CROSS-ATTENTION ENCODING MODELS REVEAL DYNAMIC SPATIOTEMPORAL ROUTING ACROSS HUMAN HIGHER VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how the brain parses actions and events from time-varying natural inputs is a central challenge in neuroscience. Recent work has used deep neural network (DNN) models to build stimulus-computable fMRI encoding models that predict single-voxel responses to complex natural videos. However, the majority of video-computable encoding models predict responses using simple linear mappings from model tokens, overlooking the spatiotemporal structure shared by video representations and neural responses. Recent cross-attention encoding models address this limitation for static images, enabling flexible stimulus-dependent weighting of image content across space. Here, we extend this framework to naturalistic video, using per-parcel cross-attention to dynamically route features from a self-supervised video model (V-JEPA-2) across both space and time, fitting this model to fMRI responses to short video clips. We compare joint spatiotemporal attention with factorized and selectively constrained alternatives, and find that joint routing improves predictions of brain responses to held-out videos across higher visual regions, most consistently in lateral and dorsal visual areas associated with dynamic motion perception. Moreover, our method provides interpretable, stimulus-specific attention maps that dynamically follow moving objects, revealing which locations and temporal moments contribute to each neural response. We further show that attention maps from parcels in different category-selective networks (face-, body-, scene-selective) differentially weight content in accordance with expected semantic selectivity. Together, this work provides a new computational framework for understanding how visual information is adaptively weighted by cortical populations during dynamic visual perception.

1 INTRODUCTION

Human observers effortlessly parse objects, events, and actions from a dynamic visual input, supported by computations in higher dorsal, ventral, and lateral visual cortex (Ungerleider & Mishkin, 1982; Pitcher & Ungerleider, 2021; Grill-Spector & Weiner, 2014). Recent work using methods such as functional magnetic resonance imaging (fMRI) along with deep neural network (DNN)-based computational models, has substantially advanced understanding of how the higher visual system processes static content such as natural scenes (Yamins & DiCarlo, 2016; Doerig et al., 2023). However, while a number of recent studies have applied similar methods to video-evoked data (e.g., Sartzetaki et al., 2025; Garcia et al., 2025; Tang et al., 2025), we still know comparatively much less about the neural basis of dynamic visual perception.

Modeling neural responses to dynamic videos is challenging due to the spatiotemporal complexity of both videos and video-evoked neural responses. Visual cortex populations, as measured with fMRI, integrate information across time nonlinearly (Hasson et al., 2008; Zhou et al., 2018; Groen et al., 2022), making it challenging to identify which temporal moment is most predictive of the measured neural response. In addition to temporal structure, natural videos are spatially complex, often including multiple objects or agents that compete for neural representation (Zoccolan et al., 2005; Reddy & Kanwisher, 2007), which similarly makes it challenging to identify which spatial region of an image is responsible for driving a neural population. Evidence suggests that neural populations in the higher visual system respond to a preferred stimulus at multiple possible locations

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

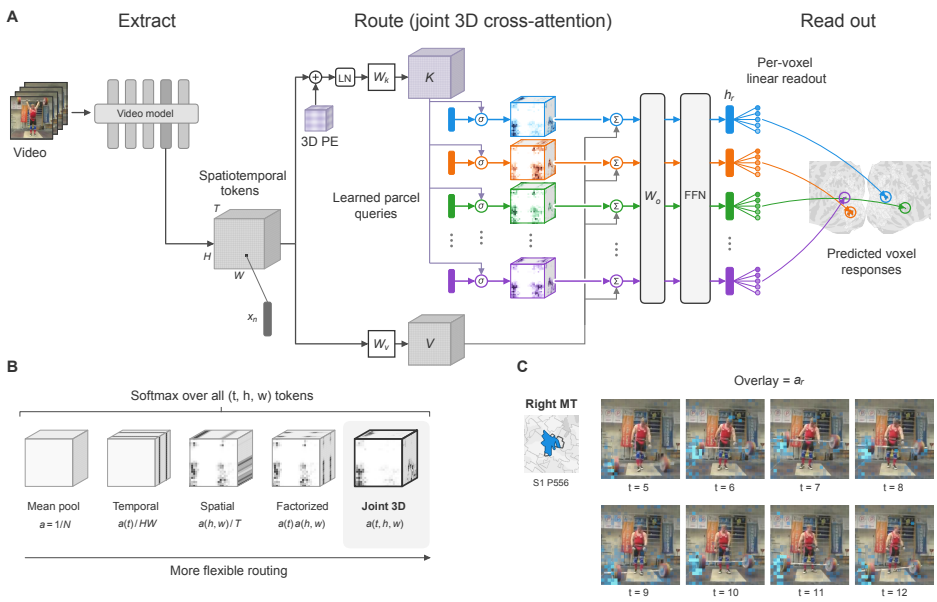

Figure 1: Cross-attention encoding model. **(A)** A video backbone (V-JEPA 2) encodes each clip into a grid of spatiotemporal tokens. A learned query for each cortical parcel attends over the tokens, and the attended feature predicts the responses of the parcel's voxels through a voxel-specific linear readout. **(B)** Constrained routing models, shown as example attention patterns over the token grid with the form of their attention weights. **(C)** Attention of the joint model for a parcel in right MT, overlaid on selected frames of a test clip of a weightlifter.

within a large spatial receptive field (Rust & DiCarlo, 2010), and their response to the stimulus may depend on other objects (Zoccolan et al., 2005; Reddy & Kanwisher, 2007; Bao & Tsao, 2018) or surrounding context (Bar, 2004). The resulting neural response properties may then be described as having spatial selectivity that is stimulus-dependent, as opposed to being driven by features within a fixed spatial receptive field. In videos, where objects can also move in time, the relative weighting of information across time may similarly depend on the input, which could result in adaptive spatiotemporal weighting that depends on characteristics of the input video.

To model such adaptive stimulus-dependent weighting of visual features by neural populations, recent work has incorporated attention operations, commonly used in Transformer neural networks (Vaswani et al., 2017; Dosovitskiy et al., 2021), into fMRI encoding models (Adeli et al., 2026; Hwang et al., 2025). While standard fMRI voxelwise encoding model approaches map from DNN features to voxel responses using a simple linear mapping (Naselaris et al., 2011), these newer approaches use a cross-attention module with a region-specific query vector to perform a more flexible routing operation that can improve model performance in higher visual regions (Adeli et al., 2026; Hwang et al., 2025). However, it is not clear whether and how such approaches can be extended to modeling brain responses to video stimuli, where spatial and temporal information may be independently and/or jointly reflected in neural response properties.

Here, we present a new cross-attention encoding model framework for predicting voxelwise fMRI activation in response to natural video clips. Using fMRI data from the BOLD Moments Dataset (BMD; Lahner et al. (2024)) and features extracted from a self-supervised video model (V-JEPA-2 (Assran et al., 2025)), we compare the predictive accuracy of models using spatial, temporal, and joint spatiotemporal cross-attention. Our contributions are as follows:

- We extend previous work by showing that a cross-attention readout leveraging spatial attention over video tokens can improve predictive accuracy of voxelwise encoding models for video-evoked fMRI data, across multiple DNN backbones.
- We show that the novel addition of joint spatiotemporal attention boosts encoding model accuracy across several higher visual regions, notably in lateral and parietal visual cortex.

- We demonstrate how joint spatiotemporal attention leads to more interpretable and accurate attention maps compared to factorized attention, tracking the movement of relevant objects across a video clip.
- We show dissociations between the attention maps derived from different functionally-localized higher visual cortex regions, with attention weights dynamically tracking the location of region-relevant video content.

2 Related Work

Neural coding during dynamic visual perception. Classic models of vision suggest that static object recognition and dynamic spatial vision are anatomically segregated into the ventral and dorsal stream, respectively (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992; Grill-Spector & Weiner, 2014), with dorsal areas (MT, MST) key to visual motion processing (Albright, 1984; Andersen, 1997). Recent work also provides evidence for a lateral visual pathway specializing in dynamic social perception (Pitcher et al., 2011; Pitcher & Ungerleider, 2021; McMahon et al., 2023), including regions along the superior temporal sulcus (STS). Other research on temporal dynamics has suggested a hierarchical posterior-to-anterior organization based on temporal receptive window size (Hasson et al., 2008; Groen et al., 2022; Zhou et al., 2018). Here, we develop a novel strategy to computationally model higher visual cortex during dynamic natural event perception.

Category-selective organization in visual cortex. Anatomically localized regions of ventral and lateral visual cortex exhibit functional selectivity for human faces (FFA, OFA, STS), human bodies (EBA), and scenes (PPA, TOS, RSC), among other categories (Epstein & Kanwisher, 1998; Kanwisher et al., 1997; Downing et al., 2006; Epstein & Baker, 2019). These regions were classically identified using isolated objects, with recent work extending this using naturalistic fMRI and computational modeling (Khosla et al. (2022); Ratan Murty et al. (2021); Luo et al. (2023), and others). However, because natural videos often include multiple categories that interact over time, modeling category-selective responses during natural video perception remains challenging. Here, we show how joint spatiotemporal attention can be used to dissociate the features driving responses in distinct category-selective networks during complex video perception.

Voxelwise fMRI encoding models. To model cortical fMRI responses during complex natural vision, forward encoding models construct a stimulus-computable function mapping from the input (i.e., a stimulus viewed by an observer) to the response of an fMRI voxel (Naselaris et al., 2011; Serences & Saproo, 2012; Wehbe et al., 2014). This is typically done by extracting stimulus features from a task-optimized DNN model, then learning a voxel-specific mapping from the DNN activations to the individual voxel's response (Naselaris et al., 2011; Yamins & DiCarlo, 2016). Benchmarking the cross-validated accuracy of DNN encoding models can be used to evaluate the alignment of DNN models and neural representations (Schrimpf et al., 2020; Wang et al., 2023a; Conwell et al., 2024; Sartzetaki et al., 2025; Doerig et al., 2023).

The voxel-specific mapping is most commonly linear (Naselaris et al., 2011), and is often preceded by spatial and/or temporal averaging to handle high-dimensional DNN features. Other approaches use a "factorized" strategy which applies a shared spatial weighting function across feature maps (Klindt et al., 2017; St-Yves & Naselaris, 2018; Lurz et al., 2021; Henderson et al., 2023), assuming separable and input-invariant spatial selectivity. Recent work replaces this with a cross-attention module that dynamically weights space based on the input (Adeli et al., 2026; Hwang et al., 2025). Here, we extend this flexible routing strategy to the spatiotemporal domain.

Video encoding and decoding models for fMRI. Large-scale video fMRI datasets (Lahner et al., 2024; Gifford et al., 2025) have spurred efforts to build predictive video encoding models based on DNN features (Sartzetaki et al., 2025; Tang et al., 2025; Höfling et al., 2026; Gamal et al., 2026; Han et al., 2024; Al-Karkari et al., 2025; Garcia et al., 2025; Pushpita & Wehbe, 2025). Other work uses video encoding models to perform video reconstruction (Chen et al., 2023; Lu et al., 2024; Yeung et al., 2025), or to synthesize an optimal video stimulus (Tang et al., 2026). Recent papers find that V-JEPA-2 features (Assran et al., 2025) achieve top predictive accuracy among video models (Tang et al., 2025; 2026; Höfling et al., 2026). Our approach is distinct in introducing a spatiotemporal cross-attention framework that yields interpretable maps of how video features are weighted over space and time by cortical populations.

3 METHODS

3.1 VIDEO CROSS-ATTENTION ENCODING MODEL FRAMEWORK

We propose a cross-attention encoding model framework for video-evoked fMRI responses (Fig. 1) with three goals. First, each cortical population should read out from a video in a stimulus-dependent way, selecting *where* and *when* in the clip to draw features from, rather than relying on a fixed receptive field or on features pooled over the video. Second, the contribution of each form of routing should be testable: spatial, temporal, factorized and joint spatiotemporal routing are constrained versions of the same attention operation, so they can be compared with the backbone, readout and training held fixed. Third, the routing should be interpretable: the attention weights form a stimulus-specific map over the video that can be compared directly with its content. Following cross-attention encoding models for static images (Adeli et al., 2026; Hwang et al., 2025), a learned query for each brain parcel attends over the token grid of a frozen video backbone, and we extend the attention to range jointly over time and space.

Below, we describe the model, the constrained comparison models and the parcel-based queries. Section 3.2 describes the data, Section 3.3 the fitting, evaluation and ridge regression baseline, and Appendix 9.13 the analyses of category selectivity in the attention maps.

Spatiotemporal token grid. A frozen video backbone maps each clip to a grid of tokens $X \in \mathbb{R}^{T\times H\times W\times D}$, with token $x_n \in \mathbb{R}^D$ at grid position (t_n, h_n, w_n) and $N = THW$ tokens per clip ($T = 16$, $H = W = 24$, $D = 1408$ and $N = 9216$ for our backbone; Appendix 9.3). Each channel is z-scored over training clips. The tokens are computed once; all learning takes place in the readout.

Cross-attention readout. Each parcel r has a learned query $q_r \in \mathbb{R}^D$ that does not depend on the stimulus. Keys and values are computed from the tokens as

$$k_n = W_k\,\mathrm{LN}\big(x_n + \alpha\,\mathrm{PE}(t_n, h_n, w_n)\big), \qquad v_n = W_v\,x_n,$$

where $W_k, W_v \in \mathbb{R}^{D\times D}$, LN is layer normalization, and PE is a fixed sinusoidal 3D positional encoding whose D channels are split into three blocks encoding t, h and w ($\alpha = 1$). Because the positional encoding enters the keys but not the values, a query can prefer locations and moments as well as content, while the values carry content only. The attention of parcel r is a single softmax over all N tokens,

$$a_{rn} = \frac{\exp\big(q_r^\top k_n/\sqrt{D}\big)}{\sum_{m=1}^{N}\exp\big(q_r^\top k_m/\sqrt{D}\big)}, \qquad \sum_n a_{rn} = 1,$$

which, reshaped to the token grid, gives the attention map $a_r(t, h, w)$ (Fig. 1). The fixed query encodes a stable preference, and the map shows where and when a clip meets it. The attended feature $z_r = \sum_n a_{rn} v_n$ is passed through an output projection and a residual feed-forward network,

$$h_r = W_o z_r, \qquad h_r \leftarrow h_r + \mathrm{FFN}\big(\mathrm{LN}(h_r)\big),$$

where the FFN is a two-layer MLP with a GELU nonlinearity and hidden width $4D$. We use a single attention head, and W_k, W_v, W_o and the FFN are shared across parcels, so parcels differ only in their query and voxel readout. Each voxel is predicted from the output h_r of its parcel by a voxel-specific linear readout, trained jointly with the attention module by gradient descent. All voxels in a parcel thus share one attention map but weight the attended features differently.

Constrained routing models. To isolate the contributions of spatial, temporal and coupled spatiotemporal routing, we compare the joint model with models that differ from it only in their attention weights. Their attention factorizes into a temporal and a spatial distribution, $a_r(t, h, w) = a_r^{\mathrm{T}}(t)\,a_r^{\mathrm{S}}(h, w)$, and each factor is uniform, fixed or routed. A uniform factor weights all frames (or locations) equally. A fixed factor is a softmax over learned logits for each parcel, a stimulus-independent receptive field that is the same for every clip. A routed factor is a softmax of the parcel's query against keys computed, with their own key projection and positional encoding, from tokens averaged over the other axis (the spatially averaged token of each frame, or the temporally averaged token at each location). We focus on four of these models:

- **Mean pool** (both factors uniform, $a_{rn} = 1/N$): every parcel reads the same clip-average feature, with no routing.

- **Temporal** (a_r^{T} routed, a_r^{S} uniform): the parcel selects moments but pools over space.
- **Spatial** (a_r^{T} uniform, a_r^{S} routed): the parcel selects locations but pools over time, as in spatial routing for static images applied to the time-averaged video.
- **Factorized** (both factors routed): the parcel selects moments and locations, but attends to the same locations at every moment, so it cannot follow moving content.

Because the joint model's attention is a single softmax over all tokens, a parcel can attend to different locations at different moments. Comparing the joint and factorized models therefore isolates the benefit of coupled spatiotemporal routing, and comparing the spatial and temporal models with mean pooling isolates the benefit of routing along each axis.

Parcel-based queries. We assign one query to each parcel of the Schaefer 1000-parcel, 7-network atlas (Schaefer et al., 2018), resampled to the voxel grid of the dataset (Appendix 9.2), which gives 423–433 queries per subject, including one for voxels without a parcel label. We judged this the appropriate granularity: fine enough for differences within and between functional regions to emerge in the attention maps, and coarse enough that each map is estimated from many voxels (Appendix 9.1).

3.2 Functional MRI video dataset

BOLD Moments Dataset. We use the BOLD Moments Dataset (BMD; Lahner et al. (2024)), in which 10 subjects watched 1,102 naturalistic 3 s video clips of everyday events during 3T fMRI (2.5 mm isotropic voxels, TR 1.75 s): 1,000 training clips shown 3 times each and 102 test clips shown 10 times each. We predict the single-trial response amplitudes released with the dataset (GLMsingle betas; Prince et al. (2022)), averaged over the repeats of each clip, in the 17.2–17.5k voxels per subject that lie in cortex with reliable responses to the videos or in the functional ROIs (preprocessing details in Appendix 9.2). We also use the per-voxel noise ceiling released with the dataset, computed for responses averaged over the 10 test repeats.

Functional ROIs. We use the 23 ROIs released with the dataset, pooled across hemispheres: retinotopic ROIs from the atlas of Wang et al. (2015), MT and parietal ROIs from the HCP multimodal parcellation (Glasser et al., 2016), and category-selective ROIs defined in each subject with a video-based functional localizer (Appendix 9.2). To summarize prediction accuracy (Fig. 2), we pool the ROIs into four groups: early visual (V1–V3, dorsal and ventral, V3ab and hV4), lateral/face/body (MT, EBA, FFA, OFA, LOC, STS), scene (PPA, RSC, TOS) and parietal (IPS0, IPS1-2-3, 7AL, BA2, PFop, PFt). We also report all modeled voxels and the voxels outside all ROIs. The groups follow a clustering of the ROIs by their responses to the videos (Appendix 9.4).

