Appendix for Same Cause; Different Effects in the Brain

Appendix A. Metric Normalization

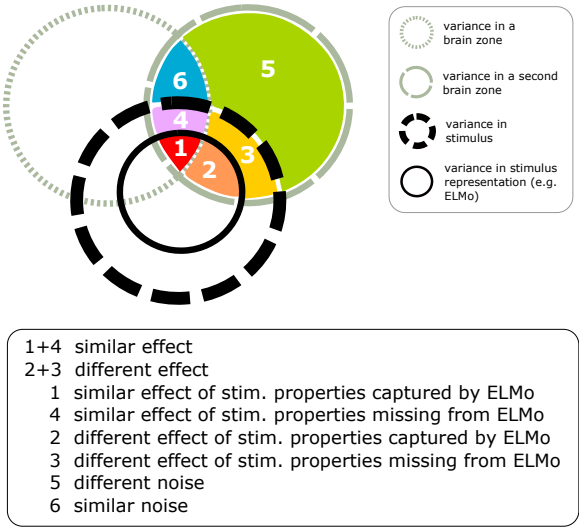

Figure 6: An illustration of possible effects that a stimulus can have on two brain zones, and the relationships with a stimulus-representation. Annotated regions **1 + 4** correspond to the stimulus having a similar effect on both brain zones, while **2 + 3** correspond to the stimulus having a different effect on the two brain zones (i.e. the stimulus overlaps with only one brain zone, not both brain zones). **1 + 2** indicate that the stimulus-representation captures the stimulus properties that cause at least part of the effect in the brain zone (**1** indicates the properties that have a similar effect on both brain zones, and **2** indicates the properties that have a different effect on both brain zones). Similarly, **3 + 4** indicate that the stimulus-representation is missing some stimulus properties that cause the specific effect in the brain zone. **6** indicates the shared noise between the brain zones, that is unrelated to the stimulus, and **5** indicates the noise that is unique to each brain zone.

The main metrics of interest defined in Section 2 are encoding model performance, zone generalization, and zone residuals. In the simple setting where all annotated regions in Fig. 6 are independent of each other, the encoding model performance is proportional to annotated regions **1 + 2**, zone generalization to region **1**, and zone residuals to regions **2 + 3** (and to the analogous regions **2 + 3** in the other brain zone). For some scientific questions, it may be more informative to normalize these metrics in different ways. For example, one may normalize the zone generalization for a target brain zone by the encoding model performance for the same brain zone to compute the proportion of $\frac{\mathbf{1}}{\mathbf{1+2}}$ (i.e. the proportion of information shared between a target brain zone and the stimulus-representation that is also shared by a second brain zone). This metric is identical to the one proposed by Toneva et al. (2020). Another type of normalization that we find informative in the current work is the inter-subject correlation (ISC), which is proportional to **1 + 2 + 3 + 4** (i.e. the information shared between a target brain zone and the stimulus). This metric can be thought of as an estimate of the maximum possible performance (i.e. the noise ceiling). A similar metric was used as an estimate of the noise ceiling by Wehbe et al. (2021), though the authors did not make

the connection to ISC explicitly. Note that the ISC across a dataset of more than two subjects is most frequently computed as the average of the pairwise ISC (i.e. the ISC for 1 of 6 subjects is the average across the ISC computed between that subject and the remaining 5 subjects). Following previous work (Hsu et al., 2004; Lescroart and Gallant, 2019), we normalize all of our metrics by the square-root of the noise ceiling, yielding normalized correlation values.

We hope that the conceptual breakdown that we present in Fig. 6 will help other researchers choose the most relevant normalization for their questions of interest.

Appendix B. Simulations

B.1. Data Generation Model

Simulating stimulus information. We generate two components that together make up all available stimulus information: $X \in \mathbb{R}^d$, which is the stimulus-representation, and $Z \in \mathbb{R}^d$, a representation of the remaining stimulus information that X does not capture. This is done by decomposing each of X and Z into four disjoint independent subsets of stimulus information: unique information that the individual brain zones respond to (X_1, X_2, Z_1, Z_2), joint information that both brain zones respond to (X_{12}, Z_{12}) and information that neither brain zone responds to (X_3, Z_3). Each X_i, Z_i, of length $\frac{d}{4}$, is independently sampled from a multivariate normal with mean 0 and a symmetric toeplitz covariance matrix with diagonal elements equal to 1. X and Z are then constructed by concatenating their four corresponding sub-components.

Simulating brain zone data. We simulate observations at two brain zones from two distinct participants using the following data generation model (motivated by Eq. 4):

$$Y_{i,P} = \alpha \times \underbrace{g_{12,P}(X)}_{\text{joint signal}} + (1-\alpha) \times \underbrace{g_{i,P}(X)}_{\text{unique signal}} + \alpha \times \underbrace{N_{i,P}}_{\text{unique noise}} + (1-\alpha) \times \underbrace{N_{12,P}}_{\text{joint noise}} \quad (7)$$

$$\text{where } N_{i,P} = \delta \times h_{i,P}(Z) + (1-\delta) \times \epsilon_{i,P} \text{ and } N_{12,P} = \delta \times h_{12,P}(Z) + (1-\delta) \times \epsilon_{12,P}.$$

Here, each $g_{i,P}(X) = \langle \theta_{i,P}, X_i \rangle$ is a linear function of the stimulus-representation that selectively acts on the corresponding X_i in X. In order to generate the necessary participant-specific parameters $\theta_{i,P} \in \mathbb{R}^{\frac{d}{4}}$, we first generate $\theta_i \in \mathbb{R}^{\frac{d}{4}}$ by independently sampling each of its components from a uniform distribution over $[0, 1)$. Each $\theta_{i,P}$ is then sampled from $\mathcal{N}(\theta_i, 0.25\mathbf{I})$ to allow for variation between participants. The same approach is used to generate each $h_{i,P}(Z) = \langle \phi_{i,P}, Z_i \rangle$ term. $\epsilon_1, \epsilon_2, \epsilon_{12} \in \mathbb{R}$ are terms that represent the information captured that is not related to the stimulus. Each ϵ_i is independently sampled from a standard normal distribution.

