BrainLMM: A Label-Free Framework for Mapping Multi-Semantic Representation in the Human Visual Cortex

Anonymous submission

Figure 1: **Cakes with text patterns.** We employ ResNet50$_{\text{CLIP}}$ as the backbone and select the top 5 images exhibiting the highest activations from a subset of voxels within the VWFA region, where the label with the highest score corresponds to "cake". As illustrated in the images, the majority of these cakes are characterized by distinct textual patterns, highlighting the strong correlation between the VWFA region and textual features.

Figure 2: **Cakes with text patterns.** We employ ViT-32_{CLIP} as the backbone and select the top 5 images exhibiting the highest activations from a subset of voxels within the VWFA region, where the label with the highest score corresponds to "cake". As illustrated in the images, the majority of these cakes are characterized by distinct textual patterns, highlighting the strong correlation between the VWFA region and textual features.

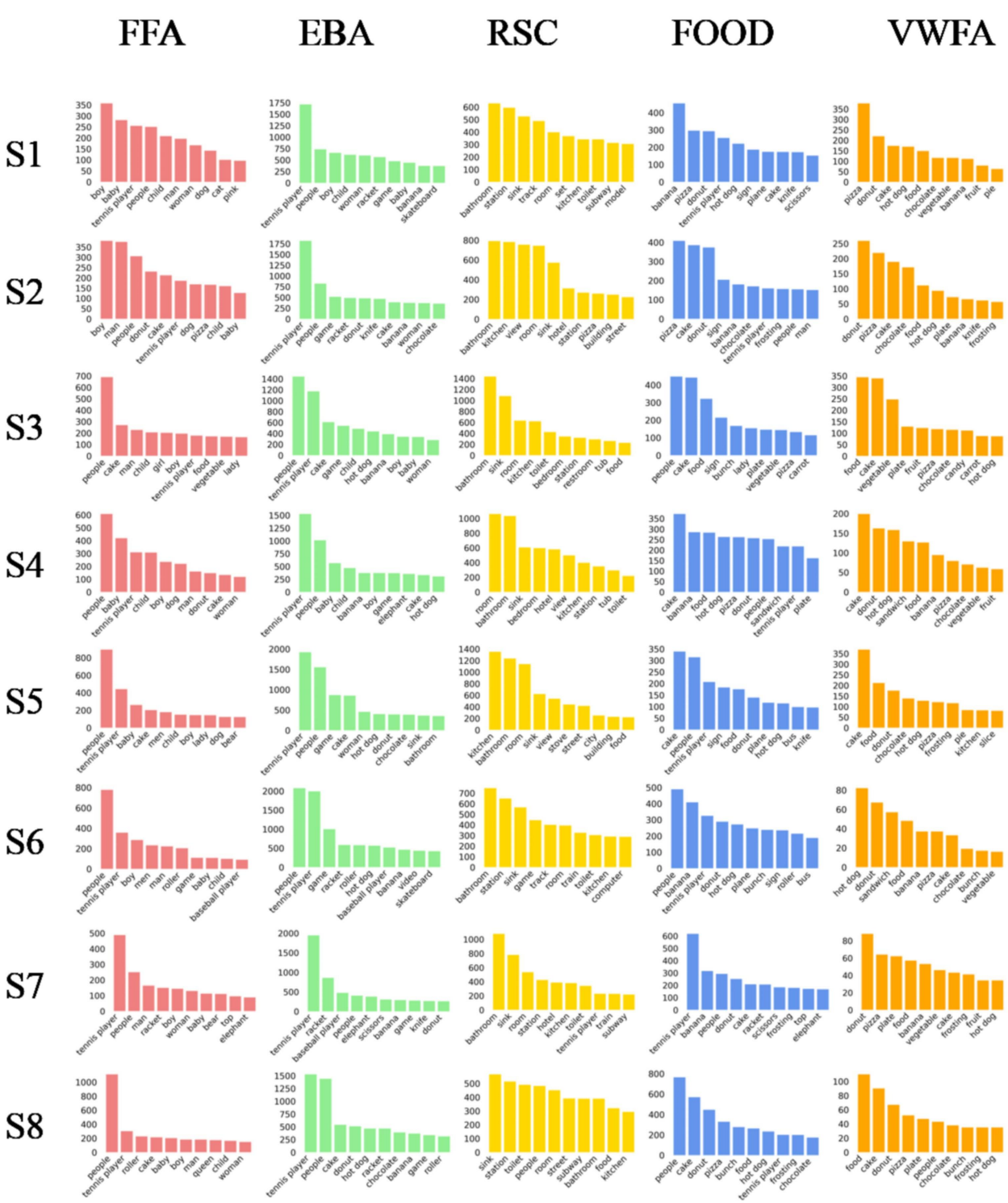

Figure 3: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NSD for $\text{ResNet50}_{\text{CLIP}}$ across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

