Supplemental Data

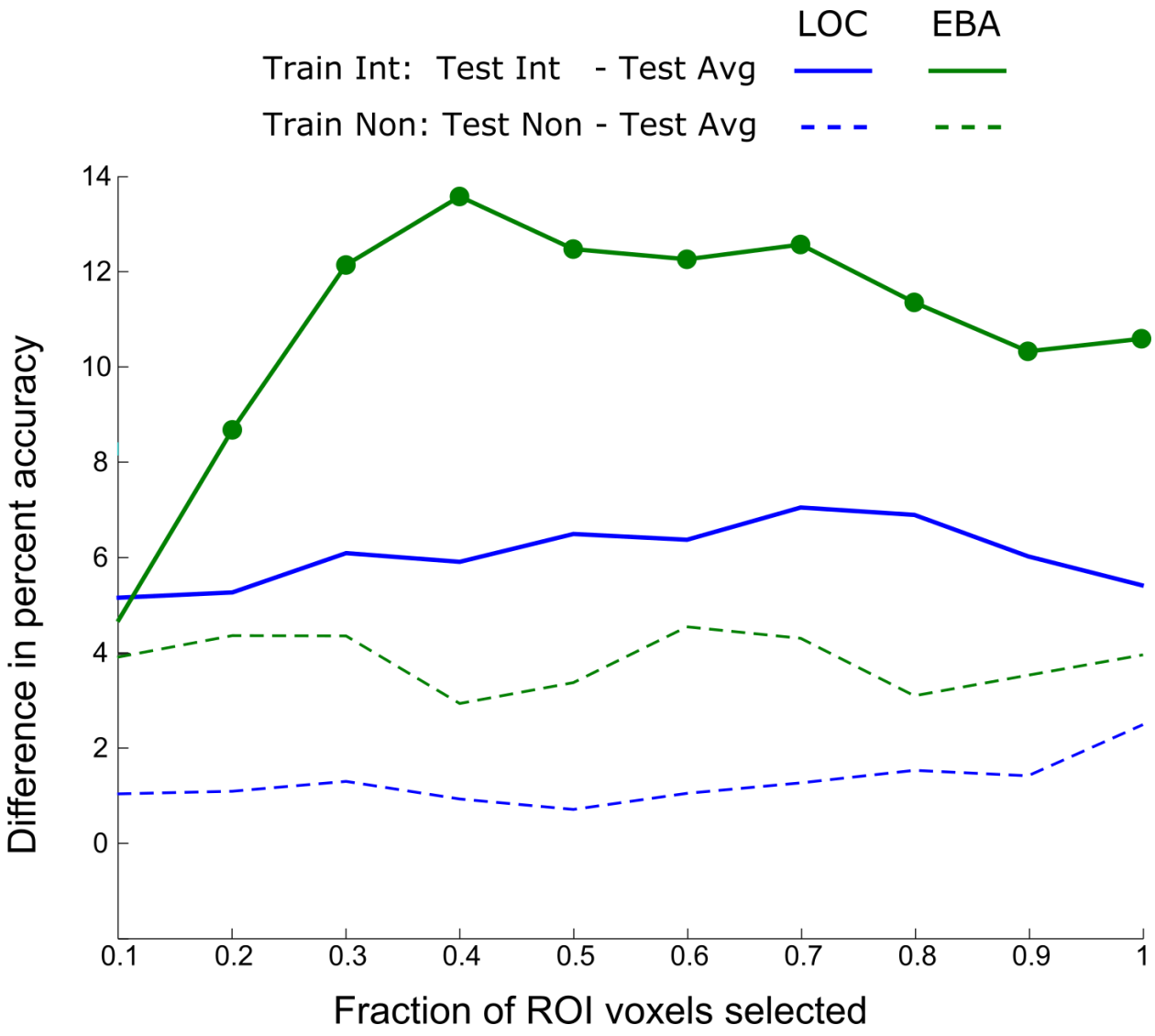

Figure S1: **Robustness of cross-decoding result to number of voxels selected**. The fraction of voxels selected for classifier training (selected based on overall visual responsiveness) did not have a major impact on the results reported in Figure 2. As long as at least 20% of the voxels in all areas were used in training, the same pattern of significant differences can be shown. Circled points are those that are significantly greater than zero ($p<0.05$ one-tailed t-test).

