Neuron, Volume *109*

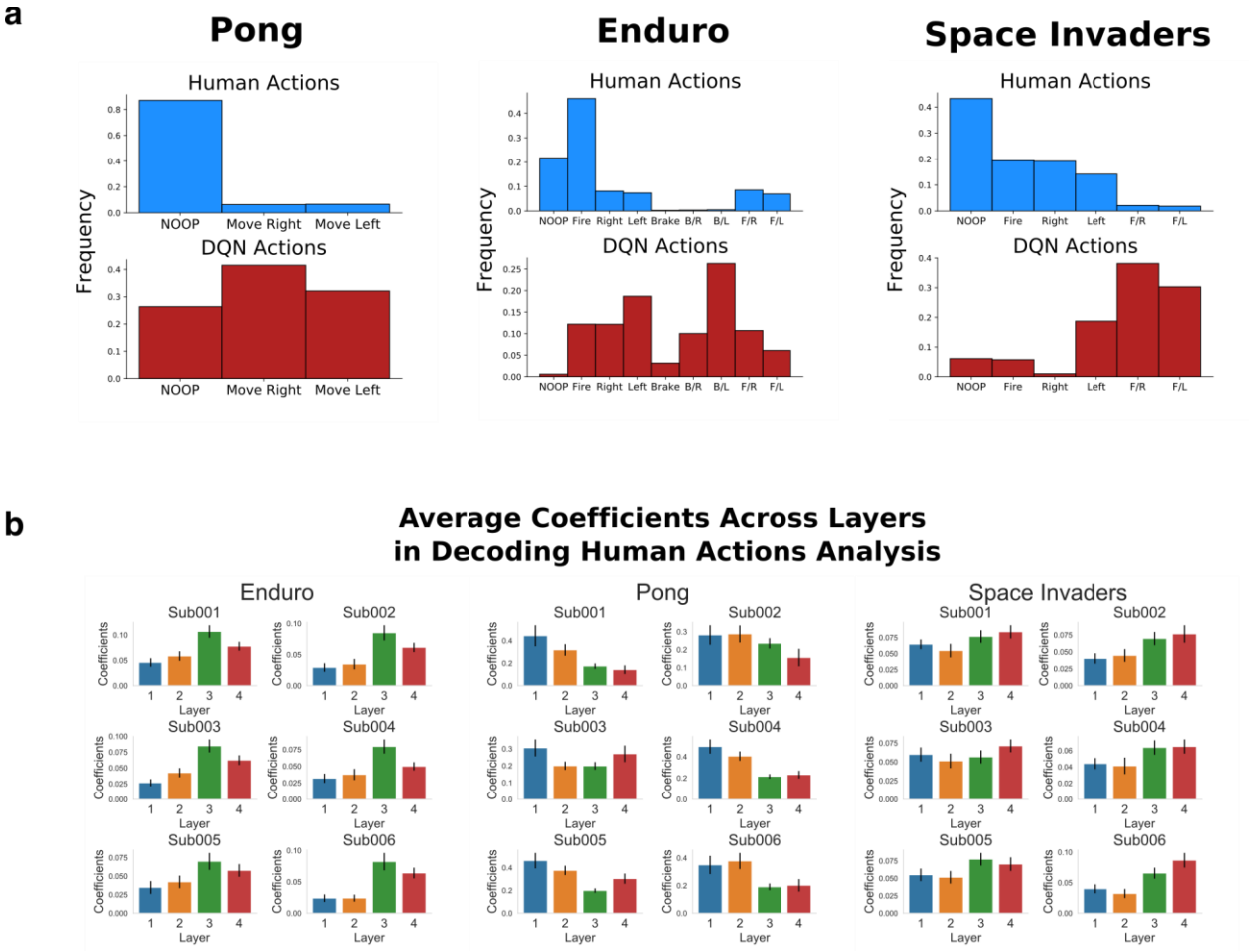
Supplemental Information

Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments

Logan Cross, Jeff Cockburn, Yisong Yue, and John P. O'Doherty

	DQN	DQN (Mnih 2015)	Sub 001	Sub 002	Sub 003	Sub 004	Sub 005	Sub 006	Human expert (Mnih 2015)	Random Agent
Pong	**17.2 (+/- 2.7)**	**18.9 (+/- 1.3)**	***-0.63 (+/-3.4)**	***0.72 (+/-2.6)**	***-1.90 (+/-4.3)**	***1.54 (+/-3.2)**	***-3.09 (+/-2.6)**	***-0.09 (+/-2.2)**	**9.3**	**-20.6 (+/- 0.91)**
Enduro	**348.0 (+/-108)**	**301.8 (+/-25)**	***230.0 (+/-37)**	***201.73 (+/-5.1)**	***202.18 (+/-3.7)**	***198.27 (+/-7.4)**	***201.55 (+/-1.8)**	***199.45 (+/-4.1)**	**309.6**	**0**
Space Invaders	**909.50 (+/-335)**	**1976 (+/-893)**	**949.23 (+/-514)**	**890.11 (+/-601)**	**1398.80 (+/-827)**	**1241.66 (+/-632)**	**777.85 (+/-280)**	**663.96 (+/-336)**	**1652**	**157.50 (+/- 135)**

