Supplementary Information for "NiCLIP: Neuroimaging contrastive language-image pretraining model for predicting text from brain activation images"

Julio A. Peraza[1], James D. Kent[2], Thomas E. Nichols[3], Jean-Baptiste Poline[4], Alejandro de la Vega[2], Angela R. Laird[1]

[1] Department of Physics, Florida International University, Miami, FL, USA
[2] Department of Psychology, University of Texas at Austin, Austin, TX, USA
[3] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[4] Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada

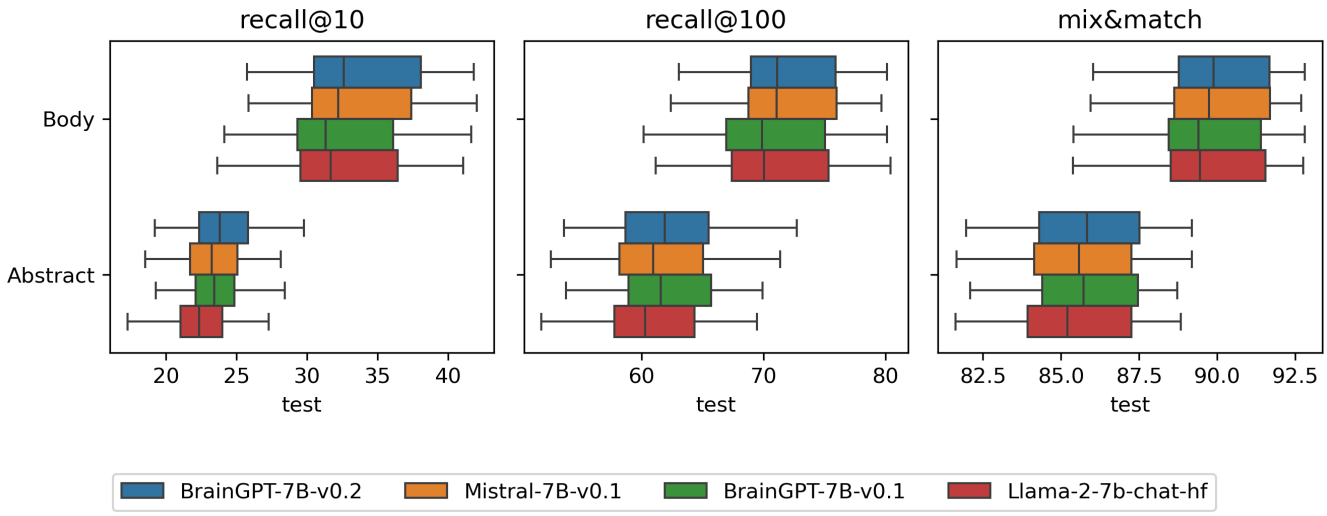


Figure S1. Text-to-Brain Association Performance Comparison. Box plots show the distribution of Recall@10, Recall@100, and Mix&Match scores obtained across cross-validation folds when training the CLIP backbone on the full article body versus the abstract only (upper and lower groups within each panel, respectively). Within each group, boxes correspond to the four large language models evaluated for text feature extraction, ordered top to bottom as BrainGPT-7B-v0.2, Mistral-7B-v0.1, BrainGPT-7B-v0.1, and Llama-2-7b-chat-hf, consistent with Table 1. Box edges span the interquartile range, the center line marks the median, and whiskers extend to the most extreme values within 1.5 times the interquartile range. Higher values indicate better text-to-brain alignment for all three metrics.

***Corresponding Author*: Dr. Angela R. Laird, Professor of Physics, Florida International University, Miami, FL, USA | Email: alaird@fiu.edu**