3.3 Model fitting and evaluation procedures

We use frozen features from layer 32 of 40 of V-JEPA 2 ViT-g/16 (Assran et al., 2025), a self-supervised video model whose features give the most accurate fMRI predictions among current large video models (Tang et al., 2025; 2026; Höfling et al., 2026); layer 32 gave the most accurate predictions in a ridge-regression sweep over the layers of several video and image backbones (Appendix 9.5). To test whether our results depend on the backbone and its depth, we also fit the joint and all constrained models to features from an early and a late layer of another video model and two image models (Appendix 9.11). Each subject is fit separately, with the mean squared error over all voxels as the loss. Every model is an ensemble of 10 members, each trained on a different random split of the 1,000 training clips into 900 clips for training and 100 for early stopping, and, following Adeli et al. (2026), the test predictions of the members are averaged voxel by voxel before evaluation. We train with AdamW and early stopping, with the same hyperparameters for all models, backbones and subjects. As a standard linear encoding model, we also fit voxelwise ridge regression to the same backbone layer, averaged over time and space into one 1408-dimensional vector per clip, with the same splits and ensembling (token extraction, optimization and ridge details in Appendix 9.3). For each voxel, we compute the Pearson correlation r between the ensemble predictions and the measured responses across the 102 test clips and report the signed $r^2 = \mathrm{sign}(r)\, r^2$.

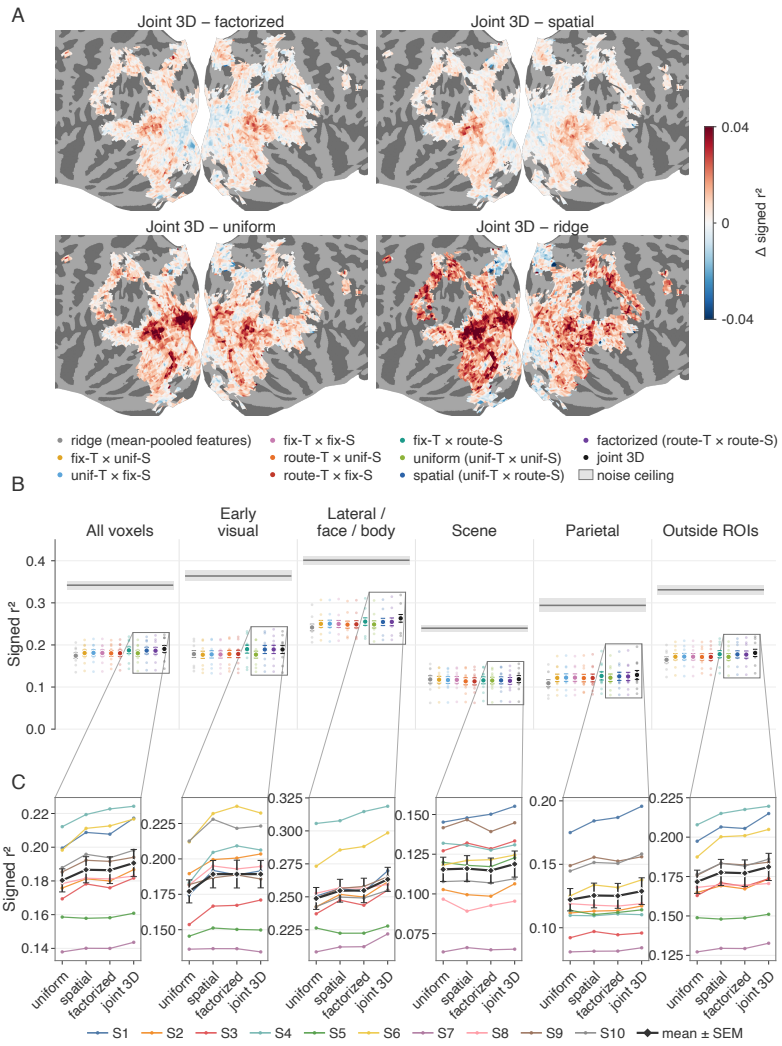

Figure 2: Comparing accuracy of joint spatiotemporal attention versus other mappings (V-JEPA 2 ViT-g, layer 32; 102 test clips). **(A)** Joint model minus the factorized, spatial and mean-pool ("uniform") models and ridge regression on mean-pooled features in voxelwise signed r^2, averaged over subjects (fsaverage flatmap, unthresholded; each subject in Appendix 9.7). **(B)** Region-mean signed r^2 of every routing configuration (T = temporal factor, S = spatial factor; each uniform, fixed or routed), the joint model and ridge regression. Large dots: mean $\pm$ SEM over subjects; small dots: individual subjects; grey band: noise ceiling (mean $\pm$ SEM). Region means include only voxels with a noise ceiling $\geq 5\%$. **(C)** Zoomed inset from (B); shows region means of the mean-pool, spatial, factorized and joint models in each subject (colored lines) and their mean $\pm$ SEM (black).

Table 1: Region-mean signed r^2 on the 102 test clips (V-JEPA 2 ViT-g, layer 32; mean over 10 subjects; voxels with noise ceiling $\geq 5\%$). The best model in each row is in bold. All routing configurations and the other backbones are given in Table 7.

Region	Ridge	Mean pool	Temporal	Spatial	Factorized	Joint	Noise ceiling
All voxels	.174	.180	.180	.187	.186	**.191**	.34
Early visual	.178	.177	.178	**.189**	**.189**	**.189**	.36
Lateral/face/body	.242	.249	.248	.255	.255	**.263**	.40
Scene	.118	.115	.114	.116	.115	**.119**	.24
Parietal	.109	.122	.121	.125	.125	**.129**	.29
Outside ROIs	.165	.172	.171	.178	.177	**.181**	.33

4 RESULTS

Comparing model performance across attention types. We first examine performance averaged across all voxels, then examine region-specific differences. Across all modeled voxels, we observed a substantial advantage for joint spatiotemporal routing, with the joint model having higher r^2 than every constrained model and than ridge regression in all 10 subjects (Fig. 2B, Table 1), including the closest, the spatial model (0.191 vs. 0.187; $t(9) = 5.85$, $p_{\text{FDR}} < 0.001$). Unless stated otherwise, we compare models with two-sided paired t-tests of region-mean signed r^2 across the 10 subjects (FDR corrected; $p_{\text{FDR}} < 0.05$; Benjamini & Hochberg (1995)); see Table 3.

The constrained models show which kind of routing produces this gain. Of the 0.010 by which the joint model exceeds mean pooling, 0.006 is recovered by routing over space alone (spatial minus mean pool: $t(9) = 5.07$, $p_{\text{FDR}} < 0.001$; see Table 4 for additional comparisons). This extends to video previous work showing an advantage of stimulus-dependent spatial routing over linear readouts for static images (Adeli et al., 2026). This pattern holds for every backbone and layer we tested, with the exception of scene-selective cortex and for the last stage of CLIP (Appendix 9.8). Routing over time alone does not improve over uniform pooling (temporal model, 0.180), and adding temporal routing to spatial routing does not improve beyond spatial routing alone (factorized model, 0.186). We also find that the benefit of spatial routing depends on the selection being driven by the stimulus: a fixed spatial distribution learned for each parcel, which acts as a static receptive field, does not exceed the performance of mean pooling (0.181; Table 7, ablations in Appendix 9.9). These results provide insight into what drives the benefit of the joint spatiotemporal routing mechanism: it allows parcels to attend to different locations at different moments in a stimulus-dependent manner.

This benefit of coupled routing differs between regions. On the flatmaps, the advantage of the joint model over the factorized and spatial models is concentrated in lateral occipitotemporal cortex (Fig. 2A). At the level of functionally-defined ROIs, the advantage is largest in lateral/face/body regions, where the joint model reaches 0.263 (66% of the noise ceiling) and exceeds the factorized, spatial, mean-pool and ridge regression models in every subject (joint minus factorized: $t(9) = 6.60$, $p_{\text{FDR}} < 0.001$; Fig. 2C; individual ROIs in Appendix 9.10). Parietal regions similarly show an advantage for joint routing over all other models (Table 3). In early visual cortex, by contrast, the spatial, factorized and joint models are tied (0.189; joint minus factorized: $t(9) = -0.03$, $p_{\text{FDR}} = 0.98$), and all three improve on mean pooling by 0.012. Stimulus-dependent selection of spatial locations thus accounts for the whole gain in early visual cortex, whereas lateral and parietal regions additionally benefit from following content over time. In scene-selective regions, we find more mixed evidence for a benefit of spatial or temporal routing. Neither the joint model nor the spatial model significantly exceed the the ridge baseline (Table 3, Table 4). This may reflect the role of scene-selective cortex in encoding global scene layout, which features averaged over the whole clip may sufficiently capture (Appendix 9.4).

The advantage of coupled routing in lateral regions is not specific to V-JEPA 2 (Appendix 9.11). With features from the late layers of a second video model (VideoMAE v2) and two image models (DINOv2 and CLIP), the joint model is again more accurate than the factorized model in lateral/face/body regions (joint minus factorized: $t(9) = 2.88$, 2.49 and 3.47; $p_{\text{FDR}} = 0.027$, 0.048 and 0.011, respectively; FDR-corrected over the 192 tests of the sweep). The pattern depends on the features, however. With features from early layers, the joint model is less accurate than models that route over space alone, and for the last stage of CLIP, ridge regression on mean-pooled features is the most accurate model in every region except early visual (Table 7).

Interpreting cross-attention maps. Fig. 3 shows the attention maps of the joint and factorized models for one parcel in each of two test clips, both in lateral/face/body regions, where coupled routing improves accuracy most. For a parcel in left EBA, the attention of the joint model stays on the body of a fencer and, after a cut to a close-up, follows the arm and sword through the swing. For a parcel in left STS, it stays on the head and body of a panda as it climbs. The factorized model cannot: its spatial map is fixed across frames and only its temporal weights change, so it concentrates its attention on a few frames and, within them, spreads it over a large region that includes the background. In the fencer clip, a single spatial map must serve both the wide shot and the close-up, although the body is in different places in the two shots (more examples in Appendix 9.12).

Cross-attention maps differentiate category-selective networks. If the attention of a parcel reflects what its voxels respond to, the attention of category-selective regions should land on their

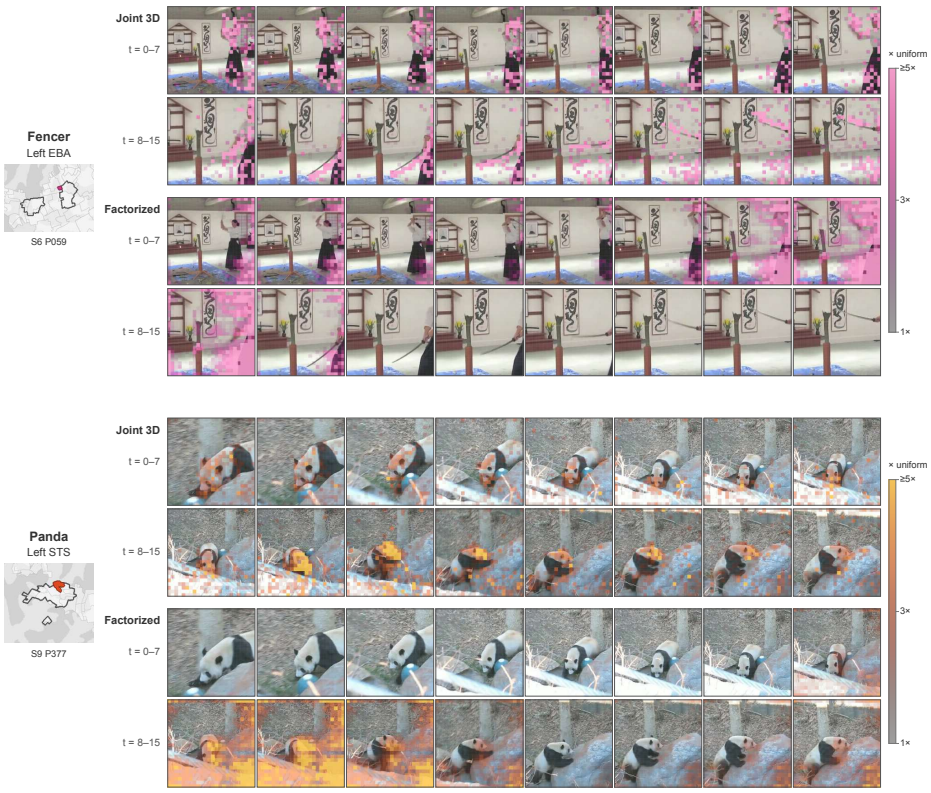

Figure 3: Joint attention follows moving content, whereas factorized attention cannot. Attention maps of one parcel in each of two test clips, for the joint model (upper two rows of each block) and the factorized model (lower two rows), over all 16 frames of the clip. Left of each block: the subject, the parcel and its location on the flatmap within its functional ROI. Attention is shown in units of the uniform level $1/N$ on the same scale for both models: tokens at or below the uniform level are transparent, and color and opacity increase up to 5 times the uniform level.

preferred category, and it should be possible to find these regions from the attention alone, without fMRI localizer data. We test both with the attention of the joint model on the 102 test clips, using the intersection over union (IoU) of each parcel's 92 most-attended tokens (the top 1%) with category masks obtained by segmenting the clips with SAM 3 (Carion et al., 2025) (Appendix 9.13). Because the raw IoU mostly reflects the size of each category (Fig. 4b), we z-score it across parcels to better isolate parcel differences (Fig. 4c). This metric indicates that FFA, OFA and EBA attend more than the average parcel to faces, bodies and animals, which usually appear together in the clips, and less to scences, whereas the scene-selective PPA and RSC show the reverse pattern. For the majority of face-, body-, and scene-selective ROIs, the highest category in this metric is the expected category (with exceptions OFA and STS, where animal and scene are slightly above face respectively).

Based on these z-scored IoU values, we compute "localizer contrasts" that quantify face, body, and scene-selectivity. As expected, the parcels with the highest contrast overlap with the corresponding ROIs above chance for all three localizers: face 0.19 (chance 0.09), body 0.14 (chance 0.02) and scene 0.28 (chance 0.10; two-sided one-sample t-tests of overlap minus chance across subjects, FDR-corrected over the three localizers, all $p_{\mathrm{FDR}} < 0.05$; Table 8). Additional supporting analysis can be found in Appendix 9.14. These results indicate that attention maps reliably differentiate the selectivity of distinct functionally-defined cortical networks.

5 Discussion

We developed a novel framework for modeling human fMRI responses to natural video clips, leveraging joint spatiotemporal cross-attention over embeddings from a self-supervised video model. Our results showed that joint attention outperformed spatial-only or factorized attention in multiple

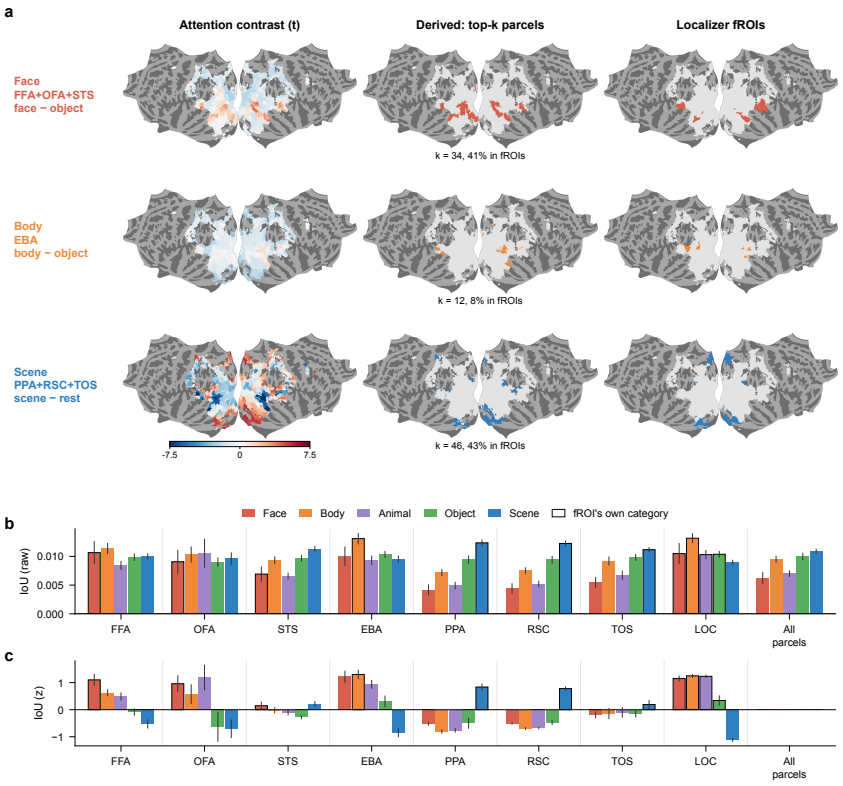

Figure 4: The attention of the joint model differentiates category-selective regions. **(A)** Localizing category-selective regions from attention alone in one subject (S8). Left: t of each localizer contrast of z-scored IoU across test clips, per parcel. Middle: the K highest-t parcels (K = number of ROI parcels) and their overlap with the ROIs. Right: the localizer-defined ROIs (parcels with at least 20% of their voxels inside). **(B)** Raw IoU of the 92 most-attended tokens (top 1%) with the masks of five categories, for eight category-selective ROIs and for all parcels (mean $\pm$ SEM over 10 subjects). **(C)** The same IoU z-scored across all parcels, so that all parcels are 0 by definition. Outlined bars mark the category that defines each ROI (for LOC, all four foreground categories).

regions of higher visual cortex, with the most consistent advantage in lateral and parietal regions. This aligns with past evidence that dorsolateral visual cortex regions play a key role in dynamic processing of motion, events and social actions (McMahon et al., 2023; Andersen, 1997; Pitcher et al., 2011), and expands upon these past findings by showing how cross-attention maps can yield interpretable tracking of the spatiotemporal features that drive cortical response on a per-frame basis.

Our results could suggest that higher visual cortex implements a mechanism analogous to spatiotemporal cross-attention over feature maps during video perception, but this remains to be empirically tested. Previous work suggests that normalization across multiple objects in images (Bao & Tsao, 2018; Zoccolan et al., 2005) and/or normalization across time (Zhou et al., 2018; Groen et al., 2022) can explain neural response properties in the visual system, and it is possible that scaled dot-product attention (Vaswani et al., 2017) approximates a similar normalization mechanism (Kozachkov et al., 2023). To evaluate this, future work should compare performance of cross-attention versus more biologically-inspired normalization models (Reynolds & Heeger, 2009; Carandini & Heeger, 2011).