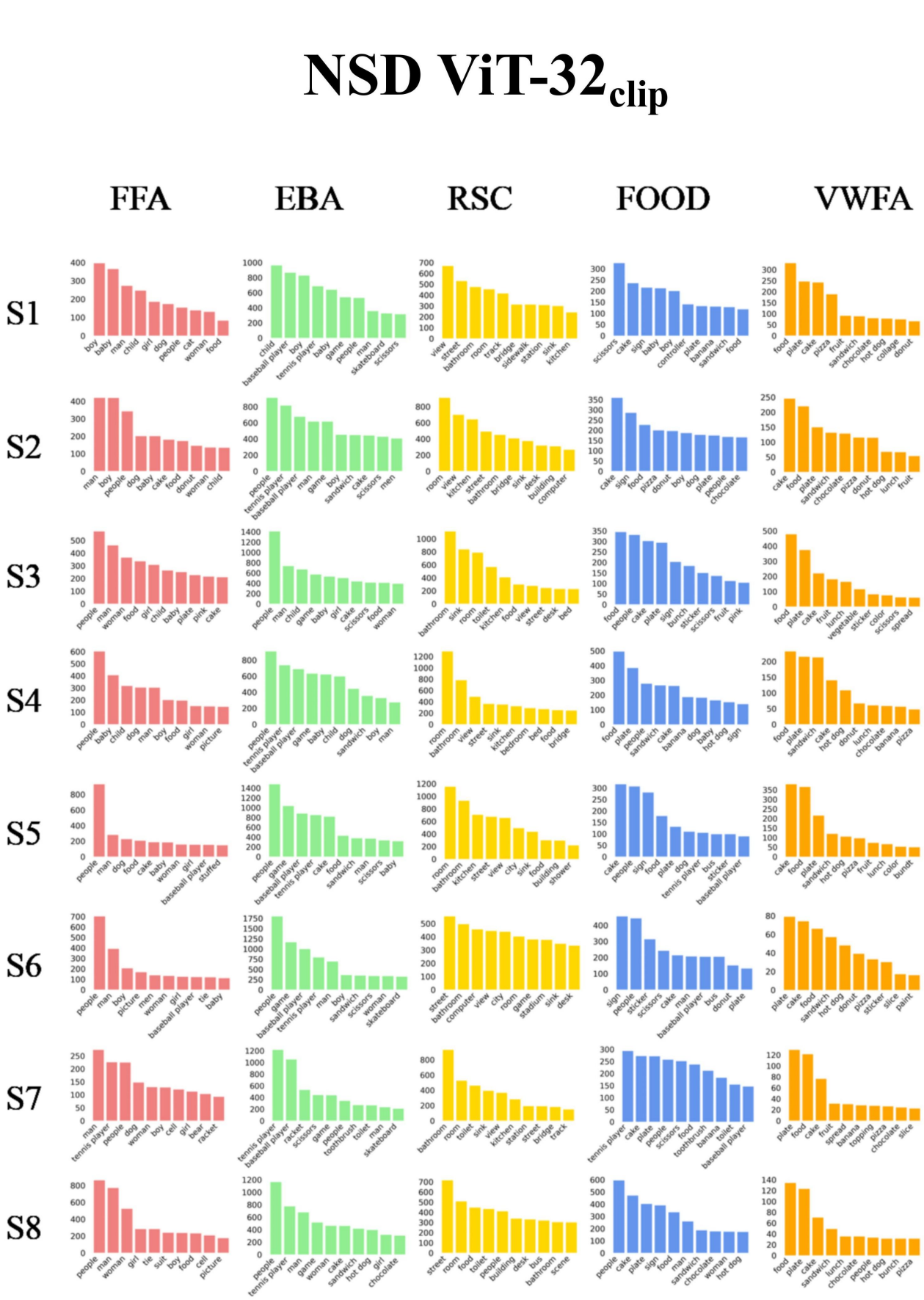

Figure 4: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NSD for ViT-32_{CLIP} across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

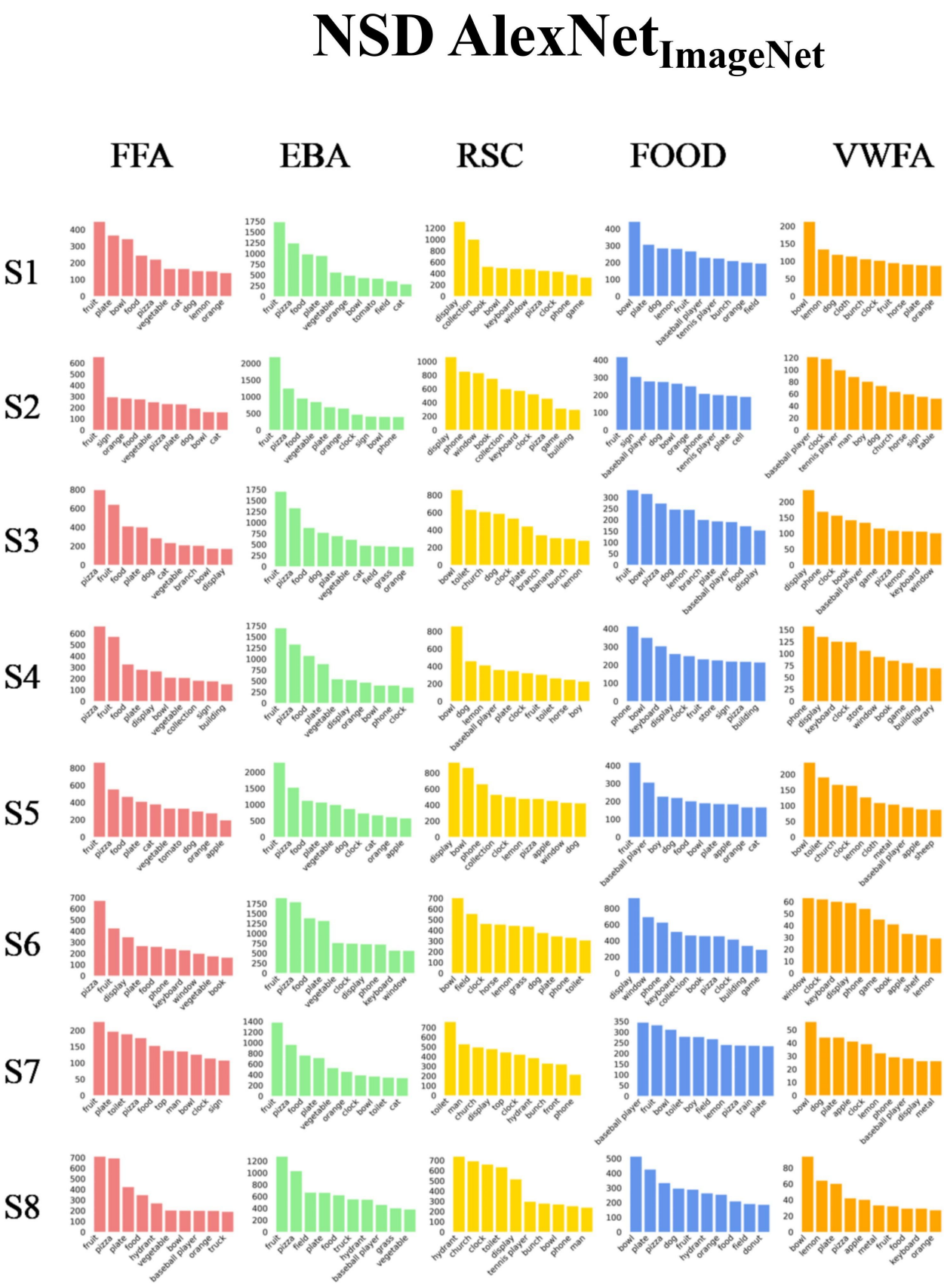

Figure 5: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NSD for AlexNet$_{\text{ImageNet}}$ across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