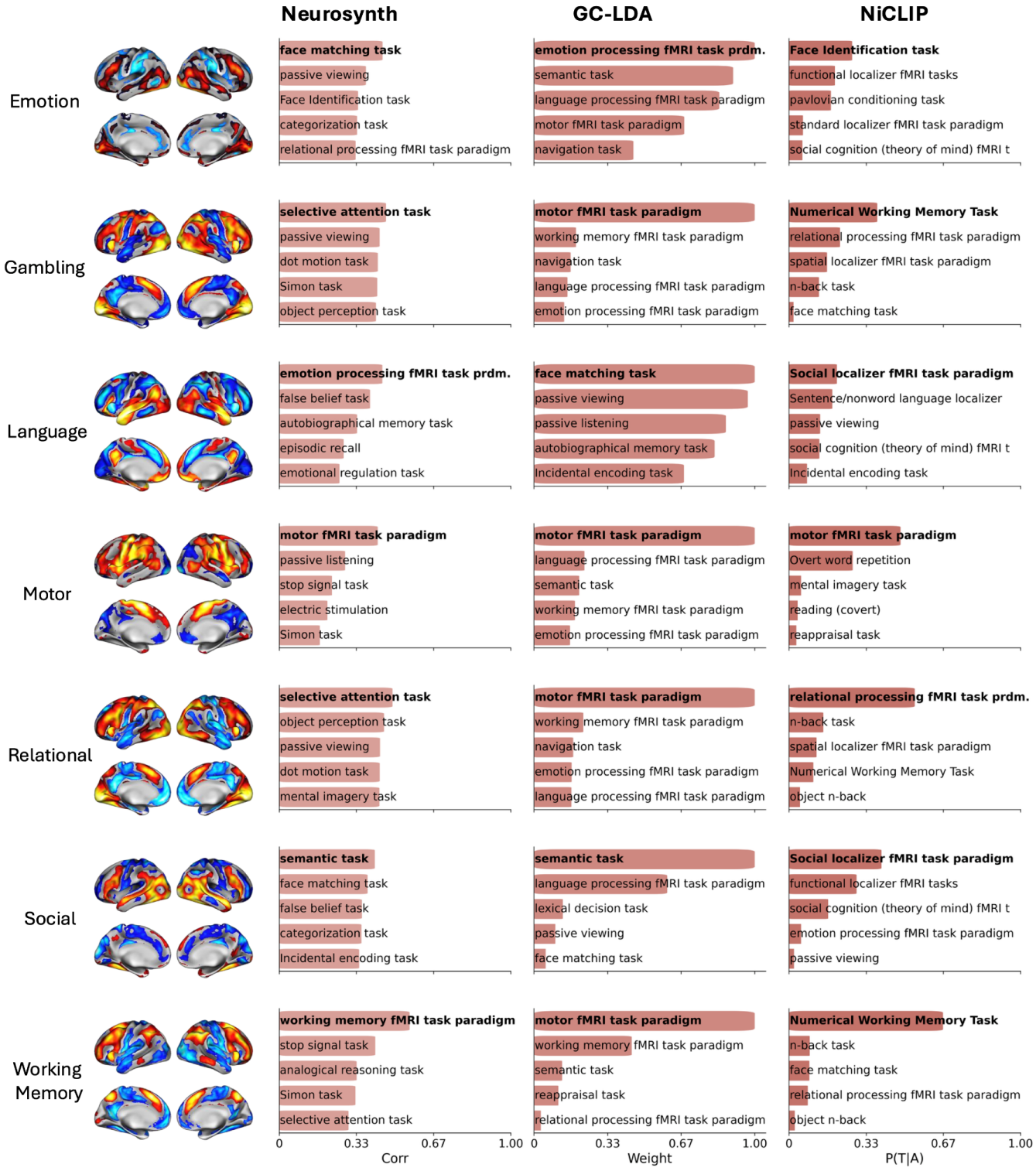


Figure S2. <u>Comparisons Between NiCLIP and Baseline Models' Task Predictions with Group-Level Maps</u>. For each of the seven HCP task domains (Emotion, Gambling, Language, Motor, Relational, Social, and Working Memory), the corresponding group-level activation map is shown on the left, followed by the top five predicted terms from Neurosynth (correlation-based decoding), GC-LDA (topic

weight-based decoding), and NiCLIP (P(T | A), the posterior probability of a task given an activation). Bars represent prediction strength for each term, and the top prediction for each method is shown in bold. This analysis complements the group-level results presented in Fig. 2, allowing a direct comparison between NiCLIP and the two baseline decoding approaches on the same activation maps.