In the data generation model above, we introduced two adjustable parameters α and δ to simulate a wide range of scenarios. Parameter $\alpha \in [0, 1]$ controls how similarly the two zones respond to the stimulus properties captured in the stimulus-representation. We designed the weightings in Eq. 7 such that the total variance of each of the following four components remains constant when varying α: the total signal (joint+unique), noise (joint+unique), joint information (signal+noise) and unique information (signal+noise). Parameter $\delta \in [0, 1]$ controls the proportion of stimulus properties that are driving the brain zones but are not captured by the stimulus-representation. The results shown in Fig. 2(Left) and Fig. 2(Right) were collected by varying α (when $\delta = 1.0$) and δ (when $\alpha = 1.0$) respectively in the simulations that were performed. This allowed us to smoothly interpolate between four inferences that we want to, but cannot, distinguish between using just encoding model performance.

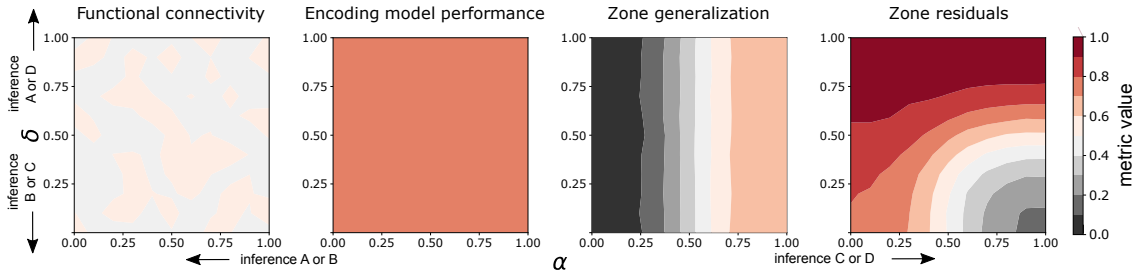

Figure 7: Plotting how each metric varies under simulations performed at different settings of α, δ.

B.2. Additional Simulation Results

In Fig. 7, we present how encoding model performance, zone generalization, zone residuals, and functional connectivity vary as we allow $\alpha, \delta \in [0, 1]$ to vary with respect to each other in Eq. 7. Since inferences A and B are characterized by both zones responding differently to the stimulus properties captured in the stimulus-representation and inferences C and D are characterized by both zones responding similarly to them, increasing α from 0.0 to 1.0 lets us adjust from the former pair of inferences to the latter. Also recall that one can separate inference A from B and inference C from D, if one has information about the extent to which both zones respond to stimulus properties not captured by the stimulus-representation. Therefore, at a high fixed α, as δ is increased from 0.0 to 1.0, we move from inference B to inference A or inference C to inference D (depending on the pair we narrowed down earlier).

Fig. 7 shows that zone generalization is the only metric of the four we consider that can be used to separate inferences A and B from inferences C and D - i.e., distinguish between brain zone data that is simulated using low and high values of α respectively at any choice of fixed δ. This aligns with our observations from Fig. 2(Left), where we concluded that looking at zone generalization can help us identify which pair of inferences we should further investigate. To know what we can precisely infer, we would need a metric that lets us distinguish between brain zones based on the extent to which they respond to stimulus properties that are not captured by the stimulus-representation - in our simulations, between brain zone data simulated at different values of δ when α is high. Fig. 7 shows that at high α, zone residuals increases with δ, and therefore, can be useful when separating inferences B and C from inferences A and D. This also aligns with our observations from Fig. 2(Right).

