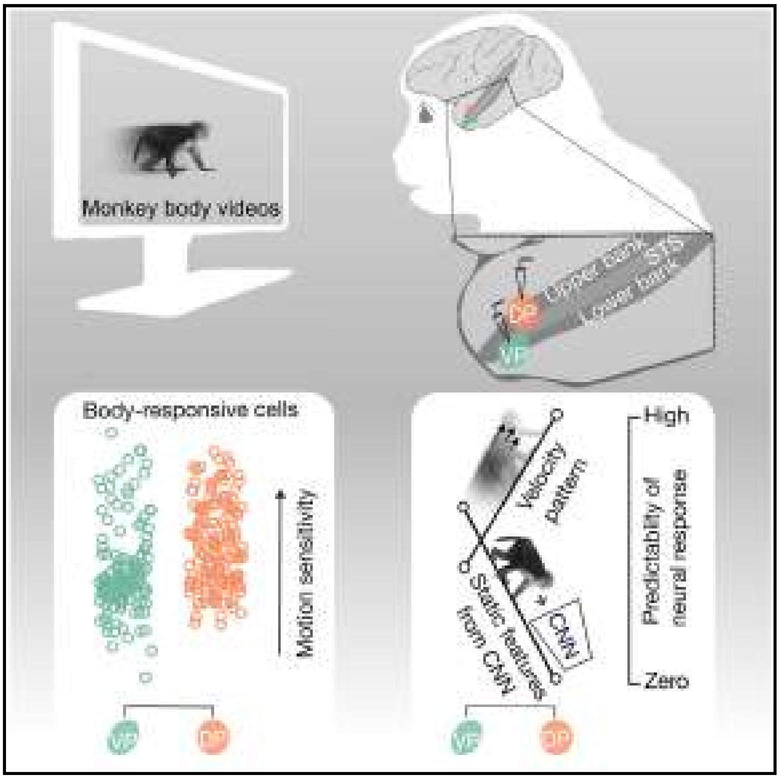
Cell Reports

Article

Bodies in motion: Unraveling the distinct roles of motion and shape in dynamic body responses in the temporal cortex

Graphical abstract



Highlights

- The selectivity of dorsal-bank STS neurons to monkey videos is mainly driven by dynamics
- The body-video selectivity of anterior IT neurons depends on dynamic and static features
- The neurons in both regions largely overlap in their motion/sequence sensitivity
- CNNs predict the selectivity for dynamic bodies in IT but not in dorsal-bank STS

Authors

Rajani Raman, Anna Bognár, Ghazaleh Ghamkhari Nejad, Nick Taubert, Martin Giese, Rufin Vogels

Correspondence

rufin.vogels@kuleuven.be

In brief

Raman et al. show that the selectivity of neurons for different moving monkey bodies is mainly driven by dynamics in the dorsal bank of the macaque STS, whereas both dynamic and static features contribute to the inferotemporal selectivity, challenging conventional views on the functional differences between these regions.

Raman et al., 2023, Cell Reports *42*, 113438
December 26, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.celrep.2023.113438

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Bodies in motion: Unraveling the distinct roles of motion and shape in dynamic body responses in the temporal cortex

Rajani Raman,[1,2,4] Anna Bognár,[1,2,4] Ghazaleh Ghamkhari Nejad,[1,2] Nick Taubert,[3] Martin Giese,[3] and Rufin Vogels[1,2,5,*]
[1]Department of Neurosciences, KU Leuven, 3000 Leuven, Belgium
[2]Leuven Brain Institute, KU Leuven, 3000 Leuven, Belgium
[3]Hertie Institute for Clinical Brain Research and Center for Integrative Neuroscience, University Clinic Tuebingen, 72074 Tuebingen, Germany
[4]These authors contributed equally
[5]Lead contact
*Correspondence: rufin.vogels@kuleuven.be
https://doi.org/10.1016/j.celrep.2023.113438

SUMMARY

The temporal cortex represents social stimuli, including bodies. We examine and compare the contributions of dynamic and static features to the single-unit responses to moving monkey bodies in and between a patch in the anterior dorsal bank of the superior temporal sulcus (dorsal patch [DP]) and patches in the anterior inferotemporal cortex (ventral patch [VP]), using fMRI guidance in macaques. The response to dynamics varies within both regions, being higher in DP. The dynamic body selectivity of VP neurons correlates with static features derived from convolutional neural networks and motion. DP neurons' dynamic body selectivity is not predicted by static features but is dominated by motion. Whereas these data support the dominance of motion in the newly proposed "dynamic social perception" stream, they challenge the traditional view that distinguishes DP and VP processing in terms of motion versus static features, underscoring the role of inferotemporal neurons in representing body dynamics.

INTRODUCTION

The visual processing of dynamic bodies is vital for reproduction, survival, and social behavior, as it conveys information about action and affect.[1,2] Previous research in the macaque visual temporal cortex found single cells selectively responding to bodies.[3,4] Using static images, monkey fMRI studies identified body-category-selective patches (body patches) in the ventral bank of the superior temporal sulcus (STS) and ventral to the STS,[3,5] both part of the inferotemporal (IT) cortex. Despite the social relevance of moving bodies, their visual processing remains poorly understood due to the focus on static images.[3] Recently, employing fMRI to map patches that are activated specifically by dynamic monkey bodies,[6] we observed patches in the dorsal-bank STS that were activated less by static images.

It has been proposed that the ventral visual stream, which includes IT, can be distinguished not only from the dorsal (parietal) stream but also from a third stream that processes dynamic social information, accentuating motion.[7] The latter "dynamic social perception" stream[7] has been linked to the human STS and likely corresponds to the dorsal bank/fundus of the macaque STS. This proposal and our recent fMRI findings[6] underscore the importance of assessing and comparing the contributions of dynamic and static features to body responses in and between IT and dorsal-bank STS, which is the aim of this study.

Single-unit studies, recording randomly in the macaque STS, showed responses to acting humans[4,8–14] and "stick" figures.[15] These studies suggested that some STS neurons respond to motion or, at least, are sensitive to the image sequence, showing less response to static images than to moving human bodies or animated stick figures.[8,15] Studies using moving stick figures[15,16] observed motion- or sequence-sensitive neurons mainly in the dorsal-bank STS, in agreement with older work that demonstrated motion selectivity in the dorsal-bank STS.[17–19] However, selective responses to static stimuli have also been observed in the dorsal-bank STS,[11,14,15,20,21] raising the question of how static and dynamic feature selectivity interact in the dorsal-bank STS. Furthermore, the contribution of motion to the responses of IT neurons to naturalistic body stimuli is unclear. Neurons responding to a hand interacting with an object were reported in the ventral middle STS,[22] but evidence for motion-related responses is lacking for other body parts and in anterior IT. We[6] found stronger fMRI activations also in the ventral STS in response to dynamic naturalistic body stimuli compared with static ones. These fMRI data raised the possibility that body dynamics also contribute to the IT neurons' responses.

Here, we conducted a comparative study of single-unit responses between the dorsal bank/fundus of the STS and the regions within and ventral to the ventral bank of the rostral STS, using video stimuli featuring moving monkeys. To increase the

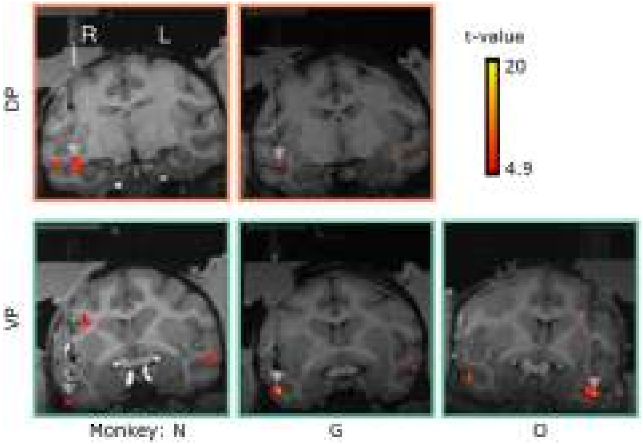


Figure 1. Targeted body patches

Top: dorsal-bank/fundus STS patches. Bottom: ventral-bank STS/IT patches. Targeted patches are indicated by arrowheads. The fMRI activations (in yellow/red colors) were obtained with the contrast dynamic monkey bodies-dynamic objects, exclusively masked by dynamic monkey faces-dynamic objects. A t score of 4.9 corresponds to $p < 0.05$, family-wise error (FWE) corrected.

probability of finding neurons that responded to naturalistic monkey videos, recordings were guided by fMRI mapping of dynamic body patches,[6] targeting patches in the anterior IT (ventral patch; VP) and the dorsal bank/fundus STS (dorsal patch; DP). We selected anterior patches because these are supposed to be related more to invariant perception than posterior patches.[3] We compared their responses to static frames of the same videos and to time-reversed versions of the videos, assessing their motion or sequence sensitivity.

We assessed the contributions of motion and static features at the population level in VP and DP using regression analysis. To assess the contribution of motion, we related differences in motion among the body videos and the neural responses. To estimate the contribution of static features to the body video responses, we related neural responses and convolutional neural network (CNN) features. Recently, CNNs have emerged as improved models of IT responses to static images.[23–28] It is unclear whether CNNs can model the selectivity for dynamic stimuli of STS neurons, particularly of DP neurons that may demonstrate a strong motion-driven response.

Overall, we expected that the responses of DP neurons to dynamic bodies would be dominated by motion and, to a lesser extent, by static features, while the VP responses would be determined by their selectivity for static body features.

RESULTS

We targeted fMRI-defined rostral STS and IT body patches (Figure 1). The targeted IT body patches were either in the anterior ventral STS (monkeys G and O; ASB in Bognár et al.[6]) or ventral to the anterior STS (monkey N; AVB in Bognár et al.[6]). We will label neurons from both ventral IT patches as VP neurons. In the main analyses, we pooled the data of both IT patches since results were similar for both patches (single-patch data are in the supplemental figures). As dorsal STS region, we targeted in two animals a body patch in the rostral dorsal bank/fundus of the STS, and these neurons are labeled DP neurons. DP corresponded to the most anterior patch in the medial upper bank/fundus of the STS in the fMRI mapping (AMUB in Bognár et al.[6]) of these two monkeys.

We selected neurons with a response to at least one body video (STAR Methods). The majority of the neurons responded on average stronger to the body videos than to the face and object videos in VP and DP (Figures 2A and 2B). The body-category selectivity was higher for VP neurons (median body-category-selectivity index [BSI; STAR Methods]: VP, 0.39, $n = 149$; DP, 0.23, $n = 175$; Wilcoxon rank-sum test $p = 4.043e-07$; Figure 2C). As reported earlier when targeting body patches,[5,29,30] many body-responsive neurons responded to some extent also to faces or objects. The neurons encoded differences in body shape and/or movement: they responded to some but not all of the 20 body videos (Figure 2E), with the effective videos differing among neurons. To quantify the (within-category) body-video selectivity, we computed for each neuron the Sparseness of the response to the 20 body videos (STAR Methods). The Sparseness can range from 0 (equal response to the body videos) to 1 (response to a single body video). The median Sparseness was high in VP (median = 0.65) and DP (median = 0.61), with no significant difference between regions (Wilcoxon rank-sum test $p = 0.556$; Figure 2D). The sparse body responses reduce the BSI, as demonstrated by the negative correlation between Sparseness and BSI (Spearman rank $\rho = -0.15$, $p = 0.009$, $n = 324$). In the main analyses, we will consider all body-responsive neurons, including the minority of neurons with low BSI, because even the latter can encode dynamic bodies, irrespective of their response to non-body stimuli.

The VP population response was relatively constant during the video, except for an initial onset response, whereas marked variations in the DP population response were present during some of the body videos (Figure S1). The latter might be related to differences in body dynamics during the video.

Responses to body dynamics and static bodies

To assess whether the neurons' responses were driven by motion (or a changing image sequence), we tested neurons in a "snapshot test" in which we presented an effective body video, a time-reversed version of that video, and 10 snapshots of the video. The effective body video ("original video") was selected based on the responses in the preceding test in which we presented the body videos. The snapshots were selected to be representative of the variety of poses and viewpoints that occurred during the video and were presented for 300 ms each in random order with an interstimulus interval of at least 1000 ms.

In both regions, some neurons responded with similar peak firing rates to the original video and a snapshot (example VP neurons in Figures 3A and 3C; example DP neurons in Figures 3D and 3F). Other neurons failed to respond to the snapshots, although they showed a sizable response when the corresponding frame occurred in the video (example neurons in Figures 3B [VP] and 3E [DP]). We quantified the difference in peak firing rate between the video and the snapshots for each neuron by a snapshot index (SSI; STAR Methods). A positive SSI corresponds to a

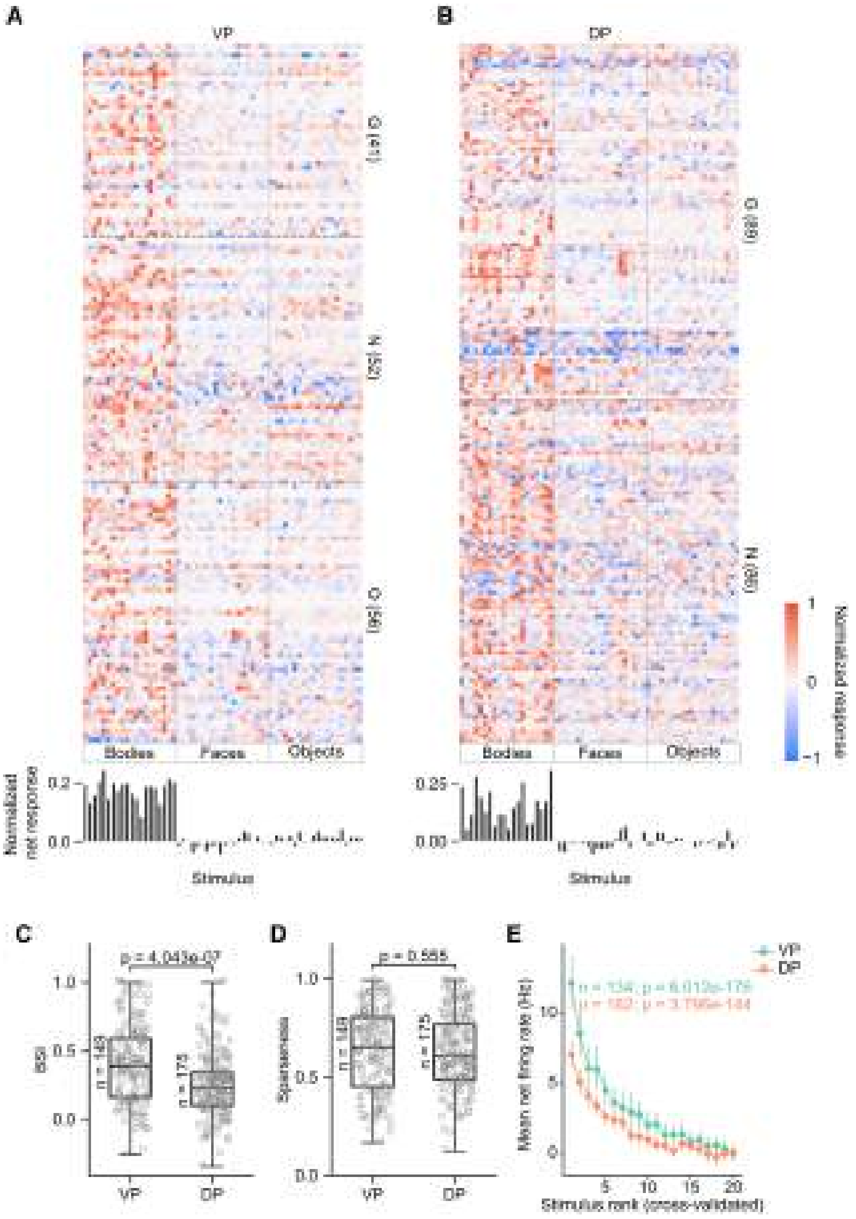


Figure 2. Stimulus selectivity of DP and VP neurons

(A) Top: response matrix for 149 cells (rows) in VP recorded from monkeys G, N, and O to videos (columns) of bodies, faces, and objects (20 videos per category). Bottom, bar plot of averaged normalized responses.