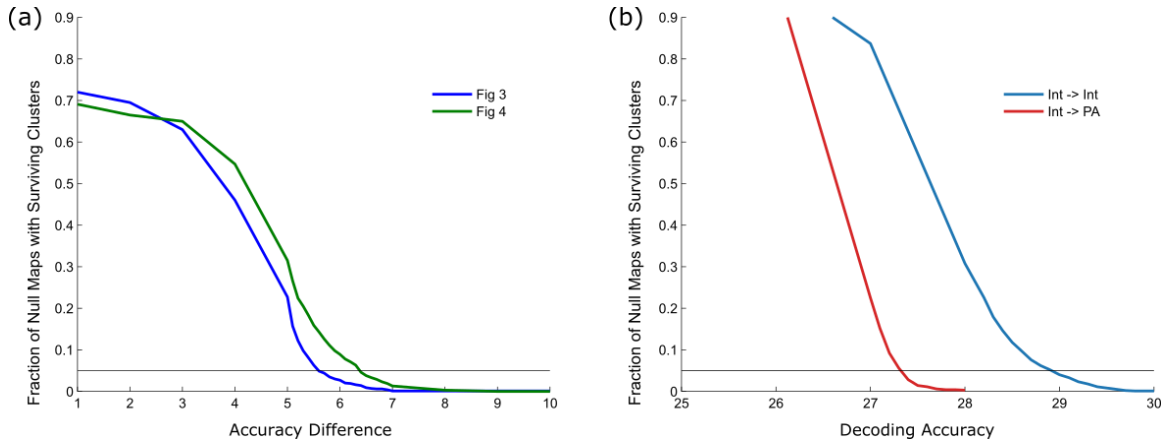

Figure S2: **Determination of decoding significance threshold**. 1,000 null searchlight maps were generated for each of the searchlight analyses, by randomly permuting the stimulus labels for each classifier and then running the decoding searchlight. A threshold was chosen for each searchlight such that fewer than 5% of the null difference maps yielded false positive clusters larger than 100 voxels. (a) For experiment 1, in which we are measuring differences between classifier accuracies, we obtain thresholds of 5.6 and 6.4. (b) For experiment 2, we are measuring 4-way decoding accuracies when training on interacting stimuli. We obtain thresholds of 29.0 when testing on interacting stimuli and 27.4 when testing on pattern averages. Since all but one of the interaction blocks from each category was used for training, only one block per category was used for testing on the interacting stimuli, while ten pattern-average blocks were available for testing; this resulted in higher variance in pure interaction decoding and thus a slightly higher group-level threshold.