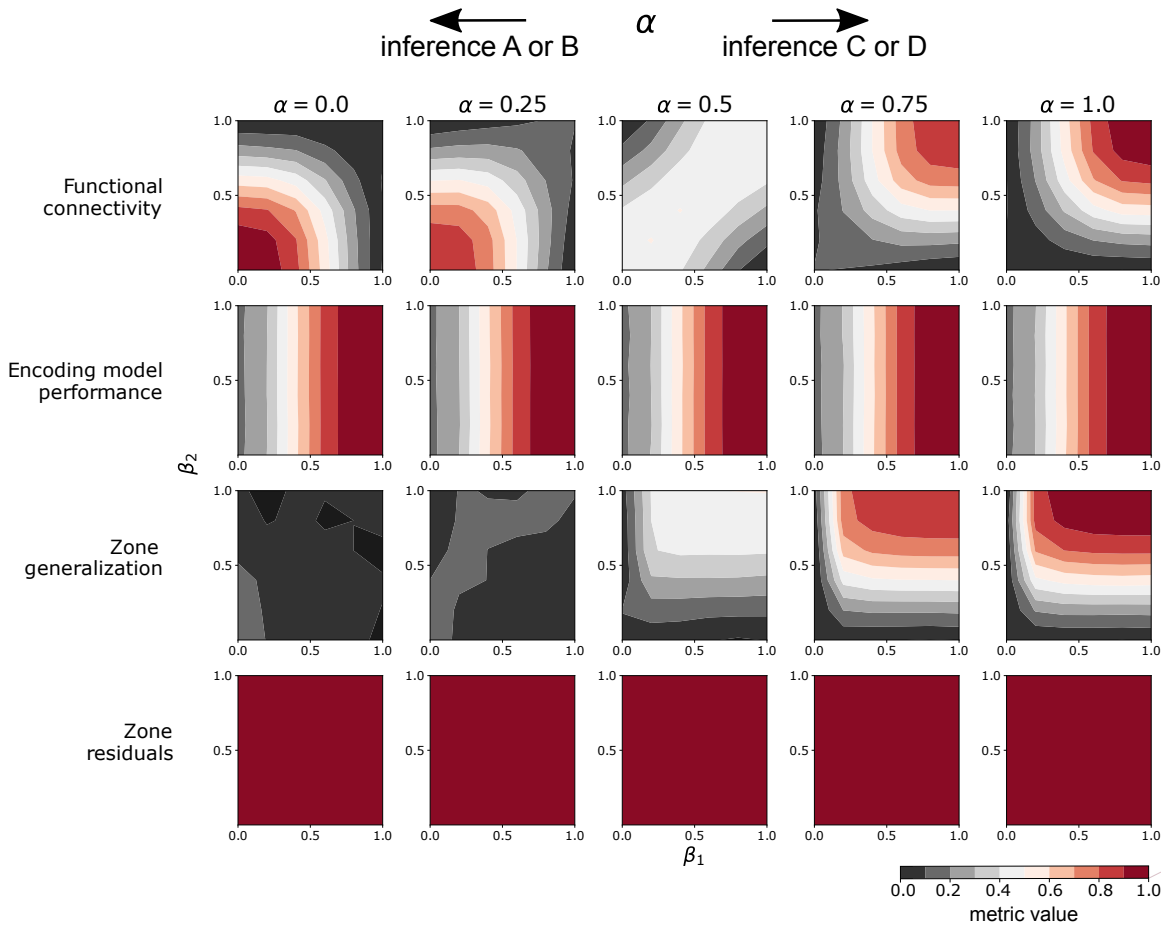

Figure 8: Extending on Fig. 2(Left), this figure shows how each metric varies under simulations performed at different signal-to-noise ratios as we vary α, β_1, β_2 when $\delta = 1.0$ is fixed.

B.3. Varying signal-to-noise ratio in simulated brain zone data

To test the limits of what each metric can tell us about the underlying relationships between a pair of brain zones, we further extend Eq. 7 to control the extent to which the activity in each zone is driven by aspects of the stimulus that are captured by (vs. missing from) the stimulus-representation:

$$Y_{i,P} = \beta_i \underbrace{[\alpha \times g_{12,P}(X) + (1-\alpha) \times g_{i,P}(X)]}_{\text{signal}} + (1-\beta_i) \underbrace{[\alpha \times N_{i,P} + (1-\alpha) \times N_{12,P}]}_{\text{noise}} \quad (8)$$

$$\text{where } N_{i,P} = \delta \times h_{i,P}(Z) + (1-\delta) \times \epsilon_{i,P} \text{ and } N_{12,P} = \delta \times h_{12,P}(Z) + (1-\delta) \times \epsilon_{12,P}.$$

Above, we introduce an additional type of parameter $\beta_i \in [0, 1]$. In this context, zone activity that is driven by stimulus properties captured by the stimulus-representation can be viewed as the signal that is retrievable by an encoding model. The remaining activity, whether stimulus-driven or not, can be viewed as the noise that an encoding model cannot explain as it only has access to the stimulus-representation.

First, we focus on the separation between the pairs of inferences A and B and inferences C and D by varying α. In Fig. 8, we show how each metric varies as we vary β_1 and β_2 for each setting of α. β_1 and β_2 control the signal-to-noise ratio in both simulated brain zones. We fix $\delta = 1.0$ here, but similar trends can be observed for other choices of fixed $\delta \in [0, 1]$ as well. We observe that encoding model performance and zone residuals do not allow us to distinguish between different α values. Note also that functional connectivity cannot be used to identify when both brain zones mostly respond to shared noise (low β_i's, low α) from when they mostly respond to shared signal (high β_i's, high α). We observe that given sufficient signal in both brain zones ($\beta_1, \beta_2 \gg 0$), zone generalization increases as α increases. However, under conditions of little to no signal, zone generalization cannot be used to distinguish between different α values. These results suggest that zone generalization is a useful metric to separate the pair of inferences A and B (low α) from the pair C and D (high α) only when the encoding models used are able to perform relatively well on the brain zones they are trained on.