(B) Top: response matrix for 175 cells and corresponding averaged normalized population response (bottom) in DP recorded from G and N. Net firing rates were averaged across the video.

(C) Distribution of body-category-selectivity indices (BSI) in VP and DP.

(D) Distribution of Sparseness in VP and DP.

(E) Average response as a function of the body video rank (STAR Methods). Error bars represent 95% confidence intervals of bootstrapped means (n = 1,000). A Friedman ANOVA showed a significant effect of stimulus rank for each region. The number of cells is indicated by n. See also Figure S1.

smaller response to the static snapshot than to the video, whereas an SSI of zero corresponds to identical peak firing rates for the video and the static presentations. Both regions showed a wide range of SSI values, but the median SSI was higher for DP than for VP (median SSI: VP, −0.01; DP, 0.18; Wilcoxon rank-sum test $p = 8.287e{-}06$; Figure 4A; individual monkey and patch data in Figure S2$A_{1\text{-}2}$). This difference in SSI between regions remained significant for neurons with BSI > 0.33 (Figure S2A_3) and when controlling for the BSI difference between regions (ANCOVA with BSI as covariate: Table S1), in line with a stronger contribution of motion to DP responses. However, we also observed VP neurons that did not respond to static snapshots, despite a strong response to the video that included the same frames (Figure 3B). In fact, 11% of VP neurons had an SSI > 0.50 (a 3-fold difference in response), whereas the same held for 17% of the DP neurons.

The majority of DP neurons showed a response to static presentations, although typically less than to the video. Ranking (using leave-one-trial-out cross-validation) of the snapshots based on the responses of each snapshot-responsive DP neuron (n = 124; STAR Methods) showed a significant effect of snapshot rank on the mean DP responses, showing selectivity for still body images (Friedman ANOVA; $p = 5.239e{-}17$; Figure 4D). The same ranking analysis showed snapshot selectivity for the snapshot-responsive VP neurons (Friedman ANOVA; $p = 3.000e{-}75$, n = 126; Figure 4D). The average response to the most effective snapshot (rank 1) was higher in VP than in DP (net average firing rate (computed with a window of 400 ms) of 17 versus 7 spikes/s), but the regions did not differ in the temporal course of the averaged responses to their most effective snapshot (Figure 4E).

Sensitivity to time reversal of body movements

The time-reversed version of the original video allowed us to assess the effect of frame sequence on the video responses. VP and DP neurons showed a range of responses to the time-reversed video. Some neurons responded with similar average firing rates to the original and the time-reversed video (Figures 3A [VP] and 3D and 3E [DP]). Other neurons showed a marked difference in response between the two sequences (Figures 3B and 3C [VP] and 3F [DP]). To quantify the difference in response between the two sequences, we computed the video reversal index (VRI; STAR Methods). A VRI of zero indicates equal

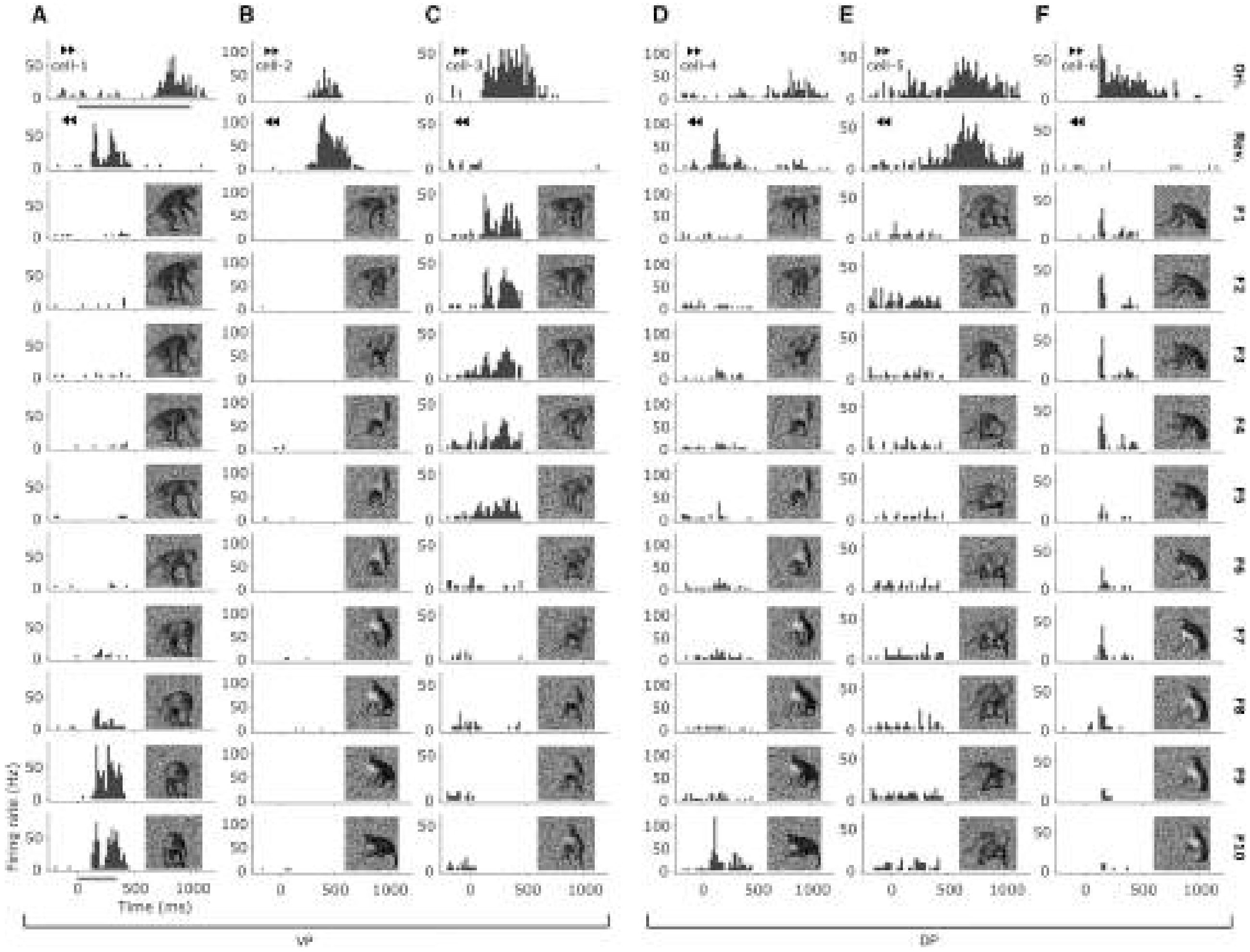


Figure 3. Responses to body video, its time reversal, and static snapshot presentations: Example neurons

Peristimulus Time Histograms (PSTHs) of (A–C) VP neurons and (D–F) DP neurons. The first row corresponds to the response to the original effective video, while the second row corresponds to its time-reversed version. Rows 3–12 correspond to the ordered snapshots taken from the video. Horizontal bar below PSTH indicates stimulus duration.

(A and D) Cells that show a response to an end segment of the original video and the corresponding last few snapshots. These neurons also responded to the first segment of the time-reversed video.

(B and E) Cells that responded to the videos but not to the snapshot.

(C and F) Cells that responded to the original video and the snapshots but failed to respond to the time-reversed video. Bin width, 20 ms.

responses to both sequences, whereas a VRI of 1 indicates a response to only one of the two sequences. The median VRI was significantly larger for DP (0.317) than for VP neurons (0.215; Wilcoxon rank-sum test; p = 3.061e−04; Figure 4B; individual monkey and patch data in Figure S2B$_{1\text{-}2}$). This difference in VRI between the two regions remained significant for neurons with BSI > 0.33 (Figure S2B$_3$) and when controlling for BSI (ANCOVA; Table S2), demonstrating a higher sequence sensitivity in DP compared with VP neurons. However, we observed VP neurons that responded to only one of the video sequences, although these consisted of the same frames, the only difference being the frame order (Figure 3C). In fact, 21% of the VP (and 38% of the DP) neurons had a VRI > 0.50, i.e., a 3-fold difference in response between the original and the time-reversed video. Furthermore, 41% of the VP neurons showed a significant difference in response between the two sequences (Wilcoxon rank-sum test; $p < 0.05$). Neurons of both regions with VRI of 1 showed an inhibitory response to the non-preferred movement (Figure S2B$_4$) without an excitatory response to the video onset.

Relating responses to dynamic and static bodies

The VRI did not correlate with the SSI (Spearman correlation; DP, $\rho = 0.041$, $p = 0.61$, $n = 146$; VP, $\rho = 0.103$, $p = 0.21$, $n = 133$; Figure 4C). Indeed, several neurons had a high VRI but were responding well to static snapshots (SSI close to 0; Figures 3C, 3F, and 4C). These neurons were sequence sensitive but did not require motion to produce a response. This raises the question of how the responses to the static images relate to the responses to these frames during the video. To assess this, we selected those neurons that showed a significant excitatory

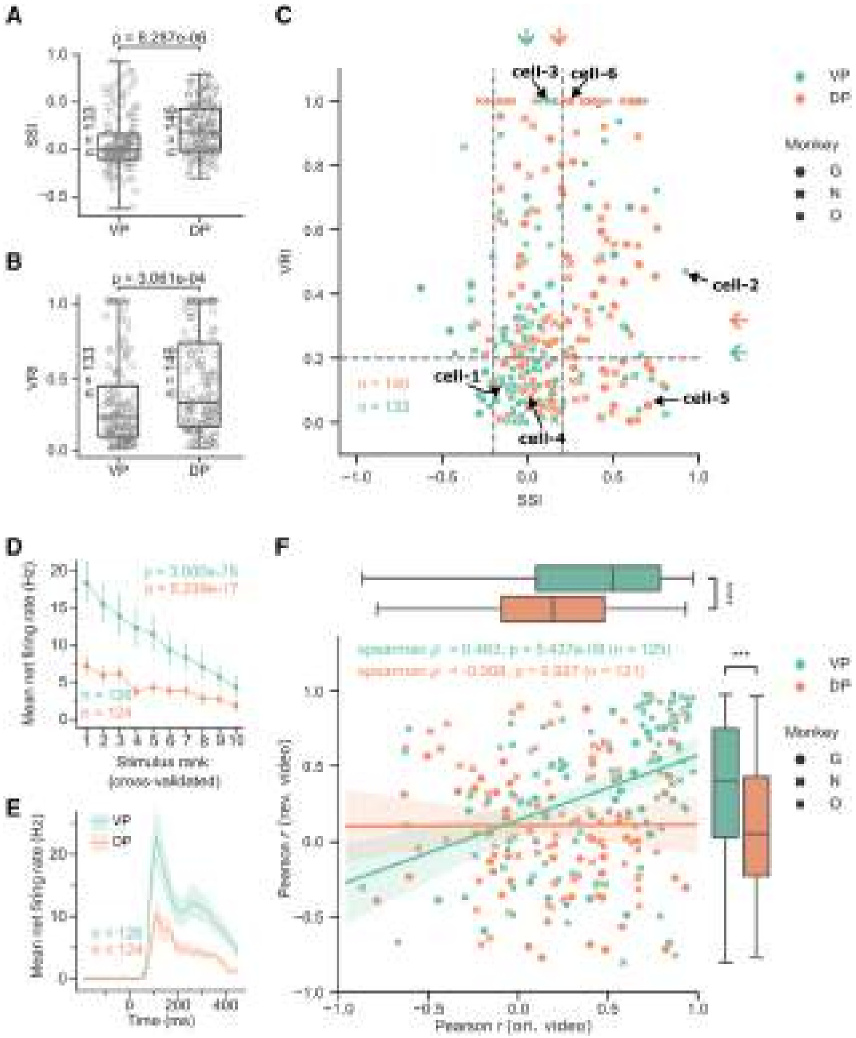


Figure 4. Response to body videos, their time reversal, and static snapshots: Population response metrics

(A) Distribution of snapshot selectivity indices (SSIs) for VP and DP. The p value is from a Wilcoxon rank-sum test comparing VP and DP.

(B) Distribution of video reversal indices (VRIs). Same conventions as in (A).

(C) Scatterplot of SSI and VRI. Round, cross, and square markers correspond to monkeys G, N, and O, respectively. Arrows at the right and top correspond to the median VRI and SSI, respectively, for DP (orange) and VP (green). Cells 1–6 correspond to (A)–(F) of Figure 3.

(D) Average firing rate of the cells that had a significant response to a snapshot as a function of snapshot rank (cross-validated). Error bars represent 95% confidence intervals of the mean (n = 1,000 resamplings; bootstrapping). The p-values correspond to Friedman ANOVA.

(E) Population PSTH for all snapshots in VP and DP. The bands indicate the 95% confidence interval (n = 1,000 resamplings).