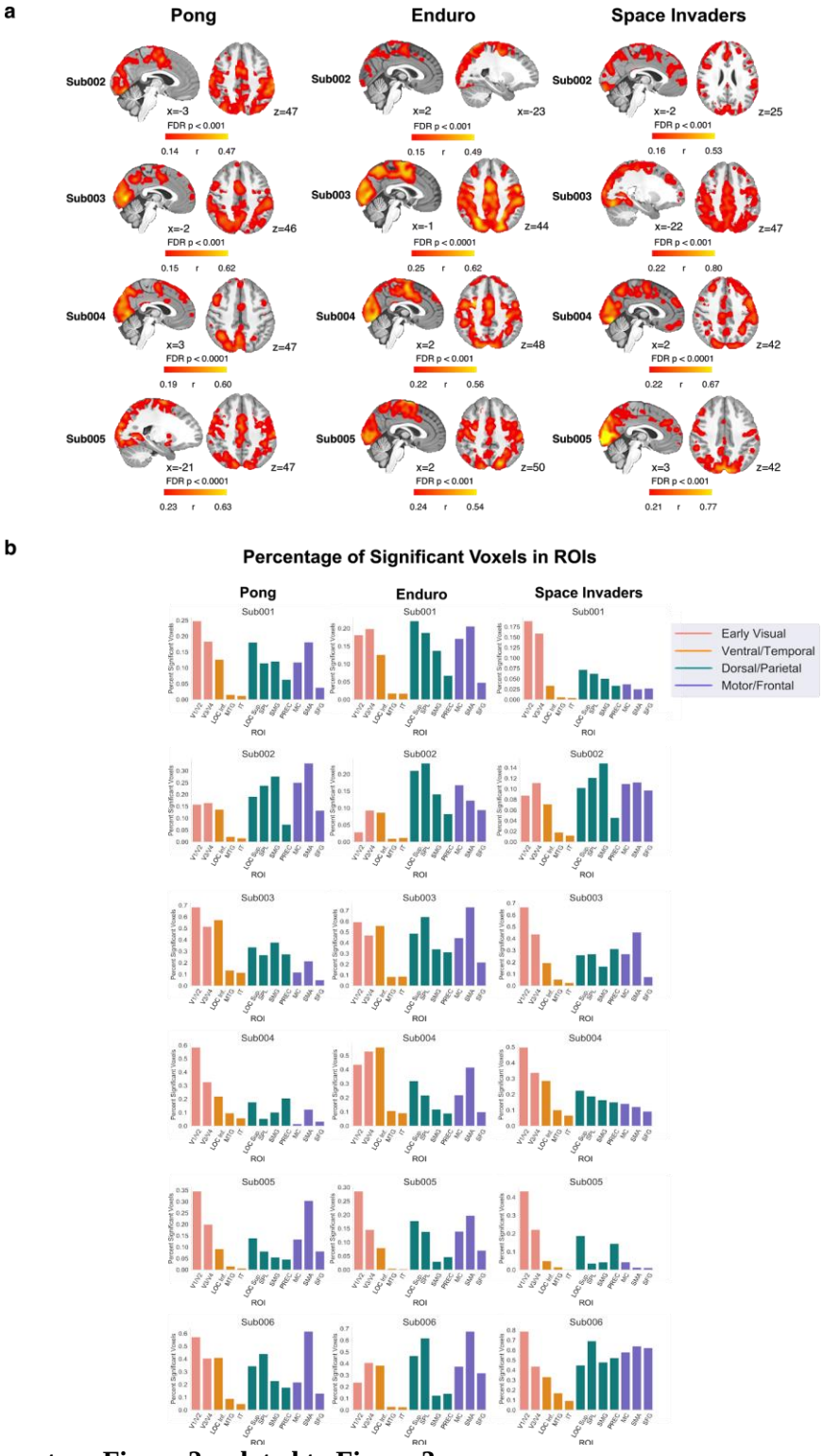
Supplementary Table 1, related to Figure 1. Game performance for all participants and the DQN agent we trained. DQN and human expert from Mnih 2015 included for reference along with a random agent. +/- refers to standard deviation of scores. For evaluation, DQN tested with ε-greedy policy (ε = 0.05) to minimize risk of evaluating an overfit policy. *Score limited by game cutting off prematurely due to time constraint of 8 minute runs.



