A single computational objective can produce specialization of streams in visual cortex

Dawn Finzi[1,2,✉], **Eshed Margalit**[3], **Kendrick Kay**[4*], **Daniel L. K. Yamins**[1,2,5,*], **and Kalanit Grill-Spector**[1,5,*]

[1]Department of Psychology, Stanford University, Stanford, CA 94305

[2]Department of Computer Science, Stanford University, Stanford, CA 94305

[3]Neurosciences Graduate Program, Stanford University, Stanford, CA 94305

[4]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455

[5]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305

[*]co-senior author

Human visual cortex is organized into dorsal, lateral, and ventral streams. A long-standing hypothesis is that the functional organization into streams emerged to support distinct visual behaviors. Here, we use a neural network-based computational model and a massive fMRI dataset to investigate why visual streams emerge. We find that models trained for stream-specific visual behaviors poorly capture brain responses and organization. Instead, a self-supervised Topographic Deep Artificial Neural Network, which encourages nearby units to respond similarly, successfully predicts brain responses, spatial segregation, and functional differentiation across streams. These findings challenge the prevailing view that streams evolved to separately support different behaviors and suggest instead that functional organization can arise from a single principle: learning generally useful visual representations subject to local spatial constraints.

processing streams | topography | vision | neural network

Correspondence: *dfinzi@stanford.edu*

A central question in neuroscience is why the visual system is spatially and functionally organized the way it is. The human visual system is comprised of extensive neural machinery, including more than two dozen visual areas organized into three anatomically distinct hierarchical processing streams, spanning the occipital, temporal, and parietal lobes [1, 2]: (1) the ventral stream, which ascends from early visual cortex (EVC: V1, V2, V3) in the occipital lobe to the inferior aspects of the temporal lobe; (2) the lateral stream, which ascends from EVC through lateral occipito-temporal cortex to the superior temporal sulcus (STS); and (3) the dorsal stream, which projects superiorly from EVC along occipito-parietal cortex. However, it remains unresolved why the visual system is organized into multiple processing streams.

The logic underlying the functional organization of the visual system has been studied for decades across empirical [3, 4, 5, 6], information-theoretic [1, 2], and computational [7, 8] approaches. Three key ideas have emerged from these investigations. First, to support complex visual behaviors, the visual system may decompose complex operations into a sequence of simpler computations distributed hierarchically across processing stages [9, 1, 2, 10, 11]. Second, to enable rapid and efficient processing, visual behaviors that are relatively independent, such as recognizing whether an approaching object is a car, gauging its distance, and predicting its trajectory, are thought to be processed in parallel by largely distinct streams [3, 4, 12, 13, 14]. Indeed, the ventral stream is hypothesized to be involved in visual recognition [3, 4] and categorization [15, 16, 17], the lateral stream in processing visual dynamics [12], actions [13], and interactions [14, 18], and the dorsal stream in spatial tasks [3, 4] such as detecting and localizing objects. Third, physical constraints, such as fitting the brain into the skull [19] and a need for fast processing, generate a bias for short-range connections [20, 8, 1, 2], which, in turn, may encourage nearby neurons to respond similarly [1, 21, 22, 23]. Indeed, nearby neurons tend to exhibit correlated responses, evident in the many topographic maps across visual cortex, ranging from retinotopic maps [24, 25, 26] in which nearby neurons in cortex process nearby regions of the visual field, to clustering of neurons that selectively respond to specific categories (e.g., faces [27, 28] or words [29]) in high-level visual cortex.

Advances in deep artificial neural networks (DANNs) over the past decade have enabled theories to be transformed into into concrete computational models that are trained with images and videos as input and tested against real neural data, revolutionizing computational and theoretical neuroscience [30, 17, 31, 32, 33, 34, 35, 36, 37, 38]. To date, the majority of vision-DANN research has focused on understanding the function of the ventral stream, which is empirically the best understood. The main insights from this research are four-fold: (i) feedforward DANNs predict responses across the ventral stream, with earlier layers in DANNs matching responses in early visual cortex (e.g., V1), and later layers matching high-level visual cortex (e.g., ventral temporal cortex, VTC, end of ventral stream) [17, 31, 36], (ii) DANNs trained on behaviorally-relevant tasks, such as categorization for the ventral stream, better predict neural responses than other DANNs [17], (iii) many DANN architectures can replicate neural responses in the ventral stream, albeit with numeric differences in performance [35, 39], and (iv) topographic organization within a visual area can emerge through wiring-constrained training [22, 40, 41, 42]. However, such research has not tackled the broader question of why the visual system organizes into streams, such as the ventral stream, in the first place. The goal of the present study is to use a DANN framework to test two main factors that have been hypothesized to contribute to the organization of the visual system into processing streams: optimization of different streams to perform different behaviors (multiple behaviors hypothesis) [3, 43, 44, 45, 46, 17], or learning useful representations under spatial constraints to minimize wiring (spatial constraints hypothesis) [19, 8, 1, 2].

Results

Computational framework to instantiate hypotheses for stream organization using deep neural networks

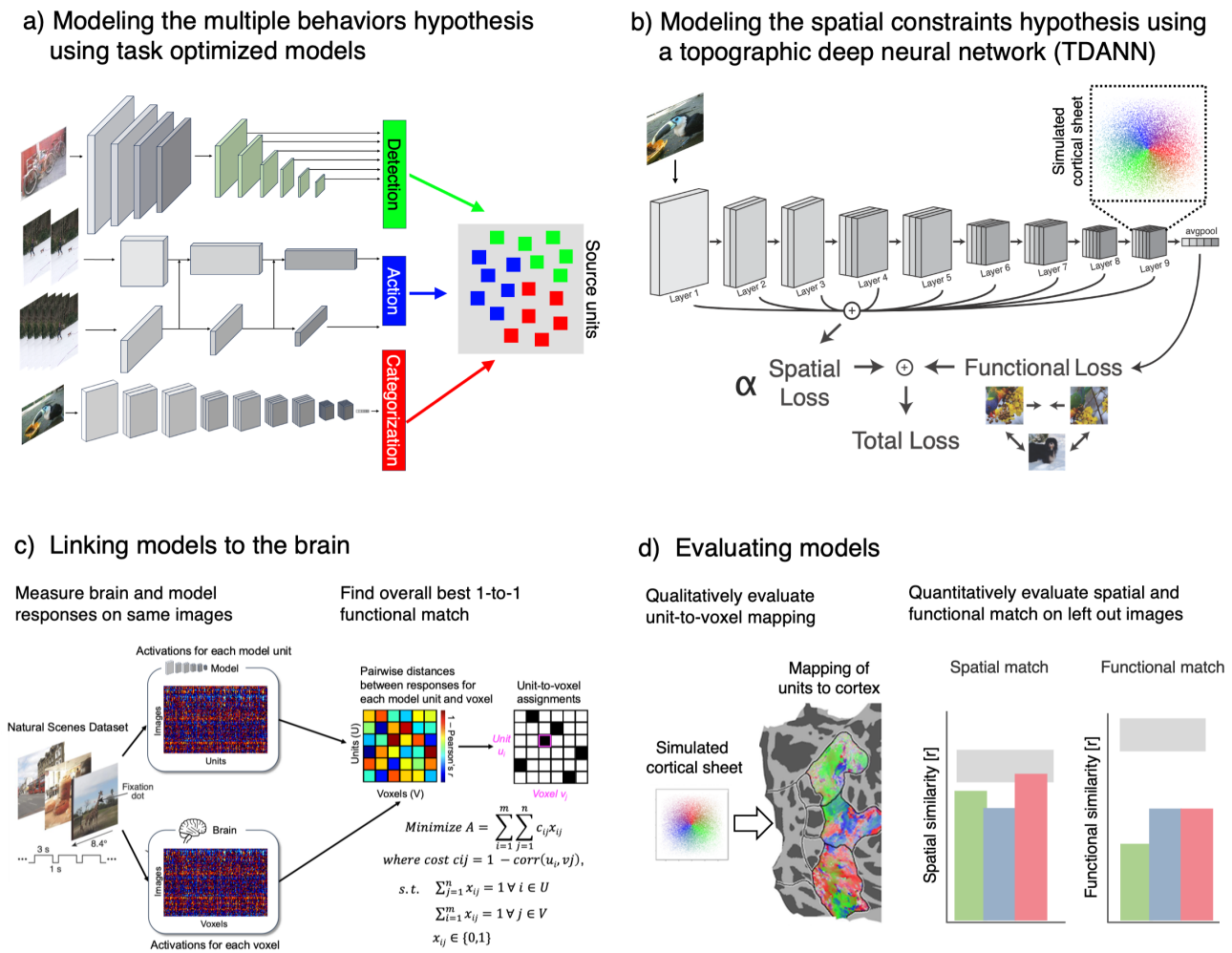

Figure 1. To test competing theories, we use two candidate model classes: Multiple behaviors and Spatial constraints. **(a) Multiple Behaviors (MB):** we implement separate DANNs each trained on a stream-hypothesized task: e.g., detect the object, determine the action, and categorize the object. To compare to the brain, we generate the end region of the MB model by sampling units from the final convolutional layer of each of the three independent DANNs, and pool them into a single combined set of candidate source units on a simulated cortical sheet that will be compared to the union of the endpoint regions of all three streams. **(b) Spatial constraints (SC) models:** we implement Topographic Deep Artificial Neural Networks (TDANNs) [41]. Different than DANNs, during training the TDANN minimizes a composite loss function that has both functional and spatial loss components, with the contribution of the latter controlled by a free parameter α. As for the MB model, to test the spatial constraints model, candidate source units from the final convolutional layer in the TDANN are compared to the union of the endpoint regions of all three streams **(c) 1-to-1 model-to-brain mapping approach.** The goal is to find the overall best functional match between each model unit and each brain voxel. 10,000s of images of natural scenes were presented to each of 8 individuals and candidate models. Responses to identical images are extracted from model units and an individual's voxels, and correlations between each unit-voxel pair are computed. Correlations are transformed into an initial cost matrix $(1 - \text{correlation})$. Using the Hungarian matching algorithm [47] we determine the optimal mapping such that each unit is assigned to a unique voxel and the average cost across all unit-voxel pairings is minimized (black: assignment). **(d) Evaluation.** Models are evaluated by determining spatial correspondence (left) and functional (right) correspondence on 515 left out images seen by all individuals. Spatial and functional correspondences are benchmarked against stream-specific noise ceilings determined by human-to-human mapping of streams using the same approach (light gray).