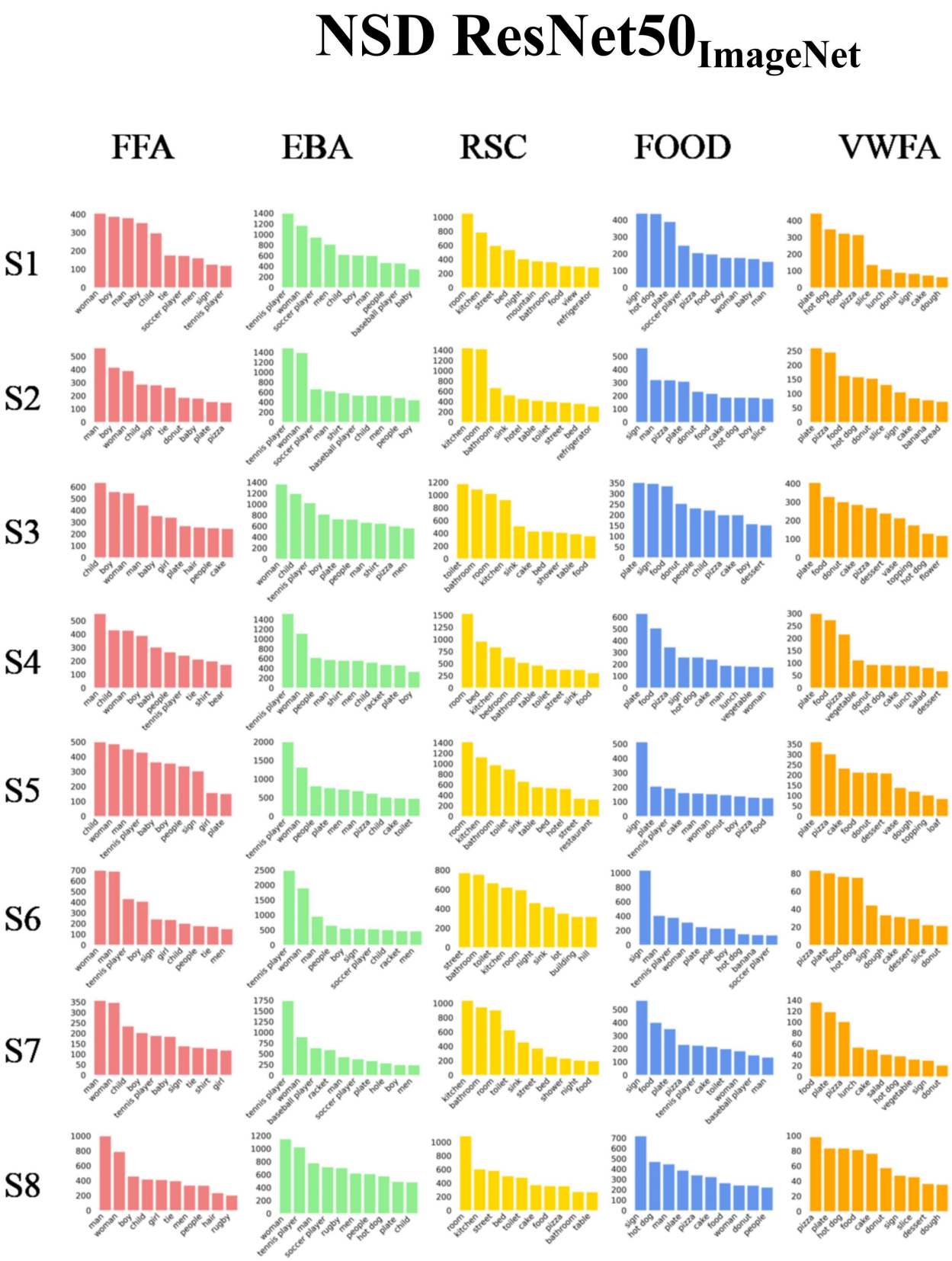

Figure 6: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NSD for $\text{ResNet50}_{\text{ImageNet}}$ across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

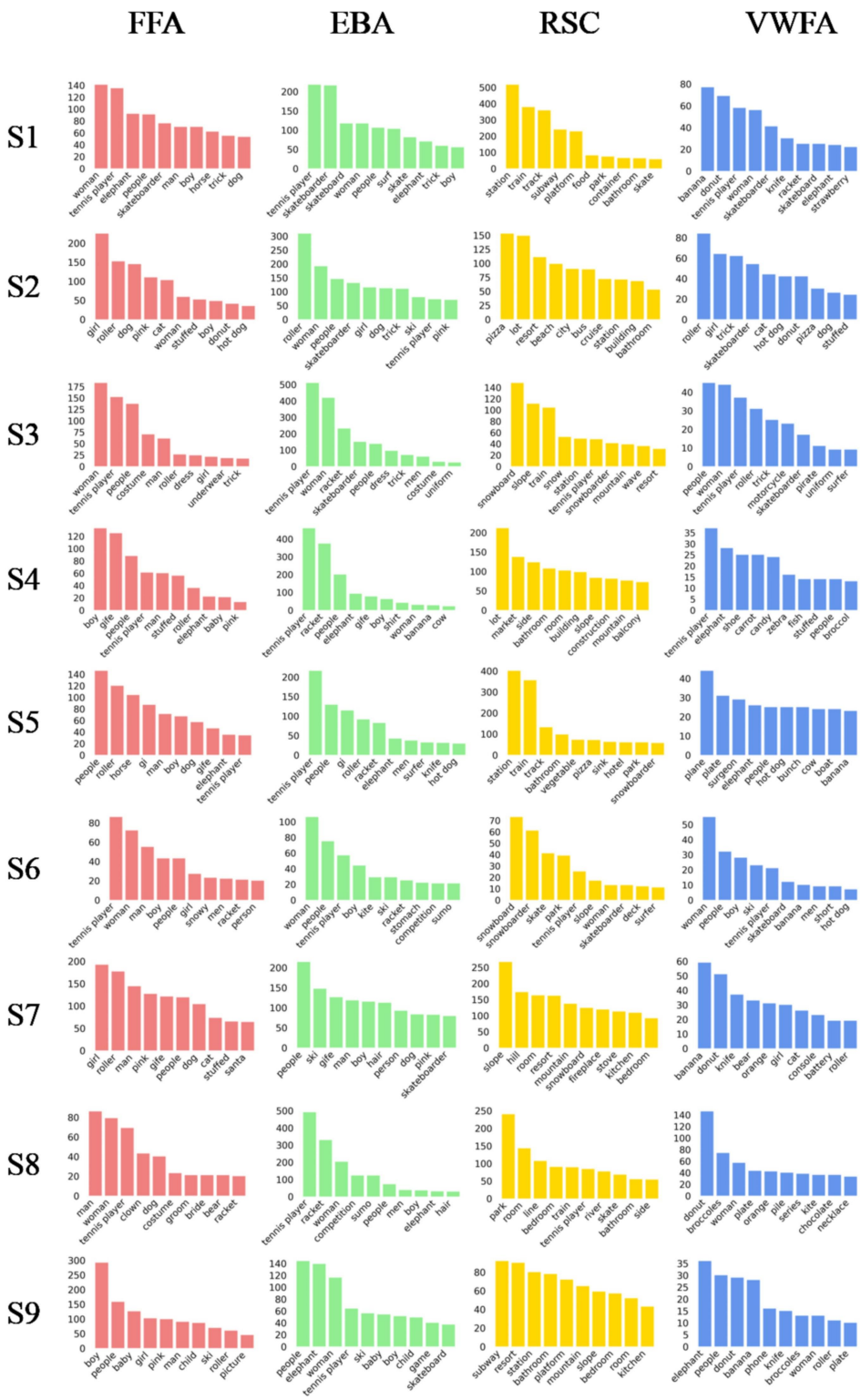

Figure 7: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NOD for ResNet50$_{\text{CLIP}}$ across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