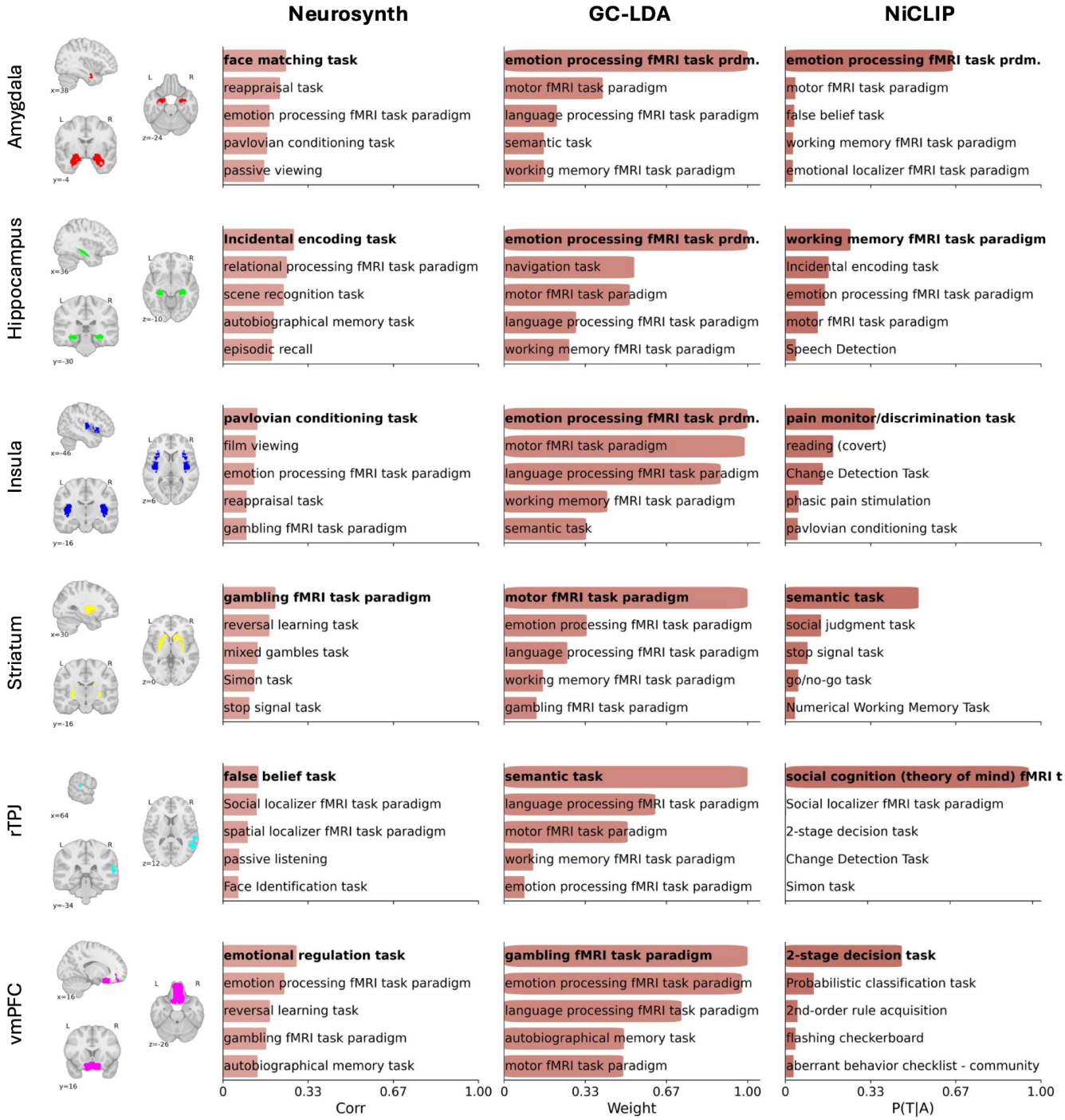


Figure S3. <u>Comparisons Between NiCLIP and Baseline Models' Task Predictions with Brain ROIs</u>. We compared task predictions across six anatomical ROIs (amygdala, hippocampus, insula, striatum,

rTPJ: right temporoparietal junction, vmPFC: ventromedial prefrontal cortex) using three decoding methods. For each ROI, the anatomical mask is shown on the left, followed by the top five predicted terms from Neurosynth (correlation-based decoding), GC-LDA (topic weight-based decoding), and NiCLIP (P(T | A), the posterior probability of a task given an activation). Bars represent prediction strength, and the top prediction for each method is shown in bold. This analysis complements the ROI-level results presented in Fig. 3, allowing a direct comparison between NiCLIP and the two baseline decoding approaches on the same regions.

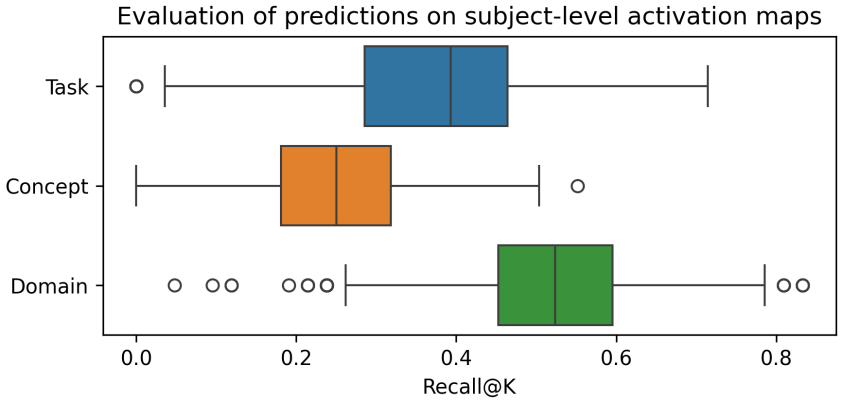


Figure S4. Distribution of Recall@K Score for Task, Concept, and Domain Predictions by NiCLIP on Subject-Level Activation Maps. Box plots show the distribution of Recall@K scores across all HCP subject-level contrast images for task, concept, and domain predictions by NiCLIP. Consistent with Table 2, task and concept decoding are evaluated with Recall@4, while domain decoding is evaluated with Recall@2, reflecting the proportion of correct labels recovered among the top ranked predictions for each activation map. Open circles denote individual subjects falling outside 1.5 times the interquartile range from the nearest box edge. Domain predictions achieve the highest median recall, followed by task and then concept predictions, although all three distributions include a subset of subjects with recall near zero.