(F) Scatterplot (with linear regression lines and 95% confidence intervals) of the correlation coefficients between the video and the snapshot responses (STAR Methods) for the original (ori. video) and time-reversed videos (rev. video). Different marker types represent different monkeys. The boxplots at the right and top summarize the distribution of the correlations for each region, and stars indicate significant differences (Wilcoxon rank-sum test) between regions (****$p < 0.0001$; ***$p < 0.001$). See also Figure S2.

response to at least one of the snapshots (split-plot ANOVA; $p < 0.05$). Then, for each selected neuron, we computed the correlation between the snapshot response and the response following the same frame during the video (STAR Methods). For the original video, the median correlation was 0.53 and 0.19 for the VP and DP neurons (Figure 4F), respectively, both significantly greater than zero (Wilcoxon test; VP, p = 3.449e−15; DP, p = 1.881e−06; individual monkey and patch data in Figure S2C_{1-2}) and significantly higher in VP compared with DP (Wilcoxon rank-sum test; p = 1.226e−05; Figure 4F). This difference between the regions was unrelated to regional differences in BSI (Figure S2C_3; ANCOVA; Table S3). This shows that one can predict the responses to the video from static snapshot responses better for VP than for DP. For the time-reversed video, the median correlation was 0.40 and 0.05 for VP and DP, respectively, both significantly larger than zero (Wilcoxon test; VP, p = 1.711e−10; DP, p = 0.0156; Figure 4F; individual monkey and patch data in Figure S2D_{1-2}).

The median correlations between the snapshot responses and those for the time-reversed video were (marginally) significantly lower than those for the original video only for VP and not for DP (Wilcoxon signed rank test; VP, p = 0.043; DP, p = 0.196; Bonferroni corrected p values). Nonetheless, the correlations between the responses to the snapshot and the frames in the original video correlated with those computed for the reversed video for VP neurons (Spearman ρ = 0.46, p = 5.437e−08, n = 125; significant for neurons with BSI > 0.33; Figure S2D_3). However, no such correlation was present for the DP neurons (Spearman ρ = −0.008, p = 0.927, n = 121; Figures 4F and S2E). This demonstrates that the response to a frame of a video can depend on the sequence in which that frame is presented, and this is to a larger extent in DP than in VP. This is in line with the stronger sequence sensitivity of DP neurons. In sum, for most VP neurons that respond to static presentations, the selectivity for static images predicts the responses to those images when presented as frames in a video. This holds less for DP neurons, showing a weaker association between responses to static images and videos.

Neural responses to dynamic bodies are predicted by velocity pattern

Next, we examined to what extent the body video responses can be explained by motion and static features. First, we

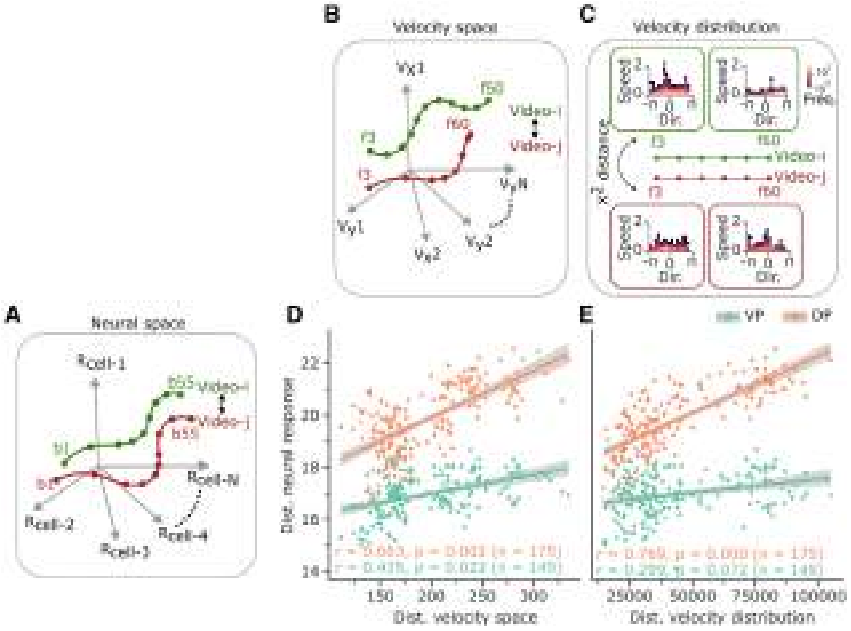


Figure 5. Correlation of velocity-based and neural distances

(A) Illustration of neural response trajectories for a pair of videos, depicted by red and green lines, in an n-dimensional space spanned by responses of n cells. The dots on the trajectories correspond to responses in successive 20 ms bins, ranging from bin 1 (b1) to bin 55 (b55) covering 1,100 ms (starting 60 ms after onset). The bidirectional arrow between videos indicates the lock-step Euclidean distance between the trajectories.

(B) Illustration of trajectories and distances of and between a pair of videos in the n (number of pixels per frame) × 2 (x and y components of the velocity per pixel) velocity space. The dots on the trajectories represent the frames (3–60). Same conventions as in (A).

(C) Illustration of the computation of the chi-squared distance between the velocity distributions of a pair of videos (red and green lines). Insets illustrate the 2D frequency distribution of velocity (speed and direction) for frames 3 and 60, with frequency (number of speed × direction combinations per bin) plotted as a heatmap (hot color map).

(D) Scatterplot of velocity space and neural distances (190 video pairs) for each patch, and linear regression lines with 95% confidence intervals. Pearson r and p values (STAR Methods) are given for VP (green) and DP (orange).

(E) Scatterplot of velocity distribution distances and neural distances. Same conventions as in (D). See also Figure S3.

computed a neural distance metric for all body video pairs, based on the lock-step Euclidean distance (STAR Methods) between the response trajectories to the videos in neural space (Figure 5A). Each dimension of the neural space corresponds to a neuron and each point in the neural space represents a response for a 20 ms bin and video. Note that the neural distance metric reflects the moment-by-moment difference in neural response between videos, unlike a distance metric computed on the responses averaged across the whole stimulus duration.

To relate the neural distances to differences in motion among the body videos, we computed two velocity-based distance metrics: one in which we computed pairwise, pixel-wise velocity differences between videos and a second in which we computed pairwise differences between the frequency distributions of the velocities. Unlike the first metric, the velocity distribution metric does not consider the spatial location of the velocity vectors but only their frequency distribution per frame and hence is a position-invariant metric. To compute the "velocity space" distances (metric 1), we defined a velocity space in which the dimensions correspond to the velocity component for a particular pixel in the horizontal or vertical axis (Figure 5B). The velocity of each pixel per frame (STAR Methods) corresponded to a point in velocity space. Then, we computed the lock-step Euclidean distance for all video pairs using the same procedure as for the neural responses, thus providing a velocity-based distance for all video pairs. For the second metric, we computed the 2D frequency distribution of velocity (direction and speed) per frame (Figure 5C; STAR Methods). Then, we computed the pairwise distance between the velocity distributions using the chi-squared distance metric (Figure 5C; STAR Methods). We thresholded the speed before computing the distances by requiring a minimum speed. The reported effects were quite robust with respect to differences in threshold speed (Figure S3B), and we report results with a threshold of 0.2 (arbitrary units). The velocity-based distance metrics correlated to some extent (Pearson r = 0.85; Figure S3A).

The neural distances for bodies correlated significantly with both velocity-based distance metrics for the DP neurons (Figures 5D and 5E; stimulus label permutation test[31]). The correlation between the VP neural distances and the velocity space distances was also significant (p = 0.022; Figure 5D), but the correlation between the VP neural distances and the velocity distributions failed to reach significance (p = 0.07; Figure 5E). The correlations for both velocity-based distances were significantly lower in VP than in DP (velocity space p = 3.600e−03; velocity distribution p < 0.001; bootstrapping neurons [1,000 resamplings]; STAR Methods). This suggests that motion contributes more to DP than to VP dynamic body responses. To assess whether the speed distribution is sufficient to determine the neural distances, we binned the velocity vector magnitude (speed), ignoring motion direction. The correlation of the pairwise speed distribution and the neural distances was significant for DP, although the correlation was lower than computed for both speed and direction (Figure S3B), suggesting a contribution of speed and, to a smaller extent, motion direction, to DP body responses.

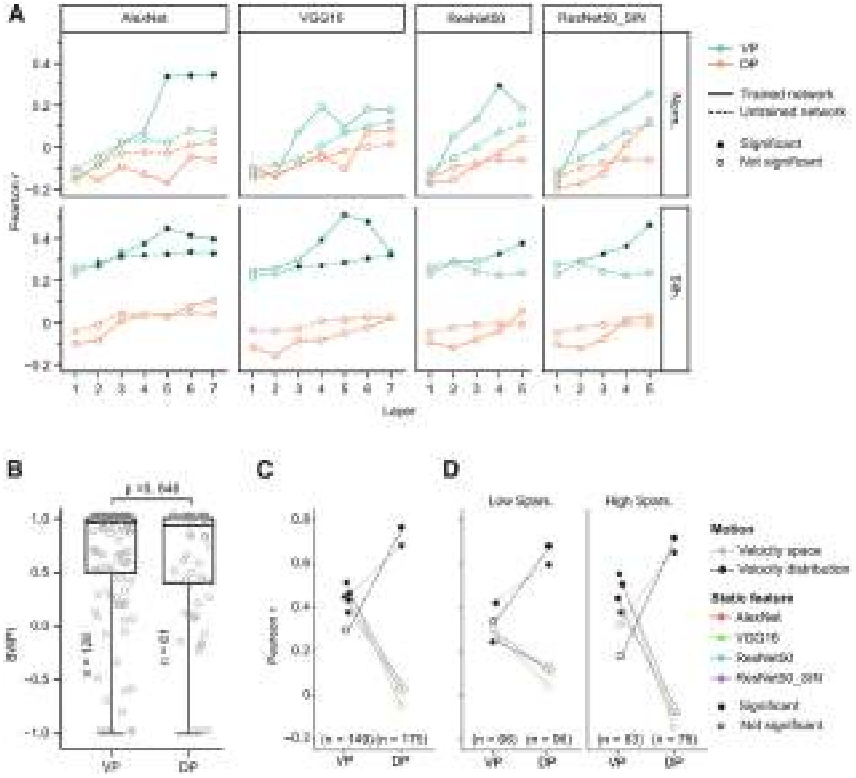


Figure 6. Correlations between CNN feature distances and neural distances

(A) The first row corresponds to correlations for the original videos with shaded, textured bodies (Norm.). The second row corresponds to correlations of the neural distances for the original videos with the CNN-feature distances for silhouette videos. A solid line corresponds to a trained network, while a dashed line corresponds to an untrained network. Solid markers depict a significant correlation ($p < 0.05$; Bonferroni corrected). For the definition of the layers, see STAR Methods. The correlations are plotted for VP and DP separately. Scatterplots are shown in Figure S4A_{1-2}.

(B) Distribution of best-worst preference index (BWPI) for responses to silhouette videos for VP and DP. The p value corresponds to a Wilcoxon rank-sum test.

(C) Summary plot of the correlations between the neural and the velocity-based distances (dashed lines) and between neural and CNN layer 5 static features distances (computed for silhouettes; colored lines). Solid markers correspond to significant ($p < 0.05$) values. Correlations are shown for VP and DP.

(D) Plots as in (C) for cells with a Sparseness below (Low Spars.) or above (High Spars.) the median Sparseness. See also Figures S4–S6.

Responses to dynamic bodies are predicted by CNN shape features in VP but not DP

To assess the contribution of static features to dynamic body responses, we presented the frames of the body videos to CNNs: AlexNet,[32] VGG16,[33] ResNet50,[34] and ResNet50_SIN.[35] We employed networks pretrained in the classification of ImageNet[36] data (AlexNet, VGG-16, and ResNet50) and stylized ImageNet[35] (ResNet50_SIN) and untrained ones as control. We computed for each video pair the lock-step Euclidean distance in the space in which each dimension corresponds to a unit of a layer (STAR Methods). We correlated the CNN-based distances for each layer with the neural distances. For VP neurons, the correlations between neural and trained CNN feature distances increased with the layer, reaching significance for AlexNet and ResNet50 (Figure 6A; stimulus label permutation test). Correlations for the untrained AlexNet were smaller than for the trained versions for the deeper layers. The correlations between trained CNN-based distances and neural distances did not differ significantly from zero for the DP neurons (Figure 6A). The correlations of CNN feature distances of layer 5 and the VP neural distances were larger than those for the DP neural distances, the difference being significant for AlexNet (bootstrapping neurons; $p < 0.001$). This provides some evidence that VP responses to dynamic bodies are more related to static features than DP responses.

Ventral middle STS body patch neurons preserve their selectivity when static natural images are transformed into silhouettes, indicating that shape features underlie their body selectivity.[37] We measured the responses of 128 VP and 61 DP body-responsive neurons to silhouette versions of the body videos. We presented to each neuron two silhouette videos, one corresponding to the original video that produced the strongest response ("best") and a second one that corresponded to an original video that produced no or a weak response ("worst"). We computed for each neuron a best-worst preference index (BWPI), which contrasted the responses to the best and worst silhouette video (STAR Methods). The best and worst silhouette videos were defined based on the responses to the original videos. A BWPI of 1 indicates no excitatory response to the worst silhouette video, while 0 corresponds to an equal response to the best and worst silhouette videos. The median BWPI was high and very similar for both regions (VP median 0.97; DP, 0.95; Wilcoxon rank-sum test; p = 0.65; Figure 6B; single-patch data and for neurons with BSI > 0.33; Figure S4B_{1-2}). Hence, the body-video selectivity was preserved for silhouette versions in both regions.

The preserved selectivity for silhouettes raised the question of whether the neural responses for the original body videos correlate with CNN responses to the silhouettes. Hence, we employed the same procedure as above to compute pairwise distances between the silhouette videos for each CNN layer and correlated these with the neural distances of VP and DP neurons for the original videos. The correlations of the VP distances with the silhouette feature distances were higher than those for the original videos (Figure 6A), despite the fact that we correlated CNN activations to silhouettes and neural responses to the grayscale videos. The CNNs that did not have a significant correlation between CNN and neural distance for the original videos showed significant correlations for the deeper layers when silhouette videos were used as input. The silhouette video distances of the deeper layers of the untrained CNNs also correlated significantly with the VP distances, although they were lower than for the trained CNNs. A parsimonious explanation for the increased correlations of VP and CNN silhouette responses is that VP neurons respond primarily to shape features that are preserved when transforming grayscale images to silhouettes. The DP responses did not show significant correlations with CNN silhouette features. Furthermore, for layer 5 of each CNN, the correlation of CNN silhouette feature distances and VP neural distances was significantly larger than those for DP (AlexNet, $p = 0.012$; VGG16, $p < 0.001$; ResNet50, $p = 2.000e{-}03$; ResNet50_SIN, $p = 8.000e{-}03$; bootstrapping neurons). This suggests that the DP responses to dynamic bodies are mainly driven by motion, whereas the VP responses are driven more by spatial features.

