-RN50 (YFCC)	YFCC-15M	15M	0.7509	0.5752	+0.1757	0.1034	0.0064
ViT	ImageNet-21k	14M	0.7510	0.5799	+0.1711	0.1028	0.0046
DINOv2	LVD-142M	142M	0.7604	0.5839	+0.1766	0.1023	0.0055
CLIP-RN50 (CC)	CC-12M	12M	0.7249	0.5701	+0.1548	0.0906	0.0088
Places365	Places365	1.8M	0.6877	0.5574	+0.1303	0.0750	0.0050
ResNet-50	ImageNet-1k	1.28M	0.6829	0.5608	+0.1221	0.0726	0.0033
VGG16	ImageNet-1k	1.28M	0.6766	0.5439	+0.1326	0.0723	0.0016
AlexNet	ImageNet-1k	1.28M	0.6710	0.5596	+0.1114	0.0672	0.0059

Table 4: **Visual feature spaces decoded from human MTL spiking.** Thirteen target spaces, each decoded from the same pseudo-population of 2,286 units over the same five interleaved folds, with every target whitened to its leading $k = 128$ directions so that identification is comparable across spaces. More details about the vision models are given in Table 1. Identification is the percentile rank of the true frame against every eligible candidate (median 6,723 after a 32 s separation buffer); chance is 0.5. *Null* is the circular-shift null (100 shifts of at least 60 s), which sits well above chance because it inherits each space's own geometry, and *Gap* is their difference; every space clears its null at the $p = 0.01$ floor of 100 shifts. λ is the calibrated signal fraction with the s.e.m. over the five folds. All values are means over the five folds, and rows are ordered by λ. The three CLIP-RN50 spaces are a ResNet-50 series sharing an architecture and a readout and differing only in their training set. The ImageNet and Places365 spaces are label-supervised, DINOv2 and DINOv3 self-supervised, the CLIP, CLIP-RN50 and SigLIP2 spaces language-supervised, and RADIOv4 is distilled from SigLIP2, DINOv3 and SAM3, so it has no training set of its own and its row is not comparable on the scale axis.

Table 5: **Decoding visual representations from the PHC and control populations.** Identification accuracy from the parahippocampal population and from size-matched control populations drawn from the rest of the MTL, using visual features extracted from the CLIP model. Visual embeddings were $k = 128$ whitened, evaluation on five cross-validation folds; values are means over folds and, for the controls, over draws, with $\pm$ giving the spread across draws within a fold. Each decode carries its own 100-shift circular null at a minimum shift of 60 s. Chance identification is 0.5.

population	units	identification	null	gap	λ
PHC	**293**	**0.7227**	**0.5531**	**+0.1695**	**0.0907**
Size-matched pool	293	0.6418 ± 0.0195	0.5568	+0.0850	0.0602
Size + rate-matched pool	293	0.6463 ± 0.0176	0.5575	+0.0888	0.0618
Amygdala	293	0.6722 ± 0.0115	0.5604	+0.1117	0.0711
Hippocampus	293	0.6206 ± 0.0155	0.5509	+0.0697	0.0527
Entorhinal	293	0.6353 ± 0.0142	0.5607	+0.0746	0.0579
Pool, 2× size	586	0.6689 ± 0.0139	0.5681	+0.1008	0.0699
Pool, 2× size, rate-matched	586	0.6768 ± 0.0140	0.5701	+0.1067	0.0728

Table 6: **Decoding performance from PHC across visual feature spaces.** Identification accuracy for PHC and firing-rate-matched control populations of equal size ($N = 293$) or twice the size ($N = 586$). Values are means across five folds and, for the controls, across matched draws. Values in parentheses are the PHC-minus-control identification margins. Low-similarity results are calculated separately within each feature space using the 20% of held-out bins with the lowest maximum similarity to any training frame.

	All held-out frames				Least-similar 20%		
Visual space	PHC	PHC null	Control $N = 293$	Control $N = 586$	PHC	Control $N = 293$	Control $N = 586$
RADIOv4	0.757	0.559	0.676 (+0.082)	0.709 (+0.049)	0.697	0.618 (+0.079)	0.640 (+0.057)
DINOv3	0.749	0.553	0.671 (+0.078)	0.705 (+0.044)	0.672	0.616 (+0.056)	0.641 (+0.031)
SigLIP2	0.742	0.557	0.663 (+0.079)	0.694 (+0.048)	0.674	0.615 (+0.059)	0.640 (+0.034)
DINOv2	0.735	0.550	0.652 (+0.084)	0.681 (+0.055)	0.650	0.599 (+0.051)	0.624 (+0.026)
CLIP	0.723	0.553	0.646 (+0.076)	0.677 (+0.046)	0.670	0.607 (+0.062)	0.633 (+0.037)
ViT	0.715	0.548	0.641 (+0.074)	0.671 (+0.044)	0.649	0.598 (+0.051)	0.621 (+0.028)
Places365	0.678	0.534	0.596 (+0.082)	0.617 (+0.061)	0.638	0.555 (+0.083)	0.560 (+0.078)
ResNet-50	0.671	0.535	0.597 (+0.074)	0.617 (+0.054)	0.622	0.553 (+0.070)	0.562 (+0.060)
VGG16	0.668	0.528	0.590 (+0.078)	0.612 (+0.057)	0.648	0.572 (+0.076)	0.593 (+0.055)
AlexNet	0.658	0.535	0.589 (+0.069)	0.610 (+0.048)	0.603	0.558 (+0.045)	0.576 (+0.028)

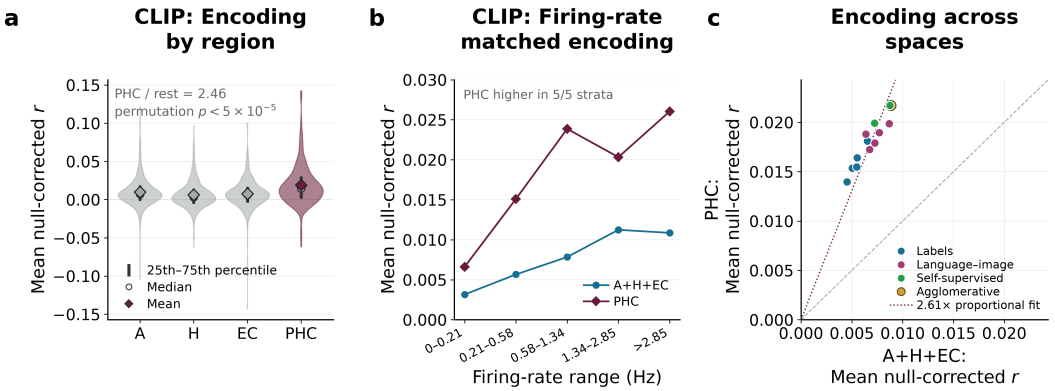

Figure 6: **Per-neuron visual encoding across MTL regions.** Encoding performance was measured as held-out Pearson r minus each neuron's mean correlation under 100 circular shifts. (**a**) Distribution of CLIP encoding performance across amygdala (A), hippocampus (H), entorhinal cortex (EC), and parahippocampal cortex (PHC); bars, circles and diamonds denote the interquartile range, median, and mean, respectively. (**b**) Mean CLIP encoding performance for PHC and pooled A+H+EC neurons within five common firing-rate ranges. (**c**) Mean PHC performance against pooled A+H+EC performance across all visual feature spaced. The dashed line denotes identity and the dotted line the $2.61\times$ proportional fit.