Other limitations of this work include the short length of the video stimuli (3s) used in the BMD dataset (Lahner et al., 2024). Expanding our framework to modeling longer movie stimuli (Boyle et al., 2025; Gifford et al., 2025), would allow investigation of temporal integration over longer time windows. The slow temporal resolution of fMRI recordings presents another limitation, so future work should also explore how this model can predict data in more temporally-resolved modalities including MEG and EEG. These extensions will lead to more comprehensive understanding of how neural populations adaptively weight time-varying information during dynamic natural vision.

6 AI USE STATEMENT

In this work, we used generative AI tools to write and debug code for model training, evaluation, statistical analysis, figure generation, and to run and monitor jobs on our computing clusters. We also used them to draft parts of the text, in particular the methods and appendix. We have not used generative AI tools to generate data, formulate the central hypothesis of the work, or identify relevant literature. SAM 3, which we use to segment the test clips, is part of the method and is described in Appendix 9.13. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

7 ETHICS STATEMENT

We used fMRI data from human participants provided as part of the Bold Moments Dataset (Lahner et al., 2024). This data collection protocol was approved by the appropriate Institutional Review Board as described in the original data paper, and was shared in an anonymized format. We do not anticipate any direct negative societal consequences associated with this basic science research.

8 REPRODUCIBILITY STATEMENT

Full details of the data preparation steps and modeling procedures we used to obtain these results are provided in the main text (3.2, 3.3) and in the Appendix. We plan to release the code upon publication.

REFERENCES

Hossein Adeli, Sun Minni, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses, 2026. URL https://arxiv.org/abs/2505.17329.

Khaled Jedoui Al-Karkari, Yingtian Tang, Martin Schrimpf, and Daniel L. K. Yamins. Do dynamics matter for neural alignment? A comparative study of video and static vision models. In *Proceedings of the Conference on Cognitive Computational Neuroscience (CCN)*, 2025. URL https://2025.ccneuro.org/abstract_pdf/Al-Karkari_2025_Do_Dynamics_Matter_Neural_Alignment_A.pdf.

Thomas D. Albright. Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, 52(6):1106–1130, 1984. doi: 10.1152/jn.1984.52.6.1106.

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, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

Richard A. Andersen. Neural mechanisms of visual motion perception in primates. *Neuron*, 18: 865–872, 1997.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Pinglei Bao and Doris Y. Tsao. Representation of multiple objects in macaque category-selective areas. *Nature Communications*, 9:1774, 2018. doi: 10.1038/s41467-018-04126-7.

Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, August 2004. doi: 10.1038/nrn1476.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

J.A. Boyle, B. Pinsard, et al. CNeuroMod data collection complete: 200h of individual fMRI across diverse naturalistic and controlled tasks to build NeuroAI models. In *Conference on Cognitive Computational Neuroscience (CCN 2025)*, 2025. Abstract.

Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, November 2011. doi: 10.1038/nrn3136. Erratum in: Nat Rev Neurosci. 2013 Feb;14(2):152.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=i913TUOvTK.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(9383), 2024. doi: 10.1038/s41467-024-53838-2.

Adrien Doerig, Rowan P. Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W. Lindsay, Konrad P. Kording, Talia Konkle, Marcel A. J. van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24:431–450, May 2023. doi: 10.1038/s41583-023-00705-w. URL https://www.nature.com/articles/s41583-023-00705-w.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

P.E. Downing, A.W. Chan, M.V. Peelen, C.M. Dodds, and N. Kanwisher. Domain specificity in visual cortex. *Cerebral Cortex*, 16(10):1453–1461, 10 2006. doi: 10.1093/CERCOR/BHJ086.

R. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, 4 1998. doi: 10.1038/33402.

Russell A. Epstein and Chris I. Baker. Scene perception in the human brain. *Annual Review of Vision Science*, 5:373–397, 2019. doi: 10.1146/annurev-vision-091718-014809.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16 (1):111–116, 2019. doi: 10.1038/s41592-018-0235-4.

Mai Gamal, Mohamed Rashad, Eman Ehab, Seif Eldawlatly, and Mennatullah Siam. Pixel-level understanding of a world in motion within a neural encoding framework. *Scientific Reports*, 16:12188, 2026. doi: 10.1038/s41598-025-34141-w. URL https://doi.org/10.1038/s41598-025-34141-w.

Kathy Garcia, Emalie McMahon, Colin Conwell, Michael Bonner, and Leyla Isik. Modeling dynamic social vision highlights gaps between deep learning and humans. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wAXsx2MYgV.

Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies, 2025. URL https://arxiv.org/abs/2501.00504.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016. doi: 10.1038/nature18933.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992. doi: 10.1016/0166-2236(92)90344-8.

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, 6 2014. doi: 10.1038/nrn3747. URL https://www.nature.com/articles/nrn3747.

Iris I. A. Groen, Giovanni Piantoni, Stephanie Montenegro, Adeen Flinker, Sasha Devore, Orrin Devinsky, Werner Doyle, Patricia Dugan, Daniel Friedman, Nick F. Ramsey, Natalia Petridou, and Jonathan Winawer. Temporal dynamics of neural responses in human visual cortex. *Journal of Neuroscience*, 42(40):7562–7580, 2022. doi: 10.1523/JNEUROSCI.1812-21.2022.

Hyewon Willow Han, Ruchira Dhar, Qingqing Yang, Maryam Hoseini Behbahani, María Alejandra Martínez Ortiz, Tolulope Samuel Oladele, Diana C Dima, Hsin-Hung Li, Anders Søgaard, and Yalda Mohsenzadeh. Investigating the role of modality and training objective on representational alignment between transformers and the brain. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. URL https://openreview.net/forum?id=t4CnKu6yXn.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550, March 2008. doi: 10.1523/JNEUROSCI.5487-07.2008.

Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, 43 (22):4144–4161, 2023. doi: 10.1523/JNEUROSCI.1822-22.2023.

Larissa Höfling, Matthias Tangemann, Lotta Piefke, Susanne Keller, Matthias Bethge, and Katrin Franke. Only brains align with brains: Cross-region alignment patterns expose limits of normative models. In *ICLR*, 2026. URL https://openreview.net/forum?id=cMGJcHHI7d.

Ethan Hwang, Hossein Adeli, Wenxuan Guo, Andrew Luo, and Nikolaus Kriegeskorte. In silico mapping of visual categorical selectivity across the whole brain. In *Advances in Neural Information Processing Systems*, 2025.

Nancy Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 6 1997. doi: 10.1523/JNEUROSCI.17-11-04302.1997.

Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32:4159–4171.e9, 10 2022. doi: 10.1016/j.cub.2022.08.009.

David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what"and "where". In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8c249675aea6c3cbd91661bbae767ff1-Paper.pdf.

Leo Kozachkov, Ksenia V. Kastanenka, and Dmitry Krotov. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34):e2219150120, August 2023. doi: 10.1073/pnas.2219150120.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N. Apurva Ratan Murty, Kendrick Kay, Aude Oliva, and Radoslaw Cichy. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15: 6241, 2024. doi: 10.1038/s41467-024-50310-3. URL https://doi.org/10.1038/s41467-024-50310-3.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. *arXiv preprint arXiv:2405.03280*, 2024.

Andrew F. Luo, Margaret M. Henderson, Leila Wehbe, and Michael J. Tarr. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Tp7kI90Htd.

Emalie McMahon, Michael F. Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, December 2023. doi: 10.1016/j.cub.2023.10.015.

Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011. doi: 10.1016/j.neuroimage.2010.07.073.

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

David Pitcher and Leslie G. Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25:100–110, 2021.

David Pitcher, Daniel D. Dilks, Rebecca R. Saxe, Christina Triantafyllou, and Nancy Kanwisher. Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage*, 56(4):2356–2363, June 2011. doi: 10.1016/j.neuroimage.2011.03.067.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick N. Kay. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, 11, 2022. doi: 10.7554/ELIFE.77599.

Subha Nawer Pushpita and Leila Wehbe. Two cortical mechanisms of audiovisual processing in the human brain. *bioRxiv*, 2025. doi: 10.1101/2025.11.05.686819. URL https://www.biorxiv.org/content/early/2025/11/06/2025.11.05.686819.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

N.A. Ratan Murty, Pouya Bashivan, Alex Abate, James J. DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, 12(5540), 2021. doi: 10.1038/s41467-021-25409-6.

Leila Reddy and Nancy Kanwisher. Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology*, 17(23):2067–2072, 2007. doi: 10.1016/j.cub.2007.10.043.

John H. Reynolds and David J. Heeger. The normalization model of attention. *Neuron*, 61(2): 168–85, January 2009. doi: 10.1016/j.neuron.2009.01.002.

Nicole C. Rust and James J. DiCarlo. Selectivity and tolerance ("Invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, 30(39): 12978–12995, September 2010. doi: 10.1523/JNEUROSCI.0179-10.2010.

Gabriel H. Sarch, Michael J. Tarr, Katerina Fragkiadaki, and Leila Wehbe. Brain dissection: fMRI-trained networks reveal spatial selectivity in the processing of natural images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Christina Sartzetaki, Gemma Roig, Cees G. M. Snoek, and Iris Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LM4PYXBId5.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020. doi: 10.1016/j.neuron.2020.07.040.

John T. Serences and Sameer Saproo. Computational advances towards linking BOLD and behavior. *Neuropsychologia*, 50:435–446, 3 2012. doi: 10.1016/j.neuropsychologia.2011.07.013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 10 2018. doi: 10.1016/J.NEUROIMAGE.2017.06.035.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Diverse perceptual representations across visual pathways emerge from a single objective. *bioRxiv*, 2025. doi: 10.1101/2025.07.22.664908. URL https://www.biorxiv.org/content/early/2025/10/08/2025.07.22.664908.

Yingtian Tang, Sogand Salehi, Ming Zhou, Amir Zamir, Leyla Isik, and Martin Schrimpf. NEvo: Neural-guided evolutionary video synthesis for dynamic visual selectivity, 2026. URL https://arxiv.org/abs/2607.02317.

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield (eds.), *Analysis of Visual Behavior*. MIT Press, 1982.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, 11 2023a. doi: 10.1038/s42256-023-00753-y.

Liang Wang, Ryan E. B. Mruczek, Michael J. Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015. doi: 10.1093/cercor/bhu277.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, 2023b.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9:e112575, 11 2014. doi: 10.1371/JOURNAL.PONE.0112575.

Daniel L. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016. doi: 10.1038/nn.4244.

Jacob Yeung, Andrew F. Luo, Gabriel Sarch, Margaret M. Henderson, Deva Ramanan, and Michael J. Tarr. Reanimating images using neural representations of dynamic stimuli. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 5331–5343, 6 2025.

Jingyang Zhou, Noah C. Benson, Kendrick N. Kay, and Jonathan Winawer. Compressive temporal summation in human visual cortex. *Journal of Neuroscience*, 38(3):691–709, January 2018. doi: 10.1523/JNEUROSCI.1724-17.2017.

Davide Zoccolan, David D. Cox, and James J. DiCarlo. Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, 25(36):8150–8164, September 2005. doi: 10.1523/JNEUROSCI.2058-05.2005.

9 APPENDIX

9.1 CHOICE OF PARCEL-BASED QUERIES

We assign one query to each parcel of the Schaefer 1000-parcel atlas (Section 3.2; Appendix 9.2). The number of queries sets a trade-off. With one query per voxel, each attention map would be estimated from a single noisy voxel; with one query per large region, a single map would be imposed on a heterogeneous population, and the regions would have to be defined in advance. Parcels sit between these extremes. Every modeled voxel belongs to exactly one parcel query, so the whole measured volume is predicted, with a single additional query for voxels without a parcel label. The parcels are fine enough for differences within and between functional regions to emerge in the attention maps. Because the atlas is defined independently of our fMRI data and functional localizers, any category selectivity that emerges in the attention maps (Appendix 9.13) is not built into the model. Finally, with about 430 queries per subject, each attention map is shared by many voxels and is therefore estimated from many voxels' responses. This follows the observation that finer query granularity improves cross-attention encoding models (Adeli et al., 2026), and the use of one query per Schaefer parcel in a whole-brain cross-attention encoder for static images (Hwang et al., 2025). Here we use the parcels in volumetric space for video, with attention over both space and time.

9.2 DATASET AND PREPROCESSING DETAILS

Preprocessing and voxel selection. We use the BMD release recommended by the dataset authors, preprocessed with fMRIPrep (Esteban et al., 2019) in MNI152NLin2009cAsym volume space at 2.5 mm. We model all voxels in the union of the dataset's group-level mask of cortex with reliable responses to the videos ("BMDGeneral" mask) and the 23 functional ROIs. The single-trial betas were estimated by the dataset authors with GLMsingle and z-scored per voxel within each scanning session. After averaging over the repeats of each clip, we z-score each voxel again separately within the training and the test clips. The second normalization puts training and test targets on the same scale, since the test responses are averaged over more repeats (10 vs. 3) and are therefore less noisy. The prediction target is thus a single repeat-averaged response per voxel and clip; we do not model the response time course within a clip. The noise ceiling released with the dataset follows the procedure of the Natural Scenes Dataset (Allen et al., 2022).

Parcellation. We use the volumetric MNI152 version of the Schaefer atlas, resample it to the BMD voxel grid with nearest-neighbor interpolation and intersect it with each subject's voxel mask. Each parcel with at least one voxel in the mask becomes a query, with the two hemispheres parcellated separately, and the voxels without a parcel label (mostly subcortical and edge voxels) are assigned to one additional query. Every modeled voxel therefore belongs to exactly one query.

Functional ROIs. The 23 ROIs are V1v, V1d, V2v, V2d, V3v, V3d, V3ab, hV4, IPS0 and IPS1-2-3 from the retinotopic atlas of Wang et al. (2015); MT, 7AL, BA2, PFop and PFt from the HCP multimodal parcellation (Glasser et al., 2016); and EBA, LOC, FFA, OFA, STS, PPA, RSC and TOS from the subject-specific functional localizer (bodies $>$ objects for EBA; objects $>$ scrambled objects for LOC; faces $>$ objects for FFA, OFA and STS; scenes $>$ objects for PPA, RSC and TOS).

9.3 IMPLEMENTATION DETAILS

Backbone tokens. We use the V-JEPA 2 ViT-g/16 model with 384 px, 64-frame input. From each 3 s clip we sample 16 frames uniformly and repeat each frame four times to fill the 64-frame input. The $2 \times 16 \times 16$ tubelets of the model give a $32 \times 24 \times 24$ token grid, and we average each pair of adjacent temporal tokens (which cover the same repeated frame), giving $16 \times 24 \times 24$ tokens of dimension $D = 1408$. Each token therefore covers one sampled frame ($\approx$0.19 s) and a 16×16 px patch.

Model size. For a typical subject the joint model has about 47M parameters, of which about 0.6M are queries, 21.8M are the shared attention and FFN block, and 24.5M are the voxel readouts.

Data splits and ensembling. The 102 test clips are held out and used only for the final evaluation. For each of the 10 ensemble members, the 1,000 training clips are split at random into 900 clips

for training and 100 for validation, with a different split and initialization for each member, and the validation clips are used only for early stopping. Because each training clip is held out for validation in about one of the 10 members, the ensemble as a whole uses all training clips. All models, including the ridge baseline, use the same 10 splits.

Optimization. We use AdamW with a learning rate of 10^{-3}, halved every 3 epochs, weight decay 0.05, batches of 8 clips, gradient-norm clipping at 1 and mixed precision. The queries are initialized from $\mathcal{N}(0, 0.02^2)$, and attention dropout of 0.1 is applied during training. Each member is trained for up to 12 epochs with early stopping on validation R^2 (patience 4), and the weights from the best validation epoch are restored (typically the third epoch). All hyperparameters are identical for the joint and constrained models, for all backbones and for all subjects.

Ridge regression. We chose mean pooling for the ridge baseline because it predicted more accurately than ridge regression on the flattened token grid after dimensionality reduction with PCA. Features are standardized using each member's training clips, and the penalty of each voxel is chosen from six values between 10^2 and 10^7 (log-spaced) by efficient leave-one-out cross-validation on the member's 900 training clips.

9.4 ROI CLUSTERING FOR THE REGION GROUPS

To group the ROIs by function rather than by a priori anatomical assignment, we clustered the 23 ROIs by their responses to the videos. For each subject, we averaged the z-scored responses over each ROI's reliable voxels (noise ceiling $\geq 5\%$), giving a matrix of 23 ROIs $\times$ 1,102 clips, and correlated the ROIs across clips. The correlation matrices were Fisher-z averaged over the 10 subjects and hierarchically clustered (distance $1 - r$, average linkage). Cutting the dendrogram into four clusters gives an early visual cluster (V1–V3, V3ab and hV4), a lateral/face/body cluster (OFA, FFA, EBA, MT, LOC and STS), a scene cluster (RSC, PPA, TOS, IPS0 and IPS1-2-3) and a parietal cluster (PFt, BA2, PFop and 7AL). This departs from a conventional ventral/lateral/dorsal division: FFA, OFA and LOC cluster with EBA, MT and STS rather than with the scene-selective ROIs, and IPS0 and IPS1-2-3 merge with each other first (distance $\approx$0.25) and then join the scene ROIs.

The region groups used in the main text follow these clusters except that IPS0 and IPS1-2-3 are placed in the parietal group because they lie in posterior parietal cortex; this exception is not data-driven. With the clustering assignment instead (both IPS ROIs in the scene group), only the scene and parietal groups change, and the joint model still outperforms the factorized, spatial, mean-pool and ridge models in both (two-sided paired t-tests across the 10 subjects, Benjamini–Hochberg FDR-corrected over these 8 tests). The differences in signed r^2 are $+0.0036$, $+0.0029$, $+0.0054$ and $+0.0073$ in the scene group ($t(9) = 4.08$, 2.50, 3.69 and 4.70; $p_{\mathrm{FDR}} = 0.007$, 0.034, 0.010 and 0.005) and $+0.0040$, $+0.0035$, $+0.0059$ and $+0.0206$ in the parietal group ($t(9) = 3.47$, 2.95, 2.64 and 6.87; $p_{\mathrm{FDR}} = 0.011$, 0.022, 0.031 and < 0.001).