The dissociation of the contribution of motion versus static features to DP and VP dynamic body responses is summarized in Figure 6C. The differences between VP and DP in the correlations between the neural distances and the velocity-based/static CNN distances remained after equating the BSI distribution of VP and DP neurons (Figure S4$D_{1\text{-}2}$) and thus did not result from DP neurons having a lower average BSI. Also, trends were similar when comparing ASB with DP neurons (Figure S4D_3).

Further analysis showed that the correlation between velocity space distances and neural distances was significant only for VP neurons with a Sparseness lower than the median of the population of neurons (Figures 6D and S4C). This effect of Sparseness on the contribution of motion to the neural distances was intensified for VP neurons with BSI > 0.33 (Figure S4C). The correlations between neural distances and velocity-based metrics were significant and similar for high- and low-sparseness DP neurons (Figures 6D and S4C). The low, non-significant correlations between velocity and distances for the high-sparseness VP neurons could be because, for those neurons, the response differences among the bodies were strongly driven by static features. This aligns with the significant correlations between the static feature and the high-sparseness VP distances for all networks (and neurons with BSI > 0.33), whereas the correlation for the static feature distances was significant only for one network for the low-sparseness neurons (Figures 6D and S4C). Alternatively, high-sparseness VP neurons may show motion pattern selectivity that is not captured by our velocity-based metrics. Notably, motion sensitivity as such, as measured by SSI and VRI, did not decrease with Sparseness (correlation between sparseness and SSI, Spearman $\rho = 0.18$ ($p = 0.04$), and VRI, $\rho = 0.05$ [$p = 0.57$]).

To assess whether motion and static features explain a common portion of the response variance of VP neurons, we employed commonality analysis (STAR Methods). This was done for the layer 5 distances of each CNN (silhouettes as input) and the velocity distribution-based distances (Figure 7A). Multiple regression produced significant correlations when using as predictors the CNN feature and the velocity distribution-based distances (stars in Figure 7A). For VP, the velocity distribution and layer 5 feature-based distances explained each a unique part of the neural distances (the slight negative commonalities for some networks reflect small negative correlations between the velocity- and the feature-based distances). For DP, only the velocity distribution-based distances explained a unique variance component of the neural distances.

We hypothesized that neurons that respond more strongly to videos than static images and neurons that are sensitive to time reversal of videos rely more on motion features than those that are insensitive to time reversal. Thus, we distinguished "static" neurons with SSI and VRI < 0.2, and "motion" neurons with SSI or VRI > 0.2 (stippled lines in Figure 5C). The commonality analysis showed a stronger contribution of the velocity-based than the CNN-feature distances for the "motion" VP neurons, while the opposite was the case for the "static" VP neurons (Figure 7B). For DP, there was a reduction of the contribution of motion for the "static" compared with the "motion" neurons, but this could be due to the smaller number of "static" DP neurons. Even for the "static" DP neurons, the CNN-feature distances showed little or no correlation with neural distances (Figure 7B).

Correlating neural distances with a spatiotemporal network

Earlier, to assess the contribution of static features, we correlated neural responses with activations of CNNs pretrained with static images, which included animals (ImageNet). We compared the neural distances also with distances computed from spatiotemporal network units pretrained to recognize human action videos (X3D[38]; STAR Methods), encoding sequence information. The significant correlations between the VP neural and the X3D distances increased with layer, yielding slightly larger values than the "static" CNNs. Although X3D-DP correlations were higher than "static" CNN-DP correlations, none reached significance (Figure S5).

DISCUSSION

We showed that some anterior ventral STS/IT (VP) neurons required motion to respond to a monkey body, while others responded to static bodies but were highly selective for the temporal sequence of images in a video of an acting monkey. Hence, the response of some VP neurons to body actions cannot be solely predicted from their selectivity to static images. In addition, other VP neurons responded equally to static presentations of bodies and the same frames during a video, and their response during the video could be predicted by their static image selectivity. Dorsal-bank STS (DP) neurons exhibited a stronger effect of motion and stronger sequence sensitivity than VP neurons. A population analysis showed that the dynamic body responses of DP neurons could be predicted from the velocity distributions present in the videos, but not from static CNN-based features. In contrast, the responses of VP neurons to the body videos were predicted by both static CNN features and velocity differences.

Our findings suggest a revision of the traditional view that distinguishes dorsal and ventral STS processing in terms of motion versus static features. First, we found that DP neurons, although dominated by motion, can respond to static stimuli. However, their response to dynamic bodies is not well predicted by the response to the same images presented statically. Second, the responses of VP neurons to dynamic bodies can be driven by motion and static shape features. Overall, our study highlights

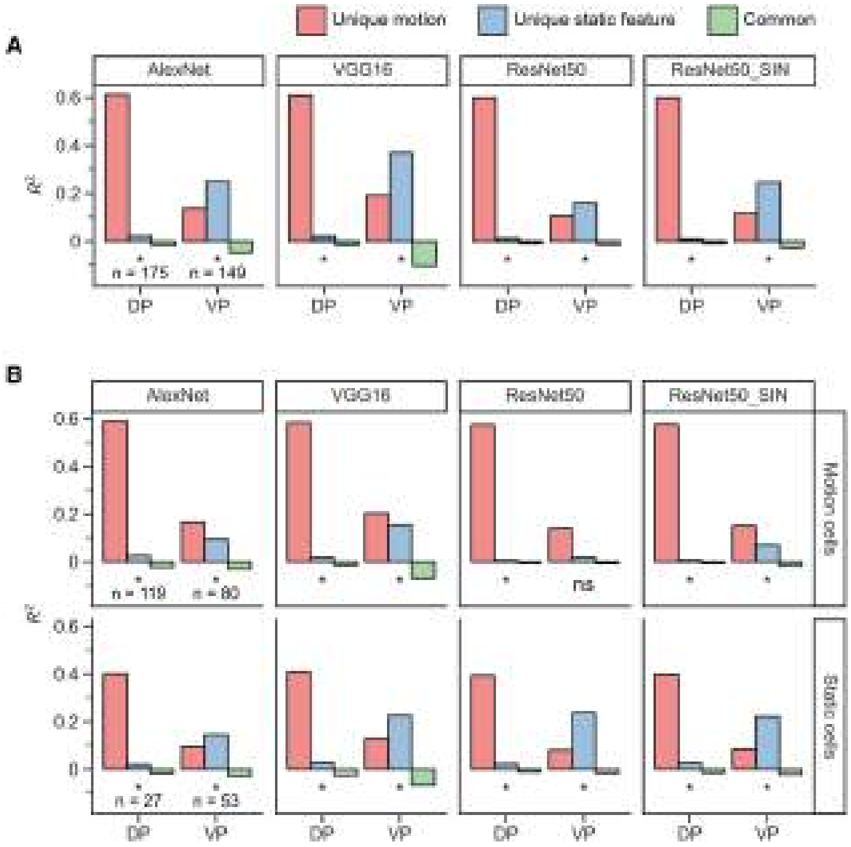


Figure 7. Commonality analysis: Contribution of motion and static features to the neural responses in DP and VP

(A) Each column shows the unique explained variance by motion (red) and static features (blue) for layer 5. The common explained variance is in green. Stars indicate a significant multiple regression correlation coefficient for the region and network (permutation test; STAR Methods).

(B) Commonality analysis plots for the "motion" and "static" cells. Same conventions as (A).

the diversity of the neural mechanisms involved in the processing of body movements and points toward the need for a more nuanced understanding of how ventral and dorsal STS neurons contribute to this process.

A recent study that recorded face-selective neurons in a dorsal STS face patch also suggested a strong contribution of motion to responses to dynamic faces.[39] However, they did not examine responses to dynamic faces in IT face patches. Furthermore, our study shows that the correlation between neural and velocity-based distances is not determined merely by differences in motion energy because neurons responded differently to videos and their time-reversed versions with the same motion energy.

The two velocity-based distance metrics yielded similar results, although the velocity distribution distances tended to show lower correlations than the velocity space distances for VP but not DP. One possible explanation is that VP neurons, especially those with a high sparseness, might be more sensitive to the spatial layout of the motion pattern, which is reduced in the velocity distribution, but this requires direct testing.

We believe the responses of the DP neurons to dynamic bodies could not be predicted by static CNN features because the responses to the body videos were strongly dominated by motion. At the single neuron level, we found a correlation between the selectivity for static presentations of body images and the responses to the same frames during the video, but these correlations were low and depended strongly on the frame sequence. This agrees with the suggestion that DP neurons are strongly driven by the motion component in the videos and that the motion response overrides their static feature selectivity.

The most effective stimuli of the DP neurons were threatening actions directed toward the observer (Video S1). Such threatening displays, which are important in the social life of monkeys, include jerky movements that drive the DP neurons effectively. In contrast, socially neutral actions such as walking and grasping elicited smaller responses. This highlights the potential importance of threat-related stimuli in driving the response of dorsal-bank STS neurons, suggesting that they may play a role in the detection of potential threats. This supports the proposal that the dorsal STS functions as a "dynamic social stream."[7]

VP neurons did not merely respond to the momentary body shape in dynamic displays. First, some VP neurons weakly responded to static bodies while responding strongly to the same frames in the context of the video. Second, some VP neurons were highly sensitive to the frame sequence. This sequence sensitivity was even present for neurons that responded to static presentations of the frames. Third, our population response analysis showed that velocity-based distances correlate with VP neural distances, suggesting the presence of motion information in anterior IT.

Russ et al. suggested that ventral face patch AM responses to 5-min-long videos differ from those to static presentations or short dynamic snippets.[40] However, the AM experiments did not control for eye fixation differences between these conditions, which makes the AM responses difficult to relate to dynamics per se. It is unlikely that the same mechanisms underlie the putative sequence effects during a long movie in AM[40] and the sequence sensitivity we observed in VP, because our movies were only 1 s long, and the dynamics we examined occurred at a shorter time scale.

Selectivity for motion-defined shapes has been reported in ventral stream areas V4[41] and IT,[42] but in those studies, motion served as a segmentation cue. Long-range motion direction selectivity has been demonstrated in V4[43] and may have contributed to the sequence sensitivity observed here in VP and DP.

Apart from short- and long-range motion, other mechanisms can contribute to the sequence sensitivity in VP and DP. One candidate is adaptation, which is prevalent in IT.[44] Adaptation can induce differences in response depending on the sequence in which effective and ineffective frames are presented, but cannot explain the strong sequence sensitivity in which no excitatory response was present for the time-reversed video. Another candidate mechanism is expectation suppression, in which responses to expected stimuli are reduced compared with unexpected stimuli.[45,46] Such a mechanism predicts a stronger response to the unexpected, less familiar, time-reversed videos than to the familiar original videos. However, this is opposite to the typical stronger response to the original than to the time-reversed video. A third mechanism is suggested by the observed association of strong sequence sensitivity and inhibition for the ineffective sequence. This is in line with models[2] that propose that sequence-selective neurons have asymmetric lateral connections with other neurons that encode individual snapshots from the motion sequence. Neurons that encode a sequence will receive excitatory input from neurons that encode the preceding snapshot of the sequence, but inhibitory input from neurons that encode the preceding snapshot of the time-reversed sequence. The VP neurons that encode snapshots can contribute to the sequence sensitivity observed in other VP neurons through such a network mechanism. Other mechanisms involving recurrent processing within temporal cortex or feedback from other regions, e.g., prefrontal cortex,[47] may underlie the sequence sensitivity. Also, the sequence sensitivity may have been inherited from middle STS body-responsive regions.[3]

Although some VP neurons showed sequence-related responses, the responses of many VP neurons to the videos were well predicted by their selectivity for static frames. Previous studies with static images showed a correlation between deep-layer CNN features and IT responses.[23–28] Here, we show that the correlation between CNN features and VP neurons extends to dynamic body stimuli. Interestingly, the correlation between CNN features and VP responses was stronger when silhouettes served as input to the CNN. This can be related to our observation that VP neurons keep their selectivity when the video is reduced to its silhouette version, suggesting that VP responds primarily to shape features. We speculate that the original images produce strong texture-driven responses in the CNN units, which reduced the correlations with the shape-driven neural responses. When using silhouettes as input to the CNN, its units will be driven by shape instead of by the absent texture features, enhancing the correlation between CNN unit activations and the shape-driven VP responses. The higher correlation with silhouette input was also present for ResNet-50-SIN,[35] which was trained on images in which the original texture of a shape was replaced by random textures, forcing the network to utilize shape for categorization. This suggests that this network still contains units that are texture selective.

We correlated the body responses with a spatiotemporal network pretrained for human action recognition, which produced numerically somewhat higher correlations compared with the "static" CNNs for both regions. Adding temporal information to the CNN did not produce a significant increase in correlations for DP, which was unexpected, because other analyses showed that motion dominated the responses of DP neurons to the videos. However, this might be due to the peculiarities of the employed CNN. Investigating spatiotemporal networks as a model for temporal cortical processing holds promise for future research.

We showed that, whereas the body responses of DP neurons are well predicted by velocity patterns, this is less the case for VP neurons, which are driven more, but not exclusively, by static shape features. The commonality analysis showed that velocity and shape features explain non-overlapping portions of the response variance of the VP neurons. Moreover, VP neurons that responded equally well to static and dynamic stimuli and/or showed low or no sequence sensitivity have a relatively stronger contribution of static features than neurons that respond more to the dynamics, suggesting that anterior IT contains a heterogeneous population of neurons that vary in their motion versus shape processing.