Next, we consider the separation between the pairs of inferences B and C and inferences A and D by keeping a high fixed α and varying δ. In Fig. 9, we show how each metric varies as we vary β_1 and β_2 for each setting of δ (when $\alpha = 1.0$). We find that encoding model performance, zone generalization and functional connectivity are not useful to distinguish between different values of δ in our simulations. We observe that given sufficient noise in brain zone 1 ($\beta_1 \ll 1$), zone residuals increases as δ increases. This suggests that the zone residuals can help us separate inferences B and C (low δ) from inferences A and D (high δ). However, in the case when β_1 is high, increasing δ does not impact the zone residuals as they are already saturated. In the case where the zone residuals are already saturated, the zone residuals enable us narrow our search down to inferences A or D.

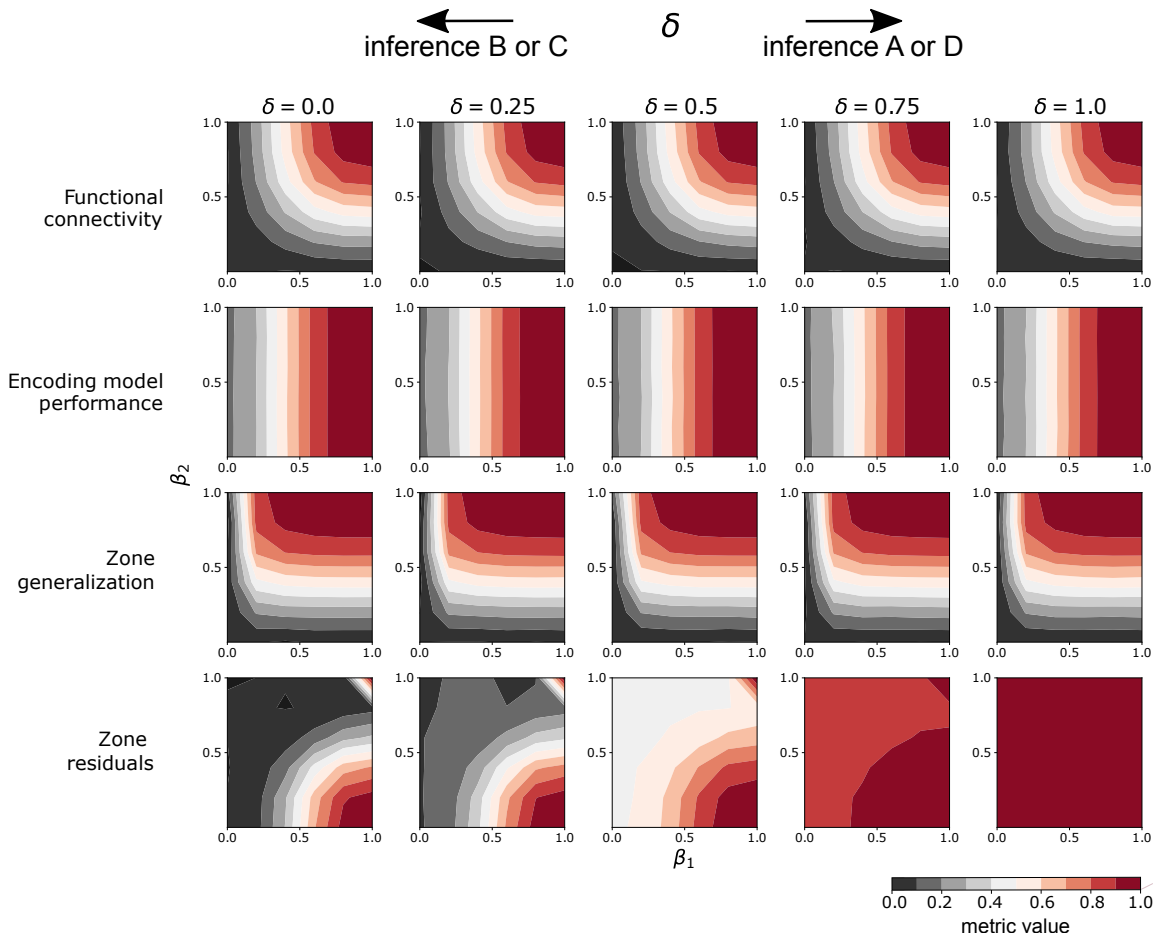

Figure 9: Extending on Fig. 2(Right), this figure shows how each metric varies under simulations performed at different signal-to-noise ratios as we vary δ, β_1, β_2 when $\alpha = 1.0$ is fixed.

B.4. RSA and Functional Connectivity Simulation Analyses

We also tested whether RSA or functional connectivity are able to infer the true underlying relationships in the same synthetic brain zone dataset described in Section 3. In Fig. 10 we extended Fig. 2 to include these RSA and functional connectivity metrics. We conducted two types of RSA: (1) between an RDM corresponding to each of the brain zones and an RDM corresponding to the synthetic stimulus-representation, and (2) between the two brain zone RDMs. What we found was that, similarly to encoding model performance in Fig. 2, all of the RSAs resulted in a flat line as we varied (1) how similarly the zones respond to the stimulus properties captured by the stimulus-representation (i.e. different values of α) (Fig. 10(Left)) and (2) the extent to which both zones respond to stimulus properties not captured by the stimulus-representation (i.e. different values of δ) (Fig. 10(Right)). The same is true of functional connectivity.