In the scene-selective ROIs alone (PPA, RSC and TOS), the joint model outperforms the factorized model in every subject ($+0.0041$ signed r^2, $t(9) = 6.16$, $p_{\mathrm{FDR}} < 0.001$, 10/10 subjects) but matches ridge regression on mean-pooled features ($+0.0007$, $t(9) = 0.42$, $p_{\mathrm{FDR}} = 0.75$, 8/10; two-sided paired t-tests, FDR-corrected over the 24 tests of Table 3). The advantage over ridge in the scene group of the clustering assignment therefore comes from the IPS ROIs, where the joint model gains substantially over ridge ($+0.018$ signed r^2 in both IPS0 and IPS1-2-3; $t(9) = 5.34$ and 5.55, both $p < 0.001$, two-sided paired t-tests, uncorrected; 10/10 subjects). This is consistent with scene-selective cortex responding to the global layout of a scene, which is already captured by features averaged over the whole clip, so that routing to specific locations and moments adds little for these ROIs.

9.5 BACKBONE AND LAYER SELECTION

We chose the backbone and layer with voxelwise ridge regression on mean-pooled features from each layer of several video and image models, fit and evaluated in the same way as the ridge baseline (Section 3.3). Table 2 gives the best layer of each backbone and its whole-brain R^2 on the test clips, averaged over all voxels in each subject, then over the 10 subjects. The V-JEPA 2 models are the most accurate, and the three V-JEPA 2 sizes are nearly tied.

Table 2: Best layer of each backbone in the ridge-regression layer sweep (mean-pooled features; test R^2, mean over all voxels in 10 subjects).

Backbone	**Best layer**	R^2
V-JEPA 2 ViT-g (Assran et al., 2025)	32	.134
V-JEPA 2 ViT-H	24	.133
V-JEPA 2 ViT-L	18	.132
V-JEPA 2 ViT-L, fine-tuned on SSv2	18	.125
VideoMAE v2 ViT-g (Wang et al., 2023b)	20	.109
V-JEPA 2.1 ViT-g **[cite]**	44	.108
Perception Encoder Spatial-L (Bolya et al., 2025)	20	.100
Perception Encoder Spatial-G	38	.098
DINOv3 ViT-7B, per frame (Siméoni et al., 2025)	32	.092

9.6 Subject-level tests for Figure 2

Table 3: Joint model minus each comparison model in region-mean signed r^2 on the 102 test clips (mean $\pm$ SEM over 10 subjects), with two-sided paired t-tests across subjects ($df = 9$). p is uncorrected; p_{FDR} is Benjamini–Hochberg corrected over all 24 tests; values below 0.05 are in bold. Positive Δ indicates that the joint model performed better. "Joint better" column indicates how many individual subjects have a numerical advantage for the joint model in each comparison. d_z: mean divided by the standard deviation of the per-subject differences.

Region	**Comparison**	**Δ signed r^2**	$t(9)$	p	p_{FDR}	**Joint better**	d_z
All voxels	factorized	$+0.0043 \pm 0.0008$	5.39	4.4e-04	**8.7e-04**	10/10	1.71
	spatial	$+0.0040 \pm 0.0007$	5.85	2.4e-04	**5.8e-04**	10/10	1.85
	mean pool	$+0.0103 \pm 0.0017$	6.16	1.7e-04	**4.5e-04**	10/10	1.95
	ridge	$+0.0162 \pm 0.0019$	8.52	1.3e-05	**1.6e-04**	10/10	2.69
Early visual	factorized	-0.0000 ± 0.0010	-0.03	0.976	0.976	5/10	-0.01
	spatial	$+0.0000 \pm 0.0009$	0.04	0.966	0.976	4/10	0.01
	mean pool	$+0.0122 \pm 0.0025$	4.84	9.2e-04	**0.001**	9/10	1.53
	ridge	$+0.0108 \pm 0.0028$	3.83	0.004	**0.006**	9/10	1.21
Lateral / face / body	factorized	$+0.0084 \pm 0.0013$	6.60	1.0e-04	**3.4e-04**	10/10	2.09
	spatial	$+0.0086 \pm 0.0011$	7.79	2.7e-05	**1.6e-04**	10/10	2.46
	mean pool	$+0.0145 \pm 0.0021$	6.80	7.9e-05	**3.2e-04**	10/10	2.15
	ridge	$+0.0216 \pm 0.0022$	10.04	3.5e-06	**8.3e-05**	10/10	3.17
Scene	factorized	$+0.0041 \pm 0.0007$	6.16	1.7e-04	**4.5e-04**	10/10	1.95
	spatial	$+0.0029 \pm 0.0011$	2.72	0.024	**0.027**	8/10	0.86
	mean pool	$+0.0033 \pm 0.0011$	3.07	0.013	**0.017**	8/10	0.97
	ridge	$+0.0007 \pm 0.0017$	0.42	0.687	0.749	8/10	0.13
Parietal	factorized	$+0.0037 \pm 0.0009$	3.99	0.003	**0.005**	9/10	1.26
	spatial	$+0.0034 \pm 0.0012$	2.93	0.017	**0.020**	9/10	0.93
	mean pool	$+0.0067 \pm 0.0022$	3.11	0.013	**0.017**	10/10	0.98
	ridge	$+0.0197 \pm 0.0028$	6.97	6.5e-05	**3.1e-04**	10/10	2.21
Outside ROIs	factorized	$+0.0040 \pm 0.0008$	4.98	7.6e-04	**0.001**	10/10	1.58
	spatial	$+0.0035 \pm 0.0007$	4.98	7.6e-04	**0.001**	10/10	1.58
	mean pool	$+0.0094 \pm 0.0017$	5.59	3.4e-04	**7.4e-04**	10/10	1.77
	ridge	$+0.0163 \pm 0.0020$	8.12	2.0e-05	**1.6e-04**	10/10	2.57

9.7 Voxelwise accuracy and differences in each subject

Figure 5 shows the voxelwise accuracy of the joint model, averaged over subjects and in each subject, and Figure 6 shows the difference between the joint model and each comparison model in each subject; Fig. 2A shows the average of these differences over subjects. All modeled voxels are shown, without a noise-ceiling threshold, on the fsaverage flatmap used in Fig. 2A. The single-subject maps

are noisier than the average, but the region-level advantage of the joint model in lateral/face/body regions holds in every subject (Table 3).

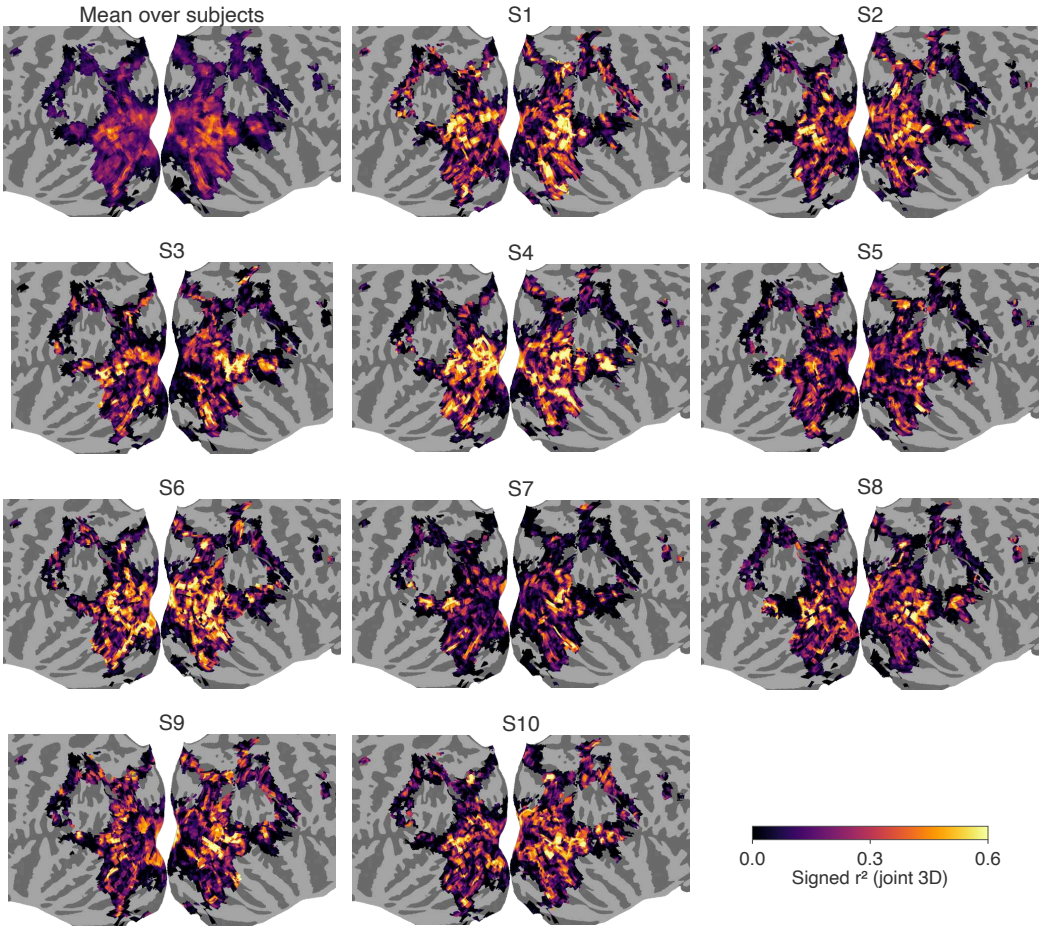

Figure 5: Voxelwise accuracy of the joint model (V-JEPA 2 ViT-g, layer 32; signed r^2 on the 102 test clips; all modeled voxels), averaged over the 10 subjects (top left) and in each subject, on the fsaverage flatmap. Values outside 0–0.6 are shown at the ends of the color scale.

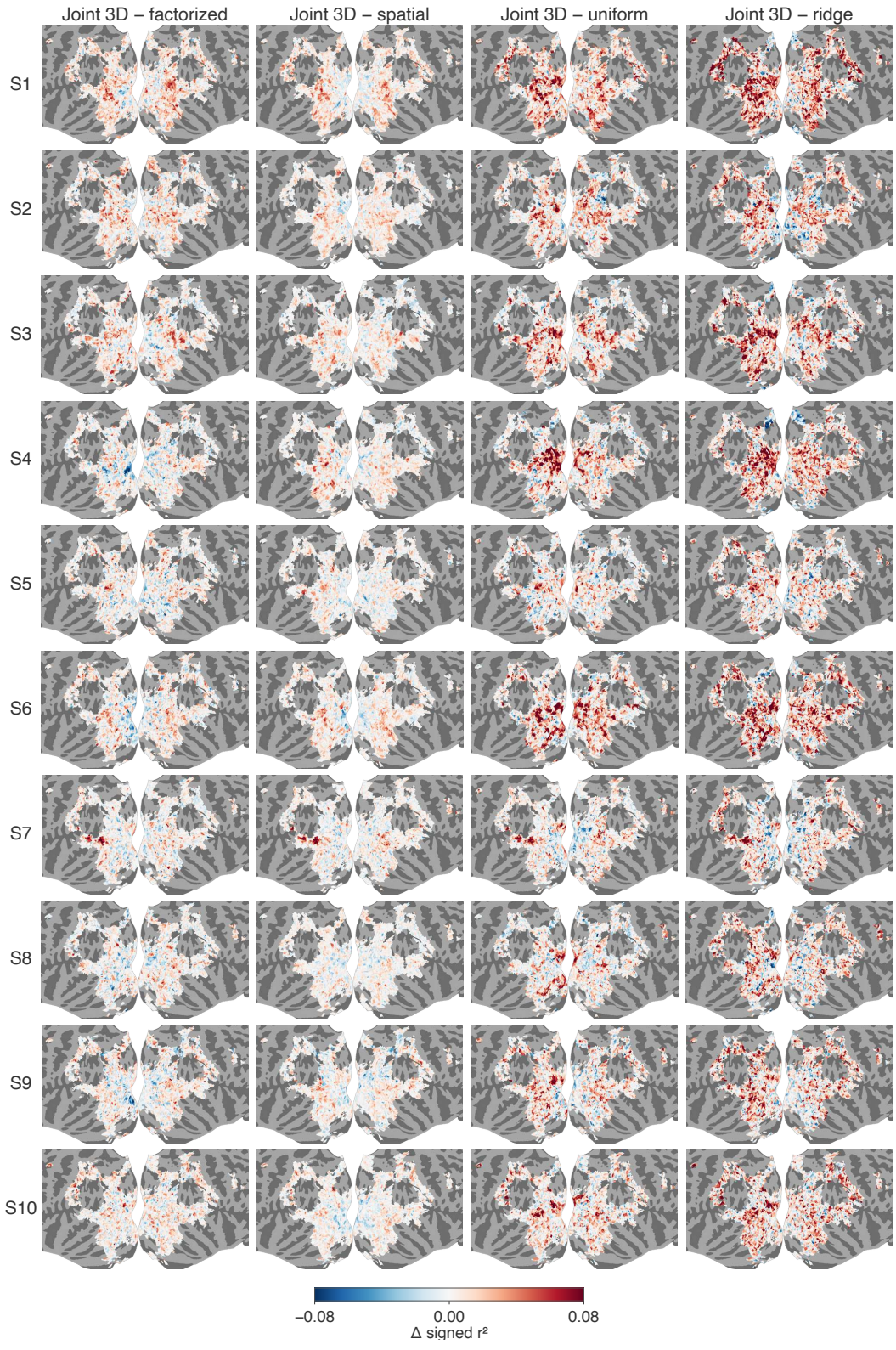

Figure 6: Joint model minus the factorized, spatial and mean-pool ("uniform") models and ridge regression on mean-pooled features in voxelwise signed r^2, in each subject (rows), on the flatmap of Fig. 2A. All panels share one color scale; differences beyond ± 0.08 are shown at its ends.

9.8 Spatial routing compared with mean pooling and ridge regression

Cross-attention encoders of static images improve on linear encoding models by routing each region's readout to stimulus-dependent spatial locations (Adeli et al., 2026). Table 4 tests the video analogue of this result, the spatial model (temporal factor uniform, spatial factor routed), against the mean-pool model and against ridge regression on mean-pooled features, for every backbone and layer of the routing sweep (Appendix 9.11). The spatial model is more accurate than both in all regions for every configuration except in scene-selective cortex, where it is not significantly different from either for V-JEPA 2 layer 32 and VideoMAE v2 layer 20, and for the last stage of CLIP, where the spatial model is less accurate than ridge regression in four of the six regions and less accurate than mean pooling in lateral/face/body regions. Of the 96 tests, 83 favor the spatial model at $p_{\mathrm{FDR}} < 0.05$.

9.9 Contributions of the readout components

Besides routing, the joint model differs from ridge regression on mean-pooled features in three ways: it is trained by gradient descent, it passes the tokens through value and output projections shared by all voxels, and it applies a feedforward network (FFN) to each parcel's representation before the voxel readouts. To separate these components, we trained three further models on V-JEPA 2 ViT-g layer 32, with the same data splits, ensembling, optimization and early stopping as the joint model (Appendix 9.3):

- **Linear readout:** the voxel readouts applied directly to the mean-pooled tokens and trained by gradient descent, the counterpart of ridge regression.
- **Uniform, no FFN:** the mean-pool model without the FFN, so that every parcel's representation is the output projection of the mean of the value projections. Because both projections are linear, this model is still linear in the mean-pooled tokens; it differs from the linear readout only in its parametrization.
- **Joint, no FFN:** the joint model without the FFN.

With ridge regression, the mean-pool model and the joint model, these form a sequence that adds one component at a time (Fig. 7, Table 5; two-sided paired t-tests across the 10 subjects, $df = 9$, FDR-corrected over the 30 tests of the table). Routing contributes the most. Without the FFN, joint routing improves on uniform attention in all six regions and in every subject ($+0.0115$ across all voxels; $t(9) = 10.51$, $p_{\mathrm{FDR}} < 0.001$), as much as it does with the FFN (joint minus mean pool, $+0.0103$; Table 3). The FFN adds less, with uniform attention ($+0.0039$; $t(9) = 5.29$, $p_{\mathrm{FDR}} = 0.001$) and with joint routing ($+0.0026$; $t(9) = 4.74$, $p_{\mathrm{FDR}} = 0.002$), and in early visual cortex it adds nothing significant in either case. Gradient descent by itself does not explain the advantage of the attention models: the linear readout is less accurate than ridge regression in every region (-0.0045 across all voxels; $t(9) = -7.11$, $p_{\mathrm{FDR}} < 0.001$). The shared projections make up this deficit and more in lateral/face/body and parietal regions, but not in early visual or scene regions.

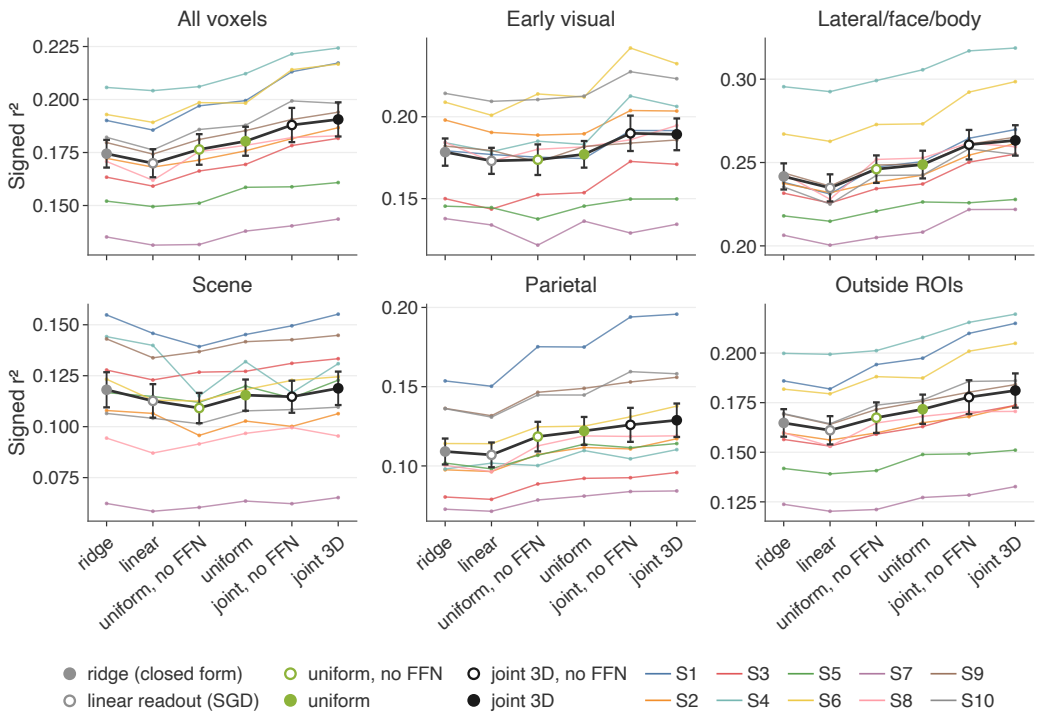

Figure 7: Contributions of the readout components (V-JEPA 2 ViT-g, layer 32; region-mean signed r^2 on the 102 test clips; voxels with noise ceiling $\geq 5\%$). From left to right: ridge regression on mean-pooled features, a linear readout of the mean-pooled features trained by gradient descent, the mean-pool model without and with the FFN ("uniform"), and the joint model without and with the FFN. Open markers: models without an FFN. Colored lines: individual subjects; markers: mean $\pm$ SEM over subjects.