Limitations of the study

The response selectivity of DP neurons for dynamic bodies was not correlated with CNN features. Since we did not examine DP responses to a wide range of static stimuli, we do not know whether the static feature selectivity of DP neurons relates to CNN features or is fundamentally different from IT. Although the DP region corresponded to the most anterior dorsomedial STS fMRI-defined body patch, it was posterior to the VP. It is unlikely that the differences in response properties between DP and VP are related to their anterior-posterior locations, as DP's motion sensitivity is in line with prior dorsal-bank STS studies[8–10,4,11–15]. Future studies should examine dynamic body responses in more posterior ventral and dorsal STS regions. Because we targeted a limited number of patches, we do not know whether our data generalize to neurons outside the recorded regions in IT and dorsal STS. Examining the relation between responses and neural networks trained on video data for monkey action recognition could offer valuable insights, but this falls beyond the scope of our study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Subjects and surgery
- METHOD DETAILS
  - ○ Stimuli
  - ○ fMRI body patch localizer
  - ○ Single-unit recordings
  - ○ Tests
  - ○ Data analysis
  - ○ Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113438.

ACKNOWLEDGMENTS

The authors thank C. Fransen, I. Puttemans, A. Hermans, W. Depuydt, C. Ulens, S. Verstraeten, K. Lodewyckx, J. Helin, and M. De Paep for technical and administrative support. This research was supported by Fonds Wetenschappelijk Onderzoek (FWO) Vlaanderen (G0E0220N), KU Leuven grant C14/21/111, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 856495).

AUTHOR CONTRIBUTIONS

Conceptualization, R.V., M.G., R.R., and A.B.; methodology, R.V., A.B., R.R., and N.T.; investigation, R.V., A.B., R.R., and G.G.N.; writing – original draft, R.V. and R.R.; writing – review & editing, R.V., R.R., A.B., and M.G.; funding acquisition, R.V. and M.G.; resources, R.V.; supervision, R.V.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 7, 2023
Revised: September 26, 2023
Accepted: October 26, 2023
Published: November 22, 2023

REFERENCES

1. de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philos. Trans. R. Soc. Lond. B Biol. Sci. *364*, 3475–3484.
2. Giese, M.A., and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. Nat. Rev. Neurosci. *4*, 179–192.
3. Vogels, R. (2022). More Than the Face: Representations of Bodies in the Inferior Temporal Cortex. Annu. Rev. Vis. Sci. *8*, 383–405.
4. Jellema, T., and Perrett, D.I. (2006). Neural representations of perceived bodily actions using a categorical frame of reference. Neuropsychologia *44*, 1535–1546.
5. Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108.
6. Bognár, A., Raman, R., Taubert, N., Zafirova, Y., Li, B., Giese, M., De Gelder, B., and Vogels, R. (2023). The contribution of dynamics to macaque body and face patch responses. Neuroimage *269*, 119907.
7. Pitcher, D., and Ungerleider, L.G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. Trends Cognit. Sci. *25*, 100–110.
8. Oram, M.W., and Perrett, D.I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J. Neurophysiol. *76*, 109–129.
9. Oram, M.W., and Perrett, D.I. (1994). Responses of Anterior Superior Temporal Polysensory (STPa) Neurons to "Biological Motion" Stimuli. J. Cognit. Neurosci. *6*, 99–116.
10. Wachsmuth, E., Oram, M.W., and Perrett, D.I. (1994). Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque. Cerebr. Cortex *4*, 509–522.
11. Barraclough, N.E., Xiao, D., Oram, M.W., and Perrett, D.I. (2006). The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. Prog. Brain Res. *154*, 135–148.
12. Jellema, T., Baker, C.I., Wicker, B., and Perrett, D.I. (2000). Neural representation for the perception of the intentionality of actions. Brain Cognit. *44*, 280–302.
13. Jellema, T., Maassen, G., and Perrett, D.I. (2004). Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey. Cerebr. Cortex *14*, 781–790.
14. Jellema, T., and Perrett, D.I. (2003). Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? Neuropsychologia *41*, 1728–1737.
15. Vangeneugden, J., De Mazière, P.A., Van Hulle, M.M., Jaeggli, T., Van Gool, L., and Vogels, R. (2011). Distinct mechanisms for coding of visual actions in macaque temporal cortex. J. Neurosci. *31*, 385–401.
16. Vangeneugden, J., Pollick, F., and Vogels, R. (2009). Functional differentiation of macaque visual temporal cortical neurons using a parametric action space. Cerebr. Cortex *19*, 593–611.
17. Bruce, C., Desimone, R., and Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. *46*, 369–384.
18. Anderson, K.C., and Siegel, R.M. (1999). Optic flow selectivity in the anterior superior temporal polysensory area, STPa, of the behaving monkey. J. Neurosci. *19*, 2681–2692.
19. Baylis, G.C., Rolls, E.T., and Leonard, C.M. (1987). Functional subdivisions of the temporal lobe neocortex. J. Neurosci. *7*, 330–342.
20. Singer, J.M., and Sheinberg, D.L. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J. Neurosci. *30*, 3133–3145.
21. Yang, Z., and Freiwald, W.A. (2021). Joint encoding of facial identity, orientation, gaze, and expression in the middle dorsal face area. Proc. Natl. Acad. Sci. USA *118*, e2108283118.
22. Barraclough, N.E., Keith, R.H., Xiao, D., Oram, M.W., and Perrett, D.I. (2009). Visual adaptation to goal-directed hand actions. J. Cognit. Neurosci. *21*, 1806–1820.
23. Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., and DiCarlo, J.J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol. *10*, e1003963.
24. Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., and DiCarlo, J.J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat. Neurosci. *22*, 974–983.
25. Ponce, C.R., Xiao, W., Schade, P.F., Hartmann, T.S., Kreiman, G., and Livingstone, M.S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. Cell *177*, 999–1009.e10.
26. Kalfas, I., Vinken, K., and Vogels, R. (2018). Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments. PLoS Comput. Biol. *14*, e1006557.
27. Kalfas, I., Kumar, S., and Vogels, R. (2017). Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. eNeuro *4*. ENEURO.0113 17.2017.
28. Raman, R., and Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. Commun. Biol. *3*, 221.
29. Popivanov, I.D., Jastorff, J., Vanduffel, W., and Vogels, R. (2014). Heterogeneous Single-Unit Selectivity in an fMRI-Defined Body-Selective Patch. J. Neurosci. *34*, 95–111.
30. Kumar, S., Popivanov, I.D., and Vogels, R. (2019). Transformation of Visual Representations Across Ventral Stream Body-selective Patches. Cerebr. Cortex *29*, 215–229.
31. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS Comput. Biol. *10*, e1003553.

32. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Commun. ACM *60*, 84–90.

33. Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at arXiv.

34. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. Preprint at arXiv.

35. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., and Wieland, B. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. Preprint at arXiv.

36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. *115*, 211–252.

37. Popivanov, I.D., Jastorff, J., Vanduffel, W., and Vogels, R. (2015). Tolerance of macaque middle STS body patch neurons to shape-preserving stimulus transformations. J. Cognit. Neurosci. *27*, 1001–1016.

38. Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. Proc Cvpr Ieee, 200–210.

39. Yang, Z., and Freiwald, W.A. (2023). Encoding of dynamic facial information in the middle dorsal face area. Proc. Natl. Acad. Sci. USA *120*, e2212735120.

40. Russ, B.E., Koyano, K.W., Day-Cooney, J., Perwez, N., and Leopold, D.A. (2023). Temporal continuity shapes visual responses of macaque face patch neurons. Neuron *111*, 903–914.e3.

41. Mysore, S.G., Vogels, R., Raiguel, S.E., and Orban, G.A. (2008). Shape selectivity for camouflage-breaking dynamic stimuli in dorsal V4 neurons. Cerebr. Cortex *18*, 1429–1443.

42. Sáry, G., Vogels, R., and Orban, G.A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. Science *260*, 995–997.

43. Bigelow, A., Kim, T., Namima, T., Bair, W., and Pasupathy, A. (2023). Dissociation in neuronal encoding of object versus surface motion in the primate brain. Curr. Biol. *33*, 711–719.e5.

44. Vogels, R. (2016). Sources of adaptation of inferior temporal cortical responses. Cortex *80*, 185–195.

45. Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proc. Natl. Acad. Sci. USA *108*, 19401–19406.

46. Kaposvari, P., Kumar, S., and Vogels, R. (2018). Statistical Learning Signals in Macaque Inferior Temporal Cortex. Cerebr. Cortex *28*, 250–266.

47. Kar, K., and DiCarlo, J.J. (2021). Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition. Neuron *109*, 164–176.e5.

48. Vanduffel, W., Fize, D., Mandeville, J.B., Nelissen, K., Van Hecke, P., Rosen, B.R., Tootell, R.B., and Orban, G.A. (2001). Visual motion processing investigated using contrast agent-enhanced fMRI in awake behaving monkeys. Neuron *32*, 565–577.

49. Ekstrom, L.B., Roelfsema, P.R., Arsenault, J.T., Bonmassar, G., and Vanduffel, W. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. Science *321*, 414–417.

50. Kolster, H., Mandeville, J.B., Arsenault, J.T., Ekstrom, L.B., Wald, L.L., and Vanduffel, W. (2009). Visual field map clusters in macaque extrastriate visual cortex. J. Neurosci. *29*, 7031–7039.

51. Leite, F.P., Tsao, D., Vanduffel, W., Fize, D., Sasaki, Y., Wald, L.L., Dale, A.M., Kwong, K.K., Orban, G.A., Rosen, B.R., et al. (2002). Repeated fMRI using iron oxide contrast agent in awake, behaving macaques at 3 Tesla. Neuroimage *16*, 283–294.

52. Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. Eur. J. Neurosci. *11*, 1239–1255.

53. Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. *73*, 713–726.

54. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The Kinetics Human Action Video Dataset. Preprint at arXiv.

55. Tao, Y., Both, A., Silveira, R.I., Buchin, K., Sijben, S., Purves, R.S., Laube, P., Peng, D., Toohey, K., and Duckham, M. (2021). A comparative analysis of trajectory similarity measures. GIsci. Remote Sens. *58*, 643–669.

56. Warne, R.T. (2011). Beyond Multiple Regression: Using Commonality Analysis to Better Understand R-2 Results. Gift. Child. Q. *55*, 313–318.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Electrophysiological data	This paper	https://osf.io/jsm5z/
Chemicals, peptides, and recombinant proteins		
Molday ION	BioPAL Inc.; Worcester, USA	http://www.biopal.com/molday-ion.htm
Experimental models: Organisms/strains		
Rhesus macaque (*Macaca mulatta*)	Biomedical Primate Research Centre https://bprc.nl/en (BPRDC), Rijswijk, the Netherlands	https://bprc.nl/en
Software and algorithms		
MATLAB R2020a	MathWorks	RRID: SCR_001622
Python Programming Language	Python	RRID: SCR_008394
PyTorch	PyTorch	RRID: SCR_018536
SPM12	SPM, University College London, UK	RRID: SCR_007037
Other		
EyeLink	SR-Research	https://www.sr-research.com
Tungsten Microelectrode	FHC	www.fh-co.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed and will be fulfilled by the Lead Contact, Rufin Vogels (Rufin.vogels@kuleuven.be).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data are available at https://osf.io/jsm5z/.
- Original code is available at https://github.com/RajaniRaman/dynamic_body as of the date of publication.
- Any additional information required to analyze the data is available from the lead contact on request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Subjects and surgery

Three male rhesus monkeys (*Macaca mulatta;* 5-6 years old) served as subjects. The monkeys were housed in pairs or triplets. The monkeys were implanted with a plastic headpost, using ceramic screws and dental cement following standard aseptic procedures and full anesthesia. They were trained to fixate continuously on a small target point for juice rewards. After the fMRI scanning, we implanted a custom-made plastic recording chamber, allowing a dorsal approach to temporal body patches. In each animal, the location of the recording chamber was guided by the fMRI body localizer. Animal care and experimental procedures complied with the regional (Flanders) and European guidelines and were approved by the local Animal Ethical Committee.

METHOD DETAILS

Stimuli

Main stimuli

We employed 60 achromatic videos: 20 dynamic body videos, 20 dynamic objects, and 20 dynamic faces. These stimuli were identical to those employed by[6] in the fMRI mapping, except for their somewhat smaller size (5 instead of 6 deg), and are described in that paper. The duration of each video was 1 sec. The dynamic body videos show a rhesus monkey performing different natural actions like grasping, picking, turning, walking, threatening, throwing, wiping, and initiating jumping. The face of the monkey was blurred, making its facial expression and identity unrecognizable. The translational component of the movement of the monkey across the display, when present (e.g. during walking), was removed and the monkey's body was centered. The 20 face videos included 12 movies of monkey faces that showed frontal face movements such as chewing, lip-smacking, fear grin, and threat. The other 8 face videos showed a moving monkey head with visible facial features, e.g. a head rotating from a frontal to a profile view. The 20 object movies included computer-rendered objects that depicted movement, e.g. an object with its parts making non-rigid movements, a rotating airplane, and cars with different motion patterns (e.g. rocking or "jumping"). For each category, the maximal extent of the centered moving stimuli fitted a 5 by 5 deg square for the singe-unit recordings. All movies were rendered with a 60 Hz frame rate. Bodies, faces, and objects were presented on top of a dynamic white noise background (size = 11 deg). The gray level of each noise background pixel was randomly sampled from a uniform distribution at a rate of 30 Hz.

In the fMRI mapping design (see below), we included also mosaic-scrambled videos, as described in[6]. These scrambled conditions were not employed to define the body patches in the present study and were not presented in the single-unit study.

Silhouette stimuli

For each body video, we prepared a silhouette version in which the pixels corresponding to the monkey were rendered black. Thus, the overall motion and shape of the monkey were preserved while the inner features of the body, its texture, and shading pattern were eliminated.

Snapshot stimuli

For each body video, we selected 10 frames, including the white noise background, that sampled different postures and views of the body during the video. Note that we sampled frames of some of the videos at an irregular temporal interval to ensure that we obtained a representative sample of the variety of postures and views present in a video and that the same posture was not presented more than once.

fMRI body patch localizer

Body patches activated by dynamic bodies were localized with fMRI preceding the recordings. The scanning procedure and details of the fMRI data analysis have been described in[6] and will be summarized here briefly.

During scanning, the monkeys sat in a horizontal sphinx position with their heads fixed in an MRI-compatible chair. The chair was positioned in front of a translucent screen on which the stimuli were projected. Eye position was monitored (120 Hz; Iscan) and the animals were performing a fixation task during scanning for a juice reward. The monkeys were scanned with a 3T Siemens Trio scanner following standard procedures[48];[49],[50]. We obtained high-resolution anatomical MRI images for each monkey in a separate session. To increase the signal-to-noise ratio[51], we injected intravenously the contrast agent Monocrystalline Iron Oxide Nanoparticle (MION) in each daily scanning session.