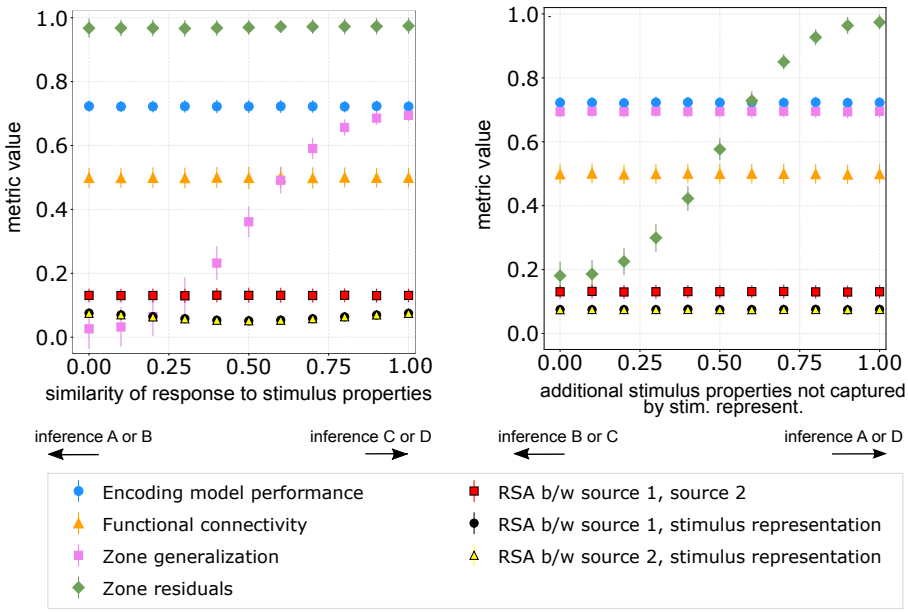

Figure 10: Extending on Fig. 2, this figure shows how different RSA-based metrics and functional connectivity vary on the same synthetic dataset.

Appendix C. Data Preprocessing

C.1. HCP

Our analyses are performed with the 3105 TRs (51 minutes and 45 seconds) suggested for analysis in the HCP documentation. These exclude rest periods and the first 6 TRs of each movie clip within a movie run. Individual-level results are presented for the six participants with the highest encoding model performance averaged over the 55 Shen atlas language ROIs.

C.2. Courtois NeuroMod

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.1.0 (Esteban et al., 2018b,a). Three participants are native French speakers and three are native English speakers. All participants are fluent in English and report regularly watching movies in English.

C.3. Other Pre-processing

The fMRI datasets and Shen atlas were provided in different template spaces and voxel sizes. We resample and register the Shen atlas (MNI27 template space, voxel size = 1 mm isotropic) to both the HCP template space (MNI152NLin6Asym, voxel size = 1.6 mm isotropic) and the Courtois NeuroMod template space (ICBM2009cNlinAsym, voxel size = 2 mm isotropic) using FSL FMRIB Linear Image Registration Tool (FLIRT) (Jenkinson et al., 2002). We perform all analyses for the two datasets in their respective template space.

We further process the ELMo embeddings before we use them as the input features to our encoding models. First, we use a Lanczos filter with the same parameters as Huth et al. to downsample the embeddings into a feature matrix where each row corresponds to a feature vector for a TR (Huth et al., 2016). Then, to reduce the dimensionality of our feature space we use principle component analysis (PCA) to select the first 10 principle components. The first 10 principle components explain 50.5% of the variance in the Courtois NeuroMod dataset and 49.9% of the variance in the HCP dataset. Next, to account for the lag in the hemodynamic response in fMRI data, we delay the feature matrix in accordance with previous work (Nishimoto et al., 2011; Wehbe et al., 2014a; Huth et al., 2016).

Appendix D. 34 Language ROI Heatmap Tick Numbers Projected on the Brain

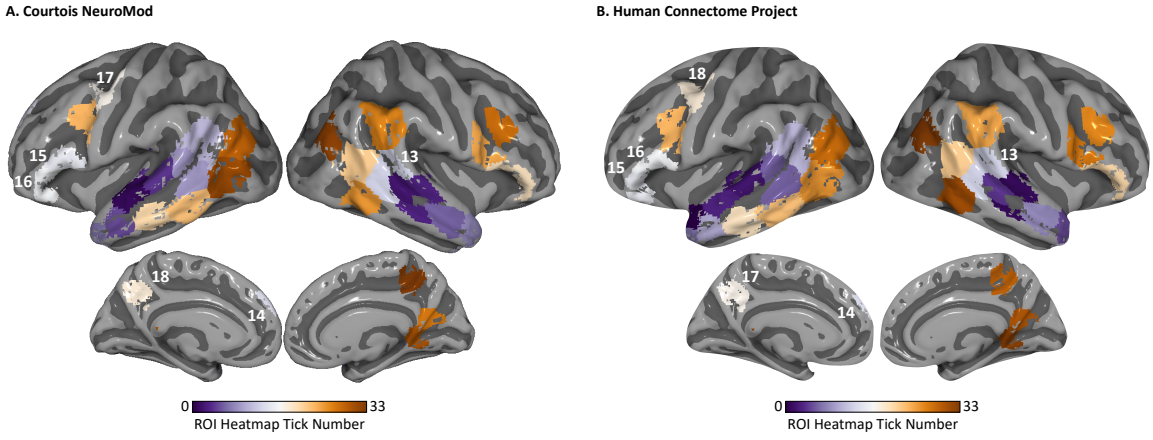

Figure 11: ROI Heatmap Tick Numbers. Signification Language ROIs on cortical surface are colored according to the tick number for each ROI in the (A) Courtois NeuroMod and (B) HCP heatmaps in Figs. 4, 12, 14, 15. The ROI are sorted from high (ROI 0) to low (ROI 33) median normalized zone generalization (average over participants).