We instantiate the multiple behaviors hypothesis using three different DANNs, each trained for a stream-hypothesized task. We reason that this will allow optimized task training and encourage learning of maximally different representations suited for each behavior. Specifically, we select tasks that are the computer vision equivalents of the behaviors each stream is thought to support - dorsal: object detection [48], lateral: action recognition [49], ventral: object categorization [50] (Fig. 1a); the union of these three DANNs is our first multiple behaviors candidate model (MB v1). We use a ResNet-50 backbone [50], as larger models achieve better task performance than shallower models with fewer parameters (e.g., ResNet-18), and previous research has shown a positive correlation between a model's task performance and its ability to predict neural responses in the ventral stream [17]. As there is less consensus on the function of the lateral stream [12, 13, 14] and recent research suggests that vision plus language models better predict responses in high-level visual cortex [51], we implemented a second multiple behaviors model (MB v2), in which the candidate model for the lateral stream is trained using contrastive language and image

pretraining (CLIP [52]), as in [51]. Additionally, we evaluated more recent, highly-performant models for the dorsal and ventral streams, using DETR and ConvNeXt for detection and categorization, respectively. Like MB v1, all MB v2 models used a ResNet-50 backbone, or equivalent in the case of ConvNeXt-T. Further, to be able to determine spatial correspondence between models and the brain, we placed model units of all models on a simulated cortical sheet, and their positions are assigned using the procedure established in [41].

We instantiate the spatial constraints hypothesis using a topographic DANN (TDANN, [41, 53]), as it implements a spatial constraint during training and has been effective in predicting the functional topography of the ventral stream [41]. In the TDANN, model units in each layer are assigned a position on a 2D simulated cortical sheet, and a spatial constraint is balanced together with contrastive self-supervised learning (SimCLR [54], Fig. 1b) using separate spatial and task losses during training. The spatial constraint encourages nearby units to have more correlated responses than distant units, and SimCLR encourages two snapshots of the same image (differing in incidental properties such as color or field of view) to have similar representations that are distinct from others [54]. We choose SimCLR because it is one of the best performing self-supervised approaches and because it generates broadly useful representations that are beneficial for a range of visual tasks [55]. We use an intermediate spatial weight, as prior research shows that this combination of training and spatial weight best replicates the functional topography of the ventral stream [41] and a ResNet-18 backbone because it has a better match between the number of model layers and areas in the brain [56]. Prior to training with the combined task and spatial losses, units positions are initialized using the same pre-optimization procedure as [41] (see Methods: Pre-optimization of positions).

As controls, we additionally implement: (i) the same MB v1 training with the stream-hypothesized tasks for each stream, but using a ResNet-18 backbone to match the architecture of the spatial constraints model, (ii) another set of spatial constraints models trained on categorization rather than SimCLR, as a large body of research suggests that DANNs trained on categorization better predict responses in the ventral stream [17, 39, 31], (iii) a ResNet-18 DANN trained with SimCLR (no spatial constraint), and (iv) a ResNet-18 DANN trained on categorization (no spatial constraint). The goal of these controls is to enable disentangling the effects of task and spatial loss components during training on models' ability to predict the functional and spatial organization of the visual system.

Framework for mapping and evaluating the functional and spatial correspondence between models and the brain

Testing the extent to which the models can recapitulate the functional organization of visual cortex into processing streams requires evaluating the functional and spatial match between model units and brain data. This necessitates: (i) measuring brain and model responses on the same stimuli to test functional correspondence, and (ii) a 1-to-1 mapping between model units and brain voxels in order to determine the spatial assignment of model units to streams. To fulfill (i) we leverage the Natural Scenes Dataset (NSD) [57]. The NSD is a massive, high-resolution, and high quality (Supplemental Fig. S1) fMRI dataset that measured responses to thousands (8000-10,000) of natural images in each of eight individuals. Importantly, the NSD generates differential responses across the dorsal, lateral, and ventral visual streams in each of the individuals, verifying that the streams are functionally distinct (Supplemental Fig. S2). To fulfill (ii), we use the Hungarian algorithm [47] to determine the unit-to-voxel pairings that have the highest functional similarity (highest correlation between responses for common set of images). Notably, this 1-to-1 mapping between units and voxels not only provides a window into spatial organization but also provides a more stringent [58, 59] and rigorous [60] testing of DANN models of the brain.

We use a two-stage approach to evaluate functional and spatial correspondence. First, using the Hungarian algorithm and the unique set of images seen by each participant, we find a 1-to-1 model-to-brain mapping between the functional responses of units in the model's convolutional end-layer to these images and the functional responses of the union of voxels in the anatomical regions of interest (ROIs) corresponding to the ends of each stream in that participant to the same images (Fig. 1c). Second, once the mapping is determined, we assess the spatial and functional correspondence between the candidate model and the brain.

We assess spatial correspondence qualitatively by visualizing where each model unit was matched onto its corresponding voxel on the cortical sheet (Fig. 1d-left), and quantitatively by measuring the spatial similarity between the topography of units on the simulated cortical sheet and that of the assigned voxels across cortex (Methods: Evaluating spatial correspondence; Fig. 1d-right). The spatial similarity metric measures the correlation between units' distances on the simulated cortical sheet and distances of their corresponding voxels on the cortical sheet. Spatial similarity is high if nearby units on the simulated cortical sheet are mapped to nearby voxels on the cortical sheet.

We assess functional correspondence quantitatively by evaluating the functional similarity (correlation) between the responses of each model unit and its paired brain voxel on an independent set of 515 left-out NSD images seen by

all eight participants; between 5-9% of the total images viewed, depending on the participant (Fig. 1d-right).

To assess the maximal performance that models can achieve under this framework, we estimate the noise ceiling of the brain data, hypothesizing that the best model for the human brain is another human brain [61]. We determine the noise ceiling by applying the same 1-to-1 mapping algorithm, using 80% of the 515 shared images for matching, and evaluating the spatial and functional correspondence between participants' brains across voxels in each stream using the remaining 20% of images (brain-to-brain noise ceiling; horizontal gray bars throughout). Results show that the majority of voxels in each stream in one participant are mapped to the correct, corresponding stream in other participants (Supplemental Fig. S3), which also validates our novel use of the 1-to-1 mapping algorithm.

Spatial constraints models provide a better functional and spatial match to the stream organization of visual cortex

Using this framework, we compare the spatial and functional correspondence between candidate models and the NSD across subjects, hemispheres, and model seeds. In total, we evaluate 1264 model-to-brain mappings.

We first visualize the topography of unit-to-voxel assignments on the cortical sheet of each participant (Fig. 2a and Fig. S4 for an example participant). Qualitatively, unit-to-voxel assignments for the multiple behaviors (MB) models are noisy, with no clustering by stream (Fig. 2a-left and Fig. S4). In other words, contrary to the prediction of the multiple behaviors hypothesis, units trained on all three tasks are mapped to all three streams and there do not appear to be substantially more units trained on the hypothesized stream behavior mapped to their corresponding stream. For example, categorization-trained units are mapped to all three streams rather than just to the ventral stream as predicted by the multiple behaviors hypothesis. In contrast, the self-supervised spatial constraints model better captures stream topography. Three distinct clusters of units on the simulated cortical sheet map to the three different streams. Further, the model-to-brain mapping is qualitatively more smooth on the cortical sheet than for the multiple behaviors model.

Quantitative analyses reveal that in all streams, MB models produce close to zero spatial similarity to the arrangement in cortex (Fig. 2b gray). In contrast, both the self-supervised trained (SimCLR, $\alpha = 0.25$, Fig. 2b purple) and categorization-trained ($\alpha = 2.5$, Fig. 2b orange) spatial constraints models produce a high spatial similarity to cortex, reaching the human-to-human noise ceiling in all streams (Fig. 2b), with significantly higher spatial correspondence than the MB models (all $ps \leq 2.0 \times 10^{-38}$, Supplemental Table 1, Fig. 2b). Individual DANNs trained with the same tasks (SimCLR and categorization, respectively) but no spatial constraints do not match the stream spatial topography. That is, task-only trained models produce close to zero spatial similarity between unit positions and their corresponding voxels in visual cortex. This suggests that the spatial constraint during training is necessary for capturing the spatial topography of streams in human visual cortex.

While our results so far suggest that the spatial constraints models captures the stream structure better than the multiple behaviors models, it is possible that the multiple behaviors models still provide a better functional match to the brain than the spatial constraints models. However, even functionally, unit-to-voxel correspondences on the left-out images for the multiple behaviors models are poor and significantly lower than the noise ceiling (all $ps \leq 2.5 \times 10^{-5}$, Fig. 2c-gray). For all streams, functional similarity is below $r = 0.16$ for MB v1 (ResNet-50), below $r = 0.15$ for MB v2 (ResNet-50), and at or below $r = 0.17$ for MB v1 (ResNet-18). In contrast, the functional correspondence of model units from the self-supervised spatial constraints model to cortex is significantly higher than the multiple behaviors models for all streams (all $ps \leq 1.6 \times 10^{-190}$, Fig. 2c-purple, Supplemental Table 2). In fact its functional match is more than double that of the best multiple behaviors models across streams (improvement, dorsal: mean 146% increase, lateral: 84%, ventral: 114%). Notably, the functional correspondence approaches the brain-to-brain noise ceiling (Fig. 2c-gray bars) in both the dorsal and ventral streams. Across all streams, the categorization-trained spatial constraints model (Fig. 2c-orange) also achieves significantly better functional correspondence than the multiple behaviors models (all $ps \leq 3.9 \times 10^{-16}$, Supplemental Tables 2) and a better functional match than the task-only categorization trained DANN. However, the functional match of the self-supervised spatial constraints model is significantly higher than the spatial constraints model trained on categorization (pairwise comparisons Bonferroni-corrected, all $ps \leq 3.4 \times 10^{-5}$). This difference between the self-supervised and categorization trained spatial constraints models is notable particularly for the ventral stream, as prior research suggests that categorization trained DANNs predict responses in the primate ventral stream [17, 55]. Thus, across models tested, the self-supervised spatial constraints model provides the best functional and spatial match to the brain (Supplemental Tables 1 and 2). These findings suggest that not only does the spatial constraint in the TDANN change the spatial layout of units, but it also changes their response properties to be more brain-like.