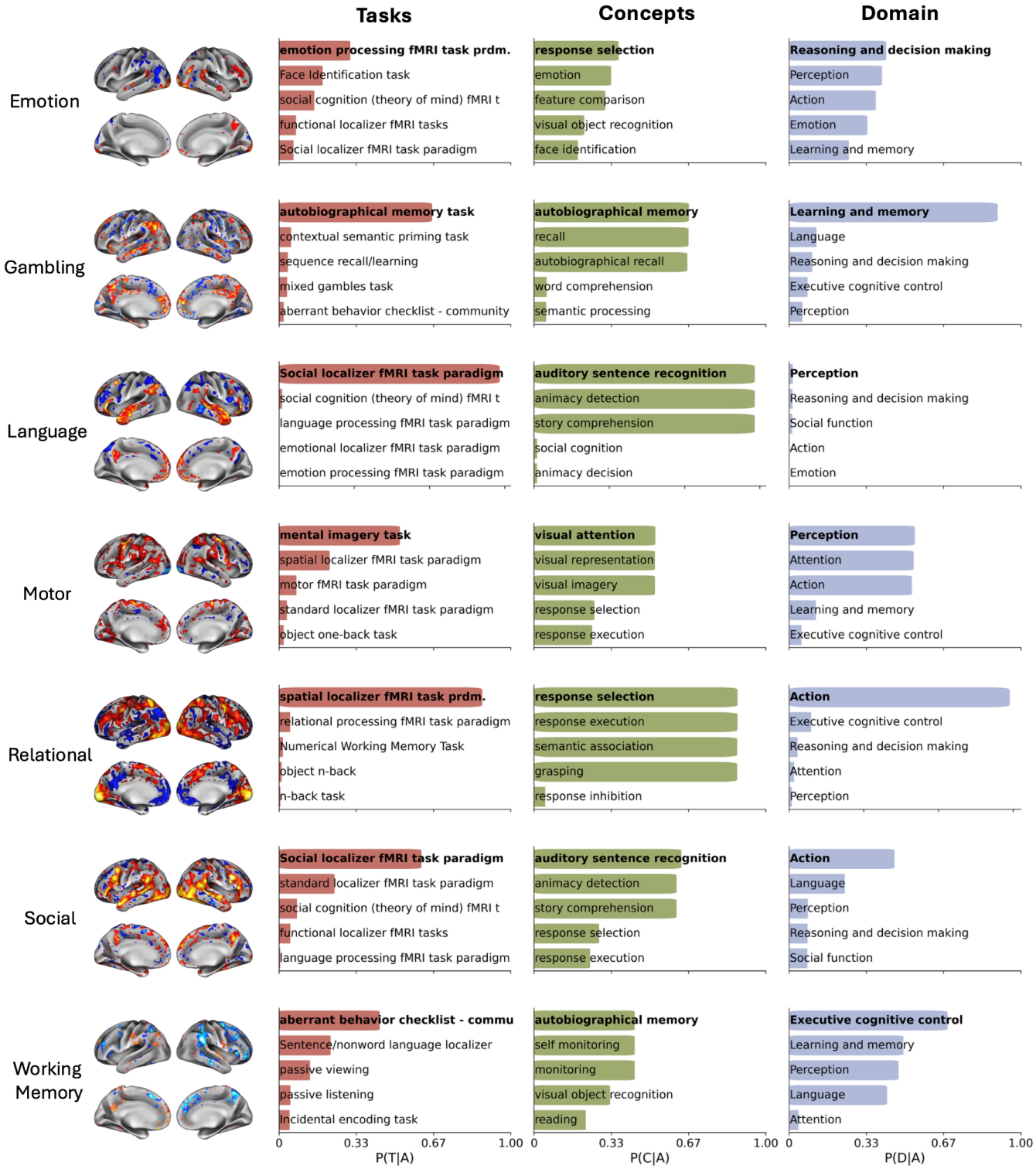


Figure S5. NiCLIP for Predicting Tasks, Concepts, and Domains from Brain Activation Patterns of Subject-Level Maps. This analysis mirrors Fig. 2 but is based on individual HCP subject-level contrast maps rather than group-level averages. For each of the seven HCP task domains (Emotion, Gambling, Language, Motor, Relational, Social, and Working Memory), a representative subject activation map is

shown alongside the top five NiCLIP predictions for task (P(T | A)), concept (P(C | A)), and domain (P(D | A)). Task labels correspond to Cognitive Atlas task names displayed for readability; the semantic embeddings used for decoding incorporate both task names and task definitions. Relative to the group-level results, subject-level predictions are noisier and less consistently aligned with the expected HCP domain, illustrating the added variability associated with individual-level decoding.