Table 4: Spatial model minus the mean-pool model and minus ridge regression on mean-pooled features, in region-mean signed r^2 on the 102 test clips (Δ in units of 10^{-3}, mean over 10 subjects; voxels with noise ceiling $\geq 5\%$). $t(9)$: two-sided paired t-test across subjects. p_{FDR}: Benjamini–Hochberg corrected over all 96 tests in the table; values below 0.05 are in bold. n_+: subjects in which the spatial model is better.

	Spatial − mean pool				Spatial − ridge			
Region	Δ	$t(9)$	p_{FDR}	n_+	Δ	$t(9)$	p_{FDR}	n_+
V-JEPA 2 ViT-g, layer 32 of 40								
All voxels	+6.3	5.07	**8.9e-4**	9/10	+12.2	8.26	**3.3e-5**	10/10
Early visual	+12.2	5.58	**4.7e-4**	10/10	+10.8	4.24	**0.003**	9/10
Lateral/face/body	+5.9	3.97	**0.004**	9/10	+13.0	8.73	**2.5e-5**	10/10
Scene	+0.5	0.37	0.734	6/10	−2.2	-1.11	0.313	5/10
Parietal	+3.3	2.38	**0.045**	7/10	+16.3	8.09	**3.7e-5**	10/10
Outside ROIs	+5.9	4.60	**0.002**	9/10	+12.8	8.25	**3.3e-5**	10/10
V-JEPA 2 ViT-g, layer 4 of 40								
All voxels	+14.3	11.29	**5.7e-6**	10/10	+18.4	18.77	**1.3e-6**	10/10
Early visual	+18.7	9.78	**1.2e-5**	10/10	+18.2	11.80	**4.9e-6**	10/10
Lateral/face/body	+14.7	9.80	**1.2e-5**	10/10	+21.2	14.05	**2.3e-6**	10/10
Scene	+10.4	5.88	**3.4e-4**	10/10	+14.1	8.34	**3.2e-5**	10/10
Parietal	+7.9	8.21	**3.4e-5**	10/10	+12.2	11.81	**4.9e-6**	10/10
Outside ROIs	+14.4	9.94	**1.2e-5**	10/10	+18.9	16.60	**1.4e-6**	10/10
VideoMAE v2 ViT-g, layer 20 of 40								
All voxels	+14.8	8.91	**2.3e-5**	10/10	+24.6	10.90	**6.9e-6**	10/10
Early visual	+20.6	6.36	**1.9e-4**	10/10	+23.4	5.22	**7.3e-4**	10/10
Lateral/face/body	+19.6	12.41	**4.0e-6**	10/10	+34.5	16.20	**1.4e-6**	10/10
Scene	+2.0	1.61	0.152	8/10	+1.4	0.76	0.483	5/10
Parietal	+8.5	4.81	**0.001**	10/10	+21.6	7.37	**7.1e-5**	10/10
Outside ROIs	+13.6	7.27	**7.8e-5**	10/10	+23.7	10.00	**1.1e-5**	10/10
VideoMAE v2 ViT-g, layer 4 of 40								
All voxels	+19.3	8.25	**3.3e-5**	10/10	+25.6	9.45	**1.5e-5**	10/10
Early visual	+32.4	6.40	**1.9e-4**	10/10	+30.6	5.76	**3.8e-4**	10/10
Lateral/face/body	+22.4	8.70	**2.5e-5**	10/10	+33.2	12.43	**4.0e-6**	10/10
Scene	+10.2	6.14	**2.5e-4**	10/10	+17.6	9.01	**2.1e-5**	10/10
Parietal	+8.6	8.54	**2.8e-5**	10/10	+16.1	14.19	**2.3e-6**	10/10
Outside ROIs	+17.3	7.21	**8.2e-5**	10/10	+23.9	8.68	**2.5e-5**	10/10
DINOv2 ViT-B/14, block 12 of 12								
All voxels	+22.9	10.71	**7.2e-6**	10/10	+27.1	9.85	**1.2e-5**	10/10
Early visual	+32.0	8.52	**2.8e-5**	10/10	+31.0	6.79	**1.3e-4**	10/10
Lateral/face/body	+27.2	13.79	**2.3e-6**	10/10	+34.2	14.33	**2.3e-6**	10/10
Scene	+9.4	11.54	**5.1e-6**	10/10	+8.6	4.50	**0.002**	10/10
Parietal	+11.5	5.46	**5.4e-4**	10/10	+19.0	7.02	**9.9e-5**	10/10
Outside ROIs	+22.0	9.59	**1.4e-5**	10/10	+26.4	8.81	**2.4e-5**	10/10
DINOv2 ViT-B/14, block 2 of 12								
All voxels	+23.0	10.77	**7.1e-6**	10/10	+25.5	13.75	**2.3e-6**	10/10
Early visual	+51.6	11.75	**4.9e-6**	10/10	+40.1	10.58	**7.7e-6**	10/10
Lateral/face/body	+20.4	10.80	**7.1e-6**	10/10	+29.1	17.66	**1.3e-6**	10/10
Scene	+6.7	4.26	**0.003**	9/10	+14.5	6.48	**1.7e-4**	10/10
Parietal	+5.5	6.65	**1.5e-4**	10/10	+11.5	12.45	**4.0e-6**	10/10
Outside ROIs	+20.8	8.83	**2.4e-5**	10/10	+23.6	12.01	**4.9e-6**	10/10
CLIP RN50, stage 4 of 4								
All voxels	+0.0	0.01	0.991	5/10	−6.7	-2.73	**0.026**	3/10
Early visual	+15.7	4.15	**0.003**	9/10	+10.0	2.65	**0.029**	8/10
Lateral/face/body	−15.2	-3.46	**0.008**	1/10	−22.2	-5.78	**3.8e-4**	0/10
Scene	+6.5	4.20	**0.003**	10/10	−5.3	-4.14	**0.003**	2/10
Parietal	+2.6	0.96	0.378	7/10	−2.9	-1.30	0.241	4/10
Outside ROIs	+0.3	0.13	0.910	5/10	−6.4	-2.97	**0.018**	3/10
CLIP RN50, stage 2 of 4								
All voxels	+28.7	9.29	**1.7e-5**	10/10	+41.9	11.62	**5.1e-6**	10/10
Early visual	+50.7	7.48	**6.5e-5**	10/10	+54.8	7.56	**6.0e-5**	10/10
Lateral/face/body	+35.0	12.71	**4.0e-6**	10/10	+53.8	15.14	**2.0e-6**	10/10
Scene	+5.2	3.35	**0.010**	10/10	+17.8	11.27	**5.7e-6**	10/10
Parietal	+12.3	7.65	**5.6e-5**	10/10	+24.6	10.38	**8.7e-6**	10/10
Outside ROIs	+25.4	8.04	**3.9e-5**	10/10	+39.1	10.95	**6.9e-6**	10/10

Table 5: Each model of Fig. 7 minus the model it extends, in region-mean signed r^2 (mean $\pm$ SEM over 10 subjects), with two-sided paired t-tests across subjects ($df = 9$). p is uncorrected; p_{FDR} is Benjamini–Hochberg corrected over all 30 tests; values below 0.05 are in bold. n_+: subjects in which the first model is better.

Contrast	Region	Δ signed r^2	$t(9)$	p	p_{FDR}	n_+
Linear readout (SGD) − ridge	All voxels	-0.0045 ± 0.0006	-7.11	5.6e-05	**4.2e-04**	0/10
	Early visual	-0.0053 ± 0.0010	-5.37	4.5e-04	**0.001**	0/10
	Lateral/face/body	-0.0069 ± 0.0012	-5.88	2.3e-04	**0.001**	0/10
	Scene	-0.0055 ± 0.0010	-5.26	5.2e-04	**0.001**	0/10
	Parietal	-0.0021 ± 0.0008	-2.53	0.032	**0.039**	1/10
	Outside ROIs	-0.0037 ± 0.0006	-6.71	8.8e-05	**4.4e-04**	0/10
Uniform, no FFN − linear readout	All voxels	$+0.0065 \pm 0.0014$	4.53	0.001	**0.002**	10/10
	Early visual	$+0.0007 \pm 0.0026$	0.27	0.790	0.790	5/10
	Lateral/face/body	$+0.0112 \pm 0.0020$	5.72	2.9e-04	**0.001**	10/10
	Scene	-0.0035 ± 0.0028	-1.24	0.248	0.265	4/10
	Parietal	$+0.0115 \pm 0.0022$	5.33	4.8e-04	**0.001**	9/10
	Outside ROIs	$+0.0063 \pm 0.0013$	4.76	0.001	**0.002**	10/10
Uniform − uniform, no FFN	All voxels	$+0.0039 \pm 0.0007$	5.29	5.0e-04	**0.001**	9/10
	Early visual	$+0.0033 \pm 0.0017$	1.88	0.093	0.104	7/10
	Lateral/face/body	$+0.0028 \pm 0.0007$	3.93	0.003	**0.005**	10/10
	Scene	$+0.0063 \pm 0.0013$	4.70	0.001	**0.002**	10/10
	Parietal	$+0.0036 \pm 0.0010$	3.51	0.007	**0.009**	8/10
	Outside ROIs	$+0.0043 \pm 0.0008$	5.50	3.8e-04	**0.001**	9/10
Joint, no FFN − uniform, no FFN	All voxels	$+0.0115 \pm 0.0011$	10.51	2.4e-06	**3.3e-05**	10/10
	Early visual	$+0.0161 \pm 0.0024$	6.77	8.2e-05	**4.4e-04**	10/10
	Lateral/face/body	$+0.0147 \pm 0.0014$	10.66	2.1e-06	**3.3e-05**	10/10
	Scene	$+0.0055 \pm 0.0010$	5.38	4.4e-04	**0.001**	10/10
	Parietal	$+0.0074 \pm 0.0016$	4.59	0.001	**0.002**	10/10
	Outside ROIs	$+0.0103 \pm 0.0010$	10.09	3.3e-06	**3.3e-05**	10/10
Joint − joint, no FFN	All voxels	$+0.0026 \pm 0.0006$	4.74	0.001	**0.002**	9/10
	Early visual	-0.0006 ± 0.0017	-0.37	0.720	0.744	4/10
	Lateral/face/body	$+0.0026 \pm 0.0011$	2.29	0.048	0.055	8/10
	Scene	$+0.0041 \pm 0.0016$	2.62	0.028	**0.035**	9/10
	Parietal	$+0.0029 \pm 0.0009$	3.34	0.009	**0.011**	9/10
	Outside ROIs	$+0.0033 \pm 0.0006$	5.55	3.6e-04	**0.001**	10/10

9.10 ROUTING IN EACH FUNCTIONAL ROI

Figure 8 repeats Fig. 2B for each of the 23 functional ROIs, and Table 6 compares the joint model with the factorized, spatial and mean-pool models and ridge regression in each ROI (two-sided paired t-tests across the 10 subjects, $df = 9$, Benjamini–Hochberg FDR-corrected over the 92 tests of the table). The ROIs follow the pattern of their region groups. In each of the eight early visual ROIs, the joint model does not differ significantly from the factorized or the spatial model, and it is better than mean pooling in seven of them. In the lateral/face/body group, it is better than the spatial model in all six ROIs and than the factorized model in all but OFA ($p_{\mathrm{FDR}} = 0.054$), with the largest gains over the factorized model in MT ($+0.016$; $t(9) = 6.54$, $p_{\mathrm{FDR}} < 0.001$) and EBA ($+0.013$; $t(9) = 8.50$, $p_{\mathrm{FDR}} < 0.001$). Among the scene-selective ROIs, the joint model is better than every comparison model in PPA, but not significantly better than any of them in TOS, and in RSC it is better only than the factorized model.

Table 6: Joint model minus each comparison model in ROI-mean signed r^2 on the 102 test clips (Δ in units of 10^{-3}, mean over 10 subjects; voxels with noise ceiling $\geq 5\%$). $t(9)$: two-sided paired t-test across subjects. p_{FDR}: Benjamini–Hochberg corrected over all 92 tests in the table; values below 0.05 are in bold. n_+: subjects (of 10) in which the joint model is better.

	Joint − factorized				**Joint − spatial**				**Joint − mean pool**				**Joint − ridge**			
ROI	**Δ**	**$t(9)$**	**p_{FDR}**	**n_+**	**Δ**	**$t(9)$**	**p_{FDR}**	**n_+**	**Δ**	**$t(9)$**	**p_{FDR}**	**n_+**	**Δ**	**$t(9)$**	**p_{FDR}**	**n_+**
Early visual																
V1v	−1.7	-1.26	0.281	5	−1.7	-1.53	0.203	3	+8.3	3.38	**0.017**	8	+3.8	1.40	0.240	6
V1d	−0.8	-0.59	0.590	4	−1.9	-2.26	0.072	3	+13.9	4.30	**0.006**	9	+12.0	3.63	**0.013**	8
V2v	−0.5	-0.40	0.717	4	−1.5	-1.84	0.130	4	+5.5	2.18	0.080	6	+2.5	0.95	0.402	7
V2d	−1.8	-1.23	0.290	4	−0.7	-0.68	0.546	4	+18.3	5.92	**0.001**	9	+18.0	5.38	**0.002**	9
V3v	+1.2	0.92	0.412	6	+0.3	0.31	0.772	5	+8.3	4.59	**0.004**	9	+5.6	3.01	**0.026**	8
V3d	−1.6	-1.10	0.339	4	+0.2	0.26	0.803	6	+18.7	4.08	**0.007**	8	+18.5	3.51	**0.015**	9
V3ab	+2.5	1.44	0.229	7	+2.7	1.62	0.177	8	+9.5	3.22	**0.021**	8	+11.7	3.34	**0.018**	8
hV4	+3.0	1.95	0.112	8	+3.9	2.25	0.072	8	+16.0	5.14	**0.002**	10	+14.0	3.47	**0.016**	10
Lateral / face / body																
MT	+15.6	6.54	**9.8e-4**	10	+12.3	6.22	**0.001**	10	+24.7	6.27	**0.001**	10	+30.7	8.31	**2.8e-4**	10
EBA	+12.7	8.50	**2.8e-4**	10	+13.3	8.19	**2.8e-4**	10	+22.3	5.73	**0.001**	10	+27.9	7.71	**3.4e-4**	10
FFA	+5.1	3.29	**0.019**	8	+5.3	3.70	**0.012**	9	+9.1	4.32	**0.006**	9	+16.6	8.16	**2.8e-4**	10
OFA	+3.7	2.49	0.054	8	+4.7	4.06	**0.007**	10	+4.8	3.28	**0.019**	8	+12.7	4.15	**0.007**	9
LOC	+9.4	4.86	**0.003**	9	+9.9	6.30	**0.001**	10	+16.9	6.17	**0.001**	10	+24.0	8.87	**2.8e-4**	10
STS	+4.7	5.00	**0.003**	10	+4.4	5.68	**0.001**	10	+6.2	5.81	**0.001**	10	+14.0	11.67	**9.0e-5**	10
Scene																
PPA	+3.5	4.54	**0.005**	10	+3.5	3.02	**0.026**	7	+6.7	4.41	**0.005**	9	+7.3	3.44	**0.016**	9
RSC	+4.5	3.93	**0.009**	9	+2.7	2.54	0.051	8	+1.3	0.96	0.402	5	−3.4	-1.72	0.156	4
TOS	+2.0	1.04	0.365	6	+1.1	0.60	0.587	7	+4.7	1.19	0.303	8	+7.4	2.41	0.060	8
Parietal																
IPS0	+4.1	1.93	0.114	7	+4.6	2.98	**0.027**	9	+11.4	4.23	**0.006**	10	+17.9	5.34	**0.002**	10
IPS1-2-3	+1.7	1.32	0.264	5	+1.2	0.74	0.510	6	+5.3	1.98	0.108	7	+17.5	5.55	**0.002**	10
7AL	+5.3	3.72	**0.012**	9	+1.9	1.33	0.261	5	+4.6	2.53	0.051	8	+21.8	8.04	**2.8e-4**	10
BA2	+3.3	2.35	0.066	8	+3.6	3.14	**0.022**	9	+5.7	2.25	0.072	7	+18.6	6.27	**0.001**	10
PFop	+3.5	2.67	**0.042**	9	+2.9	2.97	**0.027**	9	+5.6	3.17	**0.022**	8	+17.8	7.49	**3.8e-4**	10
PFt	+4.5	3.11	**0.023**	8	+5.5	2.83	**0.033**	10	+7.5	2.28	0.072	8	+25.2	4.65	**0.004**	10