Given the large empirical literature reporting functional differences across streams, we further sought to examine the multiple behaviors models by training task. We considered that our prior analyses, which summarized

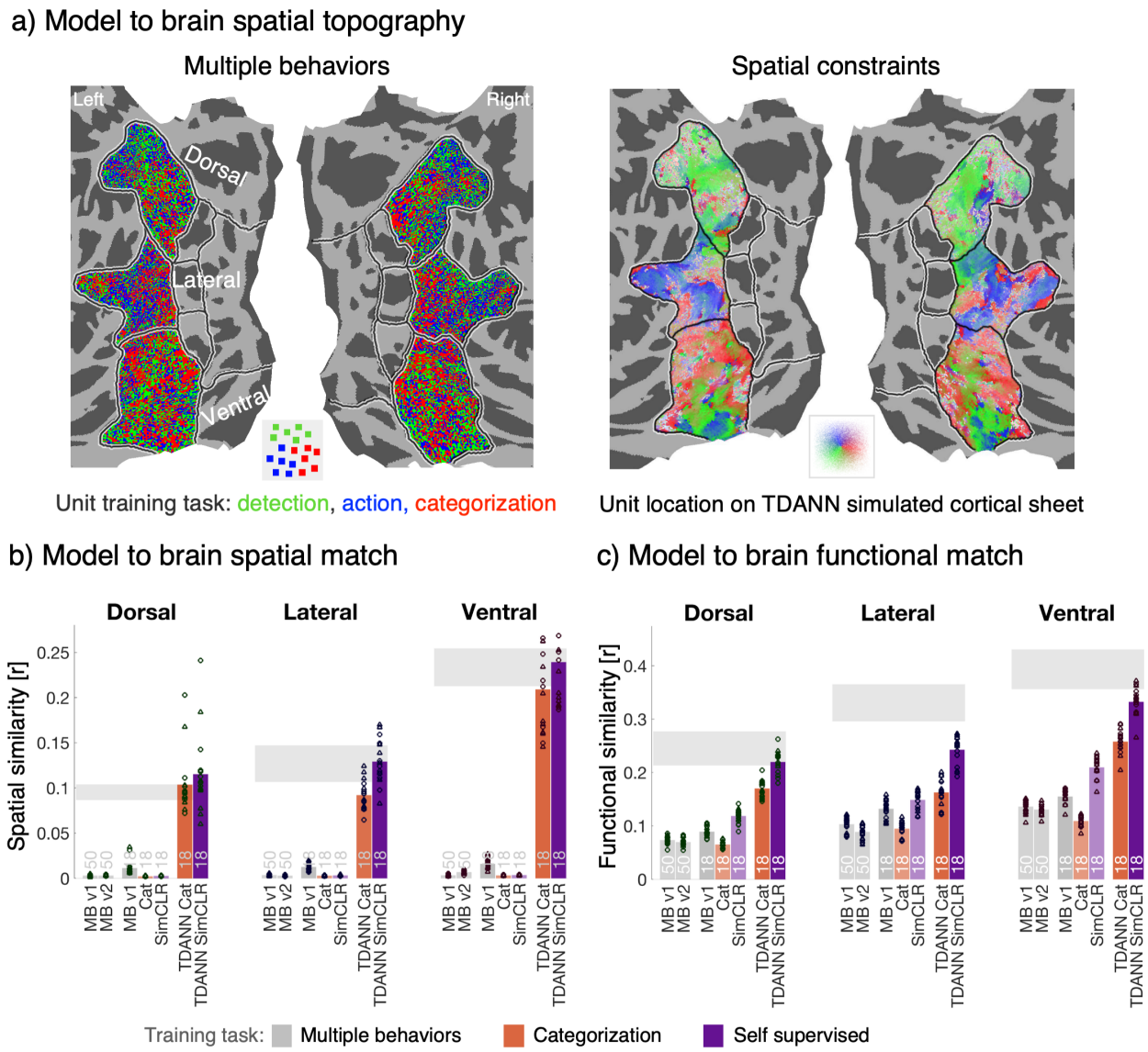

Figure 2. Spatial constraints models better match the spatial and functional organization of the human visual system into streams. (a) Unit-to-voxel mapping on an example participant's flattened cortical surface. *left*: multiple behaviors (MB v1) model, each voxel is colored based on the training task for its assigned unit. *right*: self-supervised spatial constraints model, voxels are colored by location on the simulated cortical sheet of the last convolutional layer. **(b)** Quantification of spatial correspondence. **(c)** Quantification of functional correspondence. Unit-to-voxel correlations on the test set (left-out 515 images). Each participant's data reflects the average correlation across all voxels for the corresponding stream. For (b) and (c), *colored bars*: mean across participants and hemisphere, bars are colored by training task; *numbers*: architecture backbone, 50: ResNet-50, 18: ResNet-18; *horizontal gray bars*: mean noise ceiling across participants and hemispheres ± SD. Supervised (Categorization) and self-supervised (SimCLR) spatial constraints models shown for an optimal weighting of the spatial loss ($\alpha = 2.5$ for supervised, $\alpha = 0.25$ for self-supervised). *Acronyms*: MB: multiple behaviors, v1 trained on detection, action, and categorization; v2: trained on more recently developed models: detection (DETR), language+vision (CLIP), categorization (ConvNeXt-T); TDANN: topographic deep artificial neural network, implementing the spatial constraints hypothesis. Cat: categorization.

correspondences across training tasks, might have shown reduced performance due to units being incorrectly mapped to streams. Thus, we evaluated both the spatial and functional correspondence of the multiple behaviors models by training task. The multiple behaviors hypothesis would predict that: (i) units trained on a stream's hypothesized task would be preferentially mapped to the corresponding stream compared to units trained on other tasks, and (ii) units trained on their stream-hypothesized task and assigned to their corresponding stream would demonstrate better functional correspondence than either units trained on other tasks assigned to that stream or units trained on the same task but assigned to incorrect streams.

However, our results largely do not support either prediction. For the dorsal stream, units trained on the stream-hypothesized task of detection map to dorsal voxels at a rate not much higher than chance level (Fig. 3a; mean±SD, one-sample t-tests against 33%: SSD = $34.8\% \pm 1.5$, n.s.; DETR = $35.8\% \pm 1.4$, $p = .03$; Faster R-CNN = $39.6\% \pm 3.0$, $p = 0.008$). We observe a similar failure pattern for the lateral stream (SlowFast ResNet-50 = $37.0\% \pm 0.8$, $p = 9.2 \times 10^{-5}$; CLIP = $32.5\% \pm 1.5$, n.s.; SlowFast ResNet-18 = $40.5\% \pm 1.5$, $p = 3.3 \times 10^{-5}$). For the ventral stream, units from MB v1 models trained on the stream-hypothesized task of categorization are mapped significantly more than chance to the ventral stream than other streams (ResNet-18: $42.6\% \pm 1.7$, $p = 1.5 \times 10^{-5}$; ResNet-50: $41.0\% \pm 1.4$, $p = 1.6 \times 10^{-5}$). But for MB v2 model with ConvNeXt-T, surprisingly significantly fewer than chance ConvNeXt-T units are mapped to the ventral stream, despite the categorization training ($23.3\% \pm 1.4$, $p = 2.2 \times 10^{-6}$). This divergence may reflect ConvNeXt's attention-inspired design, with larger early receptive fields, LayerNorm-based global context integration, and reduced inductive bias for hierarchical, localized feature encoding, which together may align less closely with the ventral stream's feedforward and category-selective organization (interestingly, we see a similar decrease in the number of units mapped to the ventral when implementing a ViT-B-32 base architecture control for CLIP, Supplemental Fig. S5). Nevertheless, for all MB models, the spatial match remained far below the noise ceiling of 53.8% ($SD = 1.9\%$). Functionally, there was no dissociation of streams by visual behavior: across all streams and training tasks, units trained on their stream-hypothesized tasks do not provide a better functional match than units trained on other tasks. Instead, we find that across all three streams, the categorization-trained ResNet-18 units provide the best functional match, but this functional match is poor (mean $\pm$ SD: dorsal = $0.11 \pm .008$; lateral = $0.18 \pm .009$; ventral = $0.17 \pm .01$) and significantly lower than the noise ceiling (all $ps \leq 6.9 \times 10^{-6}$).

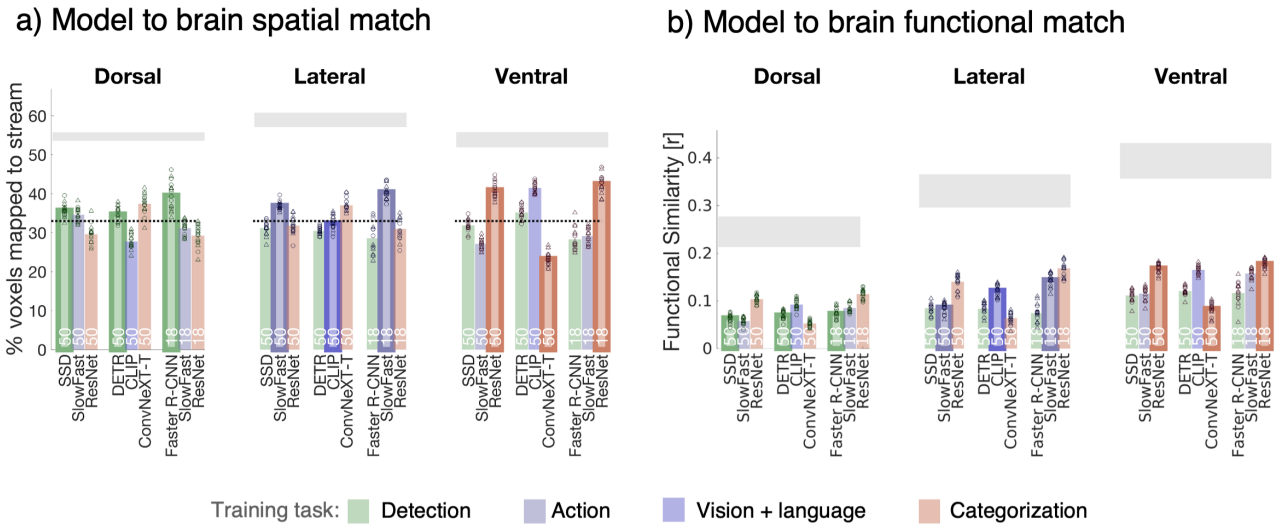

Figure 3. Stream-hypothesized task does not outperform other tasks in predicting brain spatial or functional responses. (a) Multiple behaviors model to brain spatial match: proportion of units assigned to stream by training task. Horizontal dashed line: 33 % chance level. **(b)** Multiple behaviors model to brain functional match: functional similarity of units assigned to stream by training task. For (b) and (c) colored bars: mean across participants and hemispheres; bars are colored by training task. green: detection; blue: action. lavender: vision+language; orange: categorization. Outlined bars indicate the stream-hypothesized training. horizontal gray bars: mean noise ceiling across participants and hemispheres ± SD.

Our analyses so far reveal that the multiple behaviors models do not explain the organization of cortex into streams. We find that neither highly performant models based on ResNet-50, nor models that match the TDANN backbone architecture based on ResNet-18 , recapitulate the visual system's spatial and functional organization into streams. In contrast, the spatial constraints model predicts both the functional and spatial organization of the visual system above and beyond the multiple behaviors models, suggesting a potential explanation of why the visual system is organized into processing streams.

Two key factors for stream organization: self-supervision and a balanced spatial constraint

We next ask what factors contribute to the emergence of streams in the spatial constraints model. The spatial constraints model has two key components that may affect its performance: the training task and the relative strength of the spatial constraint. Thus, we test spatial constraints models (5 seeds each), varying the training task, contrastive self-supervision (SimCLR) or supervised categorization, and the spatial weighting (α), from $\alpha = 0$, where the model is essentially a standard ResNet-18 minimizing only the task loss, to $\alpha = 25$, at which point the effect of task is dwarfed by the spatial constraint during training. Testing these models against brain data allows us to evaluate: (i) whether there is a broad or narrow range of parameters that gives rise to the stream organization, and (ii) whether similar or different parameters produce spatial and functional correspondences between models and the brain.