Appendix E. Negative Normalized Zone Generalization

We investigated why negative norm. zone generalizations arise. We have found in our empirical results that when there is a negative norm. zone generalization, the two ROIs have different positive weights on the features in the stimulus-representation. For example, some ROIs put high weights on features associated with word rate, while other ROIs put more weight on the rest of the stimulus-representation. This suggests that at least some stimulus properties affect the two ROIs differently. However, it is unclear what these ROIs respond to besides word rate. The negative norm. zone generalization suggests that the ROIs have opposite weights on the features in the stimulus-representation. This suggests that at least some stimulus properties affect the ROIs differently, however it is unclear if the stimulus-representation is incomplete and does not include all the stimulus properties that affect the ROIs similarly. Consequently, from a negative zone generalization alone it is not possible to infer if the stimulus properties affect the ROIs mostly differently (inference A) or if some stimulus properties affect ROIs differently and other properties not captured by ELMo affect the ROIs similarly (inference B). Therefore, to interpret the negative norm. zone generalization values we also need the zone residuals.

Appendix F. Normalized Zone Residuals on Two Naturalistic fMRI Datasets

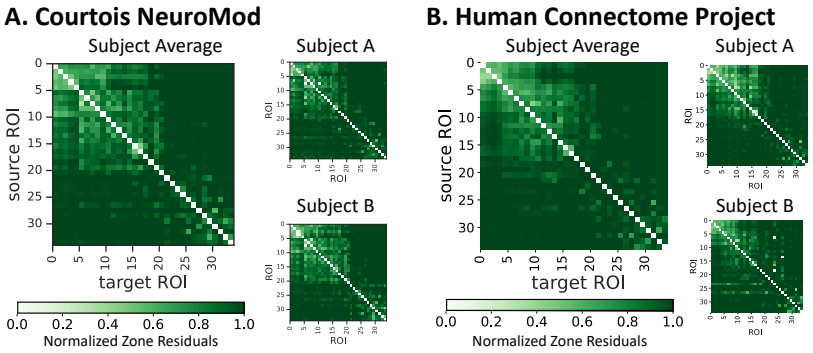

Figure 12: Zone Residuals. ROI pairs with large norm. zone residuals (dark green) are responding differently to at least some stimulus properties (inference A or D). These ROI pairs with large norm. zone residuals are consistent at the group and participant-level in both datasets.

Appendix G. Additional Participant-Level Empirical Results

We present the participant-level results for the remaining four participants in the Courtois NeuroMod dataset and four additional participants for the HCP dataset. We observe that these additional participants appear similar to the average and two representative participants presented in the main text and Appendix F.

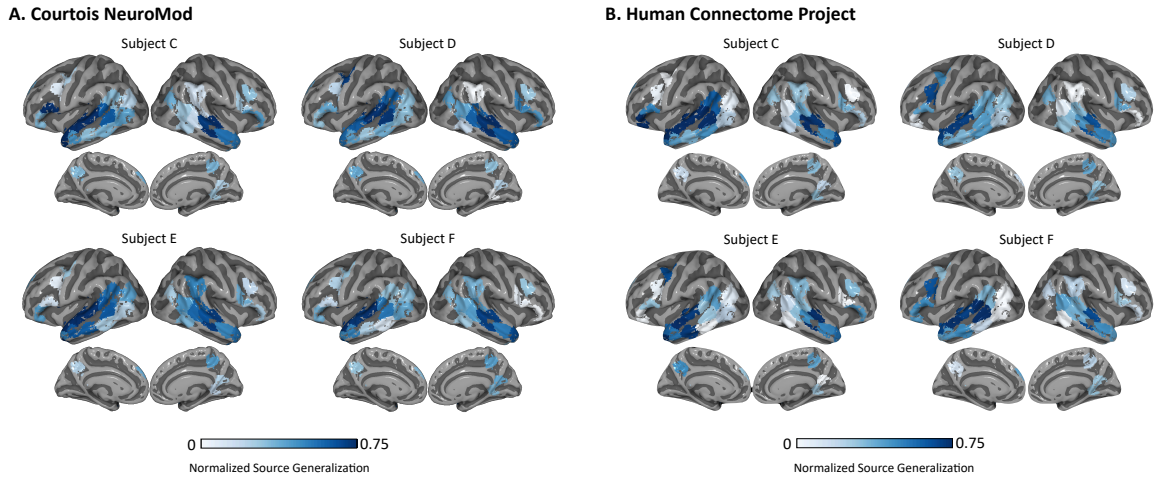

Figure 13: (Related to Fig. 3) Encoding Model Performance. Similar to Fig. 3 in the main text, this figure shows the normalized encoding model performance at 34 significantly predicted ROIs (corrected at level 0.05) for participants C-F in both the (A) Courtois NeuroMod and (B) Human Connectome Project datasets. Plots were created using the Pycortex software (Gao et al., 2015).