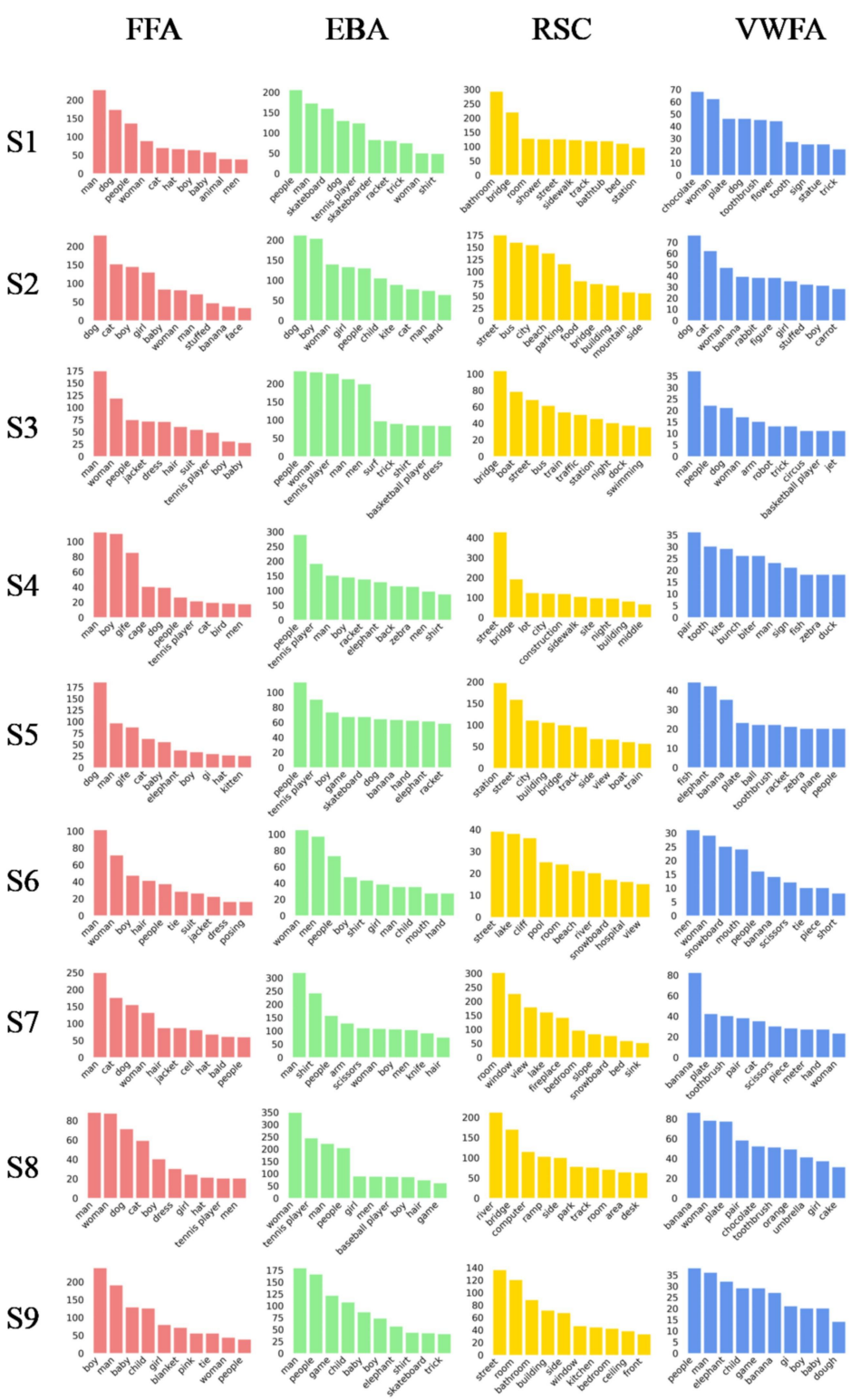

Figure 8: **The BrainLMM enables precise dissection of high-level visual categories in the selective regions.** Single semantic mapping captured via BrainLMM of all subjects in the NOD for ViT-32_{CLIP} across selective regions. The X-axis is the semantic label from the label set, and the Y-axis is the number of voxels mapped to each semantic label.

$\text{ResNet50}_{\text{clip}}$

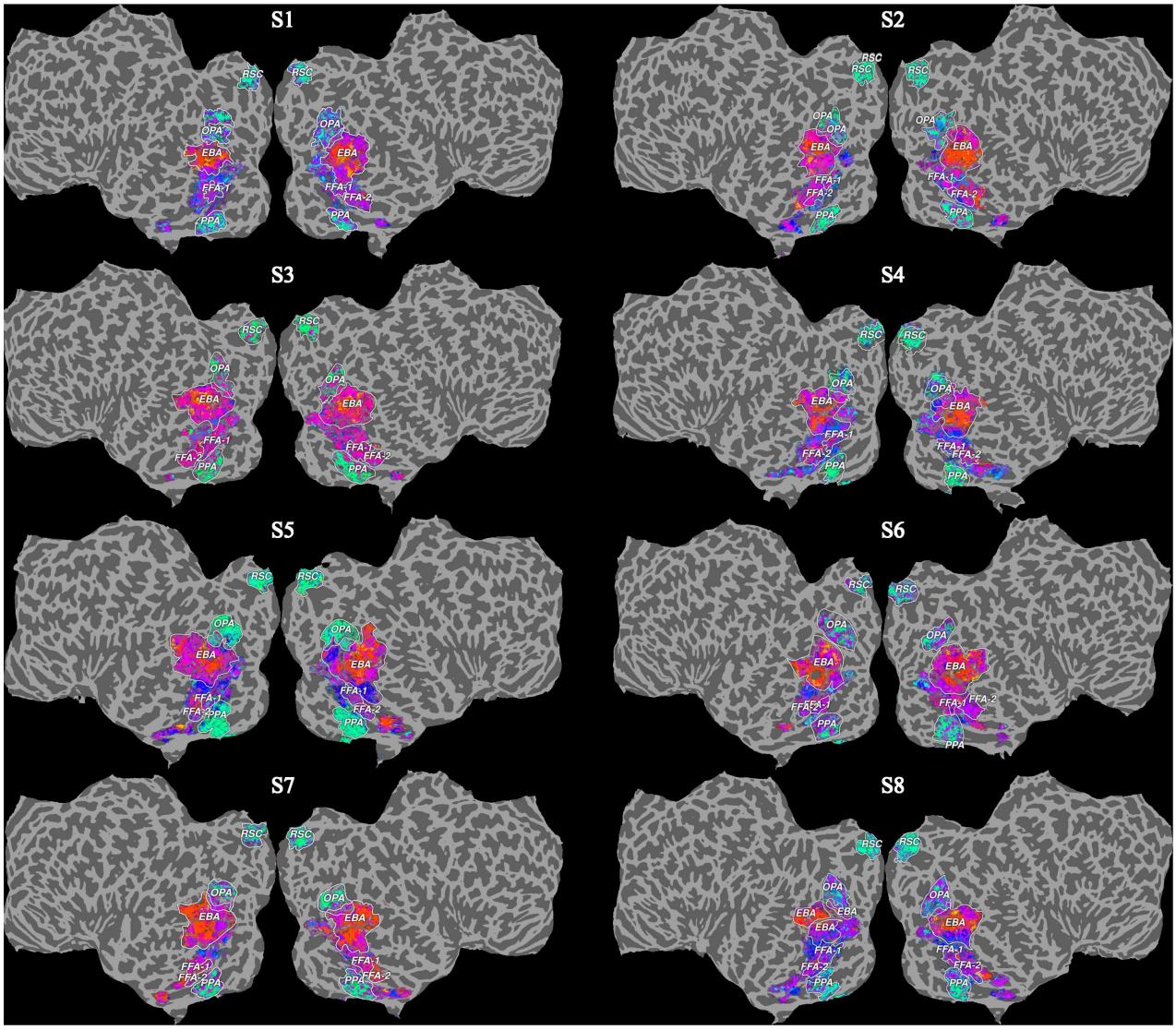

Figure 9: **Voxel-level dissection results for all subjects in the NSD dataset are obtained using $\text{ResNet50}_{\text{CLIP}}$ as the visual backbone.** Image embeddings for all stimulus images are computed through the CLIP image encoder, followed by a UMAP projection into a three-dimensional space. For each voxel, the highest-scoring label from its dissection result is embedded using the CLIP text encoder, then projected into the same UMAP space and normalized to generate RGB color values. We present a flatmap of all subjects with labeled ROIs.