We first visualized the spatial topography of spatial constraints model units mapped to their corresponding voxels on the cortical sheet as a function of training task and spatial weight (Fig. 4a). Spatial constraints models trained with no spatial weight ($\alpha = 0.0$) generate unit-to-voxel mappings which have no spatial structure for both self-supervised and supervised training tasks. I.e., when there is no spatial constraint during training, no functional topography develops in the spatial constraints model's simulated cortical sheet (Fig. 4a, Supplemental Fig. S6 and S7) and consequently the spatial correspondence is noisy and poor. This suggests that the retinotopic initialization and pre-optimization of units on the simulated cortical sheet is insufficient for the emergence of stream structure.

For spatial constraints models trained with the self-supervised contrastive task, the clearest stream structure is evident when trained with an intermediate spatial weight of $\alpha = 0.25$, with each contiguous third of the simulated cortical sheet largely mapping to a distinct stream (Fig. 4a-top); this structure is also visible using a continuous spatial gradient (Supplemental Fig. S6, see Supplemental Fig. S7 for corresponding visualizations on the model simulated cortical sheet). But when the spatial weight is increased to $\alpha = 2.5$ the stream structure is diminished. In contrast, for spatial constraints models trained on a supervised categorization task, training with a stronger spatial weight ($\alpha = 2.5$) is required for the stream structure to emerge (Fig. 3a-bottom).

Quantitative evaluation of the spatial match to the brain reveals that across all three streams, self-supervised spatial constraints models (purple) provide a better spatial match than supervised (orange) spatial constraints models (Fig. 4b-top, Supplemental Tables 3 to 5). Additionally, spatial correspondence varies significantly with the level of the spatial constraint and there is a significant interaction between the spatial constraint and training task (Supplemental Tables 3 to 5). Across all streams, self-supervised spatial constraints models achieve peak spatial correspondence at $0.25 \leq \alpha \leq 0.5$ (Fig. 4b-top), with performance reaching the human-to-human noise ceiling. In contrast, the categorization trained spatial constraints model produces peak spatial correspondence at $\alpha = 2.5$.

Quantification of the functional similarity between the spatial constraints models and the brain across training tasks and spatial weights mirrors the spatial correspondence, with significantly better functional correspondence for self-supervised than supervised training and a significant task by spatial weight interaction (Fig. 4b-bottom, Supplemental Tables 6 to 8). Across all three streams, spatial constraints models trained with self-supervision and an intermediate spatial weight ($0.25 \leq \alpha \leq 0.5$) generate the best functional match to the brain. The functional match of the best spatial constraints model approaches the noise ceiling in the dorsal and ventral streams. In contrast, across all streams spatial constraints models trained on categorization require a larger spatial weight ($\alpha = 2.5$) to functionally match responses, and even for the ventral stream do not provide a better functional match to the brain than the self-supervised spatial constraints models. Together these analyses reveal that across all three streams, self-supervised spatial constraints models with an intermediate spatial weight ($0.25 \leq \alpha \leq 0.5$) provide the best spatial and functional match to the brain, suggesting that both task and spatial constraints are needed to generate the functional organization into streams.

What might explain this high functional match between self-supervised spatial constraints models with a balanced spatial weight and the brain? One potential factor lies in the dimensionality of model and brain representations. A characteristic of visual cortex as a computational system is that the dimensionality of its representational space, that is, the effective dimensionality of encoded information, strikes a balance between efficiency and robustness [62, 63]. Information encoding is most efficient when stimulus responses are uncorrelated and high-dimensional, but most robust when responses are correlated and low-dimensional. The eigenspectra of visual cortex neural responses to natural scenes decay according to a power law with exponent $n \approx 1$ [62, 64], meaning that responses are differentiable and robust to small input perturbations, but sufficiently high-dimensional to encode complex information. Traditional DANNs, on the other hand, exhibit much higher effective dimensionality than neural responses [64, 65], and this higher effective dimensionality is thought to lead to decreased robustness to noise and vulnerability to adversarial attacks [65].

Thus, we next test if spatial constraints models that are more similar to the brain functionally, are also more similar to the brain in their effective dimensionality. We plot the relationship between functional similarity and effective

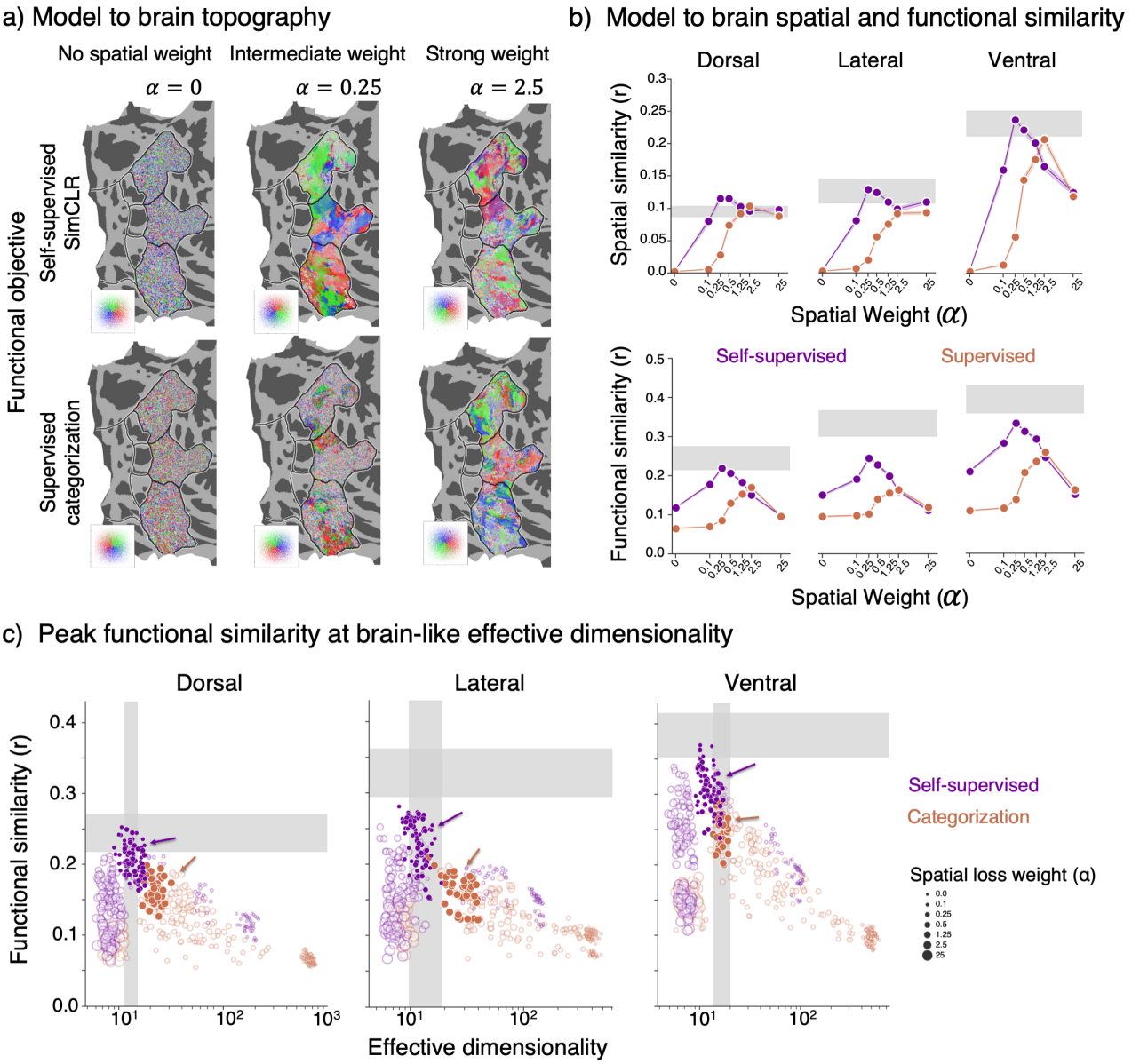

Figure 4. Both self-supervised training and a mid-weight spatial constraint are key to predicting the functional organization into streams. (a) Mapping between spatial constraints models and the brain for an example flattened right hemisphere. Voxels are colored by location on the simulated cortical sheet (see inset). Left: spatial constraints models trained with no spatial weighting ($\alpha = 0$); Middle: spatial constraints models trained with $\alpha = 0.25$, best functional similarity for self-supervised spatial constraints models; Right: spatial constraints models trained with $\alpha = 2.5$, best functional similarity for spatial constraints models trained on categorization. **(b)** Across all three streams, the spatial and functional match to brain is best for self-supervised spatial constraints models trained with $0.25 \leq \alpha \leq 0.5$. *Top row*: Spatial similarity between unit-voxel pairings per stream. Data averaged over unit-voxel pairing for lowest third of distances. *Bottom row*: average functional similarity (correlation) between unit-voxel pairings per stream. Values are averaged across model seeds and hemispheres; *shaded region*: SD across participants. *Purple*: self-supervised spatial constraints model (SimCLR); *Orange*: supervised spatial constraints model (categorization). *Horizontal gray bar*: noise ceiling, mean ± SD. **(c)** Spatial constraints models with the most brain-like functional similarity also have the most brain-like effective dimensionality. *Horizontal shaded gray bar*: brain-to-brain functional similarity, averaged across participants and hemispheres ($\pm$SD across target participants). *Vertical shaded gray bar*: effective dimensionality of brain responses by stream, averaged across participants and hemispheres ($\pm$SD across participants). *Each circle*: one model seed and participant combination, averaged across hemispheres. Circles are colored by training task, and sized by spatial weight (see legend). *Arrows, bolded circles*: models with the highest functional correspondence (b) for each of the spatial constraints training tasks ($0.25 \leq \alpha \leq 0.5$: self-supervised; $\alpha = 2.5$: categorization).

dimensionality of all spatial constraints models tested and compare to the brain. Fig. 4c-gray bars show the functional similarity (horizontal bar) vs. the effective dimensionality of the human brain (vertical bar) in each stream (mean $\pm$ SD: dorsal = 13.2 ± 1.3; lateral = 14.3 ± 3.4; ventral = 16.7 ± 1.9). We find that self-supervised spatial constraints models (Fig. 4c-purple) have lower effective dimensionality than categorization-trained spatial constraints models (Fig. 4c-orange) even as spatial constraints models trained with larger spatial weighting exhibit lower effective dimensionality in general. Comparing the spatial constraints models to the brain data reveals that across all streams the effective dimensionality of the self-supervised spatial constraints models is most similar to the brain at a spatial weight of $0.25 \leq \alpha \leq 0.5$ (Fig. 4c-purple arrows). In contrast, the categorization-trained spatial constraints models

exceed the effective dimensionality of the lateral and dorsal streams and are only similar in effective dimensionality to the human ventral stream (Fig. 4c-orange arrows). Strikingly, in all streams, the self-supervised spatial constraints models that have the most brain-like effective dimensionality also have the most brain-like functional correspondence.