Supplementary Figure 1, related to Figure 2.

a. **Distribution of actions for human participants and DQN.** For Enduro and Space Invaders, F/R and F/L correspond to the pairwise combination of fire and move right or move left respectively. For Enduro, B/R and B/L similarly correspond to the pairwise combination of brake and move right or move left.

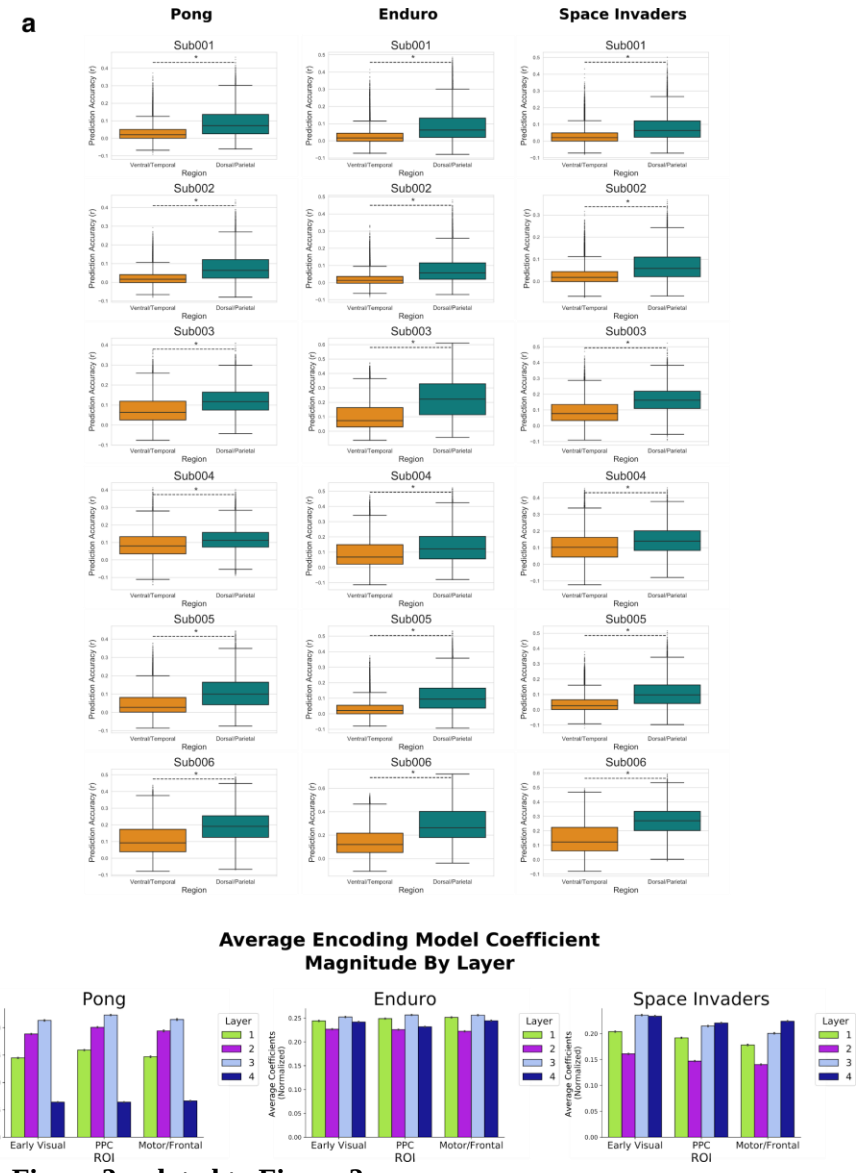
b. **Absolute value of the coefficients in the decoding human actions logistic regression model averaged across layers.** For Enduro and Space Invaders, layers 3 and 4 were the most useful for predicting human actions in every participant. For Pong, the contributions of each layer were more heterogeneous across participants with layers 1 and 2 having larger coefficients. Error bars depict SE across neurons in a layer.



Supplementary Figure 2, related to Figure 3.

a. **Encoding model results for other participants**. As in Figure 3, whole brain maps are thresholded as noted.