ViT-32$_{clip}$

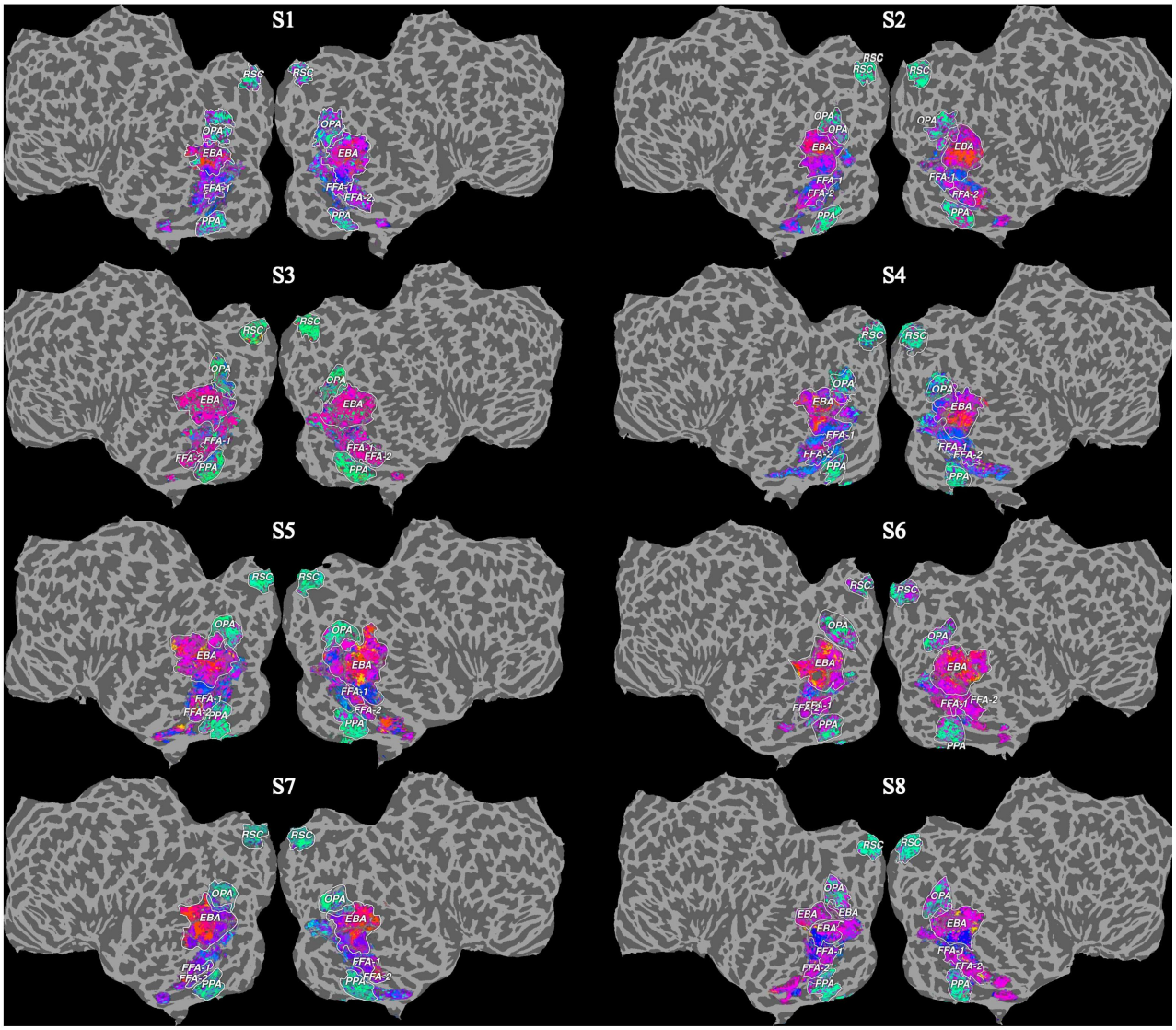

Figure 10: **Voxel-level dissection results for all subjects in the NSD dataset are obtained using ViT-32$_{CLIP}$ as the visual backbone.** Image embeddings for all stimulus images are computed through the CLIP image encoder, followed by a UMAP projection into a three-dimensional space. For each voxel, the highest-scoring label from its dissection result is embedded using the CLIP text encoder, then projected into the same UMAP space and normalized to generate RGB color values. We present a flatmap of all subjects with labeled ROIs.

$\text{ResNet50}_{\text{ImageNet}}$

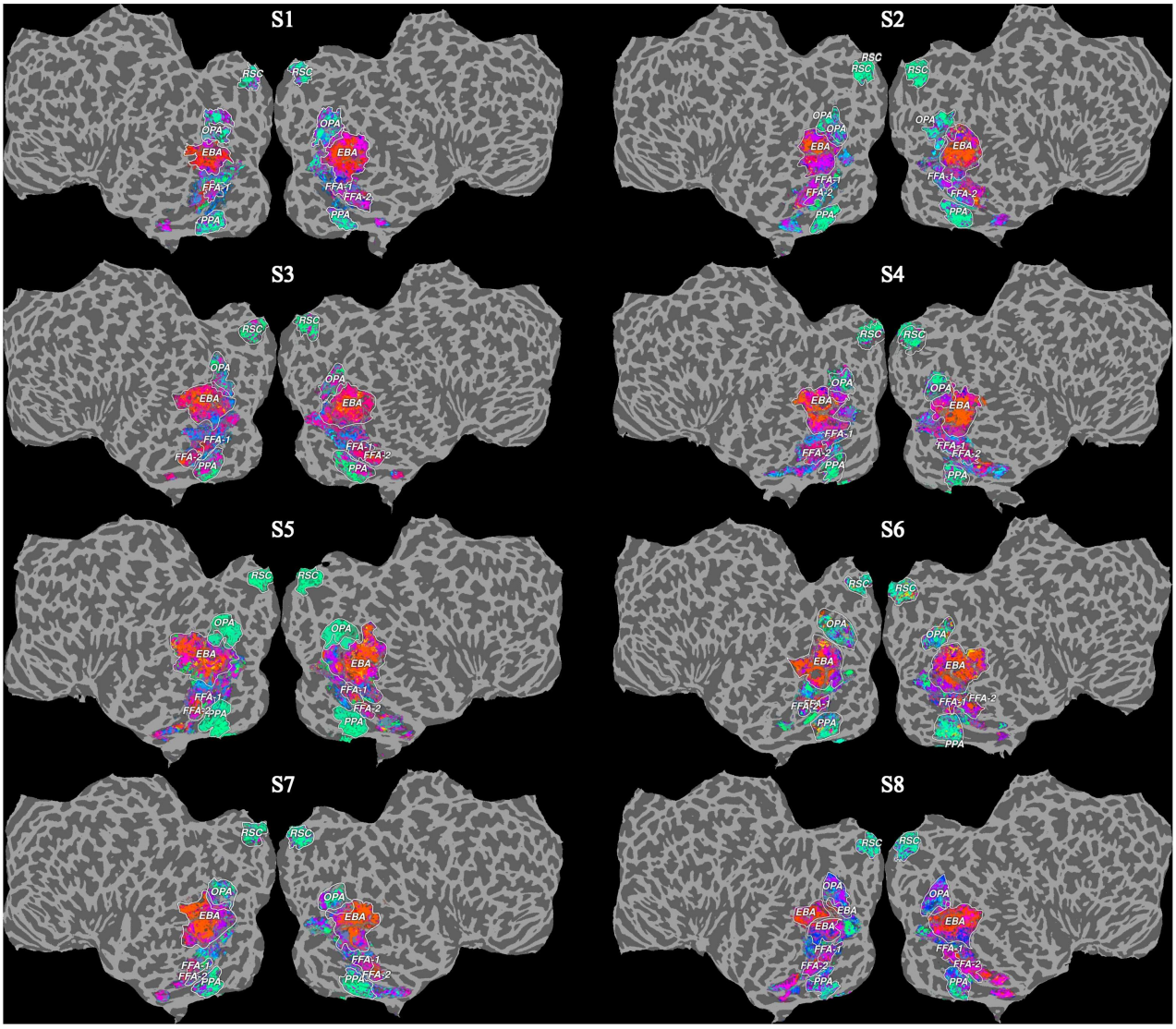

Figure 11: **Voxel-level dissection results for all subjects in the NSD dataset are obtained using $\text{ResNet50}_{\text{ImageNet}}$ as the visual backbone.** Image embeddings for all stimulus images are computed through the CLIP image encoder, followed by a UMAP projection into a three-dimensional space. For each voxel, the highest-scoring label from its dissection result is embedded using the CLIP text encoder, then projected into the same UMAP space and normalized to generate RGB color values. We present a flatmap of all subjects with labeled ROIs.