Observed functional differentiation between streams emerges organically from spatial constraints

Our findings suggest that visual processing streams can emerge in a network that learns via a single, self-supervised task, under a spatial constraint to minimize wiring. Nonetheless, empirical findings reveal multiple functional differences across visual processing streams [3, 44, 14]. Is the emergence of streams from a single training task at odds with functional differentiation across streams?

To gain initial insights into this question, we test the extent to which spatial constraints model units assigned to different streams may develop stream-relevant functional properties, focusing on key functional characteristics that have been documented to differ across streams: task specialization and receptive field properties. For task, we examined the functional differentiation of the ventral and dorsal stream, as a large body of work suggests that the ventral stream is involved in visual categorization [17, 15, 16] while the dorsal stream is more important in determining object location [3]. For receptive fields, recent evidence suggest that population receptive fields properties vary across streams. For example, population receptive fields (pRFs) in face- and body part-selective regions in the ventral stream are more central than those of face- and body part-selective regions in the lateral stream [67, 68].

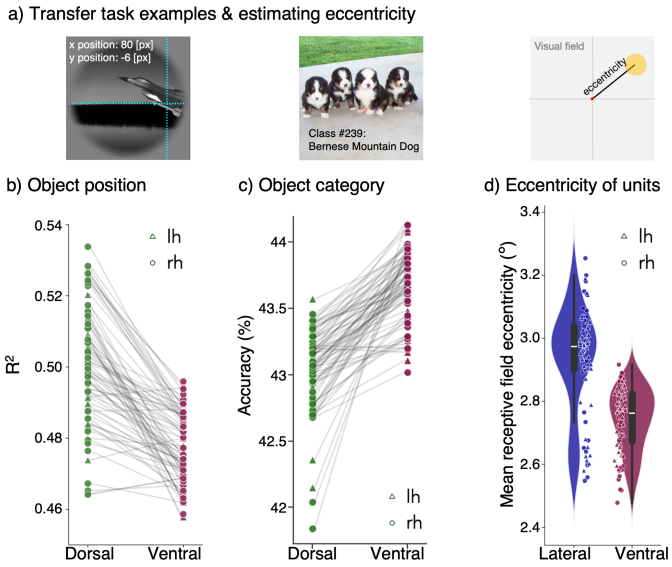

Figure 5. Functional segregation, in alignment with known stream properties, emerges from the spatial constraints model. Data shown is for self-supervised spatial constraints models trained with $0.25 \leq \alpha \leq 0.5$. Each point: a model seed, subject, and hemisphere combination. Triangles: left hemisphere; circles: right. **(a)** Examples of the two transfer tasks: object position prediction (left) and object categorization (middle). Schematic illustrating the principle of eccentricity (right). Red dot = center of the visual field. Yellow circle = receptive field. **(b)** Task transfer performance (R^2) on the object position task (location in pixels of the object's center [66]) for units assigned to either the dorsal (green) or ventral (red) stream. **(c)** Task transfer performance (percentage correct) on the 1000-way object categorization task on the Imagenet validation set for units mapped to the dorsal (green) or ventral (red) stream; **(d)** Average receptive field eccentricity for face-selective units assigned to the lateral stream (blue) or ventral (red) stream.

Specifically, we test whether spatial constraints model units assigned to the dorsal and ventral streams can determine the object's position [66] and/or the object's category (a 1000-way Imagenet categorization task [69, 54]). While position [66, 67] and category [70, 16, 71, 72] can be decoded from both dorsal and ventral streams, we hypothesize that if functional differentiation emerges in spatial constraints models, then units assigned to the dorsal stream will outperform ventral units in determining object position and units assigned to the ventral stream [15, 70], will out-perform dorsal stream units on object categorization. We find that this is indeed the case: units mapped to the dorsal stream achieve significantly higher performance than units mapped to the ventral stream on determining object position (mean $\pm$ SD: ventral = $47.8\% \pm 0.1\%$; dorsal = $50.0\% \pm 0.4\%$; $t(7) = -12.9$, $p = 4.0 \times 10^{-6}$), Fig. 5a). In contrast, units mapped to the ventral stream achieve significantly higher accuracy than units mapped to the dorsal stream on object categorization (mean $\pm$ SD: ventral = $43.7\% \pm 0.03\%$; dorsal = $43.1\% \pm 0.2\%$; $t(7) = 6.92$, $p = .00023$, Fig. 5b). This functional differentiation extends to other tasks and spatial weights (Supplemental Fig. S8).

Secondly, as recent research reported higher mean pRF eccentricity in face-selective regions in the lateral than the ventral stream [67, 73], we evaluate the mean eccentricity of receptive fields overlapping an $8°$ x $8°$ stimulus of face-selective units assigned to the lateral and ventral streams. Results show a qualitative model-to-brain correspondence (Fig. 5c): lateral face-selective units have significantly higher eccentricity than ventral face-selective units (mean $\pm$ SD: lateral = 2.93 ± 0.03; ventral = 2.74 ± 0.04; $t(15) = -10.2, p = 1.9 \times 10^{-5}$). These analyses show that even though the model was trained with a single task and spatial constraint, units assigned to different streams develop different functional properties.

Discussion

We find that a single, biologically plausible computational principle — self-supervised learning of the statistics of visual inputs under a spatial constraint that encourages nearby units to have correlated responses — can explain both the functional and spatial organization of the human visual system into processing streams. Crucially, both components, the self-supervised learning objective and an intermediate spatial weight, are needed for the stream structure to emerge. Moreover, the spatial constraint that produces the best model-to-brain spatial correspondence also produces the best model-to-brain functional correspondence and best captures the effective dimensionality of representations in the human visual system. These findings suggest that the spatial and functional dimensions of cortical organization are fundamentally intertwined, and they underscore the importance of modeling brain topography alongside function to reveal the principles underlying large-scale brain organization, unlike prevailing DANN approaches that focus solely on function [30, 17, 31, 32, 33, 34, 35, 36, 37, 38]. Taken together, our results suggest that the organization of visual streams may reflect a tradeoff between learning general-purpose representations in a self-supervised manner and conforming to the physical constraints imposed by the size and wiring limitations of the brain.

The success of our spatial constraints model builds upon recent studies of topographic map emergence in the ventral stream [22, 40, 53, 41], which suggest that wiring [22], smoothness [40], and the balancing spatial and functional constraints [41] can produce topographic organization. Our findings also align with work in mice showing that a dual-pathway model with a self-supervised predictive loss better fits visual cortex data than task-optimized models [74]. However, we show that explicit modeling of separate pathways is not necessary to reproduce the segregation of visual streams.

We were surprised to find that models based on the multiple behavioral demands hypothesis produced lower functional correspondence with human visual cortex than the spatial constraints model. This finding was particularly striking as tasks which yielded a higher functional match (e.g., categorization) improved correspondence across all streams, rather than selectively enhancing only the stream they were hypothesized to engage (e.g., the ventral stream [3, 4, 75, 12, 13, 14]). In contrast, units in the spatial constraints model, trained with a single self-supervised contrastive objective, develop distinct functional signatures aligned with the behavioral specializations of each stream. This suggests functional differentiation can emerge naturally from spatial constraints alone, without explicit optimization for separate tasks.

It is important to keep in mind that even though the self-supervised spatial constraints model outperforms multiple behaviors models, there are still some gaps between its performance and the noise ceiling, particularly in the lateral stream. Future research testing additional stimuli and tasks during training (e.g., training with videos to predict the next frame), as well as additional biological features and spatial constraints (e.g., implementing long-range connections between the ventral and dorsal streams [76, 77]) can examine what additional factors are necessary for filling in the gaps in model-to-brain spatial and functional correspondences.

Our study provides two key innovations. First, in order to determine the spatial correspondence between models units and the brain, we developed a 1-to-1 mapping algorithm that finds the best overall match between units and brain voxels in terms of functional correlation. Second, we conducted the first rigorous computational test of factors contributing to the stream organization of the visual system, evaluated against human brain data.

The mapping innovation is tied to a ongoing theoretical debate in NeuroAI as to how to evaluate if a model accurately explains the brain [61, 58, 78, 37, 79]. Recent theoretical work [58, 59] shows that commonly used metrics to evaluate models of the brain, such as linear mappings between model and brain responses, seem to be both too permissive [58, 59] and fail to distinguish between models of the brain [80, 81, 39, 59] (Supplemental Fig. S9). Specifically, Khosla and Williams (2024) show that linear mappings can obscure meaningful differences between model units and neural responses because they allow rotation, scaling, and shearing, a combination of transformations that can align representations that are functionally dissimilar. In contrast, our more stringent 1-to-1 mapping between model units and brain voxels is not only necessary to directly test spatial correspondences between models and the brain, but also distinguishes between models of the brain. Our findings underscore the necessity of using stricter metrics not only to adjudicate between putative models of the brain, but also to glean new understanding of biological systems.

Our ability to investigate the factors underlying stream organization is made possible by the recent availability of a massive fMRI dataset of brain responses to thousands of images spanning the entire visual system in multiple participants [57]. This dataset enabled rigorous quantitative comparisons of models of the human visual system with respect to an empirically derived human-to-human noise ceiling. A central theoretical insight of the present study is that a local spatial constraint that encourages nearby units to have correlated responses not only supports

faster processing [1] and reduced wiring costs [82], but also can percolate up to create broad-scale stream structure. As parallel processing streams exist in other species [75, 74], cortical systems [83, 84, 85, 86], and spatial scales [87, 88], future research can test if the same principles - balancing learning of generally useful representations while encouraging nearby units to have correlated responses – extend to other brain systems. This hypothesis can be evaluated by training a topographic deep neural network, which positions model units on a simulated cortical sheet and optimizes during training both task and spatial objectives, on diverse sensory and cognitive inputs to test if there are modality-specific or modality-general principles that lead to the emergence of parallel processing streams across the brain.