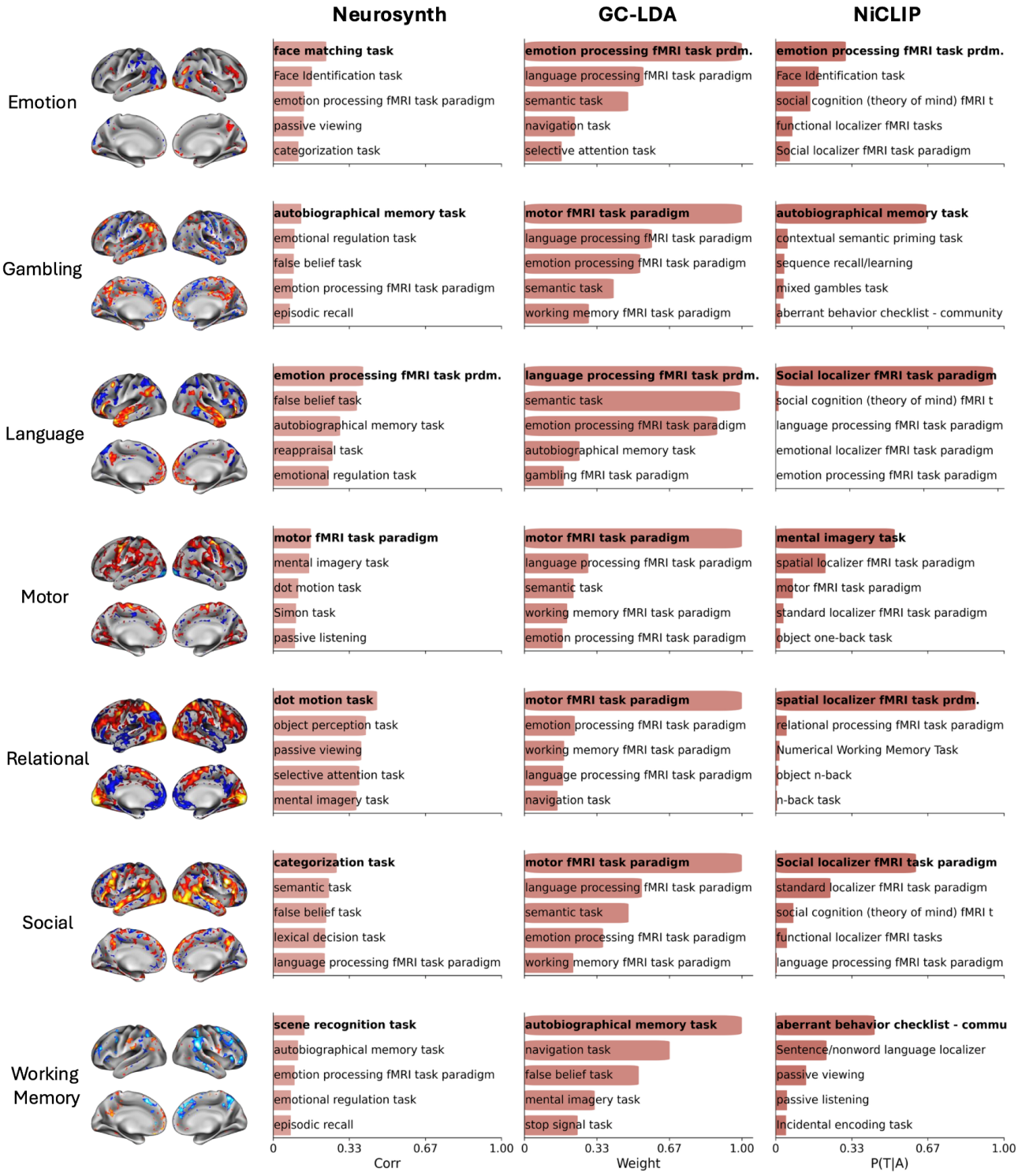


Figure S6. <u>Comparisons Between NiCLIP and Baseline Models' Task Predictions with Subject-Level Maps</u>. This analysis mirrors Fig. S2 but is based on individual HCP subject-level contrast maps rather than group-level averages. For each of the seven HCP task domains (Emotion, Gambling, Language, Motor, Relational, Social, and Working Memory), the top five predicted terms are shown for Neurosynth (correlation-based decoding), GC-LDA (topic weight-based decoding), and NiCLIP (P(T|A), the posterior probability of a task given an activation). Bars represent prediction strength, and the top prediction for each method is shown in bold. As with NiCLIP, both baseline decoders show reduced correspondence to the expected HCP domain relative to their group-level performance, reflecting the added variability of subject-level activation patterns.

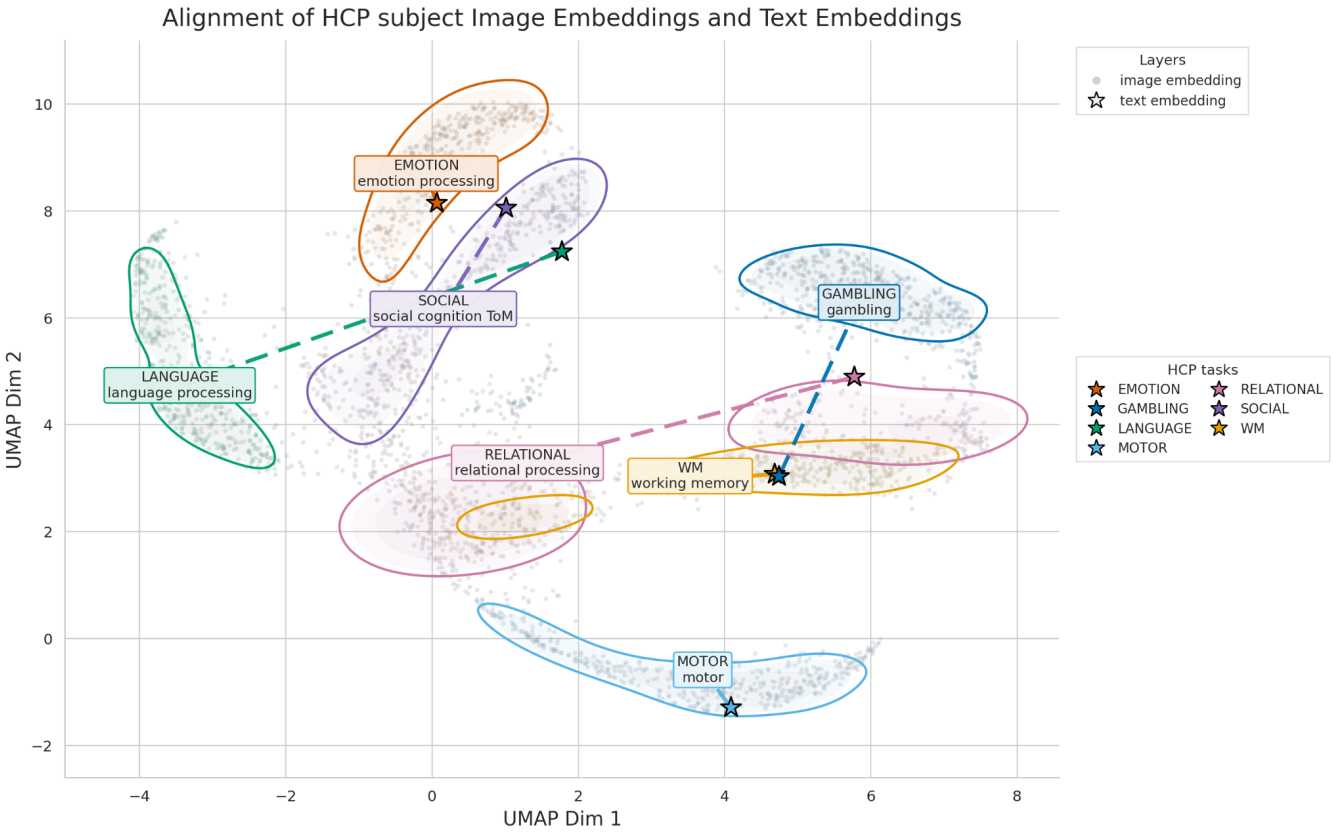


Figure S7. <u>Alignment of HCP Subject Image Embeddings and Text Embeddings</u>. NiCLIP 512-dimensional embeddings were computed for each HCP subject contrast image, then projected to two dimensions using UMAP (n_neighbors=40, min_dist=0.15, cosine metric). The UMAP manifold was fit exclusively on the image embeddings; Cognitive Atlas task text embeddings (generated by BrainGPT-7B-v0.2 from task descriptions) were projected into the same space using the fitted transform, without influencing the manifold geometry. Each grey point represents a single subject contrast image. Colored contours show kernel density estimates of the image distribution for each of the seven HCP task paradigms (i.e., emotion, gambling, language, motor, relations, social, working memory). Colored stars show the corresponding Cognitive Atlas text embedding for each task, with dashed lines connecting each label to the centroid of its image cluster. Close proximity between a star and its cluster centroid indicates strong alignment between the learned brain image representation and the text embedding of the associated task description.