$\text{AlexNet}_{\text{ImageNet}}$

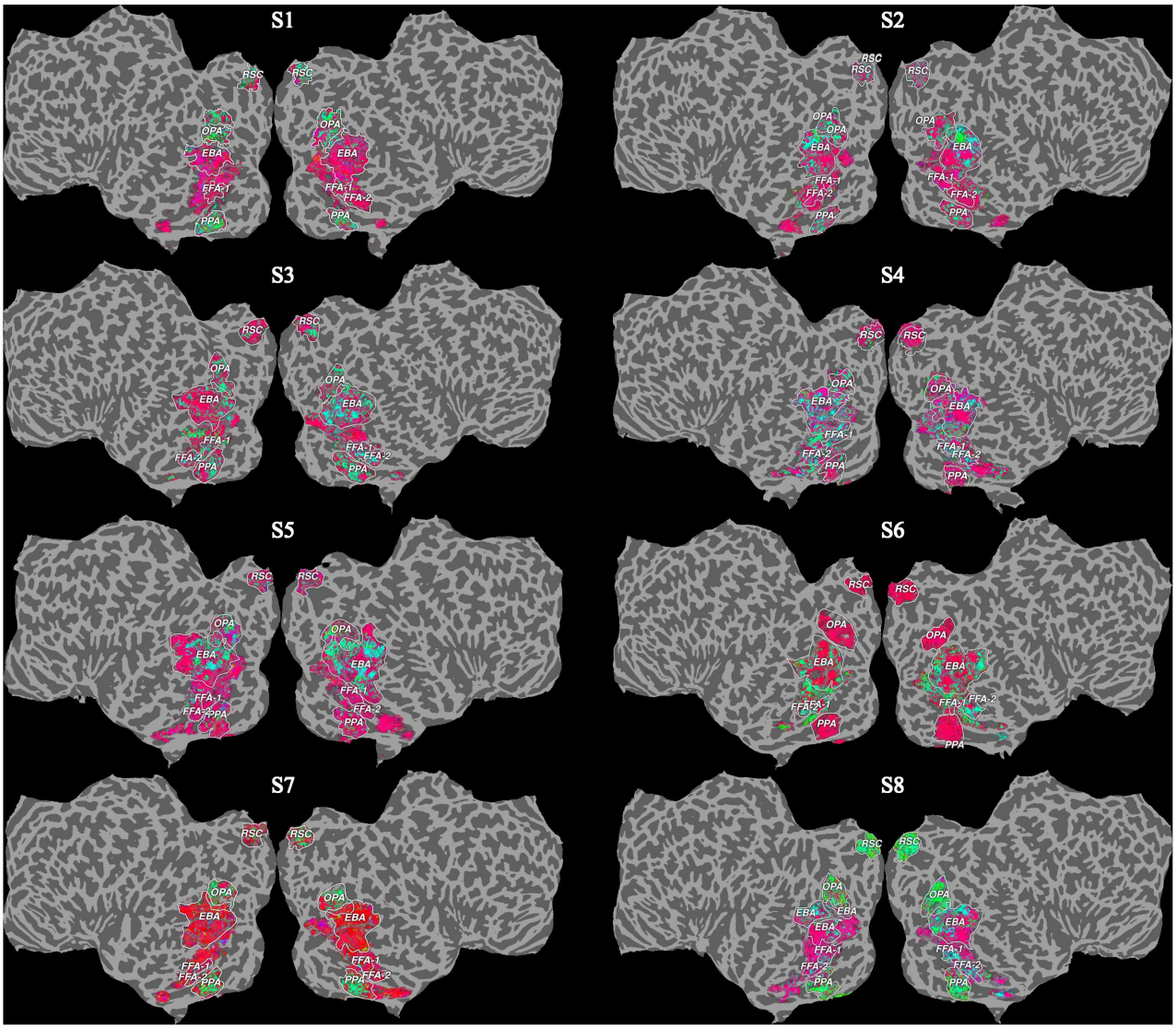

Figure 12: **Voxel-level dissection results for all subjects in the NSD dataset are obtained using $\text{AlexNet}_{\text{ImageNet}}$ as the visual backbone.** Image embeddings for all stimulus images are computed through the CLIP image encoder, followed by a UMAP projection into a three-dimensional space. For each voxel, the highest-scoring label from its dissection result is embedded using the CLIP text encoder, then projected into the same UMAP space and normalized to generate RGB color values. We present a flatmap of all subjects with labeled ROIs.