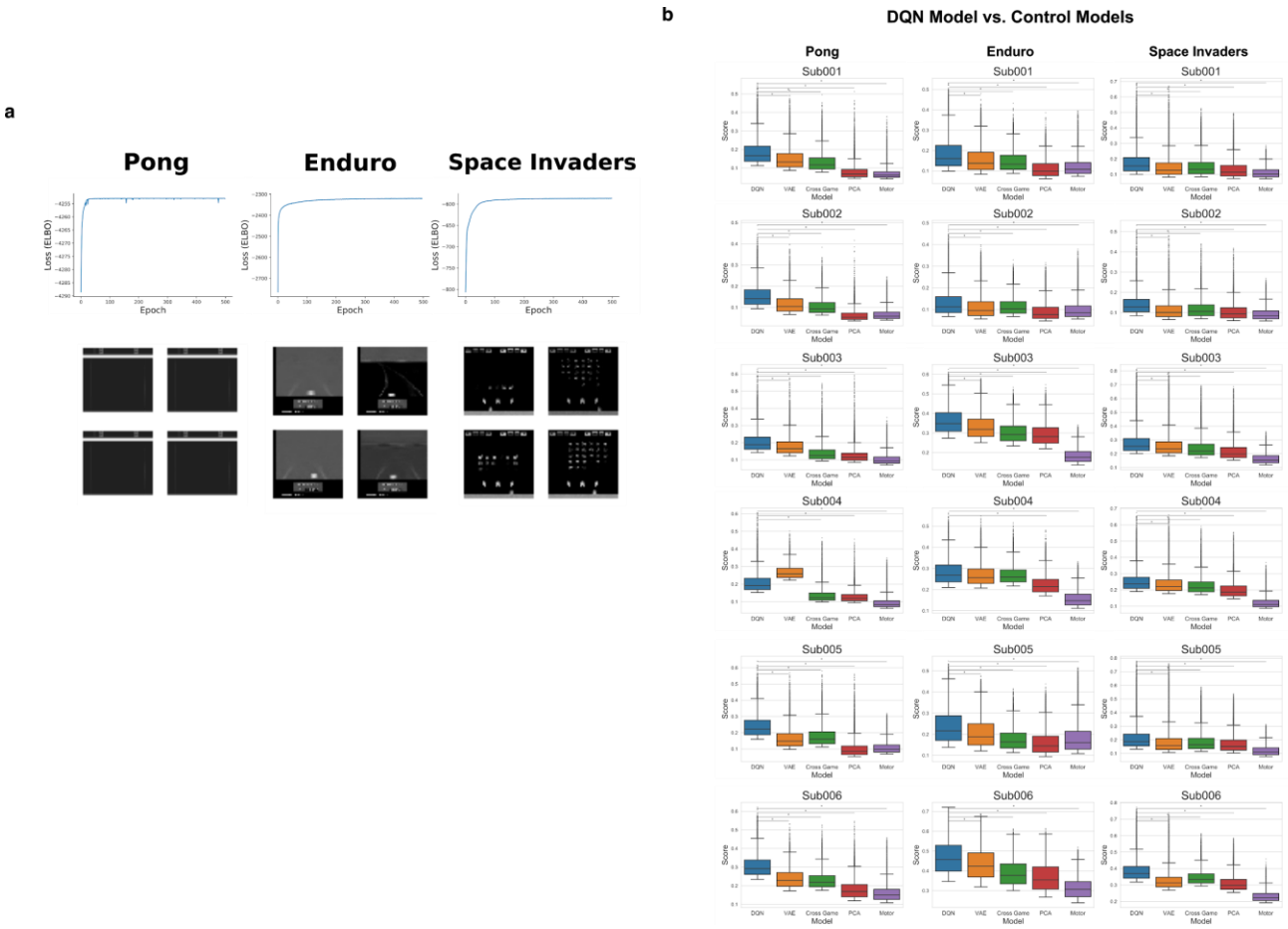
b. **Encoding model results by ROI for individual participants.** As in Figure 3e.



Supplementary Figure 3, related to Figure 3.