To localize body patches, we employed a block design: the 1-sec videos (n = 20) of a category were presented in a block back to back in random order. A run consisted of 7 conditions, each repeated twice using a palindromic sequence. The 7 conditions were: (1) a baseline fixation block in which the fixation target (size = 0.2 deg) was shown together with a dynamic white noise background, (2) moving bodies, (3) moving faces, (4) moving objects, (5) mosaic-scrambled moving bodies, (6) mosaic-scrambled moving faces, and (7) mosaic-scrambled moving objects. The order of the blocks was randomized across runs with a balanced Latin square design. The maximal extent of the stimuli was 6 by 6 deg. We used only runs in which the monkeys were fixating (fixation window size 2-3 deg) at least 90% of the run (monkey O: 38, N: 39, G: 32 valid runs). They were analyzed with a general linear model with 7 regressors (6 stimulus conditions + baseline fixation condition), plus 9 additional head-motion and eye-movement regressors per run (see for more details:[6]). We employed the following contrast to identify the body patches for the single-unit recordings: moving bodies – moving objects (threshold Family-wise error (FWE) rate corrected; $p < 0.05$), exclusively masked with moving faces – moving objects ($p < 0.001$; uncorrected). The resulting t-maps of each monkey were co-registered to their anatomical MRIs so that the body patches could be identified in each monkey's native space.

Single-unit recordings

Single-unit recordings were performed with epoxylite-insulated tungsten microelectrodes (FHC), with an impedance of around 1 MOhm, using techniques as described previously[29]. Briefly, the electrode was lowered with a Narishige microdrive into the brain using a stainless-steel guide tube that was fixed in a custom-made grid that was positioned within the recording chamber. After amplification and bandpass filtering, spikes of a single unit were isolated online using a custom amplitude- and time-based discriminator.

The position of one eye was continuously tracked using an infrared video-based tracking system (SR Research EyeLink). Stimuli were displayed on an LCD (Iiyama; 2560 x 1440 screen resolution; 1 ms GtG) at a distance of 57 cm from the monkey's eyes. The on- and offset of the stimuli was signaled by a photodiode detecting luminance changes of a small square in the corner of the display (but invisible to the animal), placed in the same frame as the stimulus events. A Digital Signal Processing-based computer system developed in-house controlled stimulus presentation, event timing, and juice delivery while sampling the photodiode signal, vertical and horizontal eye positions, spikes, and behavioral events. Time stamps of the recorded spikes, eye positions, stimulus, and behavioral events were stored for offline analyses.

In each monkey, the recording grid locations were defined so that the electrode targeted the selected body patches. Before the recordings, we performed a structural MRI in each monkey (3T Siemens Trio; Magnetization-Prepared Rapid Acquisition with

Gradient Echo (MPRAGE) sequence; 0.6 mm resolution) and visualized long glass capillaries filled with the MRI opaque copper sulfate ($CuSO_4$) that were inserted into the recording chamber grid (until the dura) at 5 positions. In addition, the recording chamber was filled with a Gadoteric Acid (Dotarem) solution to visualize the borders and orientation of the recording chamber. The functional maps of each monkey were co-registered to this structural MRI, using SPM12, and the co-registration was verified by visual examination. We could visualize guide tube tracks on structural MRIs taken after the recordings, showing that we targeted within the coronal and sagittal plane the selected body patches (Figure 1). The ventral-dorsal location of the recordings was verified in each recording session using the transitions of white and gray matter and the silence marking the sulcus between the banks of the STS.

Tests

The monkeys were performing a passive fixation task during stimulus presentation. They were rewarded with apple juice to fixate a small fixation target (size: 0.17 deg). Juice rewards were given with a fixed interval, that was titrated for each monkey, as long as the monkeys were maintaining fixation within a window of 2 by 2 deg. We examined the responses of each neuron in a "Search test" in which we presented the 20 body videos, 20 face videos, and 20 object videos in a pseudo-random order. Fixation was required in a period from 200 ms pre-stimulus to 200 ms post-stimulus onset, including the 1000 ms long stimulus presentation. A trial was aborted when the monkey interrupted fixation in this interval. In the pseudo-randomization procedure, all 60 videos were presented randomly interleaved in blocks of 60 unaborted trials. Aborted stimulus presentations were repeated within the same block in a subsequent randomly chosen trial. Neural responses of aborted trials were not analyzed. All neurons were tested with at least 3 unaborted trials per stimulus, with the large majority of neurons (90% and 93% of VP and DP neurons, respectively) tested with 5 unaborted presentations of each video. During the pre-stimulus period, a static white noise background pattern (size = 11 deg), randomly chosen from 10 patterns, was presented on top of a uniform gray background that filled the display. After the stimulus offset, only the gray background was present.

Based on the responses obtained in the Search test, we selected a body video that elicited the highest response ("best") and a body video to which the neuron did not respond ("worst"). These stimuli were employed in subsequent tests. In the Snapshot test, we presented the best body video, its time-reversed version, and 10 snapshots of this video. The duration of the snapshot presentation was 300 ms. The pre- and post-stimulus intervals were 500 ms, yielding an interstimulus interval of at least 1000 ms. During the pre-stimulus period, a static white noise background was presented. The 12 stimuli were shown randomly interleaved in blocks of 12 trials during fixation, using the same presentation schedule (except for the timings) as in the Search task. The test consisted of at least 10 unaborted presentations of each stimulus.

In another test, we presented the silhouette versions of the best and worst body video (selected in the preceding Search test for each neuron), together with other 16 silhouette videos that are not the subject of the present paper and will not be described here. The 18 videos were presented in random order using the pre- and post-stimulus time intervals as for the Snapshot test in blocks of 18 trials each. The test consisted of at least 10 unaborted presentations of each video. We tested also 30 neurons using silhouette versions of the best video, its time-reversed video, and the 10 corresponding snapshots with the Snapshot test.

Data analysis

Responsiveness and selectivity

We conducted for each neuron a split-plot ANOVA to select neurons that responded significantly to at least one of the body videos in the Search test. For each unaborted trial, the baseline mean firing rate was computed from −200 to 0 ms and the mean firing rate for the stimulus was computed from 60 to 1160 ms, with 0 representing stimulus onset. The baseline versus stimulus response was considered as a repeated-measure within-trial factor, and the 20 body videos as a between-trial factor. Cells with a significant main effect for the baseline versus stimulus activity factor ($p < 0.05$) or a significant interaction between the two factors ($p < 0.05$) were selected for further analysis. All neurons in the reported sample (n = 149 and 175 in VP and DP, respectively) had significant responses according to the ANOVA and an excitatory net response to at least one body stimulus.

We evaluated the significance of the responses of each neuron tested in the Snapshot test using a split-plot ANOVA. We computed the mean baseline and stimulus-induced firing rate for each unaborted trial of the 12 conditions. The baseline time window ranged from -200 to 0 ms, whereas for the stimulus-induced response, we employed a window of 60 to 1160 ms for the 2 videos and 60 to 460 ms for the 10 snapshot presentations. We employed the same ANOVA design and selection criteria as described above for the Search test. The responses in the Snapshot test were analyzed further only for the neurons that showed a significant and excitatory response in that test (n = 133 and 146 for VP and DP, respectively; for 8 and 22 neurons in VP and DP, respectively, the Snapshot test employed silhouettes).

Body-category selectivity index

For each responsive neuron, we compute the Body-category Selectivity Index (BSI), as follows:

$$BSI = \frac{R_b - R_{nb}}{|R_b|+|R_{nb}|}; R_{nb} = \frac{R_f+R_o}{2}$$

Where, R_b, R_f and R_o are the mean net firing rates to the body, face, and object videos, respectively, obtained in the Search test. The mean net firing rate was computed by subtracting the baseline firing rate from the firing rate in the response window (same windows as for the ANOVA; see above), averaged across trials per stimulus.

In some analyses, we selected only neurons with a BSI > 0.33, i.e. a twofold greater mean response to bodies compared to the mean response for objects and faces. In another analysis (Figure S4$D_{1\text{-}2}$), we equated the frequency distribution of DP and VP neurons, as follows. First, a histogram of the distribution of BSI (bin width = 0.07; 20 bins from minimum to maximum of the BSI) was created for both VP and DP. Then, the minimum cell count within each bin was determined by comparing the distributions of VP and DP. To equate the distribution, the surplus cells present in either VP or DP were then eliminated randomly, ensuring that the cell count matched the minimum count for that specific bin. This process resulted in populations of cells having BSI-equated distributions for VP and DP (Figure S4D_1).

Sparseness

The Sparseness of the response to the 20 body videos of each neuron was calculated as:

$$Sparseness \;=\; \frac{\left(1 \;-\; \frac{\langle r_i \rangle^2}{\langle r_i^2 \rangle}\right)}{1 \;-\; \frac{1}{20}}$$

Where $\langle . \rangle$ denotes the average and r_i is the net response of the i th body stimulus. The net firing rate was computed by subtracting the baseline firing rate from the firing rate in the response window (same windows as for the ANOVA; see above), averaged across trials per stimulus. Negative net responses were clipped to zero, as described before[52,53]. The Sparseness can range from 0 (equal response to the 20 body videos) to 1 (response to a single body video).

In some analyses, we split the neurons of each region into two groups: neurons with a Sparseness below (low- sparseness neurons) and above (high-sparseness neurons) the median of the Sparseness of the population of the neurons of both regions (VP + DP).

Selectivity for body videos and snapshots

To assess the (within-category) body-video selectivity of the responsive neurons in the Search test, we ranked for each neuron, tested with 5 unaborted trials per video, the body videos based on the net responses averaged across 4 trials per video (employing the same analysis windows as for the ANOVA). Then, the net responses of the left-out trial were stored as a function of the stimulus rank based on the four trials. This was done for each of the 5 possible groups of 4 trials, and the net responses for the left-out trials were averaged as a function of stimulus rank. We assessed the significance of the difference among the cross-validated ranked responses using a Friedman ANOVA for the DP and VP samples of neurons separately. The mean responses of the left-out trials were averaged across neurons and a 95% confidence interval of the averaged response was computed by bootstrapping neurons (1000 resamplings; percentile method). The same leave-one-trial-out cross-validation procedure was employed to assess the selectivity of the neurons for the 10 snapshots presented in the Snapshot test (using 10 times the responses averaged across 9 trials per snapshot for ranking). This was done only for neurons that responded significantly to a snapshot (Split-Plot ANOVA; 10 stimulus conditions; same windows and criteria as above).

Snapshot Selectivity Index

We compared the peak firing rate to the individual snapshots with the peak firing rate during the presentation of the body video that included these snapshots. Because the responses of the neurons could vary strongly during the video (e.g. Figure 3), averaging a response across the full video duration can underestimate the response to specific video segments. To avoid any such underestimation of the neural response to the video, we used peak firing rate instead of average firing rate as the response measure when comparing the responses of the video and snapshot presentations (same procedure as[15,16]). To compute the peak firing rates, we first convolved the spiking activity, averaged across trials for the same stimulus, with a Gaussian filter with a standard deviation of 25 ms. Then, we computed the net firing rate of the thus smoothed response to the stimulus by subtracting the smoothed baseline response. The Snapshot Selectivity Index (SSI) was computed as:

$$SSI \;=\; \frac{PR_{vs} \;-\; PR_{ss}}{|PR_{vs}|+|PR_{ss}|}$$

where PR_{vs} and PR_{fs} are the peak firing rate of the smoothed responses to the body video and the maximum peak firing rate across the ten snapshots, respectively. The response windows to find the peak firing rate for the video and snapshot stimuli were the same as those employed for the ANOVA-based significance testing.

Video Reversal Index

To quantify a cell's sensitivity to the difference between the original video and its time-reversed version, we computed:

$$VS \;=\; \frac{R_{ov} \;-\; R_{rv}}{|R_{ov}|+|R_{rv}|}$$

where R_{ov} and R_{rv} are the mean net responses to the original and time-reversed video, respectively, using the same analysis windows to compute the mean firing rate as for the ANOVA significance testing. Since the neurons differed in their preference for one of the two videos, we defined the Video Reversal Index as VRI = $|VS|$. A VRI value of 0 corresponds to no preference for the original over the time-reversed video, while a value of 1 indicates a response to only one of the two videos.

The difference between VP and DP was also significant for VS (VP: median = -0.03 (1st quartile = -0.21; 3rd quartile = 0.20); DP = 0.17 (-0.13; 0.60); Wilcoxon rank sum test p = 0.002).

Correlation between responses to the snapshots and the corresponding frames during the video

We computed the correlation between the response to each of the 10 snapshots and the corresponding frames when these were presented in the video. We computed the net average firing rate of each neuron to each snapshot and corresponding frame in the video in a 140 ms window. We determined when each neuron had its highest firing rate in response to the snapshots and used that to set the timing of the 140 ms window. To obtain this estimate of the response latency for each neuron, we averaged the net firing rate in bins of 20 ms across all 10 snapshots and identified the bin with the highest firing rate. We then used the 140 ms window around that bin to capture the neuron's response. If the window started earlier than 60 ms after the snapshot onset, the beginning of the 140 ms long window was set to 60 ms to avoid taking spikes that could not have been evoked by the snapshot or frame. Using this response window, defined per neuron, we extracted the net firing rates for the individual snapshots and the corresponding frames in the video. The Pearson correlation coefficient between the vectors (of size 10) of the responses for the snapshots and the corresponding video frames was then computed for each neuron. In some cases, a portion of the response window for a video frame fell outside of the 1160 ms long interval of responses available for the video. The responses for the snapshots/frames of those cases were removed from the vector before computing the correlation. The correlation between the responses to the snapshots and the corresponding frames during the video was computed only for those neurons that showed a statistically significant response to at least one of the 10 snapshots, which was assessed with a split-plot ANOVA.

CNN modeling: Networks trained with static images

We employed four instances of three distinct CNN architectures, namely AlexNet, VGG16, and ResNet50, along with an additional ResNet50 architecture trained on stylized ImageNet (SIN) called ResNet50_SIN. As a control, we also included an untrained version of these networks. The weights for AlexNet (AlexNet_Weights.IMAGENET1K_V1), VGG-16 (VGG16_Weights.IMAGENET1K_V1), ResNet50 (ResNet50_Weights.IMAGENET1K_V1), and the untrained version were imported from TorchVision in PyTorch. The random weights in the untrained networks of PyTorch were drawn from a Gaussian distribution, except for the weights from AlexNet, which were drawn from a uniform distribution.