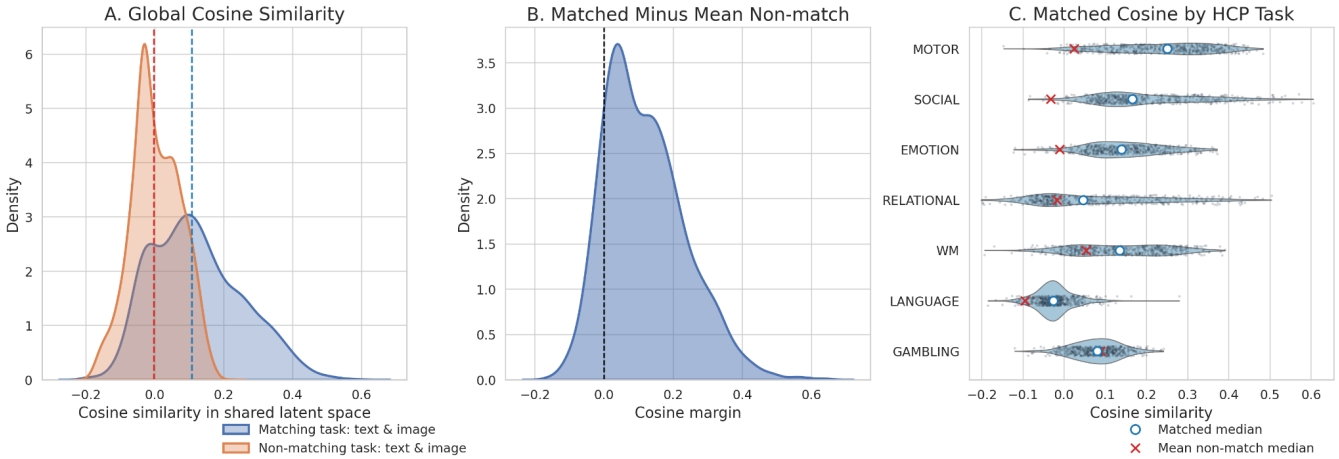


Figure S8. Cosine Similarity Between Matching Image and Text Embedding. For each HCP subject contrast image, BrainCLIP embeddings were computed and compared against BrainGPT-7B-v0.2 text embeddings for all Cognitive Atlas task terms via cosine similarity. (A) Global distributions of cosine similarity for matched text–image pairs (image compared against the embedding of its true task label) versus the mean cosine similarity across all non-matching task texts. Dashed lines indicate medians. (B) Distribution of the cosine margin — the difference between matched cosine similarity and the mean non-match cosine similarity — across all subject images. Values greater than zero indicate that a brain image is on average more similar to its correct task text than to other tasks in the vocabulary. (C) Per-task distributions of matched cosine similarity (violin plots with individual subject points), with matched median (open circle) and mean non-match median (red cross) overlaid. Tasks are ordered by mean cosine margin. MOTOR and SOCIAL show the strongest and most consistent text–image alignment, while GAMBLING shows the weakest separation from the non-match baseline.