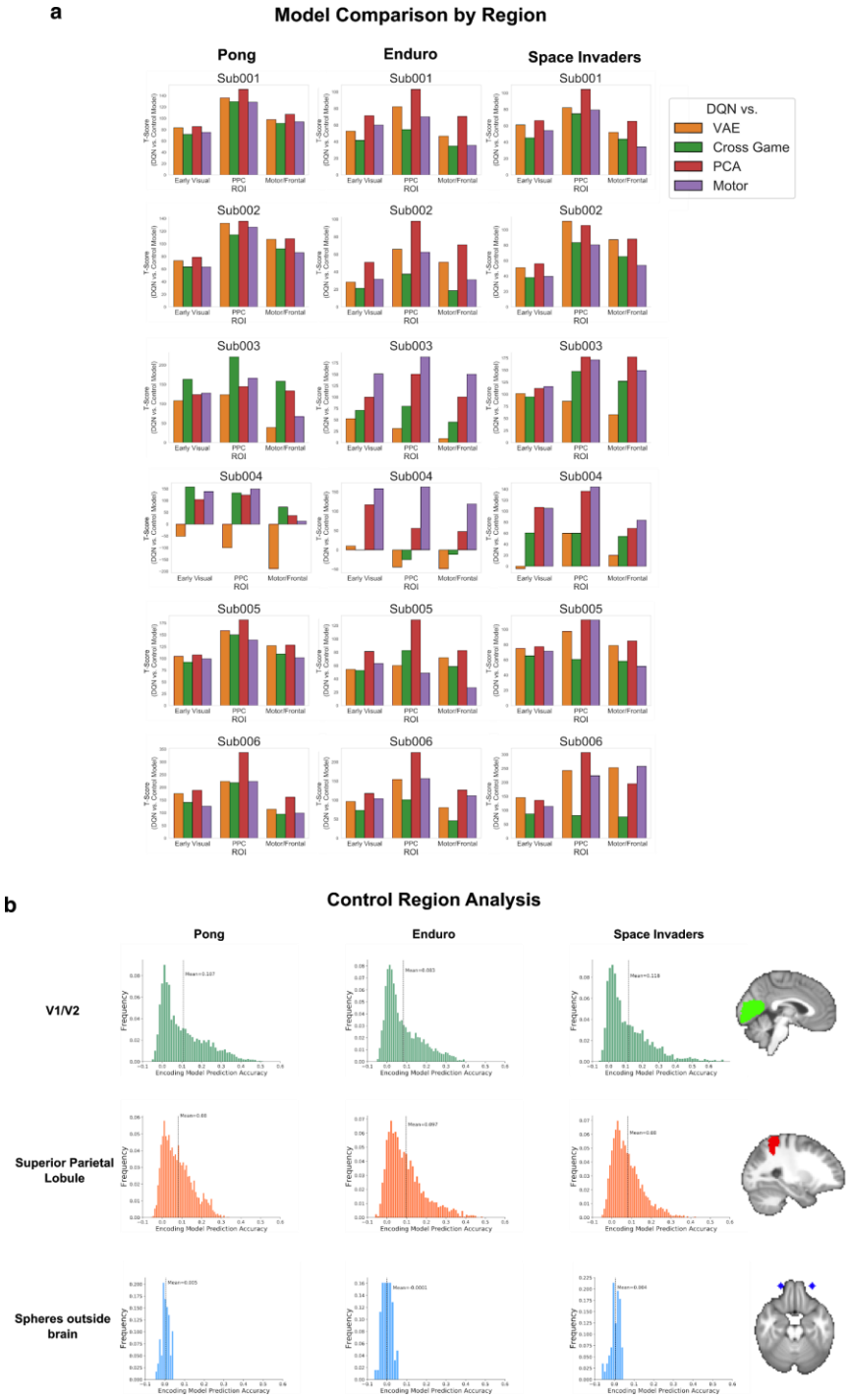
a. **Encoding model prediction accuracies are higher in dorsal visual stream than in ventral visual stream.** Prediction accuracies (Pearson r) for regions of interest in the dorsal visual stream and/or parietal lobe (superior lateral occipital cortex, superior parietal lobule, supramarginal gyrus, and precuneus) and in the ventral visual stream and/or temporal lobe (inferior lateral occipital cortex, middle temporal gyrus, inferior temporal lobe). Prediction accuracies are significantly higher in dorsal stream/parietal lobe ROIs for all subjects and all games (two-sample T-test, $P < 1e\text{-}10$, signified by *).

b. **Average coefficient magnitude by layer.** Absolute value of the coefficients by layer in the encoding model analysis averaged across participants. Each layer has 100 coefficients corresponding to 100 principal components of that layer. Error bars reflect SEM across voxels in all participants. Early visual ROI includes V1-V4; PPC includes LOC superior, superior parietal lobule, supramarginal gyrus, precuneus; Motor/Frontal includes motor and premotor cortex, SMA, and superior frontal gyrus.



Supplementary Figure 4, related to Figure 4.