References

1. Semir Zeki and Stewart Shipp. The functional logic of cortical connections. *Nature*, 335(6188):311–317, 1988.
2. David C Van Essen, Charles H Anderson, and Daniel J Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992.
3. Leslie G Ungerleider and Mortimer Mishkin. Two cortical visual systems. analysis of visual behavior. *Analysis of Visual Behavior*, pages 549–586, 1982.
4. Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.
5. Brian A Wandell, Serge O Dumoulin, and Alyssa A Brewer. Visual field maps in human cortex. *Neuron*, 56(2): 366–383, 2007.
6. Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27(1):649–677, 2004.
7. David Marr. Vision: A computational investigation into the human representation and processing of visual information, 1982.
8. Robert A Jacobs and Michael I Jordan. Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience*, 4(4):323–336, 1992.
9. Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.
10. Maximilian Riesenhuber and Tomaso Poggio. Models of object recognition. *Nature neuroscience*, 3(11): 1199–1204, 2000.
11. James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8): 333–341, 2007.
12. Kevin S Weiner and Kalanit Grill-Spector. Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychological Research*, 77(1):74–97, 2013.
13. Moritz Wurm and Alfonso Caramazza. Action and object representation in the ventral" what" stream. *PsyArXiv*, 2021.
14. David Pitcher and Leslie G Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 2020.
15. James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 (5539):2425–2430, 2001.
16. Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.
17. 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.
18. Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152, 2017.
19. Mark E Nelson and James M Bower. Brain maps and parallel computers. *Trends in neurosciences*, 13(10): 403–408, 1990.
20. David C van Essen. A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature*, 385(6614):313–318, 1997.
21. Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
22. Nicholas M Blauch, Marlene Behrmann, and David C Plaut. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119(3):e2112566119, 2022.

23. Jascha Achterberg, Danyal Akarca, DJ Strouse, John Duncan, and Duncan E Astle. Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nature Machine Intelligence*, 5(12):1369–1381, 2023.

24. David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

25. Brian A Wandell and Jonathan Winawer. Imaging retinotopic maps in the human brain. *Vision research*, 51(7): 718–737, 2011.

26. Michael J Arcaro and Margaret S Livingstone. A hierarchical, retinotopic proto-organization of the primate visual system at birth. *Elife*, 6:e26196, 2017.

27. Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

28. Nancy Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170, 2010.

29. Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

30. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

31. 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.

32. Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

33. Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

34. Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

35. Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007, 2020.

36. Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

37. Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20, 2023.

38. Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, 2023.

39. Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022.

40. Fenil R Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

41. Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14): 2435–2451, 2024.

42. Zejin Lu, Adrien Doerig, Victoria Bosch, Bas Krahmer, Daniel Kaiser, Radoslaw M Cichy, and Tim C Kietzmann. End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions. *Nature Human Behavior*, 2025.

43. Leslie G Ungerleider and James V Haxby. 'what'and 'where'in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165, 1994.

44. Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, 1983.

45. Patrick Mineault, Shahab Bakhtiari, Blake Richards, and Christopher Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, 34, 2021.

46. H Steven Scholte, Max M Losch, Kandan Ramakrishnan, Edward HF de Haan, and Sander M Bohte. Visual pathways from the perspective of cost functions and multi-task deep neural networks. *cortex*, 98:249–261, 2018.

47. James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

48. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

49. Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

50. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

51. 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*, pages 1–12, 2023.

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

53. Hyodong Lee, Eshed Margalit, Kamila M Jozwik, Michael A Cohen, Nancy Kanwisher, Daniel LK Yamins, and James J DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020.

54. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 2020.

55. Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

56. Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

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

58. Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

59. Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 326–341. PMLR, 2024.

60. Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, 2021.

61. Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

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

63. Larry F Abbott, Kanaka Rajan, and Haim Sompolinsky. Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. *The dynamic brain: an exploration of neuronal variability and its functional significance*, pages 1–16, 2011.

64. Nathan CL Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022.

65. Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLoS computational biology*, 20(1):e1011792, 2024.

66. Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.

67. Dawn Finzi, Jesse Gomez, Marisa Nordt, Alex A Rezai, Sonia Poltoratski, and Kalanit Grill-Spector. Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nature communications*, 12(1):2278, 2021.

68. Jewelia K Yao, Justin Choo, Dawn Finzi, and Kalanit Grill-Spector. Visuospatial computations vary by category and stream and continue to develop in adolescence. *bioRxiv*, pages 2025–01, 2025.

69. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

70. Maryam Vaziri-Pashkam and Yaoda Xu. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, 37(36):8767–8782, 2017.

71. Rory Sayres and Kalanit Grill-Spector. Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of neurophysiology*, 100(1):249–267, 2008.

72. Amanda K Robinson, Genevieve L Quek, and Thomas A Carlson. Visual representations: Insights from neural decoding. *Annual Review of Vision Science*, 9(1):313–335, 2023.

73. Magdalena W Sliwinska, Caitlin Bearpark, Julia Corkhill, Aimee McPhillips, and David Pitcher. Dissociable pathways for moving and static face perception begin in early visual cortex: Evidence from an acquired prosopagnosic. *Cortex*, 130:327–339, 2020.

74. Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34:25164–25178, 2021.

75. David C Van Essen and Jack L Gallant. Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1–10, 1994.

76. Jason D Yeatman, Kevin S Weiner, Franco Pestilli, Ariel Rokem, Aviv Mezer, and Brian A Wandell. The vertical occipital fasciculus: a century of controversy resolved by in vivo measurements. *Proceedings of the National Academy of Sciences*, 111(48):E5214–E5223, 2014.

77. Kevin S Weiner, Jason D Yeatman, and Brian A Wandell. The posterior arcuate fasciculus and the vertical occipital fasciculus. *Cortex; a journal devoted to the study of the nervous system and behavior*, 97:274, 2017.

78. Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Is it that simple? linear mapping models in cognitive neuroscience. *bioRxiv*, page 438248, 2021.

79. Jenelle Feather, Meenakshi Khosla, N Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.

80. Dawn Finzi, Daniel LK Yamins, Kendrick Kay, and Kalanit Grill-Spector. Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 Conference on Cognitive Computational Neuroscience*, 2022.

81. Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 2021.

82. Dmitri B Chklovskii, Thomas Schikorski, and Charles F Stevens. Wiring optimization in cortical circuits. *Neuron*, 34(3):341–347, 2002.

83. Josef P Rauschecker and Biao Tian. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22):11800–11806, 2000.

84. Takahiro Osada, Akitoshi Ogawa, Akimitsu Suda, Koji Nakajima, Masaki Tanaka, Satoshi Oka, Koji Kamagata, Shigeki Aoki, Yasushi Oshima, Sakae Tanaka, et al. Parallel cognitive processing streams in human prefrontal cortex: Parsing areal-level brain network for response inhibition. *Cell Reports*, 36(12), 2021.

85. Judith Tanné-Gariépy, Eric M Rouiller, and Driss Boussaoud. Parietal inputs to dorsal versus ventral premotor areas in the macaque monkey: evidence for largely segregated visuomotor pathways. *Experimental Brain Research*, 145(1):91–103, 2002.

86. Simone Vossel, Joy J Geng, and Gereon R Fink. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, 20(2):150–159, 2014.

87. BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

88. Richard F Betzel and Danielle S Bassett. Multi-scale brain networks. *Neuroimage*, 160:73–83, 2017.

Materials and methods

Code and data availability

Original code for this study is available at https://github.com/dawnfinzi/spacestream. Code to use and train TDANNs can found at https://github.com/neuroailab/TDANN. The neural data analyzed in this study comes from the Natural Scenes Dataset (NSD) [1] available at http://naturalscenesdataset.org/. Processed data relevant to this study can be found at https://osf.io/qy32x.

Training: Neural network architectures and training tasks

Multiple behaviors models. To test the *multiple behavioral demands hypothesis*, we instantiate model versions of the hypothesis. Each 'multiple behaviors' model is comprised of three sub-models, each trained on a task corresponding to a hypothesized functional goal of a stream. We start by testing the classical view of processing streams using three sub-models, each trained on one of object categorization, action recognition, and object detection (MB v1). We chose these tasks as these are the computer vision equivalents of the proposed behaviors each stream is thought to support (ventral: what is it, lateral: what is it doing, dorsal: where is it). For the object categorization model, we used a ResNet-50 [2] trained on object categorization on ILSVRC-2012 (ImageNet Large-Scale Visual Recognition Challenge [3]). For the action recognition model, we used the SlowFast model architecture [4], which is a dual-pathway network with a 3D ResNet backbone trained on the Kinetics-400 video dataset [5]. Finally, for the object detection model, we used a Faster R-CNN [6] trained on MS-COCO [7]. All three sub-models use ResNet-50 as their base architecture.

Secondly, we test an alternative (MB v2) using more recent candidate models and replacing action recognition as the hypothesized lateral stream task with fusing language and visual representations, given the lateral stream's proximity to key language areas such as Wernicke's Area and recent work showing Contrastive Language-Image Pretraining (CLIP) models better predict responses in the high-level visual cortex [8]. In this version, for the ventral stream and object categorization, we use the ConvNeXT model [9], again trained on ImageNet, as this next-generation model shows improved categorization performance, as well as being one of the top-performing models on Brain-Score [10]. As mentioned, for the lateral stream, we test the OpenAI trained CLIP [11], where an image encoder and a text encoder are jointly trained to predict the correct pairings of a batch of (image, text) training examples. Finally, for object detection, we use the DETR model [12] trained on MS-COCO. DETR is comprised of a CNN vision backbone and a transformer encoder/decoder architecture. We chose DETR based on the improved task performance [12] and its ubiquitous use in the field. As in the previous version of the multiple behaviors model, we use ResNet-50 or equivalent as a base architecture for all sub-models. For ConvNeXT, this corresponds to the ConvNeXT-T architecture, which is explicitly based on ResNet-50 [9].

Finally, we implement an additional control using ResNet-18 as the base architecture to more closely match the TDANN architecture, and the number of visual areas in primate cortex [13, 14]. As MB v1 slightly outperformed MB v2 in functional correspondence, this is the version we chose to re-implement using ResNet-18 as a base. For the object categorization and action recognition sub-models, the only change was from ResNet-50 to ResNet-18. However, for the ResNet-18 object detection control model, we used a single-stage object detection network, SSD [15], for greater correspondence with the other multiple behaviors models, and trained on Pascal VOC (2007 and 2012) [16] instead of MS-COCO [7]. While MS-COCO is considered state-of-the-art for object detection and Pascal VOC is an older, smaller dataset, switching to Pascal VOC avoided any potential confound with the images used for the neuroimaging data collection and removed the possibility that the low spatial correspondence for the Resnet-50 based multiple behaviors models was in part due to an artificially inflated match between the object detection model and non-dorsal voxels with high SNR.

We randomly subsampled an equal number of units from layer 4.1 or its equivalent from each network (ignoring any "visually non-responsive" units that did not respond to any of the images) so that the total number of units was equal to the total number of voxels. In order to allow for the most direct comparison against spatial constraints models, the units sampled from the task-trained models were assigned random positions on a two-dimensional simulated cortical sheet. We then followed the same pre-optimization procedure as in Initialization of model unit position: *Stage 2* in order to be able to fairly apply the same mapping algorithm.