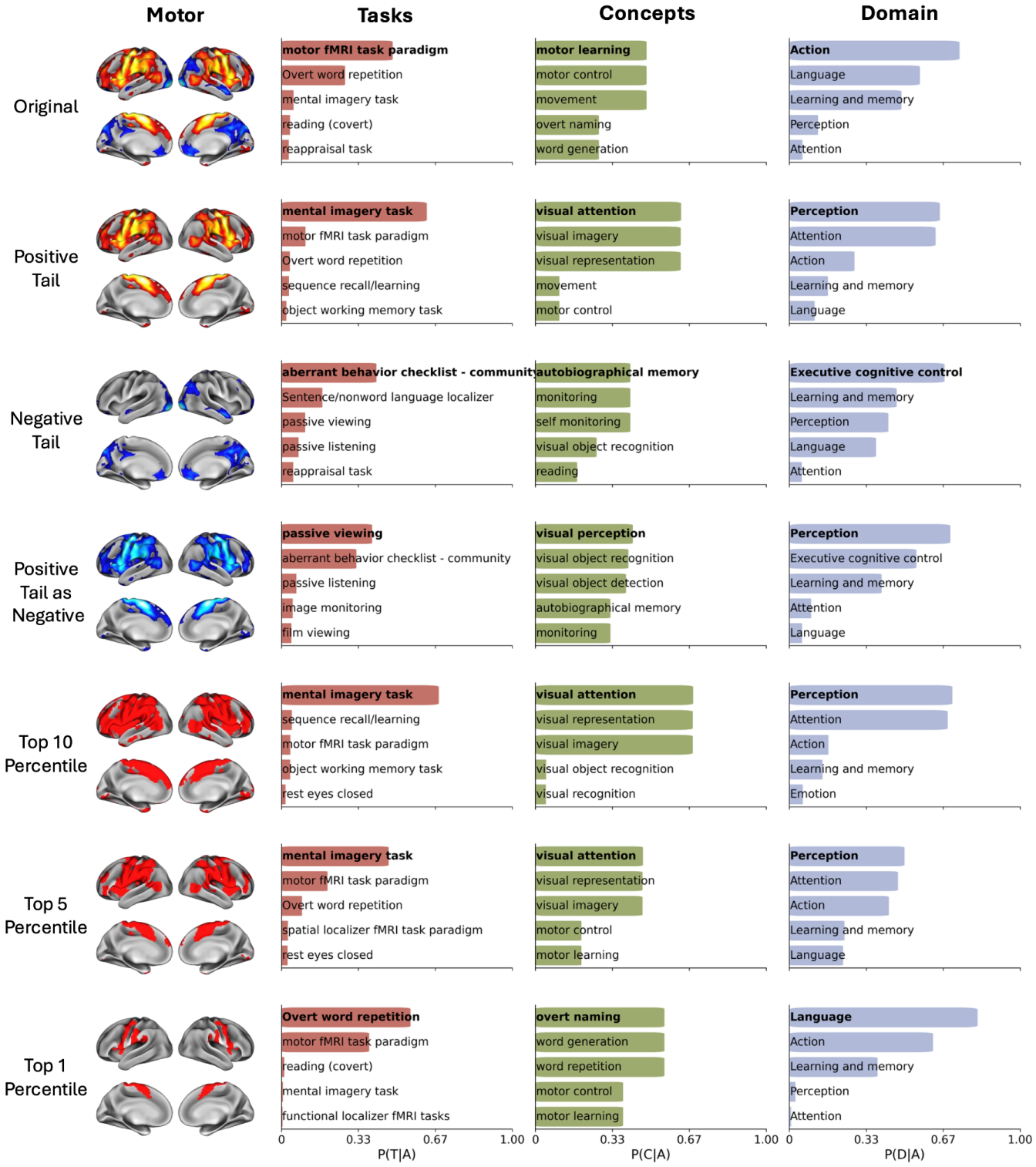


Figure S9. <u>NiCLIP for Predicting Tasks, Concepts, and Domains from Different Variations of the Same Motor Group-Level Maps</u>. We created seven variations of the HCP motor group-level activation map to probe NiCLIP's sensitivity to signal properties: the original map, a map retaining only the positively activated region (Positive Tail), a map retaining only the negatively activated region (Negative

Tail), the positive tail with its sign flipped (Positive Tail as Negative), and three ROI maps built from the top 10, 5, and 1 percentiles of positive activation. For each variation, the resulting brain map is shown alongside the top five NiCLIP predictions for task (P(T | A)), concept (P(C | A)), and domain (P(D | A)), sorted by posterior probability magnitude. The Negative Tail and Positive Tail as Negative maps failed to recover a motor-related task among the top predictions, whereas the Top 10 and Top 5 Percentile maps included the motor fMRI task paradigm among the top three predictions without matching the original map's top two predictions. The Positive Tail map more closely approximated the original prediction pattern, and the Top 1 Percentile map was the only variation that reproduced the original map's exact top predictions, differing only in the ranking of the top two tasks.

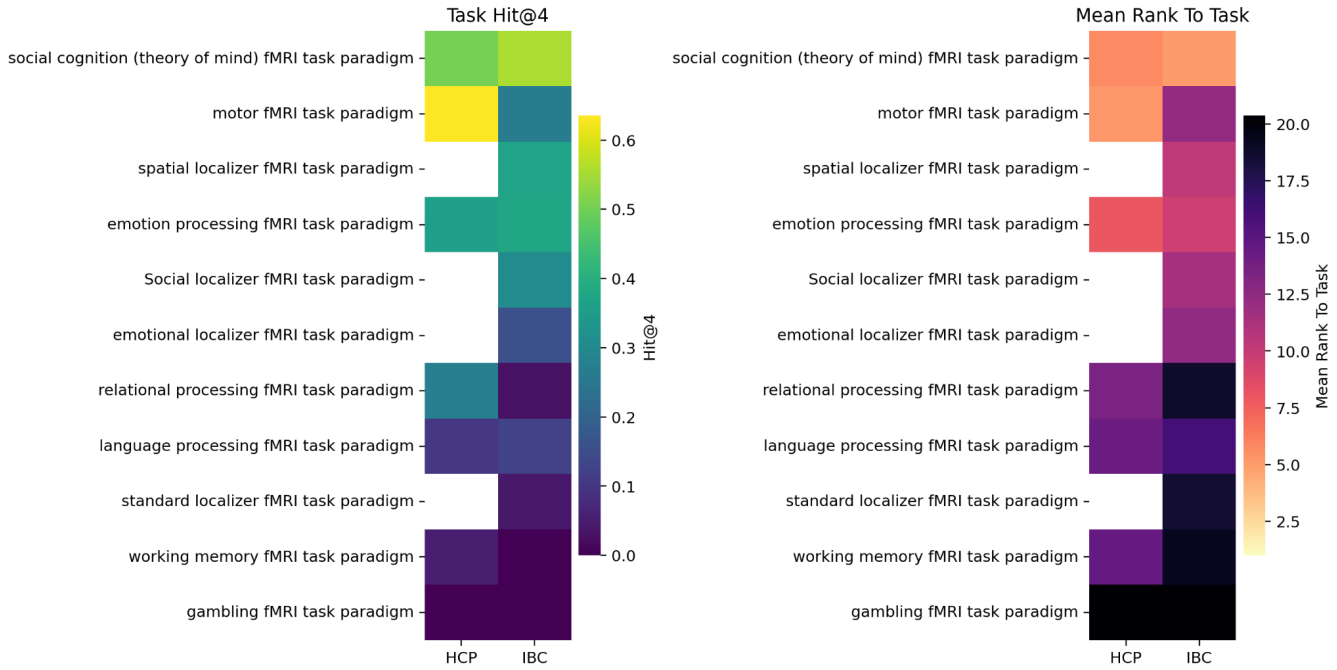


Fig S10. <u>Task-Level Decoding Accuracy Across HCP and IBC Datasets</u>. Decoding was evaluated by retrieving the top-ranked terms from the full Cognitive Atlas vocabulary for each brain image. (Left) Hit@4: proportion of images for which the correct task term appeared in the top 4 retrieved terms. (Right) Mean rank of the correct task term across all images. Tasks are ordered by mean Hit@4 across datasets. Social cognition and motor tasks are decoded most reliably, while gambling shows the weakest performance, consistent across both datasets.