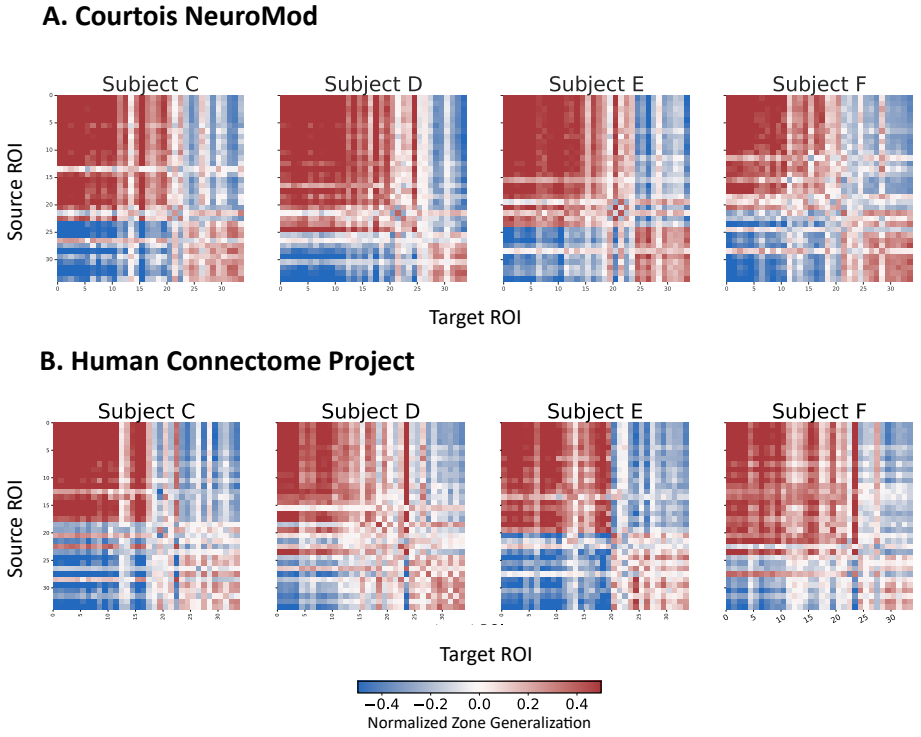

Figure 14: (Related to Fig. 4) Zone Generalization. Similar to Fig. 4 in the main text, this figure shows the normalized zone generalization for participants C-F in both the (A) Courtois NeuroMod and (B) Human Connectome Project datasets. ROI pairs with high normalized zone generalization (red) are consistent across participants C-F in both datasets. They are also consistent with the group level and participants presented in the main text.

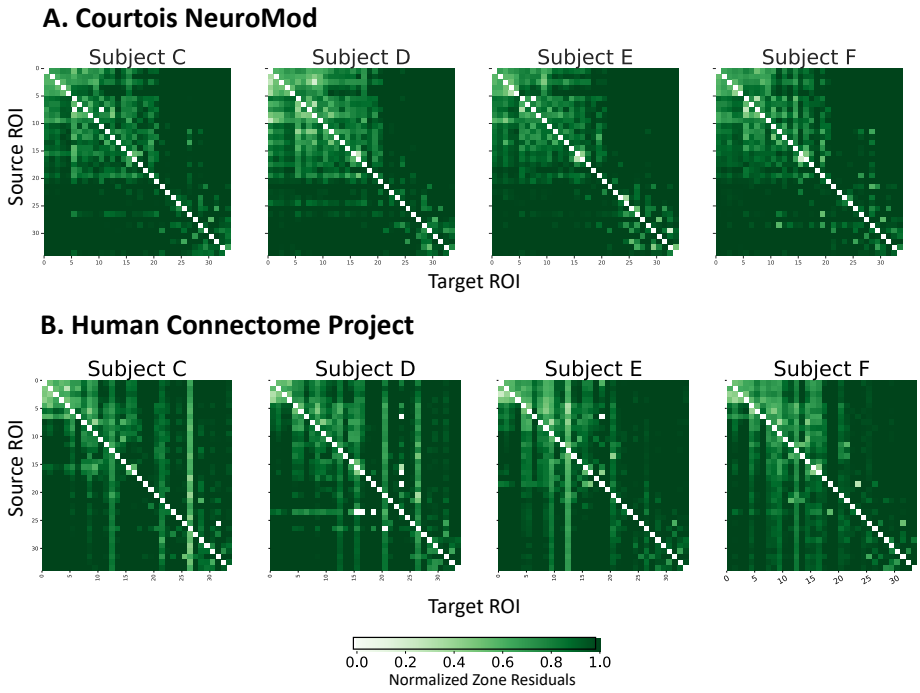

Figure 15: (Related to Appendix Fig. 12) Zone Residuals. Similar to Appendix Fig. 12, this figure shows the normalized zone residuals for participants C-F in both the (A) Courtois NeuroMod and (B) Human Connectome Project datasets. ROI pairs with high normalized zone residuals (dark green) are consistent across participants C-F in both datasets. They are also consistent with the group level and participants presented in Appendix Fig. 12.