Spatial constraints models. We implemented the *spatial constraints hypothesis* using a Topographic Deep Artificial Neural Network (TDANN) architecture. The TDANN is based on the ResNet-18 model architecture, with two key differences: (1) model units are assigned positions on a 2D simulated cortical sheet and (2) the model is trained to jointly minimize a spatial and a task loss. All TDANN models were built using the ResNet-18 [2] base architecture (from the *torchvision* implementation) and trained building on the VISSL framework [17]. ResNet-18 was chosen

because it has been shown to achieve strong task performance, accurately predict neuronal responses across the visual system [10], and has roughly comparable number of layers to stages (areas) in the primate visual system [18, 14]. Models were trained for 200 epochs using the ILSVRC-2012 [3] training set, with each model being trained from five different random initial seeds. We optimized the network parameters using stochastic gradient descent with momentum (with γ set to 0.9), a batch size of 512, and a learning rate initialized to 0.6, which then decayed according to a cosine learning schedule [19] Models were trained using a self-supervised contrastive objective "SimCLR" [20] or a supervised 1000-way object categorization task.

Initialization of model unit position Prior to training with spatial and task losses, model units in each layer were assigned fixed positions in a two-dimensional simulated cortical sheet specific to that layer. The size of the cortical sheet in each layer, and the size of the "cortical zone" used during training (computation of the spatial loss is restricted to units within the same cortical zone), was determined by the presumed correspondence, based on previous work comparing convolutional neural networks (CNNs) and the primate visual system [21, 22, 23], between model layers and human visual areas (see [13] for further details). Positions were then assigned in a two-stage process.

Stage 1: Retinoptopic initialization As each layer convolves over the outputs of the previous layer, the resulting responses are structured into spatial grids. To maintain this inherent organization, we assigned each model unit to a specific area of the simulated cortical sheet that aligns with its spatial receptive field.

Stage 2: Pre-optimization of positions In CNNs, filter weights are shared between units at different locations, which means that local updates to one unit affect all units with the same filter weights. This global coordination constraint makes it challenging to achieve local smoothness when the units are arbitrarily positioned. To address this, a pre-optimization of unit positions was necessary to identify a set of positions that enables learning smooth cortical maps. We spatially shuffled the units of a pre-trained model on the cortical sheet, so that nearby units had correlated responses to a set of sine grating images. The use of sine gratings is based on studies that show that propagating retinal waves drive development of the visual system in the womb in primates and other mammals [24, 25, 26, 27]

The spatial shuffling works as follows: 1) Randomly select a cortical zone. 2) Compute pairwise response correlations for all units in the zone 3) Select a random pair of units, and swap their locations in the cortical sheet. 4) If swapping positions decreases local correlations (measured as an increase in the Spatial Loss function described below), undo the swap. 5) Repeat steps 3-4 500 times. 6) Repeat steps 1-5 10,000 times.

Loss functions We trained the TDANN models using a weighted sum of two types of loss functions: a task loss, which served to encourage learning of visual representations and a spatial loss, which encourages local correlations in responses to visual inputs. Optimization on the total loss function leads to both successful visual representation learning and minimization of inter-layer wiring length [13].

Spatial loss The spatial loss function encourages nearby model units on the simulated cortical sheet to be correlated in their responses to the training stimuli. Specifically, SL_l is the spatial correlation loss computed for the l-th layer and SL_l is computed on a given batch by randomly sampling a local cortical zone and calculating for pairs of units, (1) correlation (Pearson's r) between the response profiles, ($\overrightarrow{R}$), and (2) the stabilized reciprocal Euclidean distances ($\overrightarrow{D}$):

$$\overrightarrow{D} = \frac{1}{(1+\overrightarrow{d})} \tag{S1}$$

where $\overrightarrow{d}$ is the vector of pairwise cortical distances. These two terms are then related as follows:

$$\mathrm{SL}_l = 1 - \mathrm{Corr}\left(\overrightarrow{R}, \overrightarrow{D}\right) \tag{S2}$$

such that SL_l is minimized when nearby units have correlated responses to the training stimuli.

Task loss The task loss (TL) is computed from the output of the final model layer. We tested two candidate TLs: supervised object categorization cross-entropy loss [28] and the self-supervised SimCLR objective [20]. The SimCLR objective is a contrastive loss function which works by creating two "views", or augmentations, of each image in a batch, using random cropping, horizontal flips, color distortion, and Gaussian blur. These views are passed to the network and the final layer outputs are passed through a 2-layer multi-layer perceptron (MLP), producing a low-dimensional representation of each view which serves as the input to the loss function. The SimCLR loss function then attempts to maximize the similarity of representations for two views of the same source image, while pushing that representation away from all other images in the batch.

Overview of training In sum, the model is trained on Imagenet [3] to minimize this total loss, which is the sum of the weighted spatial loss for each layer and the task loss as follows:

$$\text{Total Loss} = \text{TL} + \alpha \sum_{l \in \text{layers}} \text{SL}_l \tag{S3}$$

where α is the weight of the spatial loss component (fixed across all layers), and SL_l is the spatial correlation loss computed for the l-th layer.

The total model training process consists of 6 steps:

1. The model is trained using the task loss only.
2. Positions are initialized to preserve coarse retinotopy in each layer (Stage 1).
3. Positions are pre-optimized in an iterative process that preserves retinotopy while bringing together units with correlated responses to sine gratings (Stage 2).
4. After pre-optimization, positions are permanently frozen.
5. All network weights are randomly re-initialized.
6. The network is trained to minimize the total loss.

Testing: Evaluating theories

Brain data. As our neural comparison, we used the Natural Scenes Dataset (NSD) [1], a high-resolution fMRI dataset that densely sampled responses to up to 10,000 natural images in each of eight individuals over the course of 32-40 scan sessions. Full details on data collection and processing can be found in Allen et al. [1]. Briefly, scanning was conducted at 7T using gradient-echo EPI at 1.8-mm isotropic resolution with whole-brain coverage. Images were taken from Microsoft's COCO image database [7], square cropped to 425 pixels x 425 pixels, and presented at fixation at a size of 8.4° x 8.4° for 3 seconds with 1 second gaps in between images. Data were preprocessed using one temporal resampling (to correct for slice timing differences) and one spatial resampling (to correct for head motion, EPI distortions and gradient non-linearities), resulting in upsampled 1.0mm resolution (temporal resolution 1.0 s). Single-trial beta weights were estimated using a general linear model approach designed to optimize the quality of single trial betas (GLMsingle [29]).

Specifically, we used the single-trial responses from the NSD public data release (beta version b3). We additionally z-scored the betas across images for each voxel and session, averaged across 3 trial repeats, and used cortex-based alignment to align all data to the fsaverage surface. Throughout reporting of the results, we refer to the brain units of measurement as "voxels" for interpretability, though they are more technically "vertices" as we are using the fsaverage preparation.

ROIs We defined regions of interest (ROIs) for early, intermediate, and high-level visual cortex for each of the three streams based on a combination of anatomical landmarks, noise ceiling estimates (Supplemental Fig. S1), and a constraint to roughly match the number of voxels per stream. We focus only on the high-level visual ROIs for the purposes of this paper and compare the end point of the models to the end points of the processing stream in each brain. However, the full details for drawing all seven (one early, three intermediate and three higher-level) ROIs are included here for completeness and because the high-level ROIs share boundaries with the intermediate ROIs. ROIs were drawn on the fsaverage surface as follows:

Early visual cortex ROI: The early visual cortex ROI was drawn as the union of the V1v, V1d, V2v, V2d, V3v and V3d ROIs from the Wang retinotopic atlas [30]. Additionally, V2v and V2d, as well as V3v and V3d, were connected such that the part of the occipital pole typically containing foveal representations was included in the ROIs.

Intermediate ROIs: Three intermediate ROIs were drawn corresponding to each of the three streams: ventral, lateral and dorsal. All three ROIs border the early visual cortex ROI on the posterior side. The intermediate ventral ROI was drawn to reflect the inferior boundary of hV4 from the Wang atlas [30] and includes the inferior occipital gyrus (IOG), with the anterior border of the ROI drawn based on the anterior edge of the inferior occipital sulcus (IOS). The intermediate lateral ROI was drawn directly superior to the intermediate ventral ROI, with the superior and anterior borders determined as the LO1 and LO2 boundaries from the Wang atlas [30]. The intermediate dorsal ROI was drawn directly superior to that to include V3A and V3B from the Wang atlas.

Higher-level ROIs: Three higher-level ROIs were drawn for each of the ventral, lateral and dorsal streams, bordering their respective intermediate ROIs on their posterior edges. The ventral ROI was drawn to follow the anterior lingual

sulcus (ALS), including the anterior lingual gyrus (ALG) on its inferior border and to follow the inferior lip of the inferior temporal sulcus (ITS) on its superior border. The anterior border was drawn based on the midpoint of the occipital temporal sulcus (OTS). The lateral ROI was drawn such that the higher-level ventral ROI was its inferior border and the superior lip of the superior temporal sulcus (STS) was used to mark the anterior/superior boundary. The rest of the superior boundary traced the edge of angular gyrus, up to the tip of the posterior STS (pSTS). The dorsal ROI was drawn to reflect the boundary of the lateral ROI on its inferior edge and to otherwise trace the borders of and include the union of IPS0, IPS1, IPS2, IPS3, IPS4, IPS5 and SPL1 from the Wang retinotopic atlas.

These ROIs, defined by the authors, correspond to the "streams" ROIs included in the NSD public data release. The three higher-level ROI were then trimmed using the prepared noise ceiling maps for beta version b3 and the fsaverage surface [1]. The noise ceiling estimates represent the amount of variance contributed by the signal expressed as a percentage of the total amount of variance in the data, for the average of responses across three trial presentations. An approximate cutoff of 10% was used to guide trimming of the higher-level ROIs, such that we were left with reduced ROIs where all voxels had a noise ceiling $\geq 10\%$ theoretically predictable variance. These ROIs were contiguous and roughly matched in size (number of voxels per ROI right hemisphere: dorsal = 6688, lateral = 6839, ventral = 5638; left hemisphere: dorsal = 6182, lateral = 5849, ventral = 6126).

Model data. For each of our DANN models, we extracted features in response to the same set of 73,000 total NSD images seen by participants in the scanner. Features were extracted from convolutional layer 4.1 or equivalent. This layer was chosen as past work shows that it has the best functional correspondence to higher-level visual areas, including on the NSD [14, 13, 31]. For the SlowFast network, features were averaged over the temporal dimension. This results in a matrix of features of the form *number of images* x *number of units*, where the number of units is 7 x 7 x 512 = 25,088, i.e. the total number of units in layer 4.1, for the TDANN models, and 19,164 total subsampled units for the multiple behaviors models (6,388 units per model), which matches the maximum number of voxels per hemisphere. Thus, for all models, the number of source units was of the same order of magnitude as the number of target voxels [32, 33].

1-to-1 mapping between model units and brain voxels. To evaluate the model correspondence to the brain we use a 1-to-1 mapping algorithm mapping between model units and brain voxels. We reason that units in a neural network model abstract neural computations and thus may be a good model for the aggregated neural response of a voxel (i.e., in the Goldilocks zone of computational abstraction [34]). To test this reasoning and validate the approach we first developed the algorithm between brain-to-brain, for two reasons. First, in the case of brain-to-brain mapping we are mapping voxels in one brain to voxels in another brain so this tests by definition a mapping between corresponding structures. Second, if the stream organization is not recovered in the 1-to-1 mapping applied between brain-to-brain then we would not be able to test the experimental hypotheses between model and brain. Once satisfied with the mapping algorithm in the brain-to-brain case, we used that algorithm with as few modifications as possible to map model-to-brain.