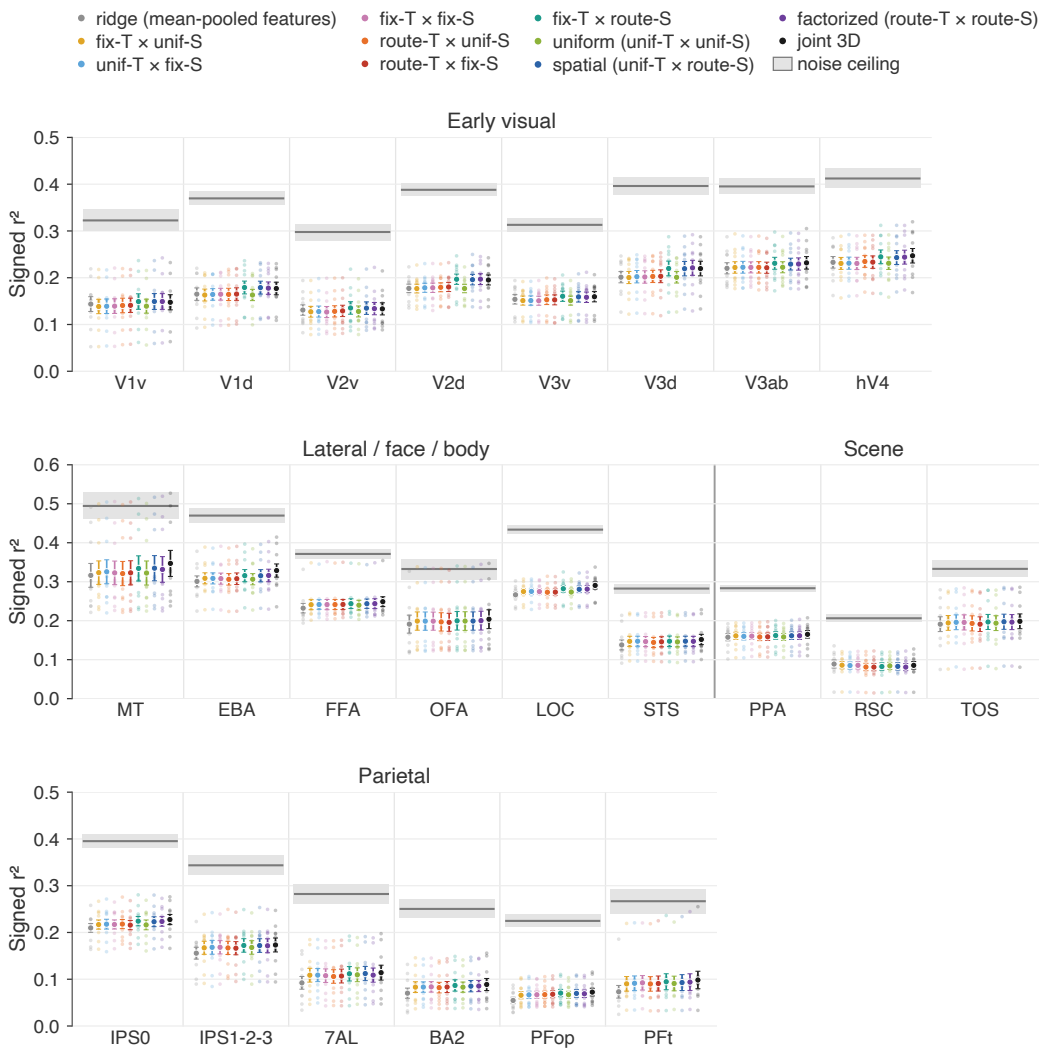

Figure 8: Region-mean signed r^2 of every routing configuration, the joint model and ridge regression in each functional ROI (V-JEPA 2 ViT-g, layer 32; 102 test clips; voxels with noise ceiling $\geq 5\%$), grouped by region as in Fig. 2. Large dots: mean $\pm$ SEM over subjects; small dots: individual subjects; grey band: noise ceiling (mean $\pm$ SEM).

9.11 ROUTING SWEEP

Table 7 gives the region-mean signed r^2 of every routing configuration and of ridge regression for two layers of each of four backbones, all fit and evaluated as for V-JEPA 2 layer 32 (Section 3.3). Figure 9 summarizes the table as the difference between the joint model and the factorized, spatial and mean-pool models and ridge regression. Figures 10–16 show every configuration other than V-JEPA 2 layer 32 in the format of Fig. 2B,C.

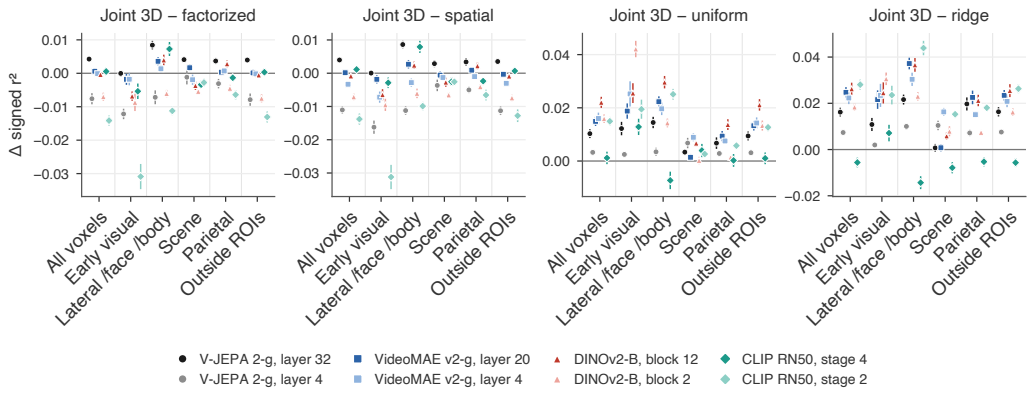

Figure 9: Joint model minus the factorized, spatial and mean-pool ("uniform") models and ridge regression on mean-pooled features of the same layer, in region-mean signed r^2 on the 102 test clips, for every backbone and layer of the routing sweep (mean ± SEM over 10 subjects; voxels with noise ceiling $\geq 5\%$).

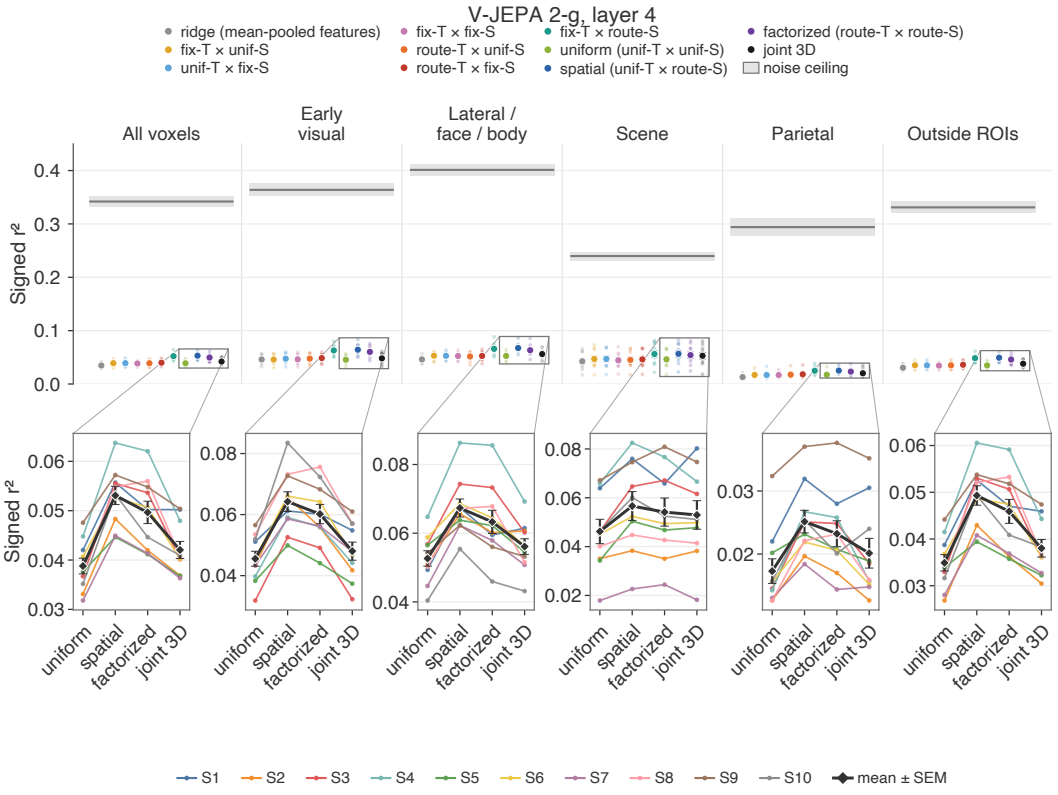

Figure 10: Fig. 2B,C for V-JEPA 2 ViT-g, layer 4 of 40: region-mean signed r^2 of every routing configuration, the joint model and ridge regression (top), and of the mean-pool, spatial, factorized and joint models in each subject (bottom).

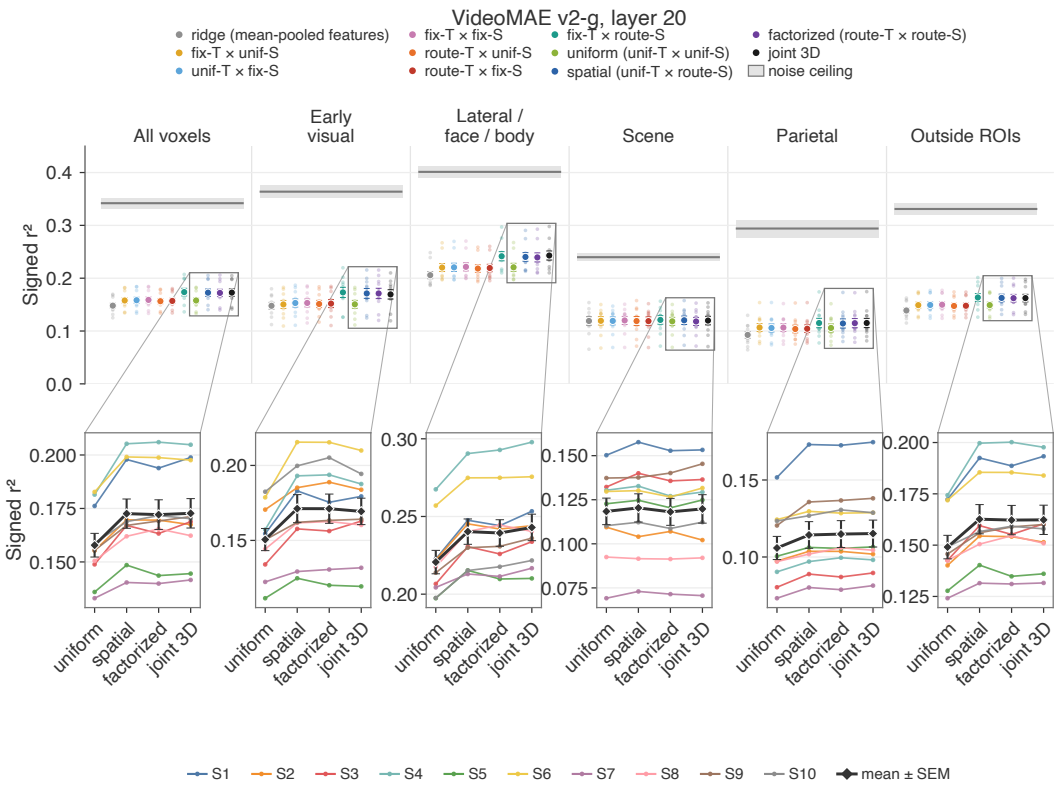

Figure 11: As Fig. 10, for VideoMAE v2 ViT-g, layer 20 of 40.

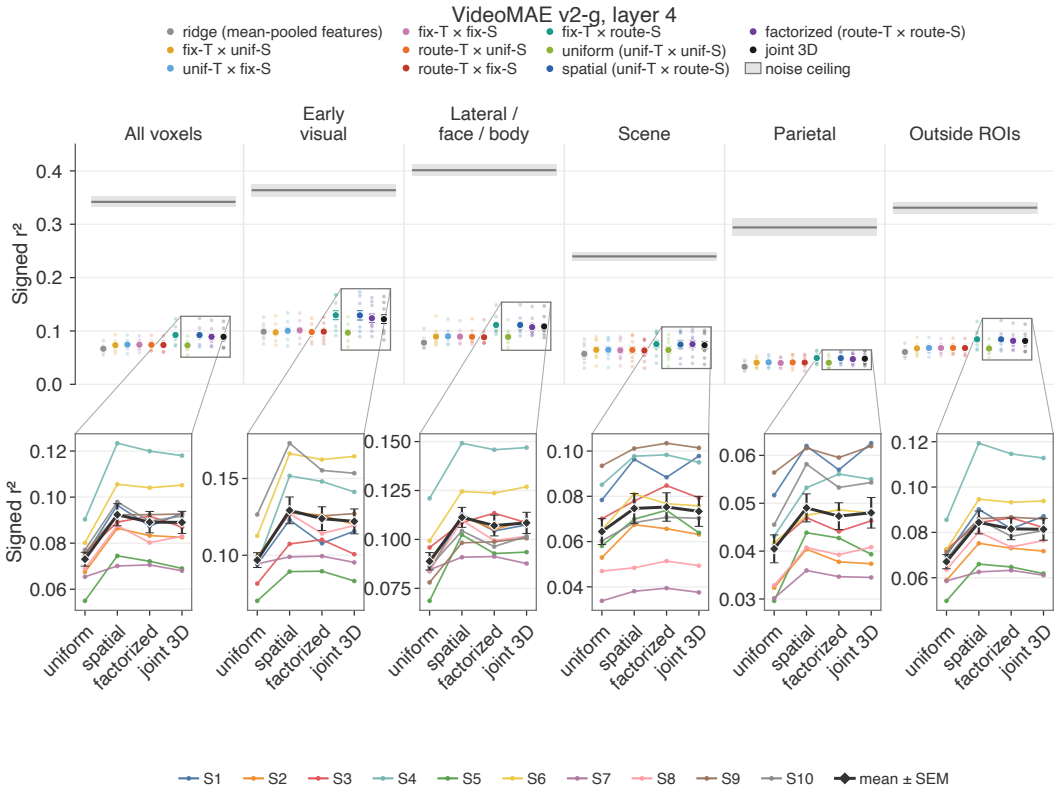

Figure 12: As Fig. 10, for VideoMAE v2 ViT-g, layer 4 of 40.

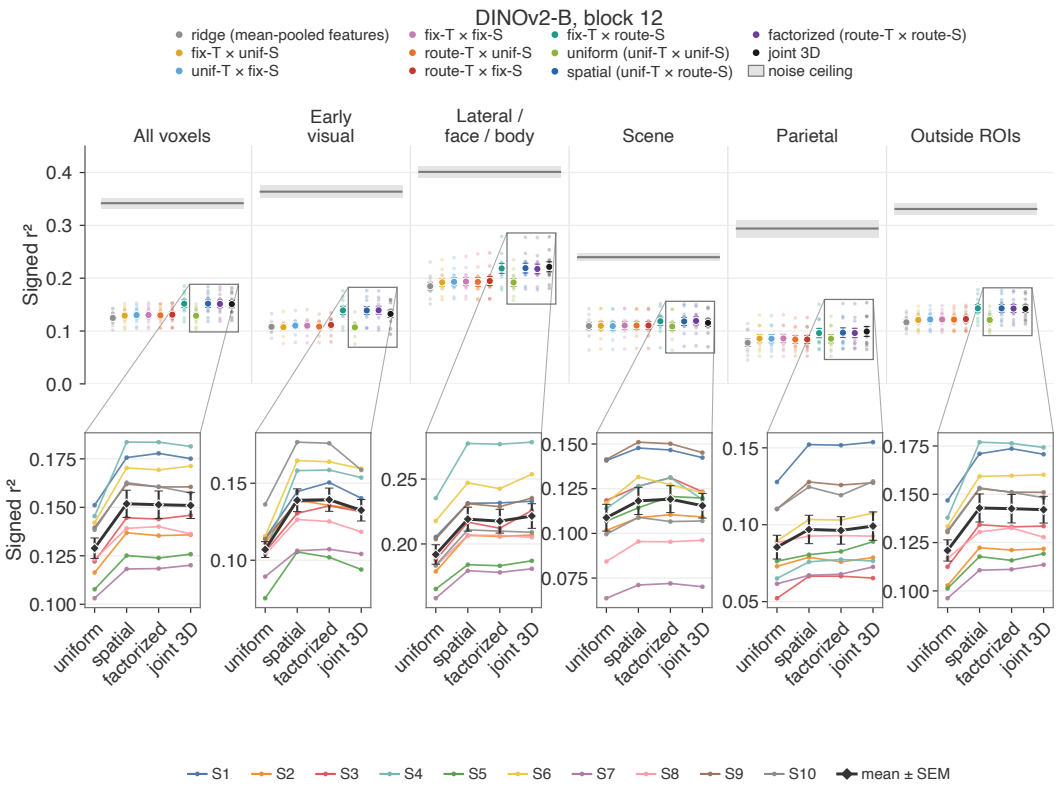

Figure 13: As Fig. 10, for DINOv2 ViT-B/14, block 12 of 12.

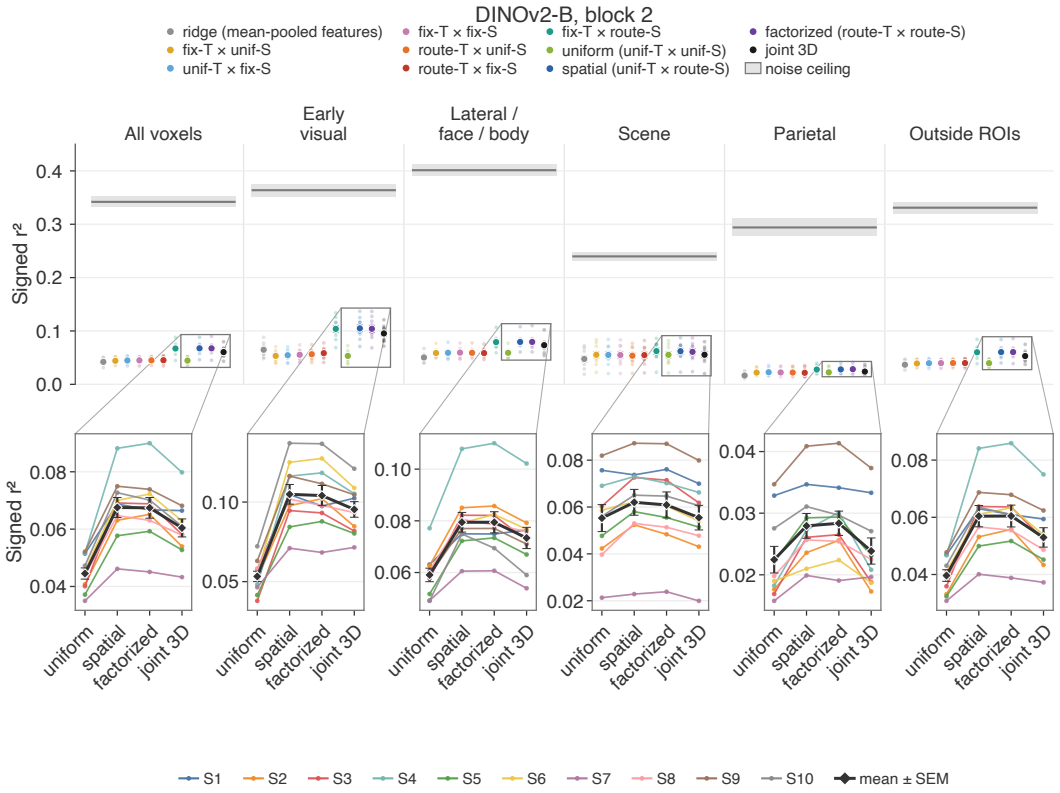

Figure 14: As Fig. 10, for DINOv2 ViT-B/14, block 2 of 12.