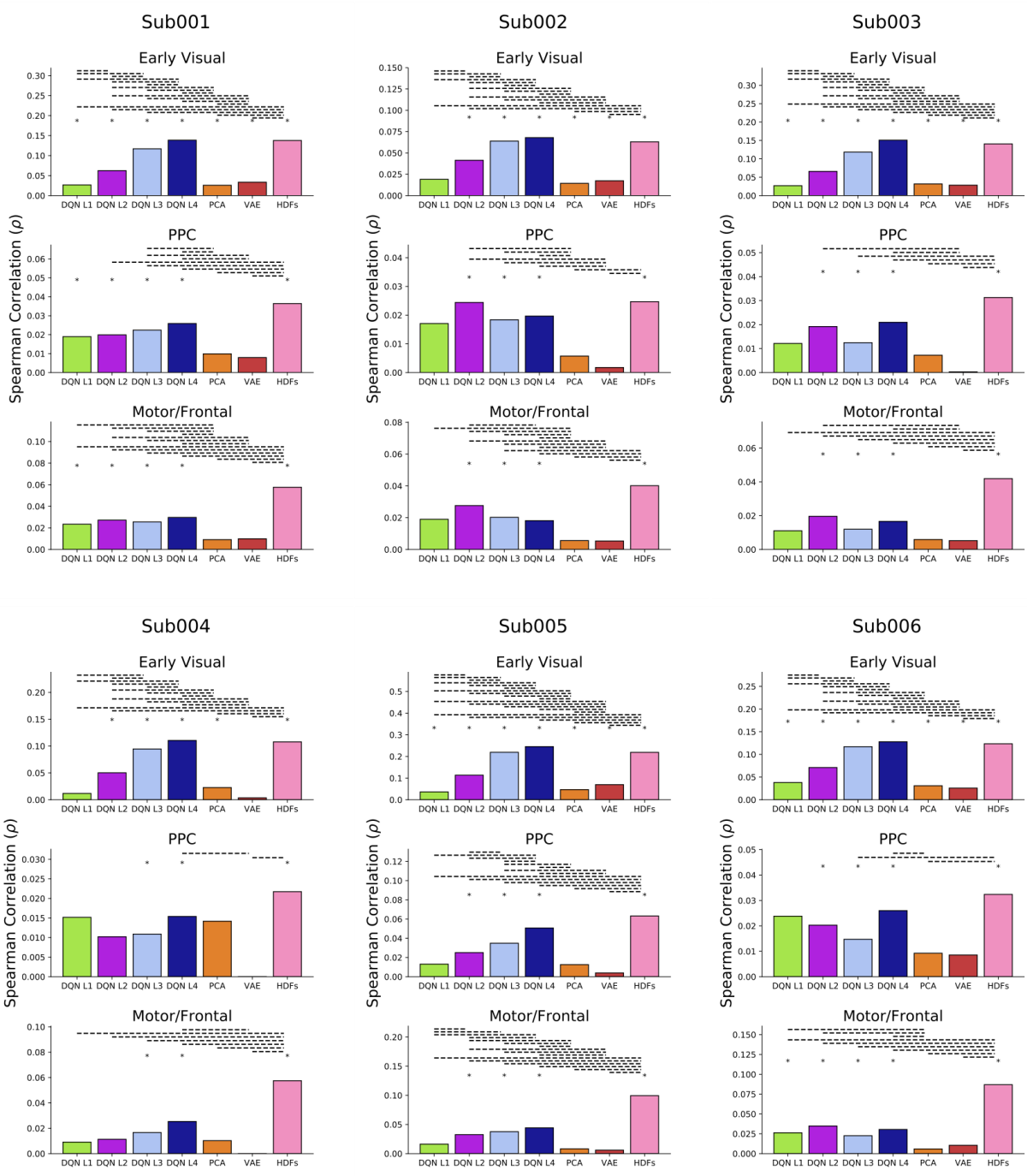
- **a. VAE Control Model.** Top row shows the training performance for the first 500 training epochs. Models were trained to maximize the evidence lower bound (ELBO) on the log-likelihood of the data. Bottom row: examples of VAE generated outputs. Images are generated by sampling latent vectors from a Gaussian distribution and inputting these samples into the VAE decoder.
- **b. Control models prediction accuracy distributions.** Boxplots show distributions of the prediction accuracies in the upper 20th percentile for each model. Outlier points represent the voxels with the highest prediction accuracies in the model.The DQN outperforms all other models in every game and participant other than sub004 (shown $P < 1e\text{-}10$ with * symbols).