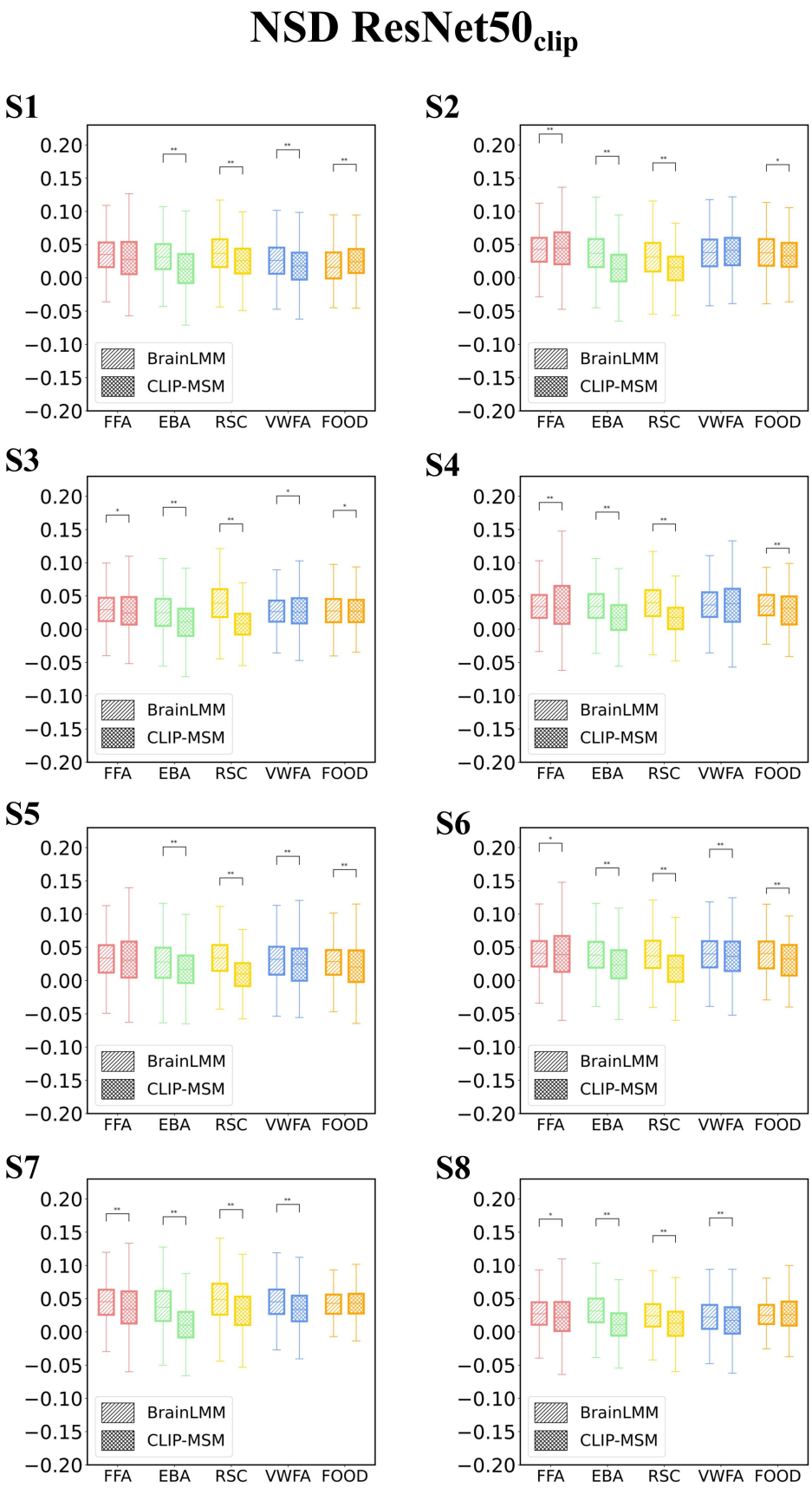

Figure 13: **Performance evaluation of BrainLMM and CLIP-MSM for semantic map capture using the NSD dataset.** We evaluate the performance of both methods by mapping the label of each voxel to its corresponding text embedding using the text encoder. ResNet50$_{\text{CLIP}}$ utilizes its respective text encoders, in line with the original CLIP implementation. A higher cosine similarity between the text embeddings and voxel-wise weights signifies a stronger semantic alignment with the selective regions. As demonstrated, BrainLMM provides a more precise semantic mapping compared to CLIP-MSM ($^{*}P < 0.05$, $^{**}P < 0.001$, paired t-test).

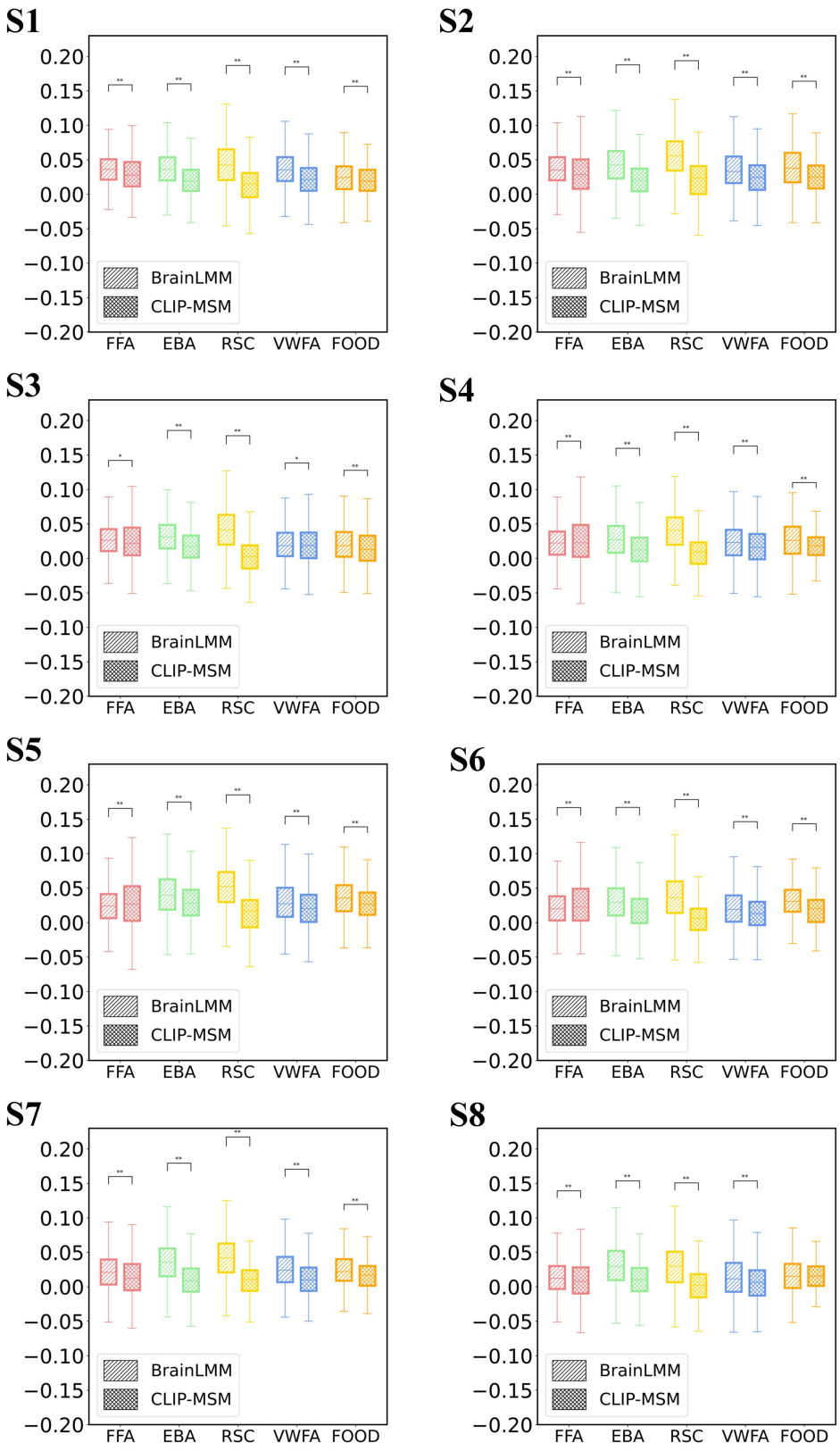

Figure 14: **Performance evaluation of BrainLMM and CLIP-MSM for semantic map capture using the NSD dataset.** We evaluate the performance of both methods by mapping the label of each voxel to its corresponding text embedding using the text encoder. ViT-32_{CLIP} utilizes its respective text encoders, in line with the original CLIP implementation. A higher cosine similarity between the text embeddings and voxel-wise weights signifies a stronger semantic alignment with the selective regions. As demonstrated, BrainLMM provides a more precise semantic mapping compared to CLIP-MSM ($^{*}P < 0.05$, $^{**}P < 0.001$, paired t-test).