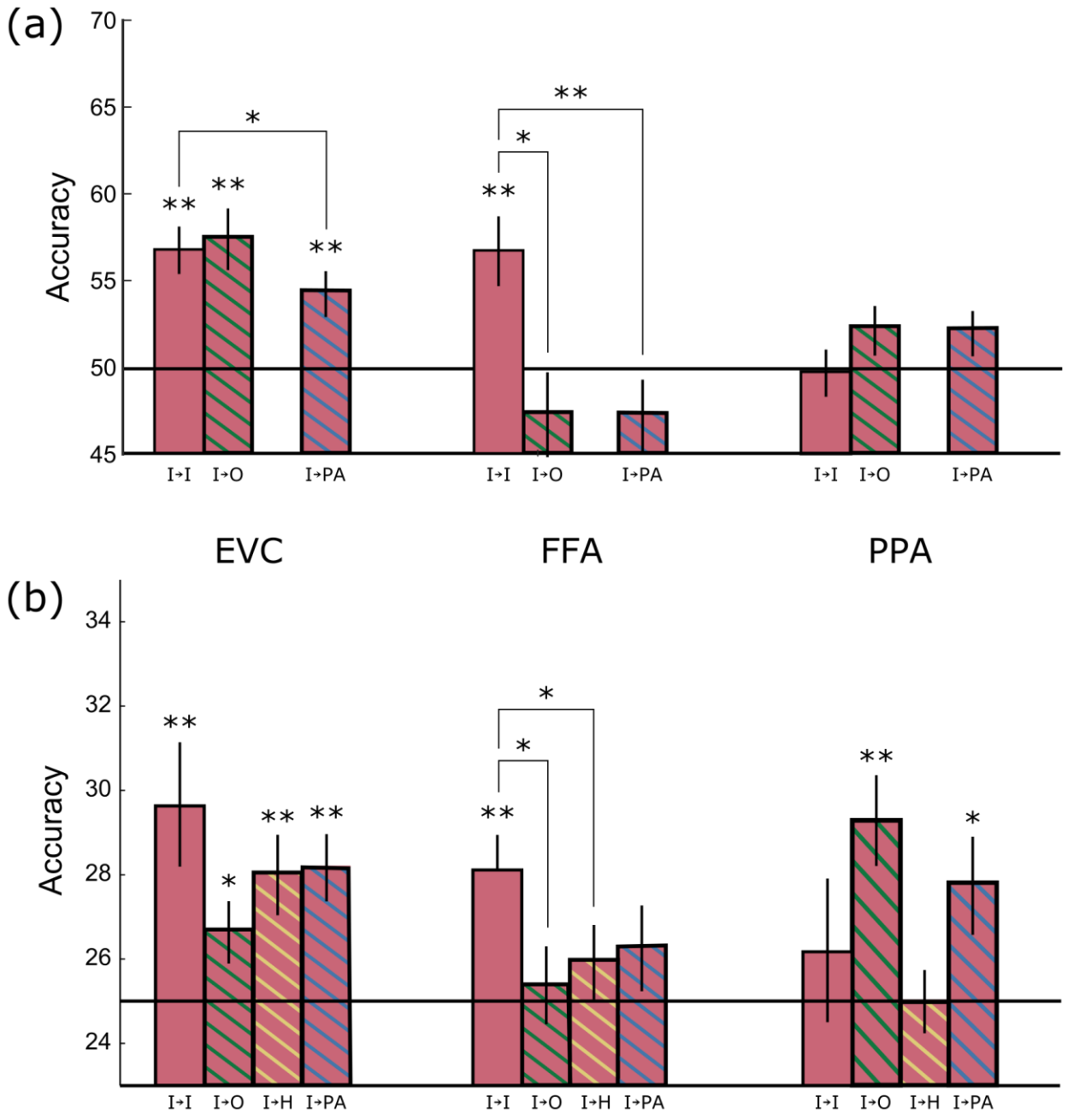

Figure S3: **Results in early visual cortex, FFA, and PPA.** (a) In Experiment 1, the nonlinearity of Interaction decoding was also tested for early visual cortex (EVC, average of results in V1, V2, ventral V3, and hV4), FFA, and PPA. EVC showed significant decoding when training on interactions and testing on all three conditions (interactions t_9=5.24, p<0.001; objects t_9=4.36, p<0.001; pattern averages, t_9=3.41, p=0.004; one-tailed t-test). There was a statistically significant but small drop when cross-decoding to pattern averages (t_9=2.04, p=0.036, one-tailed t-test) but not objects (t_9=-0.46, p=0.67, one-tailed t-test), indicating a small amount of nonlinear category representation. However, this drop was significantly smaller than that observed in EBA (Figure 2) (t_9=2.85, p=0.019, two-tailed t-test). FFA showed significant decoding only for pure interaction decoding (interactions t_9=3.48, p=0.003; objects t_9=-1.14, p=0.85; pattern averages,

t_9=-1.36, p=0.90; one-tailed t-test), with large drops in cross-decoding (to objects, t_9=2.75, p=0.011; to pattern averages, t_9=2.96, p=0.008; one-tailed t-test), indicating that its category representations are highly dissimilar from isolated objects or pattern averages. PPA failed to significantly decode interactions in any condition (interactions t_9=-0.24, p=0.59; objects t_9=1.58, p=0.07; pattern averages, t_9=1.61, p=0.07; one-tailed t-test). Note that there is no I→H condition in experiment 1, since the stimuli did not include category-specific human poses. (b) The same analysis was run in Experiment 2, and now cross-decoding was also applied to isolated humans (which had class-specific poses, unlike Experiment 1). Early visual cortex significantly decoded interaction classes, and this decoder successfully extended to isolated objects, isolated humans, and their pattern average (interactions t_{11}=3.23, p=0.004; objects t_{11}=2.31, p=0.021; humans t_{11}=3.23, p=0.004; pattern averages t_{11}=4.12, p<0.001; one-tailed t-test) with no significant drops (to objects t_{11}=1.78, p=0.052; to humans t_{11}=1.39, p=0.10; to pattern averages t_{11}=1.01, p=0.15; one-tailed t-test) indicating that the learned category representation is largely explained by the individual human and object components. FFA showed a very similar pattern of results to Experiment 1, with significant decoding only for pure interaction decoding (interactions t_{11}=4.10, p<0.001; objects t_{11}=0.42, p=0.34; humans t_{11}=1.08, p=0.15; pattern averages, t_{11}=1.27, p=0.12; one-tailed t-test), with drops in cross-decoding to isolated objects or humans (to objects, t_{11}=2.57, p=0.013; to humans, t_{11}=2.43, p=0.017; to pattern averages, t_{11}=1.59, p=0.060; one-tailed t-test). Finally, PPA shows an unusual pattern of results, in which the interacting decoder is not above chance but can successfully decode stimuli containing object information (interactions t_{11}=-0.72, p=0.24; objects t_{11}=4.10, p=0<0.001; humans t_{11}=-0.015, p=0.51; pattern averages, t_{11}=2.42, p=0.017; one-tailed t-test), indicating that there is noisy information about the isolated object present in the response to interaction images. Pure interaction decoding is significantly above chance when we restrict ourselves only to the PHC1/2 visual field maps in posterior PPA (Figure 7).