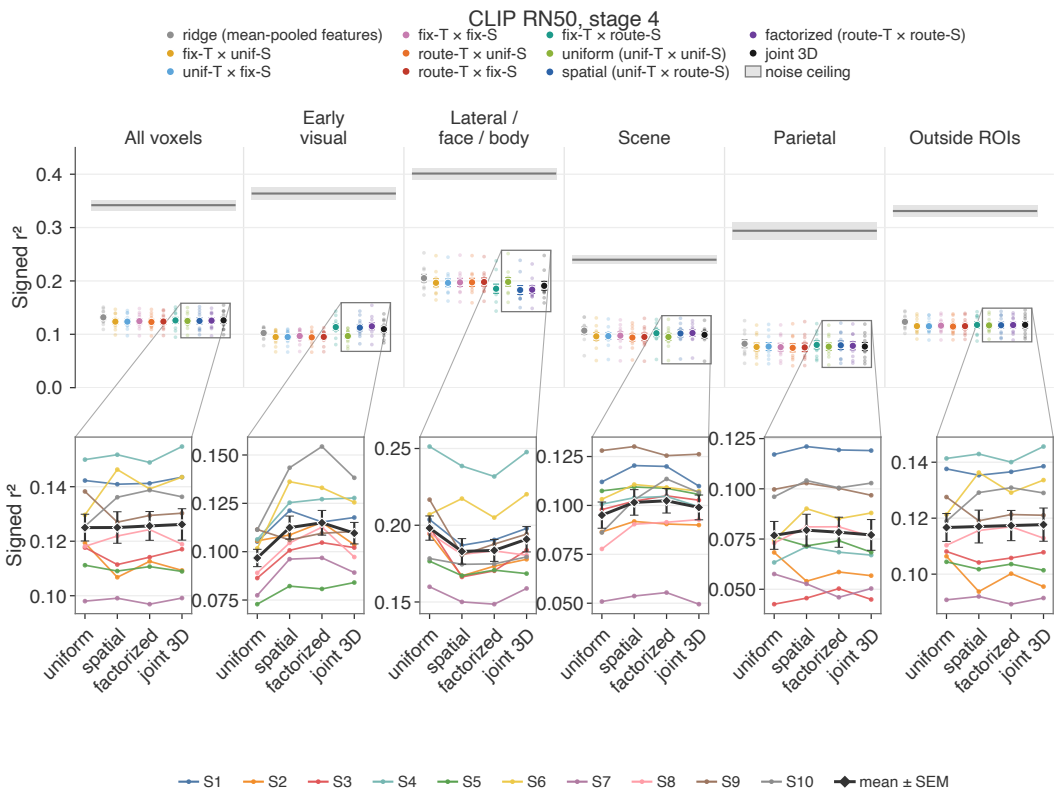

Figure 15: As Fig. 10, for CLIP RN50, stage 4 of 4.

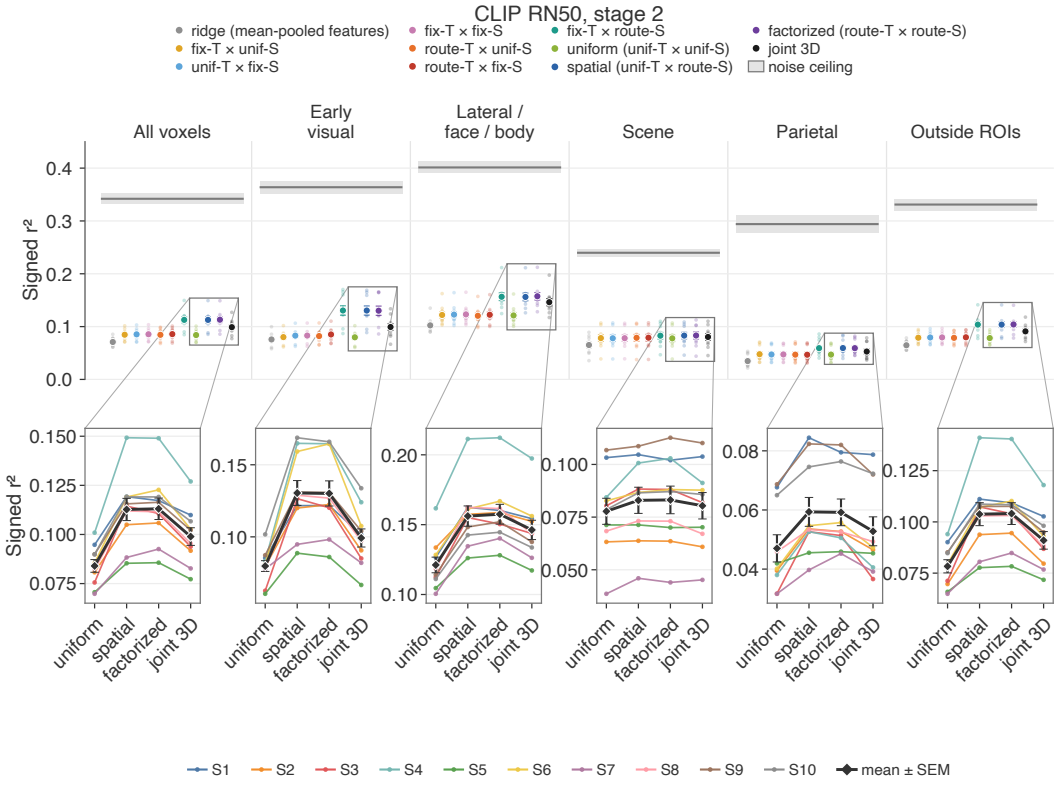

Figure 16: As Fig. 10, for CLIP RN50, stage 2 of 4.

Table 7: Region-mean signed r^2 on the 102 test clips (mean over 10 subjects; voxels with noise ceiling $\geq 5\%$) for every routing configuration and backbone. The factorized models are labelled by their temporal factor (column groups) and spatial factor (U = uniform, F = fixed, R = routed), where a fixed factor is a learned, stimulus-independent distribution for each parcel. Temporal uniform with spatial uniform is the mean-pool model, temporal uniform with spatial routed is the spatial model, temporal routed with spatial uniform is the temporal model, and temporal routed with spatial routed is the factorized model. Ridge is ridge regression on the mean-pooled features of the same layer. The best model in each row is in bold.

		Temporal uniform			**Temporal fixed**			**Temporal routed**			
Region	**Ridge**	**U**	**F**	**R**	**U**	**F**	**R**	**U**	**F**	**R**	**Joint**
V-JEPA 2 ViT-g (Assran et al., 2025), layer 32 of 40											
All voxels	.174	.180	.181	.187	.181	.181	.187	.180	.180	.186	**.191**
Early visual	.178	.177	.178	.189	.177	.177	**.190**	.178	.179	.189	.189
Lateral/face/body	.242	.249	.250	.255	.250	.250	.255	.248	.249	.255	**.263**
Scene	.118	.115	.117	.116	.117	.117	.116	.114	.114	.115	**.119**
Parietal	.109	.122	.122	.125	.122	.122	.126	.121	.122	.125	**.129**
Outside ROIs	.165	.172	.172	.178	.172	.172	.178	.171	.172	.177	**.181**
V-JEPA 2 ViT-g, layer 4 of 40											
All voxels	.035	.039	.039	**.053**	.039	.039	.052	.039	.040	.050	.042
Early visual	.046	.046	.047	**.064**	.046	.046	.063	.048	.049	.060	.048
Lateral/face/body	.046	.053	.053	**.067**	.053	.052	.066	.052	.053	.063	.056
Scene	.043	.046	.047	**.057**	.047	.044	.056	.045	.046	.054	.053
Parietal	.013	.017	.017	**.025**	.017	.017	**.025**	.018	.018	.023	.020
Outside ROIs	.030	.035	.035	**.049**	.035	.035	.048	.035	.036	.046	.038
VideoMAE v2 ViT-g (Wang et al., 2023b), layer 20 of 40											
All voxels	.148	.158	.158	.173	.158	.159	**.174**	.156	.157	.172	.173
Early visual	.148	.151	.153	.171	.150	.153	**.173**	.151	.152	.171	.169
Lateral/face/body	.206	.221	.221	.240	.220	.221	.242	.218	.219	.239	**.243**
Scene	.119	.118	.119	.120	.119	.120	**.121**	.119	.119	.118	.120
Parietal	.093	.106	.106	.114	.107	.107	**.115**	.104	.104	**.115**	**.115**
Outside ROIs	.139	.149	.149	.163	.149	.150	**.164**	.147	.148	.162	.162
VideoMAE v2 ViT-g, layer 4 of 40											
All voxels	.067	.073	.074	**.092**	.073	.074	**.092**	.074	.074	.089	.089
Early visual	.099	.097	.100	**.129**	.097	.101	**.129**	.098	.099	.124	.122
Lateral/face/body	.078	.089	.090	**.111**	.089	.089	**.111**	.089	.088	.107	.108
Scene	.057	.064	.065	**.075**	.065	.064	**.075**	.064	.063	**.075**	.073
Parietal	.033	.040	.041	**.049**	.041	.040	**.049**	.041	.041	.047	.048
Outside ROIs	.061	.067	.068	**.084**	.067	.068	**.084**	.068	.068	.082	.081
DINOv2 ViT-B/14 (Oquab et al., 2024), block 12 of 12											
All voxels	.125	.129	.130	**.152**	.129	.130	**.152**	.130	.131	.151	.151
Early visual	.108	.107	.110	**.139**	.107	.110	**.139**	.108	.112	**.139**	.132
Lateral/face/body	.185	.192	.193	.219	.192	.194	.218	.193	.195	.218	**.222**
Scene	.110	.109	.109	.118	.110	.111	**.119**	.110	.111	**.119**	.115
Parietal	.078	.085	.085	.097	.086	.086	.096	.084	.084	.096	**.099**
Outside ROIs	.117	.121	.122	**.143**	.121	.122	**.143**	.122	.123	**.143**	.142
DINOv2 ViT-B/14, block 2 of 12											
All voxels	.042	.045	.045	**.068**	.044	.045	.067	.045	.045	.067	.060
Early visual	.065	.053	.055	**.105**	.053	.055	.104	.056	.058	.104	.095
Lateral/face/body	.050	.059	.059	**.079**	.059	.059	**.079**	.059	.059	**.079**	.073
Scene	.048	.055	.055	**.062**	.055	.055	**.062**	.054	.055	.061	.055
Parietal	.016	.022	.023	**.028**	.022	.022	**.028**	.022	.022	**.028**	.024
Outside ROIs	.037	.040	.040	**.060**	.039	.040	**.060**	.040	.040	**.060**	.053
CLIP RN50 (Radford et al., 2021), stage 4 of 4											
All voxels	**.132**	.125	.124	.125	.124	.125	.126	.123	.124	.126	.126
Early visual	.102	.097	.095	.112	.095	.097	.113	.094	.095	**.115**	.110
Lateral/face/body	**.205**	.198	.196	.183	.196	.197	.185	.197	.198	.184	.191
Scene	**.107**	.095	.097	.102	.096	.097	.102	.094	.095	.102	.099
Parietal	**.082**	.077	.077	.079	.076	.076	.080	.075	.075	.078	.077
Outside ROIs	**.123**	.117	.115	.117	.115	.116	.118	.115	.115	.117	.118
CLIP RN50, stage 2 of 4											
All voxels	.071	.084	.085	**.113**	.085	.086	**.113**	.084	.086	**.113**	.099
Early visual	.076	.080	.083	**.130**	.080	.083	**.130**	.082	.085	**.130**	.099
Lateral/face/body	.102	.121	.123	.156	.122	.123	.156	.120	.122	**.157**	.146
Scene	.065	.078	.078	**.083**	.078	.078	**.083**	.079	.079	**.083**	.080
Parietal	.035	.047	.047	**.059**	.048	.047	**.059**	.047	.047	**.059**	.053
Outside ROIs	.065	.078	.079	**.104**	.079	.080	**.104**	.079	.080	**.104**	.091

9.12 Additional examples of attention maps

Figures 17–19 show nine further test clips in the format of Fig. 3, for parcels in right MT (golfer, cat and weightlifter), left EBA (surfer, skier and child on stairs), left STS (toddlers and child in snow) and right FFA (eagle). We chose the clips by eye as examples of moving people and animals; they illustrate the attention maps and are not a quantitative test. For the golfer, the surfer, the skier, the toddlers and the child on stairs, the attention of the joint model stays on the moving person across frames, and for the cat it moves with the head and paw as the cat turns in its basket. The factorized model instead concentrates its attention on a few frames and, within them, on fixed parts of the frame that often include the background, such as the basket. For the weightlifter, the joint model attends mostly to the barbell as it is dropped and rolls, rather than to the lifter, whereas the factorized model spreads its attention over a large region on the left of the frame. In the eagle and child-in-snow clips, where the attended head moves little, the two models attend to similar locations.

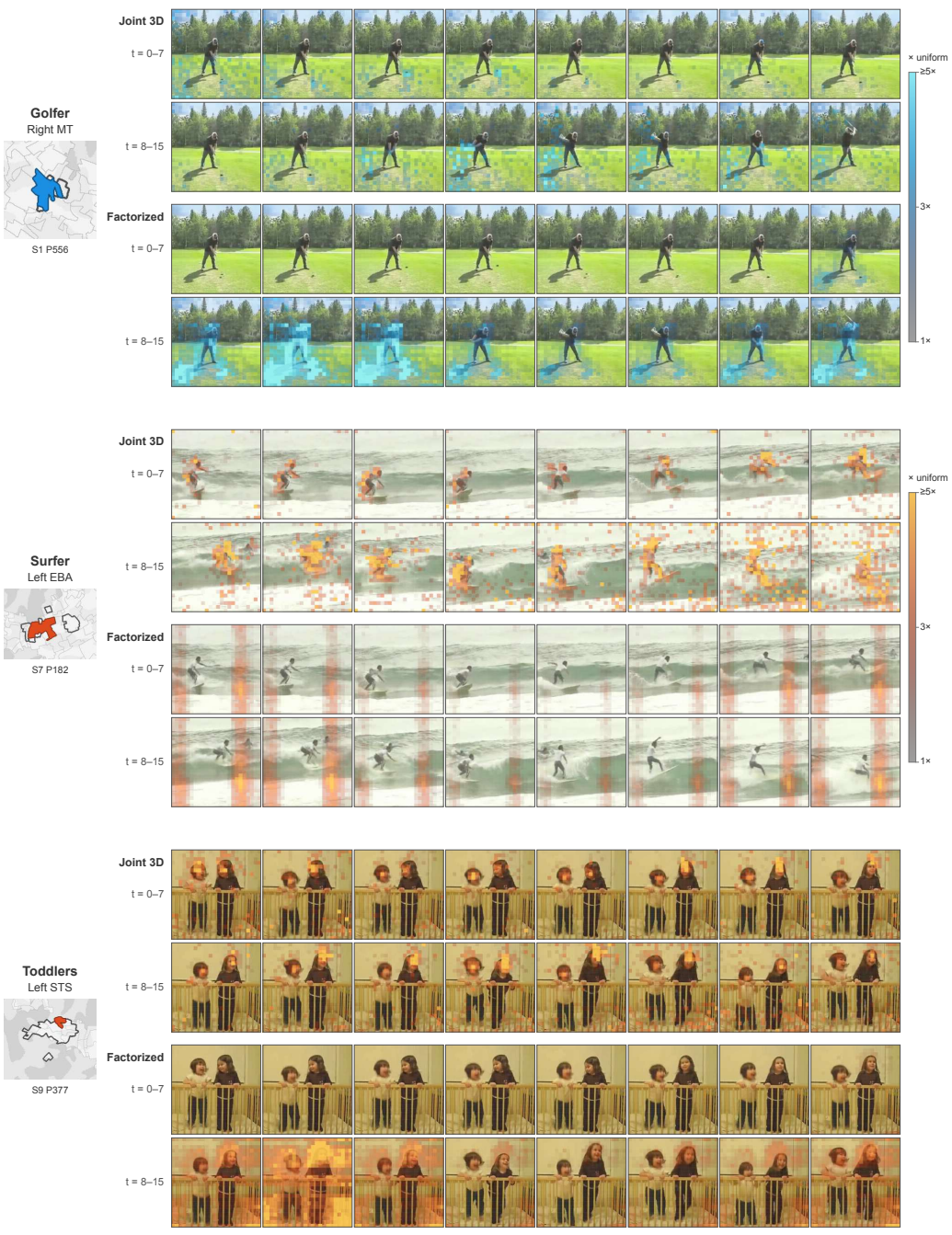

Figure 17: Further examples of attention maps in the format of Fig. 3: joint model (upper two rows of each block) and factorized model (lower two rows) over all 16 frames of each clip, in units of the uniform level $1/N$, from transparent at or below the uniform level to saturated at 5 times the uniform level.