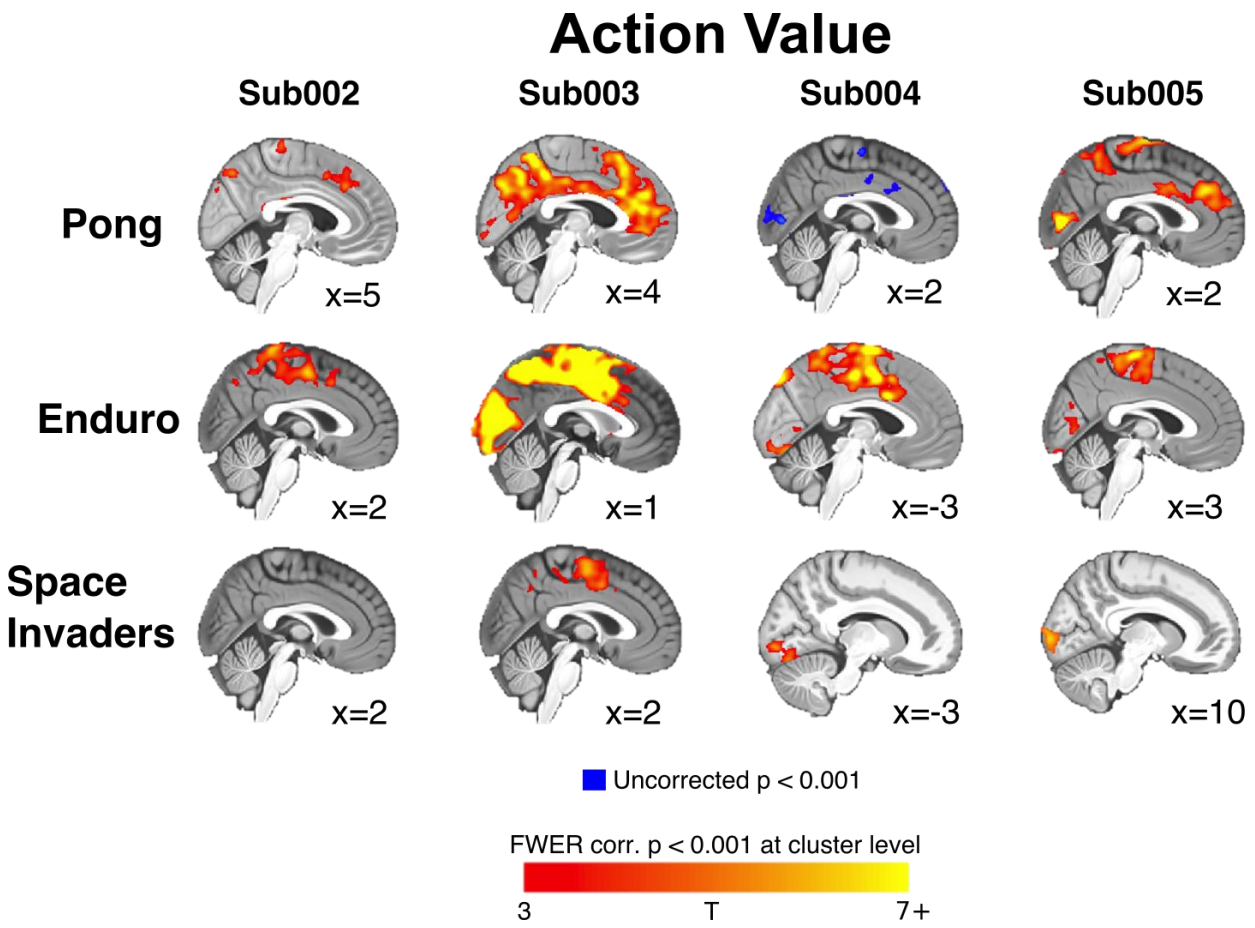
Supplementary Figure 5, related to Figure 4.

a. **DQN vs. control models by region for individual participants**. As in Figure 4B.