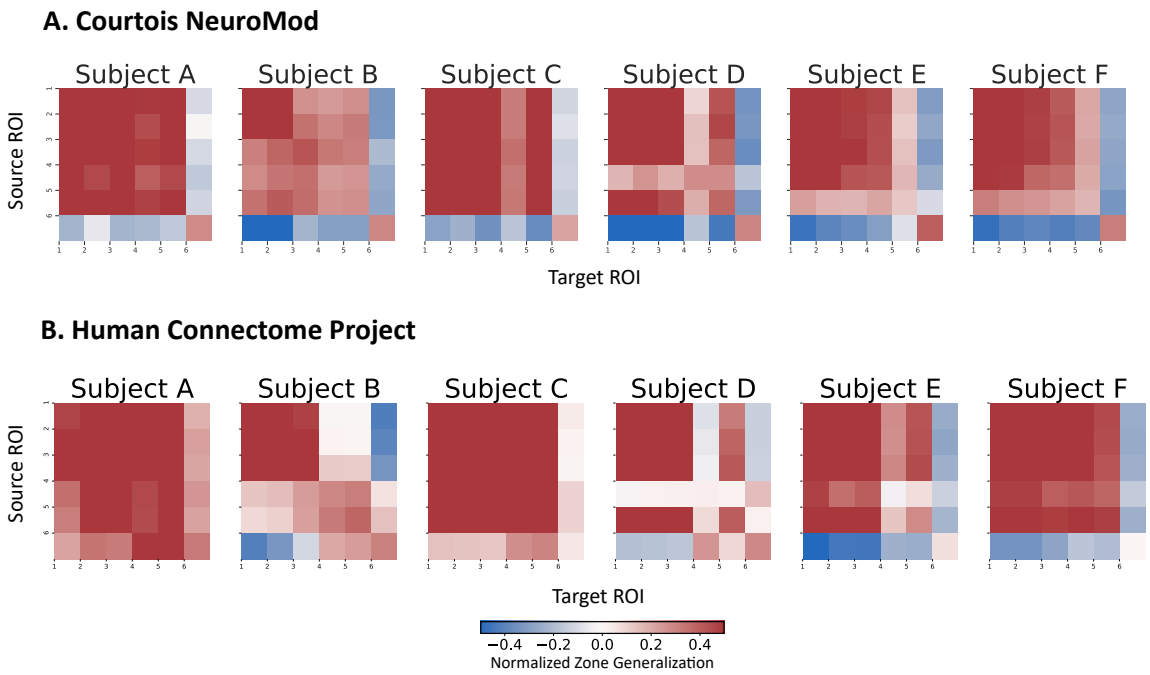

Figure 16: (Related to Fig. 5) Proposed Framework Example Participant-Level Zone Generalization. Similar to Fig. 5 in the main text, this figure shows the normalized zone generalization for the six ROIs in the example using the proposed framework for participants A-F in both the (A) Courtois Neuromod and (B) Human Connectome Project datasets. The ROI pairs with high normalized zone generalization (red) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.

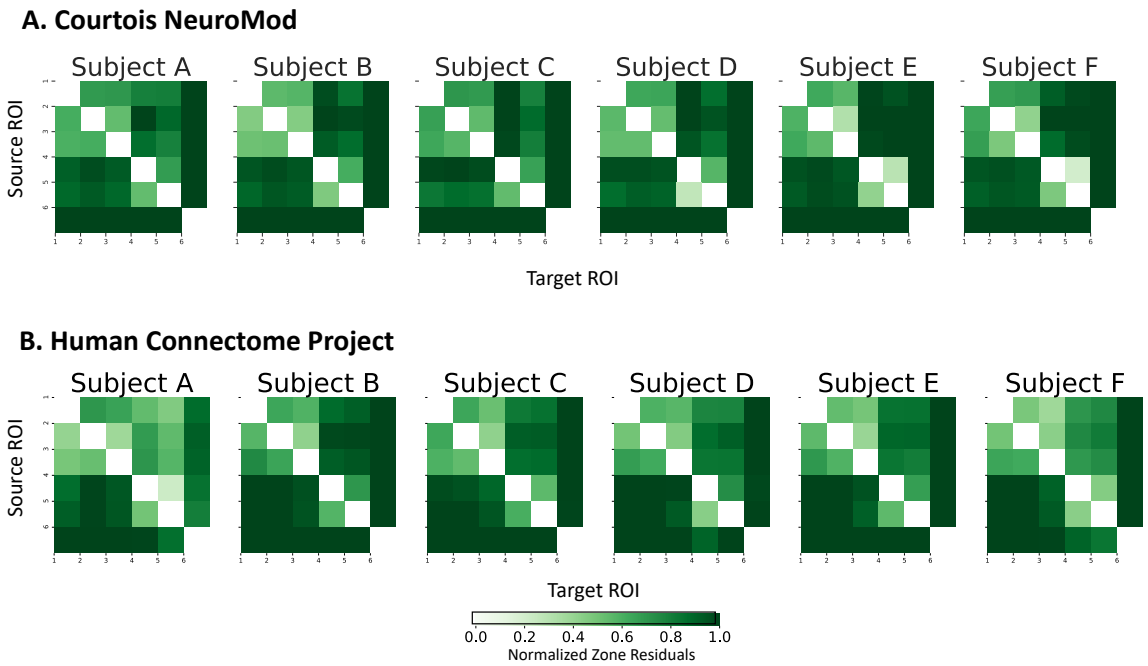

Figure 17: (Related to Fig. 5) Proposed Framework Example Participant-Level Zone Residuals. Similar to Fig. 5 in the main text, this figure shows the normalized zone residuals for the six ROIs in the example using the proposed framework for participants A-F in both the (A) Courtois Neuromod and (B) Human Connectome Project datasets. The ROI pairs with high normalized zone residuals (dark green) are consistent across participants A-F in both datasets. They are also consistent with the group level presented in the main text.