To examine the responses of CNN units, we presented frames from the body videos (20 X 60 frames) with a grey background (RGB value = 128) to the networks. We did not include the white noise background, since we wanted to compute the activations to the body images per se. Additionally, we obtained the response to a version of the stimuli in which the monkey's body was reduced to a silhouette. The frames were pre-processed by rescaling and subtracting the mean and division by the standard deviation of the ImageNet data.

For AlexNet models, we present the data for all seven ReLU layers, whereas for VGG16, we present the data for ReLU layers 1.2, 2.2, 3.3, 4.3, 5.3, 6, and 7, with layers 6 and 7 being the fully connected layers. For the ResNet50 architectures, we obtained responses from the ReLU layers (relu, layer1.2.relu_2, layer2.3.relu_2, layer3.5.relu_2, layer4.2.relu_2) available at the end of each of the five "stages", which we label in the Results as layers 1 to 5. We removed the units of a layer that did not respond to any of the 1200 frames. Plots of the number of responding units (features; in log units) in each layer of the networks are presented in Figure S6.

CNN modeling: Networks trained with human action videos

We also included a pre-trained spatiotemporal network, X3D[38] (X3D-M; available at https://github.com/facebookresearch/SlowFast/blob/main/MODEL_ZOO.md) The X3D model was pretrained on the Kinetics-400 dataset[54], which encompasses videos featuring 400 distinct human action classes, designed for action classification tasks. This spatiotemporal 3D network (1 temporal and 2 spatial dimensions) expands in the temporal dimension (frame sequences), while also encompasses optimization of spatial, 2D hyper-parameters. Notably, X3D evolves from the foundational 2D structure of ResNet, roughly retaining its stages. The layers 1 through 5 (Figures S5 and S6) correspond to the 'stages' 1 through 5, mirroring the organization of the ResNet architecture. The activations of responding units were extracted as described above for the other CNNs. Plots of the number of responding units (features; in log units) in each layer of the network are shown in Figure S6.

Velocity estimation

We estimated the pixel-wise velocities of each body video (60 frames of the size 210 x 210 pixels) using the Lucas Kanade derivative of Gaussian filter optic flow algorithm implemented by the opticalFlowLKDOG Matlab function with the same parameter settings as in[6]. Because we aimed to compute the velocities of the bodies, we removed the white noise background and replaced it with a gray background. We obtained a pixel-wise map of the x and y components of the velocity vector for 58 frames of each video, resulting in a tensor (58 X 210 X 210 X 2) per video. For the first two frames, the algorithm does not produce a valid optic-flow measure, explaining why we had measures for 58 frames.

Pairwise between-video distances

Distance measure: Lock-step Euclidean distance. To obtain pairwise between-video trajectory distances in an N-dimensional neural, velocity, and CNN feature space, we calculated the lock-step Euclidean distance[55] between every pair of body videos (V_i and V_j) as:

$$L(V_i, V_j) = \sqrt{\sum_{m=1}^{M} d_2^2\left(v_i^m, v_j^m\right)}; i \neq j, v \in \mathbb{R}^{\mathrm{N}}$$

Where, $d_2(v_i^m, v_j^m) = \sqrt{\sum_{n=1}^{N} (v_i^m(n) - v_j^m(n))^2}$ is the Euclidian distance between body video V_i and V_j in an N-dimensional space at the point $m \in \{1, 2, ..M\}$ of a temporal trajectory. N corresponds to the number of cells, pixels (multiplied by 2), and units in a layer, for the neural, optic-flow, and CNN feature distances (see Table S4), respectively. M corresponds to the number of bins and frames in the case of the neural and velocity/static feature space, respectively (see Table S4). In the case of the neural trajectories, we considered the binned responses in an interval ranging from 60 to 1160 ms post-stimulus onset.

In simpler terms, we created a matrix of $M \times N$ responses/values for each video, which we then flattened into a vector. We computed the Euclidean distance between the vectors for each video pair, resulting in 190 distance values (20* 19 /2) for the 20 body videos. To calculate the neural distances, the net responses of each neuron to the videos were normalized by its maximum peak net response (bin width = 20 ms) across the body videos.

Distance measure: Chi-square distance between velocity distributions

For each of the 58 frames of which we had velocity vectors, we created a 2-dimensional frequency distribution of speed and direction, with a bin width of 0.5 (arbitrary units) for the speed axis and $\pi/8$ for the direction axis. The speed axis ranged from a low-speed threshold of 0.2 for the reported data in the Results (see Figure S3B for results with other thresholds) to 7.7, which is the maximum speed observed among all frames in all videos, while the direction axis ranged from $-\pi$ to π. Once we created the two-dimensional frequency distribution for each frame, we flattened it into a vector and concatenated all the vectors from all frames (in order) of a video to create a grand vector of size K = 13050 (15 speed bins x 15 direction bins x 58 frames). This grand vector represents a grand frequency distribution for the entire video while capturing the temporal pattern. We then computed the chi-square distances for all 190 pairs of videos V_i and V_j as follows:

$$X^2(V_i, V_j) = \frac{1}{2} \sum_{k=1}^{K} \frac{\left(V_i^k - V_j^k\right)^2}{\left(V_i^k + V_j^k\right)}$$

Best-worst preference index

For each neuron that was tested with the silhouette versions of the videos, we computed a best-worst preference index (BWPI):

$$BWPI = \frac{R_{best} - R_{worst}}{|R_{best}| + |R_{worst}|}$$

where R_{best} and R_{worst} are the net average firing rate for the silhouette version of the original stimuli that elicited the best and worst response in the Search test, respectively. The same analysis windows as for the ANOVA were used.

Statistical analysis

Significance tests for the correlation analyses of pairwise neural and model distances

We permuted stimulus labels[31] to determine the significance of the correlation coefficient between pairwise neural and velocity / CNN feature distances. To do so, we created a distance matrix of the neural pairwise distances. Next, we permuted the labels of the matrix by randomly reordering the rows and columns, i.e., the stimulus labels. We then computed the correlation between the permuted neural distance matrix and the pairwise distances corresponding to the velocity-based or CNN features, using the upper off-diagonal values of the matrix. We repeated this process of permutation and correlation computation 1000 times to generate a null distribution of correlation coefficients. From this null distribution, we obtained the percentile (Pc) of the observed correlation coefficient. We used the Pc to compute the two-tailed p-value as:

$$p = \begin{cases} 2 \times \dfrac{100 - Pc}{100}; Pc \geq 50 \\ 2 \times \dfrac{Pc}{100}; Pc < 50 \end{cases}$$

If the p-value was less than 0.05, we rejected the null hypothesis.

To assess the significance of the difference between the correlations of two distance matrices obtained for DP and VP, we utilized bootstrapping by resampling with replacement the cells of VP and DP. We computed the correlation coefficient between neural distances for the resampled cells and velocity-based or static CNN feature distances and subtracted the correlation coefficient obtained for the resampled VP from that of the resampled DP. We repeated this process 1000 times to obtain a distribution of differences in correlation values between VP and DP. We then computed the percentile of a value of 0 from this distribution and computed the p-value (two-tailed) based on the percentile with the same method as described above.

Commonality analysis

We utilized a multiple regression-based commonality analysis[56] to determine whether motion and static features account for a shared portion of the response variance in VP (and DP) neurons, or instead, whether they contribute uniquely to the neural response variance. We obtained the explained variance R^2_{M+F} through multiple regression of neural distances from the velocity and static feature distances as predictors. To isolate the unique contribution of motion and static features, we subtracted the explained variance R_F^2, corresponding to static features, obtained by regression using the feature distances as a single predictor, and the explained variance R_M^2 corresponding to motion, respectively from R_{M+F}^2 . The common explained variance R_C^2 was then calculated as: $R_C^2 \ = \ (R_M^2 + R_F^2) - R_{M+F}^2$. To assess the significance of R_{M+F}^2 , we computed the p-value (one-tailed) as: $p \ = \ \frac{100 - Pc}{100}$, where Pc is the percentile of the observed R_{M+F}^2 relative to the null distribution obtained using stimulus label permutation of the neural distances as described above. Significant R_{M+F}^2 values are indicated by stars in Figure 7.

Cell Reports, Volume *42*

Supplemental information

Bodies in motion: Unraveling the distinct roles of motion and shape in dynamic body responses in the temporal cortex

Rajani Raman, Anna Bognár, Ghazaleh Ghamkhari Nejad, Nick Taubert, Martin Giese, and Rufin Vogels

Supplementary Information

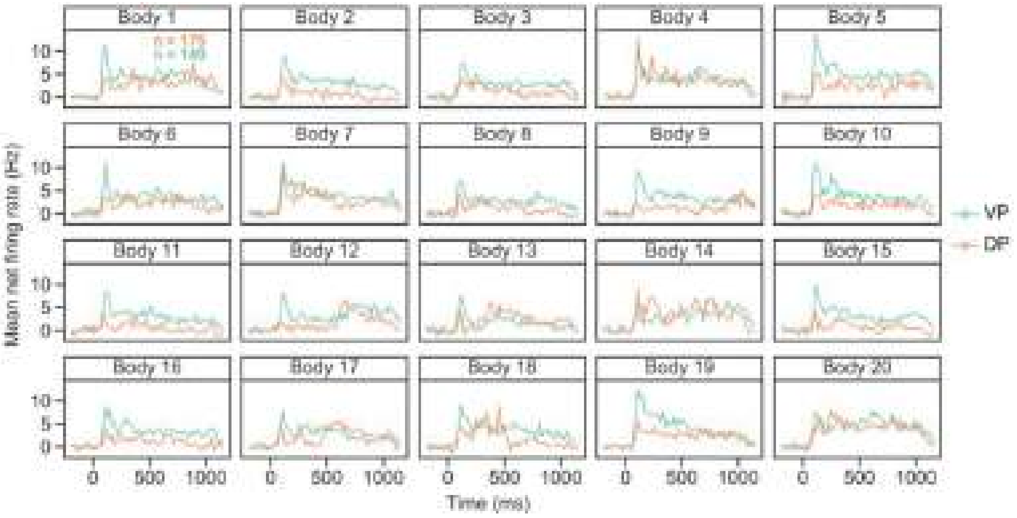


Figure S1. Mean population PSTHs in VP (green) and DP (red) to 20 body videos. Related to Figure 2.

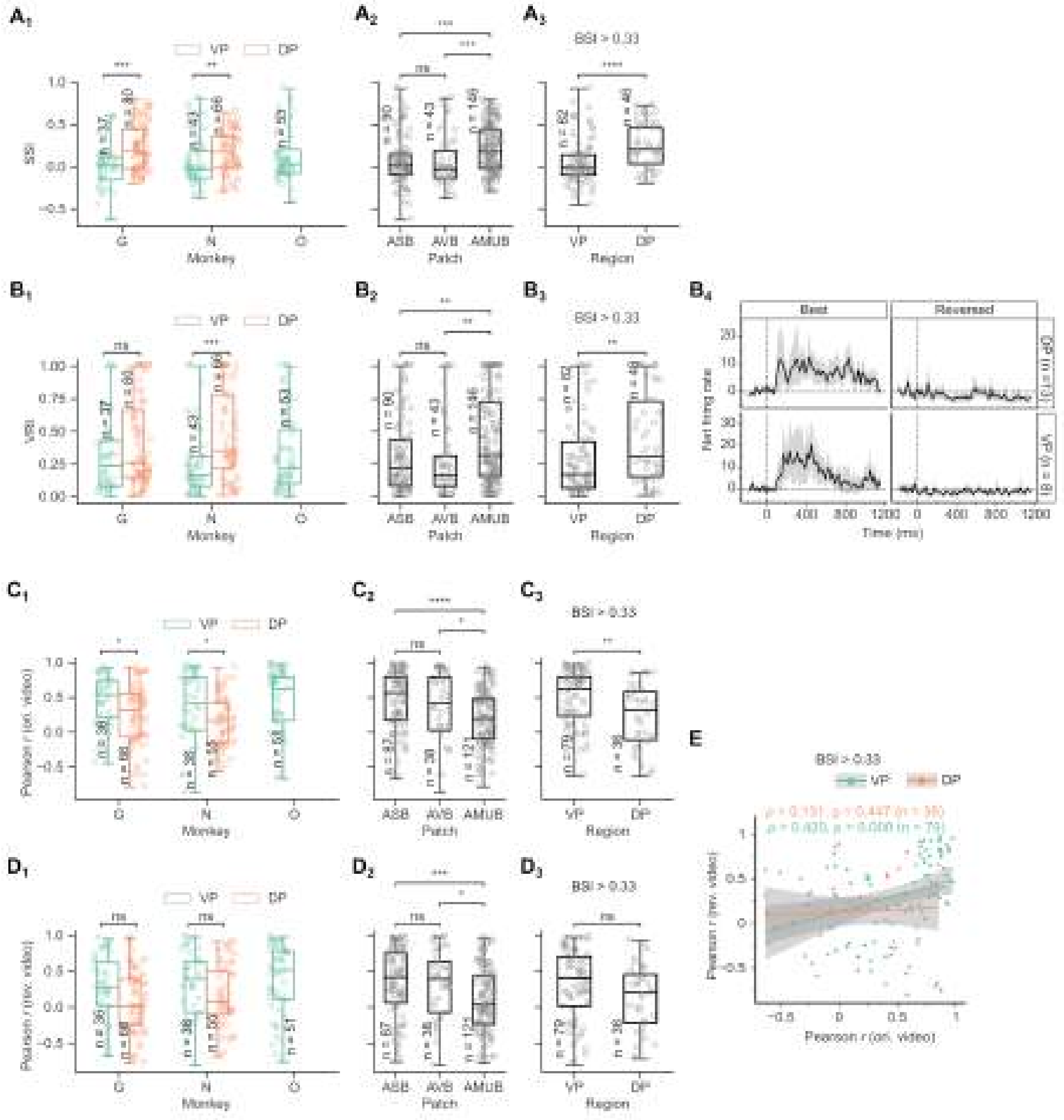


Figure S2. Additional analyses of motion/sequence sensitivity. Related to Figure 4.