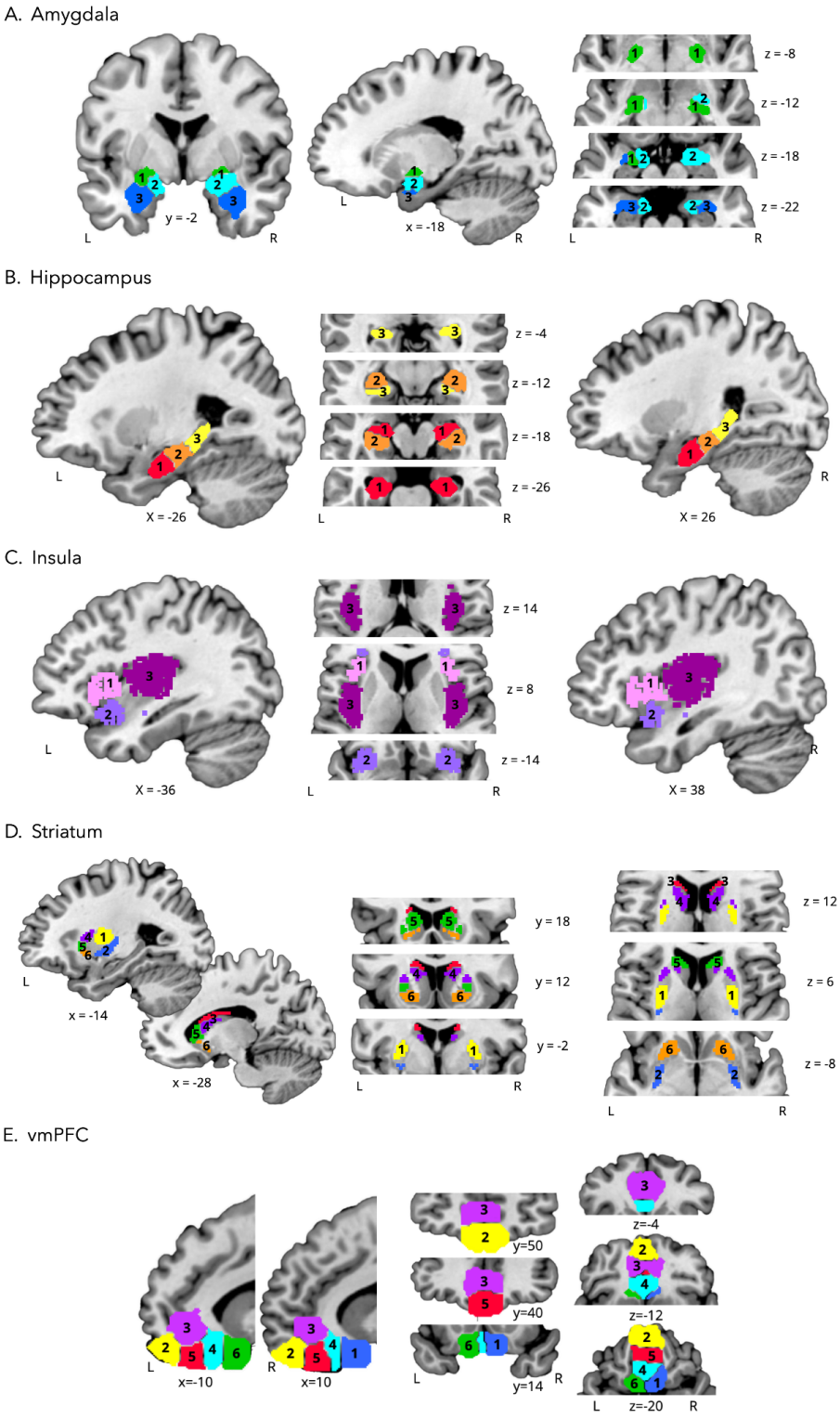


Figure S11. <u>Meta-Analytic Regions of Interest</u>. Existing meta-analytic parcellations for the regions of interest were used. The amygdala ROI (Panel A; 3 seeds per hemisphere) was taken from Bzdok et al. and included the centromedial nuclei group (1-green), superficial nuclei group (2-cyan), and laterobasal nuclei group (3-blue). The hippocampus ROI (Panel B, 3 seeds per hemisphere) was taken from Plachti

et al. and included the anterior (1-red), intermediate (2-orange), and posterior (3-yellow) seeds. The insula ROI (Panel C, 3 seeds per hemisphere) was taken from Chang et al. and included the dorsoanterior (1-pink), ventroanterior (2-light orchid), and posterior (3-dark magenta) seeds. The striatum ROI (Panel D, 6 seeds per hemisphere) was taken from Liu et al. and included the caudal (dorsal part) (1-yellow), caudal (ventral part) (2-blue), dorsal (3-red), dorsolateral (4-purple), rostral (5-green), and ventral (6-orange) seeds. The vmPFC ROI (Panel E, 6 seeds) was taken from Chase et al. and included the right subgenual anterior cingulate cortex (sgACC) (1-blue), anterior medial (2-yellow), central dorsal (3-purple), posterior (4-cyan), central ventral (5-red), and left subgenual sgACC (6-green) seeds. ROIs are available from the ANIMA database (https://anima.fz-juelich.de).