For each pair of subjects (source subject mapped to target subject), we used the single-trial z-scored betas for each source voxel and each target voxel in response to the 515 images shared across all subjects (80% used for assignment, 20% used for evaluation) and computed the correlations between each pair of voxels. This correlation matrix was then transformed into a cost matrix (1 – correlation) and assignment was done on the basis of this cost matrix using the Hungarian algorithm [35], a combinatorial optimization algorithm which solves the assignment problem in polynomial time. This 1-to-1 mapping algorithm, which was purely based on functional correlation, performed significantly above chance in assigning voxels to the correct streams (chance = 33%. Mean $\pm$ SE, 1-sample t-test. Left hemisphere: dorsal = 0.519 ± 0.006, $t(7) = 31.4, p = 8.6 \times 10^{-9}$); lateral = 0.559 ± 0.007, $t(7) = 28.6, p = 1.6 \times 10^{-8}$; ventral = 0.528 ± 0.006, $t(7) = 29.5, p = 1.3 \times 10^{-8}$. Right hemisphere: dorsal = 0.575 ± 0.004, $t(7) = 63.5, p = 6.3 \times 10^{-11}$); lateral = 0.620 ± 0.008, $t(7) = 35.9, p = 3.3 \times 10^{-9}$; ventral = 0.549 ± 0.008, $t(7) = 27.0, p = 2.5 \times 10^{-8}$). Consequently, we apply this 1-to-1 mapping in the unit-to-voxel case with only one change. In the unit-to-voxel case, we were no longer restricted to the shared set of 515 images as we could extract responses from the models to an arbitrary number of images. As such, we leveraged the full set of unique images (up to 9485 per individual) to link models to individual brains. The 515 shared images were then reserved for evaluating theories.

Total models tested. We evaluated 5 instances initialized with different random seeds per each of the spatial constraints models across 2 training tasks (SimCLR and categorization) and 7 levels of spatial weightings (α), as well as 3 versions of the multiple behaviors models, each version comprised of 3 candidate sub-models. Each model was then mapped to the 2 hemispheres for each of the 8 participants, totaling 1264 model-to-brain mappings tested.

Evaluating spatial correspondence. We evaluated the spatial correspondence in two ways. First, as our main spatial metric to compare the multiple behaviors models and the spatial constraints models directly, we calculated a spatial similarity metric. The spatial similarity metric measures for each unit-to-voxel pairing, the normalized distances on the model cortical sheet (from that unit to other units) and normalized distances on the actual cortical surface (for the assigned voxel to the other units' assigned voxels). For each unit, this metric is calculated across only the closest 33% of units, to simulate stream boundaries, which has the additional benefit of discounting high distances where there are few pairs. This results in two vectors, where each element in the vector is the normalized distance between a pair of units (first vector), or between their assigned voxels (second vector), and the correlation between these two vectors is calculated. This calculation is then repeated for all unit-to-voxel pairings and averaged. The same calculation was performed in the brain-to-brain case to determine the actual cortical level of spatial similarity. We report the average spatial similarity across unit-to-voxel pairings.

For the multiple behaviors models, we could additionally evaluate spatial correspondence by calculating the percentage of voxels for each stream that were assigned to the multiple behaviors model trained using that stream's corresponding task. If the hypothesized tasks mapped perfectly to their candidate streams, we would expect to see, for example, 100% correspondence between dorsal and the sub-model trained for object detection, 100% correspondence between lateral and action recognition, and 100% correspondence between ventral and object categorization. Chance performance is 33%.

Additionally, for visualization purposes only, in the case of spatial constraints models we divided the TDANN simulated cortical sheet into three sections (candidate stream partition scheme), assuming a log-polar transform, and calculated the highest percentage of voxels that match the partition scheme across candidate partitions (reflection and $5°$ rotations). We could then use that color scheme to visualize back on the brain, for each voxel, the location of that voxel's assigned unit on the simulated cortical sheet.

Evaluating functional correspondence. To evaluate functional correspondence between candidate models and the brain, we report the 1-to-1 correlations calculated on the left-out set of 515 shared images, using the unit-to-voxel assignments determined by the mapping procedure. Each unit-to-voxel correlation was normalized by the individual voxel noise ceiling (r) of that assigned voxel (see [1] for information on the calculation of the intra-individual voxel noise ceilings in NSD). 1-to-1 correlations were calculated on an individual subject and hemisphere basis for each of the candidate models. The voxel-to-voxel assignments were used to calculate the overall inter-individual i.e. brain-to-brain noise ceiling (correlations evaluated on test set of 20% of the shared images, averaged across 5 splits for each source and target subject combination).

Linear regression. To compare the 1-to-1 mapping results to the commonly used mapping method of linear regression between model units and the brain [22, 36, 10] (Supplemental Fig. S8), we also calculated model to brain correspondence by regressing model responses from the final convolution layer onto individual voxel responses using ridge regression. As in [10], to decrease computational costs without sacrificing performance, we first projected unit activations into a lower dimensional space using a subsample of the ImageNet validation images and retained the first 1000 PCs. Performance was evaluated on a left-out test set (8/9 train, 1/9 test, shared images excluded, 10 splits) for each subject separately. Test R^2 for each voxel is normalized by the individual voxel noise ceiling (R^2). To evaluate the upper-bound model performance given the shared variance across subjects, we again calculated a brain-to-brain noise ceiling for each stream by repeating the same ridge regression procedure as for model-to-brain but instead using all other subjects' responses to predict the left-out subject (80/20 train-test split using the set of 515 shared images, 10 splits).

Effective dimensionality

To calculate effective dimensionality, i.e. the dimensionality of the space of how information is represented by a system (also referred to as the latent dimensionality), we considered the responses of the subset of units assigned to each stream by subject. Using the unit activations (in the case of the models) or z-scored betas (in the case of the human subjects), we computed the eigenspectrum of these responses to the MSCOCO images used in NSD. Following [37] and [38], we computed effective dimensionality from the eigenvalues (λ) as:

$$ED = \frac{\left(\sum_{i=1}^{N} \lambda_i\right)^2}{\sum_{i=1}^{N} \lambda_i^2} \qquad \textbf{(S4)}$$

where N is the number of eigenvectors. Intuitively, if the eigenspectrum decays slowly, that means there are many informative dimensions, and the ED, which in words is simply the squared sum of the eigenvalues over the sum of squares of the eigenvalues, will be high. On the other hand, if the eigenspectrum decays rapidly, meaning that information is largely encoded in only a few dimensions, then the ED will be low.

Task transfer

We tested performance of self-supervised spatial constraints model units mapped to the dorsal and ventral stream on new tasks - position prediction and object categorization, respectively - associated with each stream (results for the lateral stream and additional tasks in Supplementary Fig. S8). We refer to this as task transfer performance as the model was not trained on any of these tasks and model weights were frozen. Performance on the transfer task was tested for the self-supervised spatial constraints models which best match the brain ($0.25 \leq \alpha \leq 0.5$ in Fig. 5, full results across spatial weightings in Supplementary Fig. S7) and each hemisphere of each subject.

Position task. We evaluated the performance of spatial constraints model units assigned to each stream on predicting the vertical and the horizontal locations in pixels of an object center in an image, using the stimulus set from Hong et al., which has also been used in the evaluation of neural network models of the mouse [40] and primate [36, 41] visual systems. This stimulus set consists of 5760 gray-scale images of 64 distinct objects chosen from one of eight categories (animals, boats, cars, chairs, faces, fruits, planes, tables) placed on randomly chosen, realistic background scene images. Object position, pose and size in this stimulus set varied at different levels from no variation, to medium variation and high variation levels.

For each TDANN model and individual subject, smaller "stream models" were created for each of the three streams by selecting the 5000 units assigned to that stream with the highest correlations. We extracted activations from these units and reduced the dimensionality of the activations to 1000 dimensions using principal components analysis (PCA). We used Ridge regression, with the regularization parameter, α, cross-validated from

$$\alpha \in [0.01, 0.1, 1, 10] \tag{S5}$$

to predict the vertical and the horizontal locations in pixels of the object center in the image. We performed five-fold cross-validation on the training split of the no- and medium-variation image subsets, consisting of 3200 images, and computed performance on the test split of the high-variation set consisting of 1280 images. Ten different category-balanced train-test splits were randomly selected. We report R^2 on the high-variation test set, averaged across the 10 splits.

Categorization task. To evaluate the performance of TDANN units assigned to each stream on a downstream categorization task, we used the 1000-way ImageNet object categorization task [3]. For each model and individual subject, smaller "stream models" were created for each of the three streams as in the position prediction task. A single linear layer was then trained directly from the outputs of those units. The linear layer was trained for 28 epochs of the ILSVRC-12 ImageNet training set (1,281,167 images) with a batch size of 1,024 and a learning rate which was initialized to 0.04 and decreased by a factor of 10 every eight epochs. We report the top-1 performance on the held-out validation set (50,000 images).

Model unit selectivity and receptive field properties

Previous studies reported that voxels in face-selective regions in the lateral stream are more peripheral than those in face-selective regions of the ventral stream [42]. To test if this functional feature was also present in the spatial constraints model, we identified face-selective units in each stream and then estimated the eccentricity of their receptive fields. To identify face-selective units in the spatial constraints model, we used the functional localizer (fLoc) stimulus set [43]. fLoc contains stimuli from five categories, each with two subcategories consisting of 144 images each. The categories are faces (adult and child faces), bodies (headless bodies and limbs), written characters (pseudowords and numbers), places (houses and corridors), and objects (string instruments and cars). These stimuli have been previously used to localize and describe category-selective responses in human higher visual cortex in fMRI studies [43, 42, 44]. Selectivity was assessed by computing the t-statistic over the set of functional localizer stimuli and defining a threshold above which units were considered selective.

$$t = \frac{\mu_{\text{on}} - \mu_{\text{off}}}{\sqrt{\frac{\sigma^2_{\text{on}}}{N_{\text{on}}} + \frac{\sigma^2_{\text{off}}}{N_{\text{off}}}}}, \tag{S6}$$

where μ_{on} and μ_{off} are the mean responses to the "on" categories (adult and child faces) and "off" categories (all non-face categories), respectively, σ^2 are the associated variances of responses to exemplars from those categories, and N is the number of exemplars being averaged over. As in fMRI experiments, units with $t > 3$ were classified as face-selective. For each unit, eccentricity was then calculated based on the unit's (x,y) position from the center of the $7x7$ filter, converted to degrees of visual angle (by multiplying by $8°$ input stimulus / 7 grid size). From there, we divided these units based on which stream they were assigned to and report the mean eccentricity across face-selective units for each model seed x subject x hemisphere combination.