*Distribution (box plots) of SSI (A1) for each monkey and region, (A2) for each targeted patch, and (A3) for the cells with BSI > 0.33. Distribution (box plots) of VRI (B1) for each monkey, (B2) for each targeted patch, and (B3) for the cells with BSI > 0.33 in each region. *** p < 0.001; ** p < 0.01; * p < 0.05; p values of Wilcoxon rank sum test. N corresponds to the number of cells. BSI > 0.33 corresponds to a twofold greater average response to bodies compared to faces and objects. (B4) Average response (and bootstrapped 95% confidence intervals) of neurons that showed a significant difference between the original and time-reversed video, and have VRI = 1. For each neuron, we labeled the video version that produced the highest response of the two as "best" and the other as "reversed". The first and second columns correspond to the mean net response to the "best" and "reversed" body video, respectively. The top and bottom rows are DP and VP neurons, respectively. Note the inhibition for the time-reversed video. The median BSI was 0.30 (1st quartile: 0.25; 3rd quartile: 0.45) and 0.30 (0.02-0.46) for these DP* *and VP neurons. (C1) Distribution of the correlation coefficient between the static snapshot response and the response to the same snapshot presented during the original video for each monkey and region, (C2) for each patch, and (C3) for the cells with BSI > 0.33 in both regions. (D1) correlations for the time-reversed video for each monkey and region, (D2) for each patch, and (D3) for the cells with BSI > 0.33 in both regions. P-values of Wilcoxon rank sum test: *** p < 0.001; ** p < 0.01; * p < 0.05. (E) Scatter plot (with linear regression lines and corresponding 95% confidence intervals) of the correlation coefficients between the video and snapshot responses for the original (ori. video) and time-reversed videos (rev. video; as in Figure 4F, main text) for neurons with BSI > 0.33.*

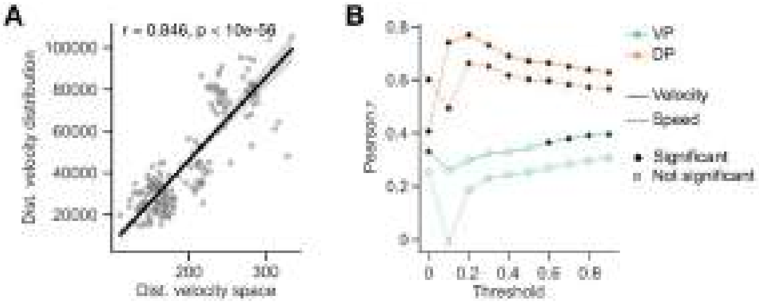


Figure S3. Velocity-based distance metrics. Related to Figure 5.

(A) Scatter plot (with regression line and 95% confidence interval) of velocity space and velocity distribution distances. Note that both velocity-based distances did not correlate with the difference in body area (number of pixels) between the corresponding video frames of a video pair (velocity space: r = -0.11 (p = 0.11); velocity distribution r = -0.07 (p = 0.30)). (B) Correlation between neural and velocity distribution distances (speed and direction; solid line) and between neural distances and speed-only distribution distances (dashed line), respectively, as a function of the threshold value for the minimum speed.

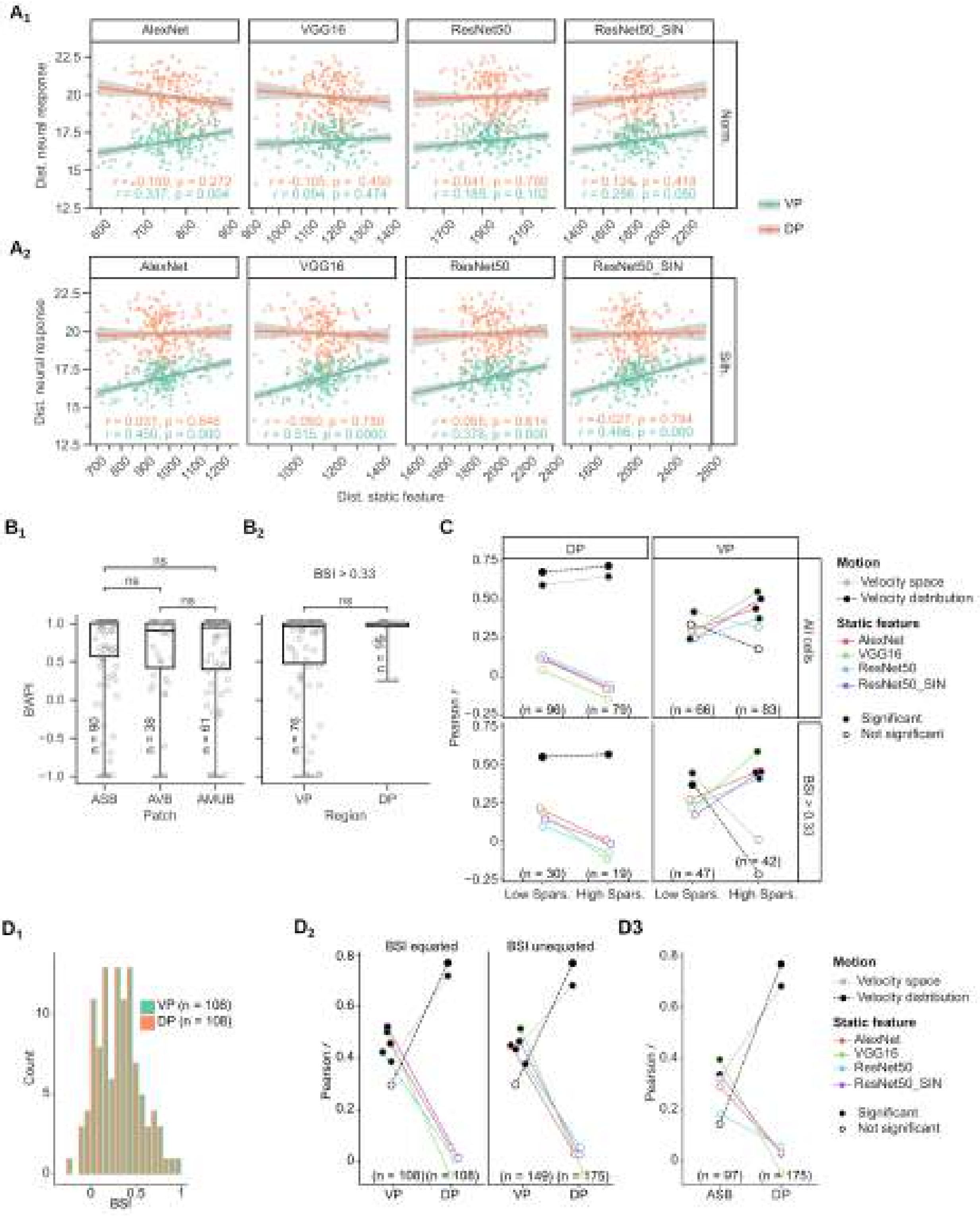


Figure S4. Relating static features, motion features and neural responses to dynamic videos. Related to Figure 6.

(A1) Scatter plots (with regression lines and 95% confidence interval) of neural and static feature distances for layer 5 of each CNN to original (shaded, and textured; "Norm.") videos, and (A2) to Silhouette ("Silh.") videos. (B1) Distribution (box plots) of Best-Worst Preference Index (BWPI) for responses to silhouette videos for each patch, and (B2) for the cells with BSI > 0.33. Statistical significance between patches or regions was tested with a Wilcoxon rank

sum test. (C) Correlation between velocity-based distances and neural distances (dashed lines), and between CNN layer 5 static feature and neural distances (computed for silhouettes) for the cells with low Sparseness (below median Sparseness of all neurons) and high Sparseness (above median (0.62)). The rows correspond to the regions (DP and VP) and the columns correspond to all cells (top) and the ones with BSI > 0.33 (bottom). VP pools ASB (median Sparseness: 0.67 (1st quartile: 0.45; 3rd quartile: 0.83)) and AVB (0.61 (0.45; 0.74)) neurons. Similar results were obtained for neurons with BSI > 0.2. Conventions as in Figure 6C, main text. (D1) Equated distribution of BSI for both regions (Methods). (D2) Correlations between the neural and velocity-based distances (dashed lines), and between neural and CNN layer 5 static features distances (computed for silhouettes; colored lines) for the BSI equated sets of neurons (left). The right plot shows the correlations for the original, non-equated distributions (same plot as Figure 6C, main text) for comparison. (D3) Correlations between the neural and velocity-based distances (dashed lines), and between neural and CNN layer 5 static features distances (computed for silhouettes; colored lines) for the targeted ASB patch neurons and DP. The same conventions as in Figure 6C, main text.

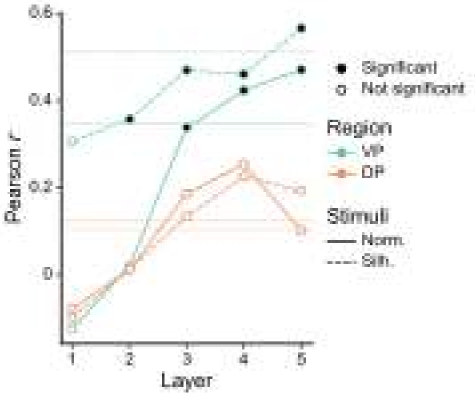


Figure S5. Correlation between neural distances and distances computed from X3D features, a spatiotemporal network trained with human body videos for action recognition. Related to Figure 6.

Horizontal lines correspond to the maximum correlation value obtained across all layers of all static CNN networks, i.e. the highest correlation obtained with the 4 CNNs trained with ImageNet data. All VP and DP neurons were included.

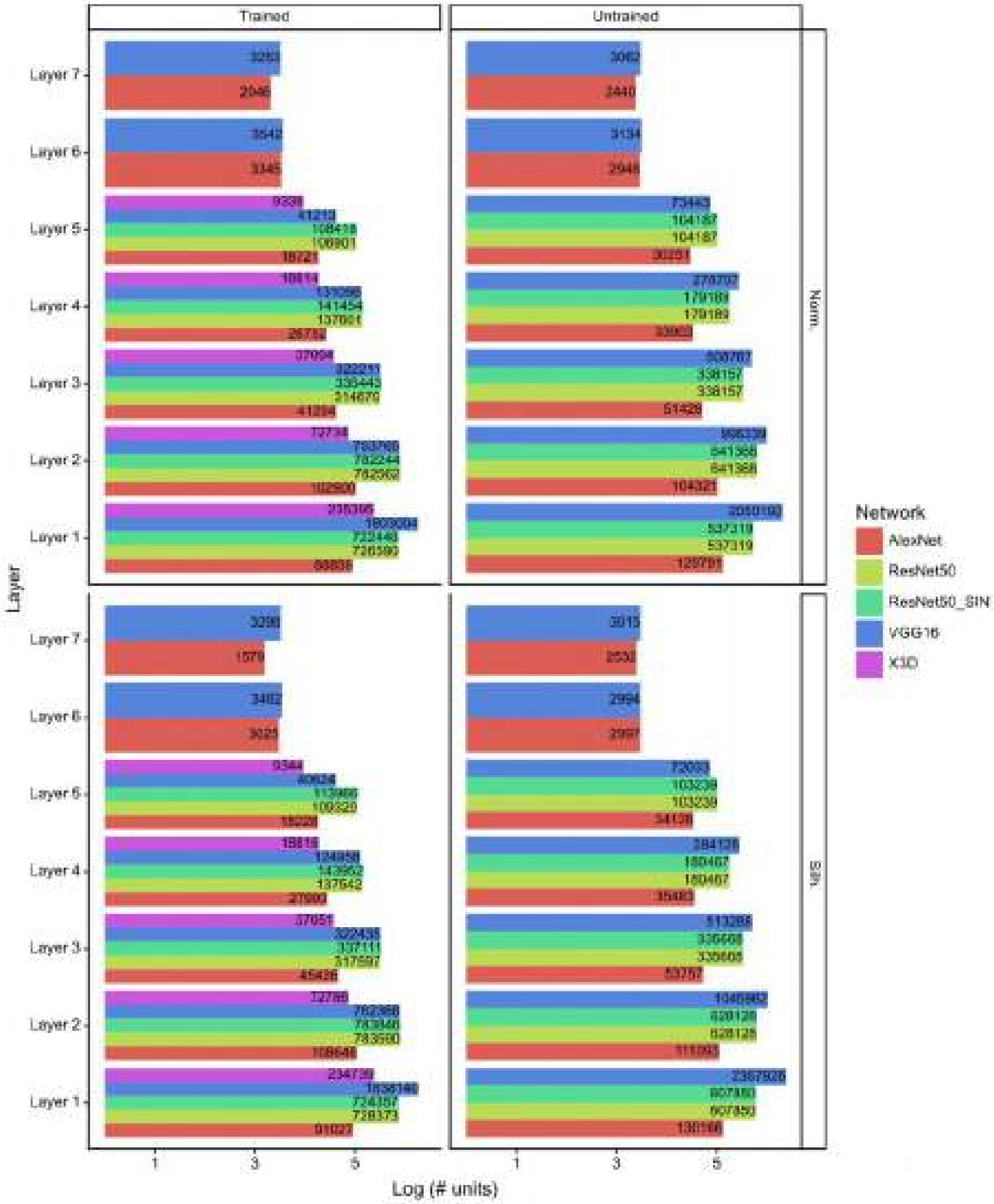


Figure S6. Number of responding units in each layer of the CNNs. Related to Figure 6.

Plots of the number of responding units (features; in log units) in each layer of the networks. Numbers in the bars correspond to the actual number of units. Columns correspond to the trained and untrained networks, while rows correspond to the original (Orig.) and silhouette (Silh.) stimuli that served as input to the networks.

Table S1. ANCOVA: Dependent variable = SSI, Between factor = Region (VP vs. DP), Covariate = BSI

Source	*SS*	*DF*	*F*	*p*
Region	1.226488	1	15.868826	0.000087
BSI	0.018430	1	0.238452	0.625714
Residual	21.331795	276		

Table S2. ANCOVA: Dependent variable: VRI, Between factor= Region (VP vs. DP), Covariate = BSI

Source	*SS*	*DF*	*F*	*p*
Region	0.976634	1	9.822762	0.001909
BSI	0.128162	1	1.289026	0.257212
Residual	27.441462	276		

Table S3. ANCOVA: Dependent variable = Pearson *r* (ori. video), Between factor = Region (VP vs. DP), Covariate = BSI

Source	*SS*	*DF*	*F*	*p*
Region	2.166592	1	12.470263	0.000495
BSI	0.759137	1	4.369367	0.037630
Residual	42.218986	243		

Table S4: Dimensionality of neural, velocity and CNN feature space.

	N	M
Neural space (Fig. 5A)	# of cells: 149 (VP), 175 (DP)	# of time bins: 55 (20 ms bins)
Velocity space (Fig. 5B)	2 * # of pixels: 2 * 210 *210	# of frames: 58
CNN feature space	# of responsive units in a layer	# of frames: 60