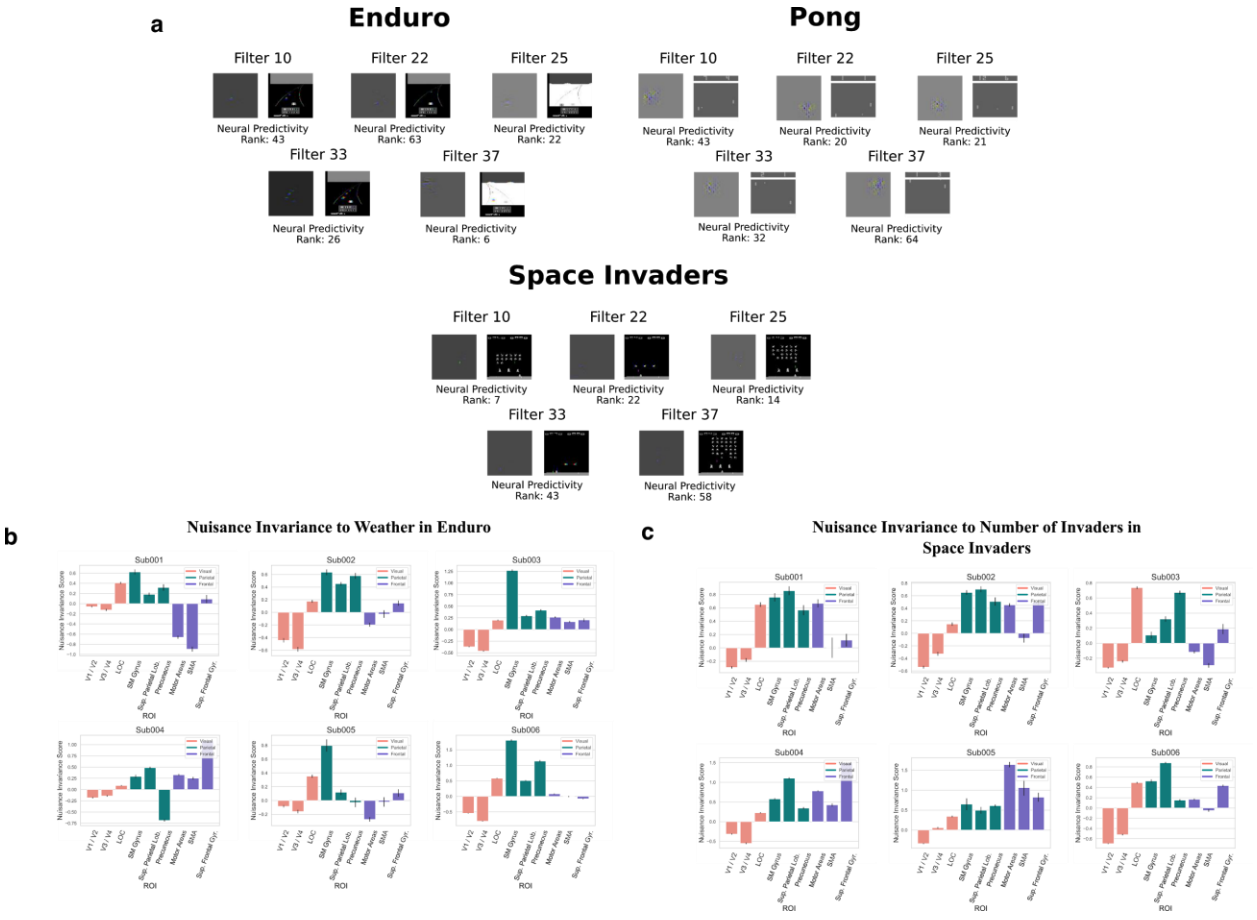
b. **Control Region Analysis.** To rule out the possibility that the encoding model analysis is picking up on motion related artifacts or other nuisances that affect the whole fMRI image, we ran the encoding model pipeline on two spheres of air directly outside of the brain (anterior) for sub001. The distribution of scores were around zero for every game and no voxels had significant prediction accuracies. The distribution of scores for V1/V2 and the superior parietal lobule are shown for comparison.



Supplementary Figure 6, related to Figure 5. Representational Similarity Analysis on fMRI data for Pong for individual subjects. As in Figure 5C. Asterisks (*) above bars indicate significance (block permutation tests, P < 0.01, FWER corrected for multiple comparisons). Dotted lines above bars indicate significant differences between models (block permutation tests, P < 0.01, FWER corrected for multiple comparisons).



Supplementary Figure 7, related to Figure 6. Action value results for other participants. Neural encoding of action value in premotor/motor areas. As in Figure 6B, whole brain maps thresholded at $p < 0.001$ (FWER corrected, cluster-level).



Supplementary Figure 8, related to Figures 7 and 8.

a. **Visualization of example DQN layer 3 filters using guided backpropagation.** Five randomly selected filters for each game are visualized for each game (as in Figure 7E), along with their Neural Predictivity rank. Images to the right of each example represent one image from the human gameplay data that activates the filter very highly. Gray images to the left of each example represent which parts of the pixel space affect the activation of the filter the most from this input image. Red, green, and blue colors reflect pixels that changed across the frames in the input.

b. **Nuisance invariance to weather/time of day in Enduro in individual participants.** Results depicted as in Figure 8B. The Motor Areas ROI includes both the primary motor cortex and premotor cortex.

c. **Nuisance invariance to number of invaders on the Screen in Space Invaders in individual participants**. Results depicted as in Figure 8C.