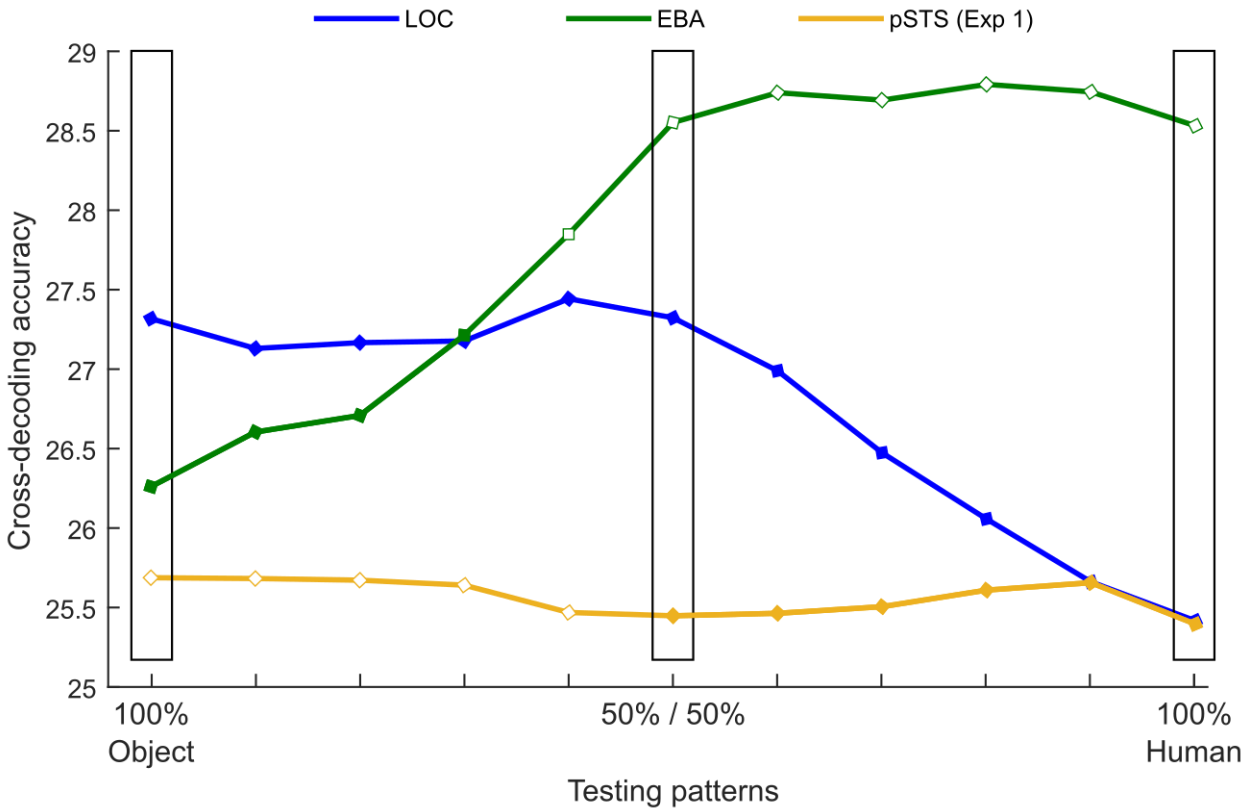

Figure S4: **Experiment 2 results for weighted pattern averages.** In the main text (Figure 6), a classifier is trained on responses to interacting stimuli and then tested on isolated objects or humans, or equally-weighted pattern averages of objects and humans (boxed points). Based on previous work (Baeck et al. 2013), it is possible that cross-decoding accuracy could be improved by using an unequal weighting of the object and human patterns. Sweeping the mixing fraction from object patterns alone to human patterns alone, however, shows that no unequal weighting provides substantially higher accuracy than an equal weighting (all $p>0.25$, one-tailed t-test). Filled points denote accuracies that are significantly ($p<0.05$, one-tailed t-test) below the accuracy when tested on full interactions.