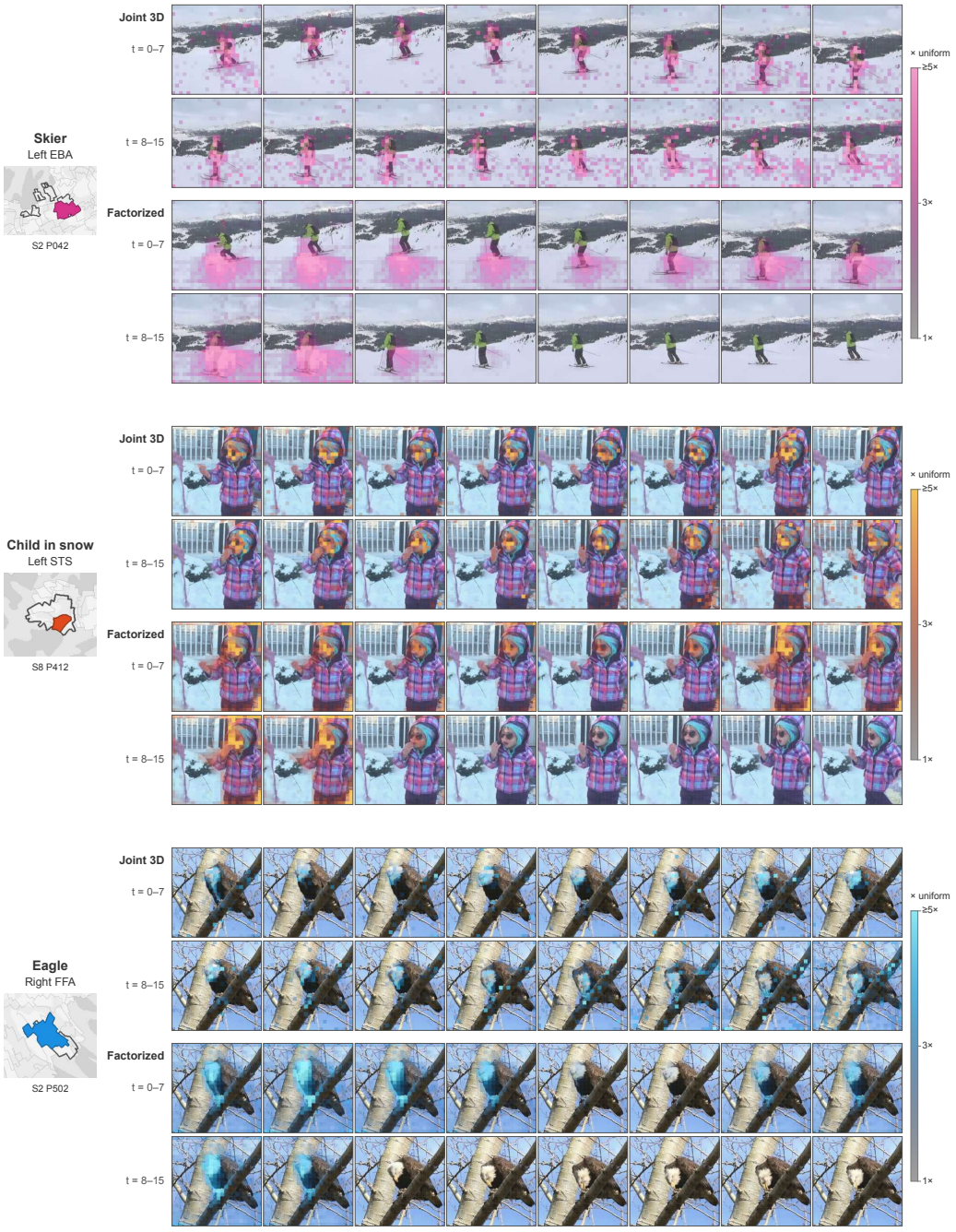

Figure 18: Further examples of attention maps, as in Fig. 17.

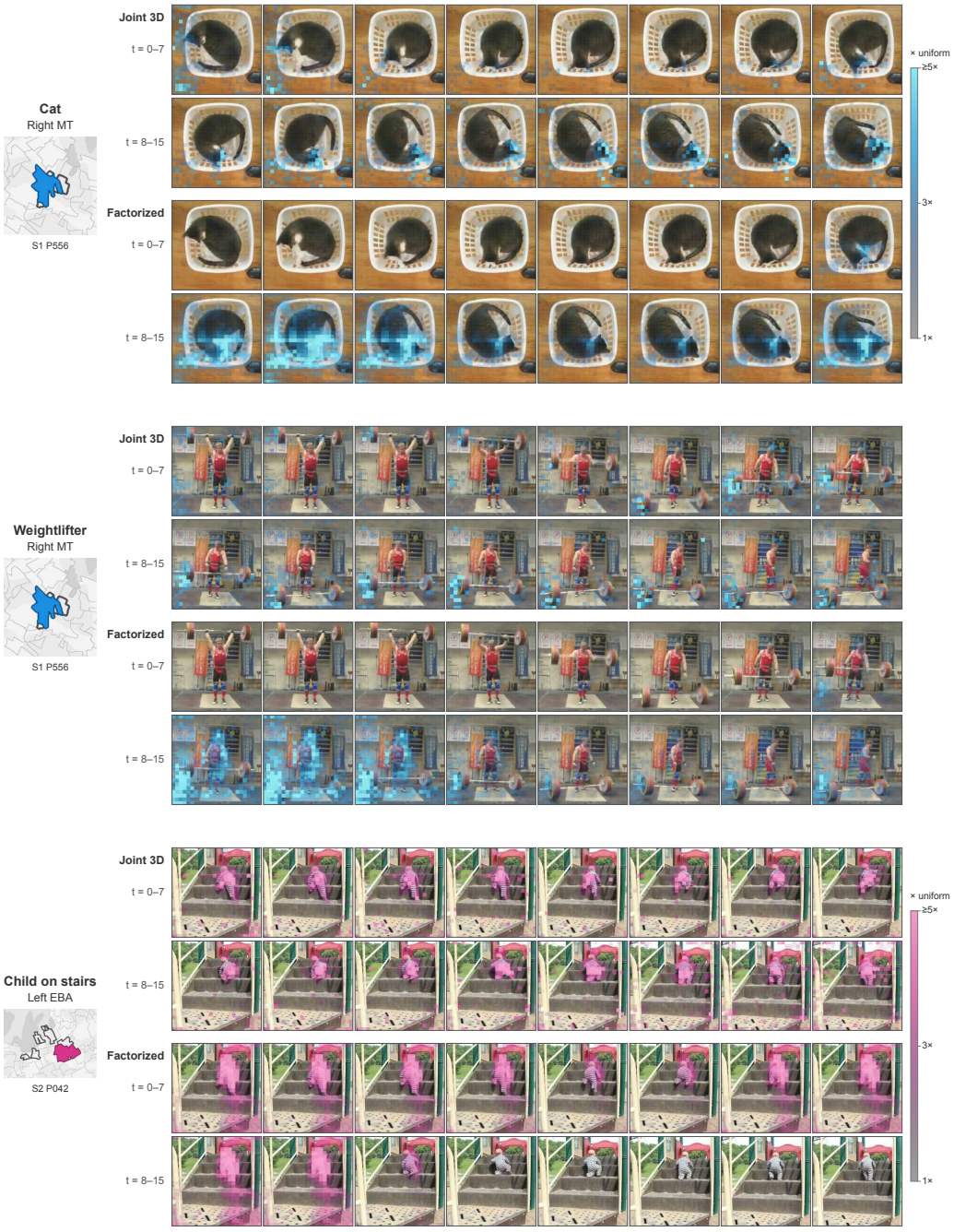

Figure 19: Further examples of attention maps, as in Fig. 17.

9.13 CATEGORY SELECTIVITY AND LOCALIZATION FROM ATTENTION MAPS

All analyses of category selectivity use the attention maps of the joint model (V-JEPA 2 ViT-g, layer 32) on the 102 test clips, with one model per subject trained as described in Section 3.3. We assign a parcel to a functional ROI if at least 20% of its voxels fall in that ROI. Because each parcel has a single attention map, the whole parcel is used, so a parcel can belong to more than one ROI and may also contain voxels outside the ROI.

Category masks. For each test clip we take 16 frames evenly spaced through the clip, one for each temporal position of the token grid, and segment them with SAM 3 (Carion et al., 2025), which returns all instances matching a text prompt; we keep detections with a confidence of at least 0.35. We use 28 prompts and merge them into mutually exclusive categories: face (face), body (person, hand), animal (animal, animal face), food (food), object (car, vehicle, ball, toy, musical instrument, tool, utensil, phone, book, furniture, bag, text, sign) and scene (building, tree, sky, grass, road, water, mountain, wall, floor). A pixel claimed by several categories is assigned to the first in this order. The order runs from the most to the least specific category, so that a face is not also counted as part of the person, and a hand holding a phone is not counted as an object. Pixels matched by no prompt are unlabeled. The masks are computed at 64×64 pixels per frame and area-averaged onto the 24×24 token grid, which gives the fraction $M_c(t, h, w) \in [0, 1]$ of every token covered by category c. Because the categories are exclusive, these fractions sum to 1 over the categories and unlabeled. The unlabeled pixels are mostly scene surfaces that the scene prompts miss (for example snow, dirt paths, plain walls and courts), so when we score the known functional ROIs we merge scene and unlabeled into a single place category.

Category selectivity of known functional ROIs. We measure where a parcel reads in a clip by the 92 tokens (the top 1% of the $16 \times 24 \times 24$ grid) that it attends to most, the same fraction as the 2,000 most-attended pixels of each image used by Adeli et al. (2026). Restricting to the most-attended tokens measures where the parcel reads, rather than the diffuse tail of its attention that is spread over the whole clip. The IoU of category c is the intersection over union of these tokens with the binary mask of c (tokens at least half covered), as in analyses of spatial selectivity in fMRI-trained networks (Sarch et al., 2023). We average it over the clips in which c covers between 1% and 90% of the clip: when a category covers very few tokens, its IoU is determined by those few tokens and dominated by noise, and when it fills nearly the whole clip, any 92 tokens overlap it, so the IoU carries no information about where the parcel attends. The raw IoU mostly reflects the size of each category (Fig. 4b). A place mask covers about 6,500 tokens, so 92 tokens can reach an IoU of only about 0.014 even for a parcel that attends only to places, whereas a face mask covers about 550 tokens and is hard to hit. We therefore z-score the IoU across parcels: the score of a functional ROI for category c is the mean over its parcels minus the mean over all parcels, divided by the standard deviation across parcels. Because every parcel sees the same clips and masks, the size of each category is the same for all parcels and cancels, and the score asks whether the ROI puts more of its attention on the category than a typical parcel does. We report the raw and z-scored IoU for the eight category-selective ROIs (FFA, OFA, STS, EBA, PPA, RSC, TOS and LOC), and the raw IoU of all parcels for reference, for five categories (face, body, animal, object and place) as the mean $\pm$ SEM over subjects (Fig. 4b,c).

Localizing category-selective regions from attention. Functional localizers contrast the responses of each voxel to different stimulus categories, for example faces and objects. We apply the same logic to the attention maps: for each parcel we measure how much attention it puts on each category in each clip, form the localizer contrasts from these measures, and ask which parcels show the contrast consistently across clips.

Contrasts. In each clip, we compute the IoU of each parcel's 92 most-attended tokens with the mask of each category, as above, and leave it undefined for a clip in which the category covers less than 1% or more than 90% of the clip. Here the scene category is the scene mask alone, without the unlabeled pixels. Within each subject, we z-score the IoU of each category across all clips and parcels where it is defined. The z-scoring puts the categories on a common scale, so that a contrast is not dominated by the category whose IoU is largest or varies most, which depends on its size. It also sets 0 to the average parcel on the average clip, so that a contrast measures a parcel's preference relative to other parcels rather than a preference that all parcels share (for example, if faces draw attention from every parcel). From the z-scored IoU we form, in each clip, contrasts that mirror the localizers used to define the functional ROIs: face minus object for FFA, OFA and STS, and

body minus object for EBA. For PPA, RSC and TOS we contrast scene with the mean of face, body, animal and object. The scene localizer compares scenes with isolated objects, but in natural video every foreground item appears within a scene, so the analogous contrast is between scene and every kind of labeled foreground, not objects alone. A contrast is undefined in a clip if any of its terms is undefined; for the scene contrast, the mean is taken over the foreground categories that are defined. For each parcel and contrast, a one-sample t-test across clips then gives a t value, treating clips as independent replicates, as trials are in a localizer. Parcels with fewer than 5 clips in which the contrast is defined are left unscored (Fig. 4a).

Scoring. To compare these attention-derived maps with the functional ROIs, we pool the parcels of the ROIs that each localizer defines (face: FFA, OFA and STS; body: EBA; scene: PPA, RSC and TOS), with parcels assigned to ROIs as above. Let K be the number of these ROI parcels. We take the K scored parcels with the highest t and measure the overlap as the fraction of them that belong to the ROIs (Fig. 4a shows the t map, the top-K parcels and the ROIs of each contrast for one subject, S8). Setting the size of the derived region equal to the size of the ROI avoids choosing a t threshold and makes the overlap equal to both the precision and the recall of the derived region. If the ranking were unrelated to ROI membership, the number of ROI parcels among the top K would follow a hypergeometric distribution, with an expected overlap equal to the fraction of scored parcels that belong to the ROIs (0.02–0.10). We therefore test the overlap in each subject with a one-sided hypergeometric test (uncorrected), and across subjects with a two-sided one-sample t-test of the overlap minus chance ($df = 9$), with p-values corrected over the three localizers with the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995). As a threshold-free complement, we report the area under the ROC curve (AUC) of t for ROI versus non-ROI scored parcels, which measures the separation of the two groups over the whole ranking.

Detailed results. The raw IoU mostly reflects the size of each category: faces, which are small, have the lowest IoU in most ROIs and in the average parcel, although the face IoU of FFA (0.0107) is well above that of all parcels (0.0063; Fig. 4b). After z-scoring (Fig. 4c), FFA, OFA and EBA put more attention than the average parcel on faces, bodies and animals ($z \approx 0.5$–1.3) and less on places; because these categories usually appear together in the clips, an area that attends to one of them also tends to attend to the others. PPA and RSC show the reverse pattern (place $z = 0.84$ and 0.75, all foreground categories negative). LOC is high for faces, bodies and animals ($z = 1.15$–1.25), lower for objects ($z = 0.34$) and strongly negative for places ($z = -1.24$), consistent with a preference for foreground content in general. In STS and TOS, all categories are close to the average parcel ($|z| \leq 0.28$). In the localization, the overlap is significant in 5, 3 and 8 of the 10 individual subjects for the face, body and scene localizers (one-sided hypergeometric test, $p < 0.05$, uncorrected), and ROI parcels rank above other parcels across the whole ranking, not only at the top (AUC 0.61–0.77; Table 8). Fig. 4a shows the subject with the highest mean overlap across the three localizers (S8; face 0.41, body 0.08, scene 0.43).

9.14 Spatial-prior control for the localization

The IoU of a parcel with a category can differ between parcels even if no parcel's attention follows the content of the clips, because each parcel's attention has a spatial prior and categories occupy typical positions in the frame. The IoU is computed from the most-attended tokens, and which tokens are most attended depends on the prior. Z-scoring does not remove this confound: unlike the size of a category, which is the same for all parcels and cancels, the prior differs between parcels. To measure this contribution, we repeat the localization with each clip's attention scored against the category masks of a different test clip. The pairing is a single random permutation of the 102 test clips in which no clip is paired with itself, drawn once and used for every parcel and subject. The mismatched masks keep each parcel's spatial prior and the typical layout of the categories, but remove any correspondence between the attention and the content of the clip. Table 8 compares the localization with the correct and the mismatched masks; the two are compared with a two-sided paired t-test across subjects ($df = 9$), FDR-corrected over the three localizers.

For faces, the mismatched masks give an overlap at chance (0.10, chance 0.09; $t(9) = 0.25$, $p_{\mathrm{FDR}} = 0.81$), and the correct masks give a significantly larger overlap ($t(9) = 3.13$, $p_{\mathrm{FDR}} = 0.036$), so the localization of the face-selective ROIs depends on the content of each clip. For bodies, the mismatched-mask overlap (0.08) is not significantly above chance ($t(9) = 2.00$, $p_{\mathrm{FDR}} = 0.11$), but the correct masks do not exceed it significantly either ($t(9) = 1.70$, $p_{\mathrm{FDR}} = 0.18$). For scenes,

the mismatched masks give nearly the same overlap as the correct masks (0.26 vs. 0.28; $t(9) = 1.11$, $p_{\mathrm{FDR}} = 0.30$) and are themselves well above chance ($t(9) = 3.74$, $p_{\mathrm{FDR}} = 0.014$), so the localization of the scene-selective ROIs mostly reflects the parcels' spatial priors rather than attention to the scenery in each clip.

Table 8: Localization of category-selective ROIs from z-scored IoU contrasts with the correct masks and with the masks of a different clip (joint model, 10 subjects, 102 test clips; face: FFA, OFA and STS; body: EBA; scene: PPA, RSC and TOS). Overlap: fraction of the top-K parcels that belong to the ROIs, mean (SD) over subjects; chance is the fraction of scored parcels in the ROIs. $t(9)$ and p_{FDR}: two-sided one-sample t-test of overlap minus chance across subjects, Benjamini–Hochberg corrected over the three localizers within each condition. Subjects: number with a one-sided hypergeometric $p < 0.05$ (uncorrected). Correct − mismatched: two-sided paired t-test of the overlaps across subjects, FDR-corrected over the three localizers.

		Correct masks					**Mismatched masks**				**Correct − mismatched**		
Localizer	**Chance**	**Overlap**	$\boldsymbol{t(9)}$	$\boldsymbol{p_{\mathrm{FDR}}}$	**AUC**	**Subjects**	**Overlap**	$\boldsymbol{t(9)}$	$\boldsymbol{p_{\mathrm{FDR}}}$	**Subjects**	$\boldsymbol{t(9)}$	$\boldsymbol{p}$	$\boldsymbol{p_{\mathrm{FDR}}}$
Face	0.09	0.19 (0.11)	2.81	0.021	0.61	5/10	0.10 (0.05)	0.25	0.81	0/10	3.13	0.012	0.036
Body	0.02	0.14 (0.09)	4.02	0.005	0.76	3/10	0.08 (0.09)	2.00	0.11	3/10	1.70	0.12	0.18
Scene	0.10	0.28 (0.12)	4.85	0.003	0.77	8/10	0.26 (0.14)	3.74	0.014	8/10	1.11	0.30	0.30