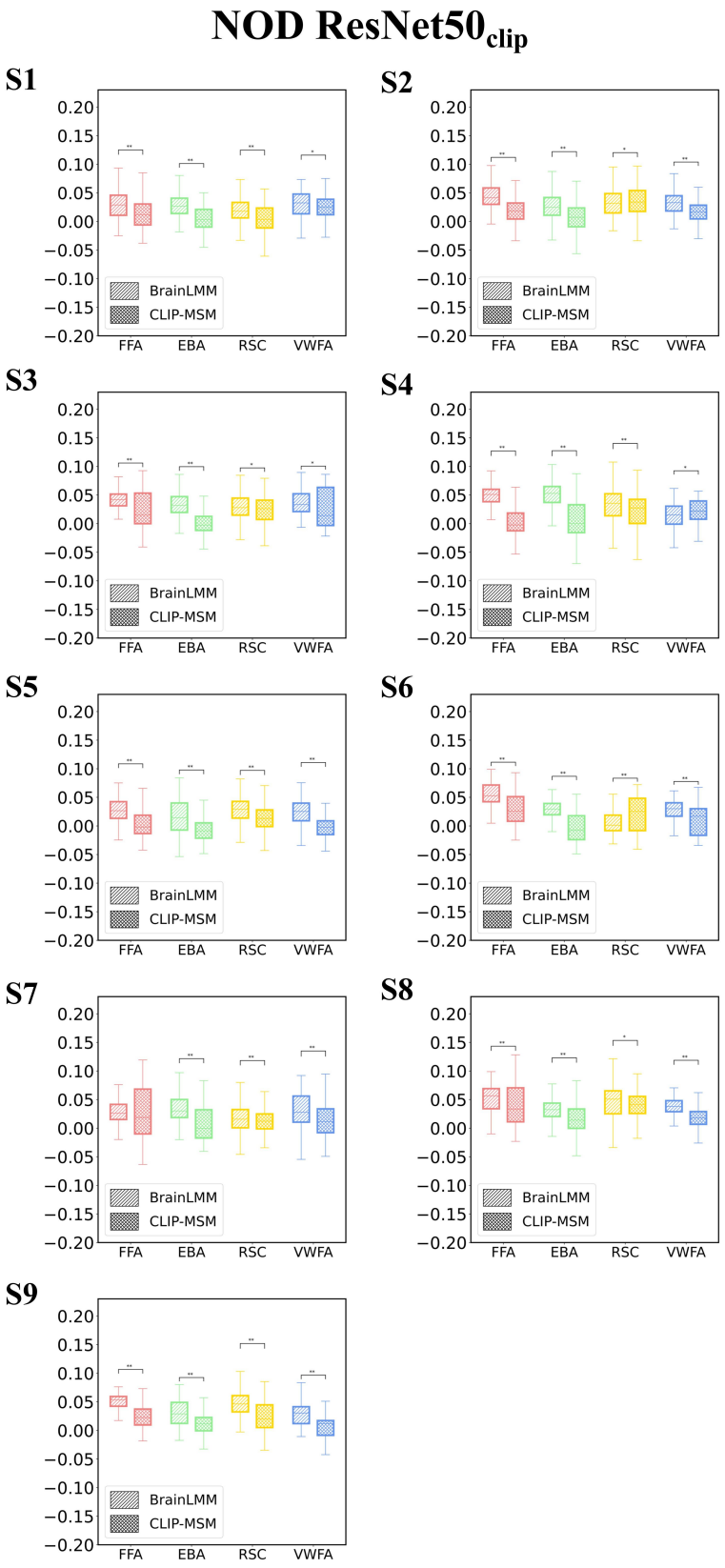

Figure 15: **Performance evaluation of BrainLMM and CLIP-MSM for semantic map capture using the NOD dataset.** We evaluate the performance of both methods by mapping the label of each voxel to its corresponding text embedding using the text encoder. ResNet50$_{\text{CLIP}}$ utilizes its respective text encoders, in line with the original CLIP implementation. A higher cosine similarity between the text embeddings and voxel-wise weights signifies a stronger semantic alignment with the selective regions. As demonstrated, BrainLMM provides a more precise semantic mapping compared to CLIP-MSM ($^{*}P < 0.05$, $^{**}P < 0.001$, paired t-test).

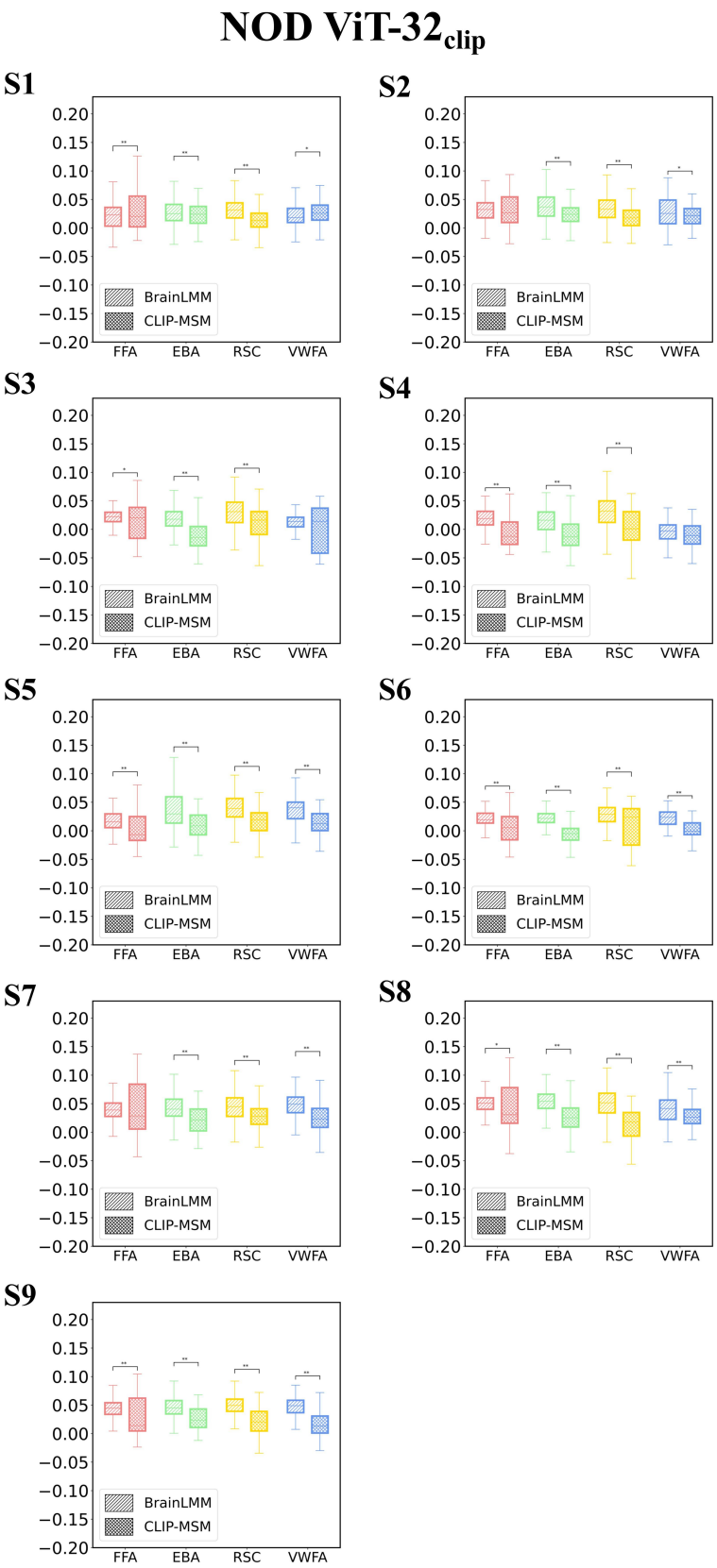

Figure 16: **Performance evaluation of BrainLMM and CLIP-MSM for semantic map capture using the NOD dataset.** We evaluate the performance of both methods by mapping the label of each voxel to its corresponding text embedding using the text encoder. ViT-32_{CLIP} utilizes its respective text encoders, in line with the original CLIP implementation. A higher cosine similarity between the text embeddings and voxel-wise weights signifies a stronger semantic alignment with the selective regions. As demonstrated, BrainLMM provides a more precise semantic mapping compared to CLIP-MSM ($^{*}P < 0.05$, $^{**}P < 0.001$, paired *t*-test).

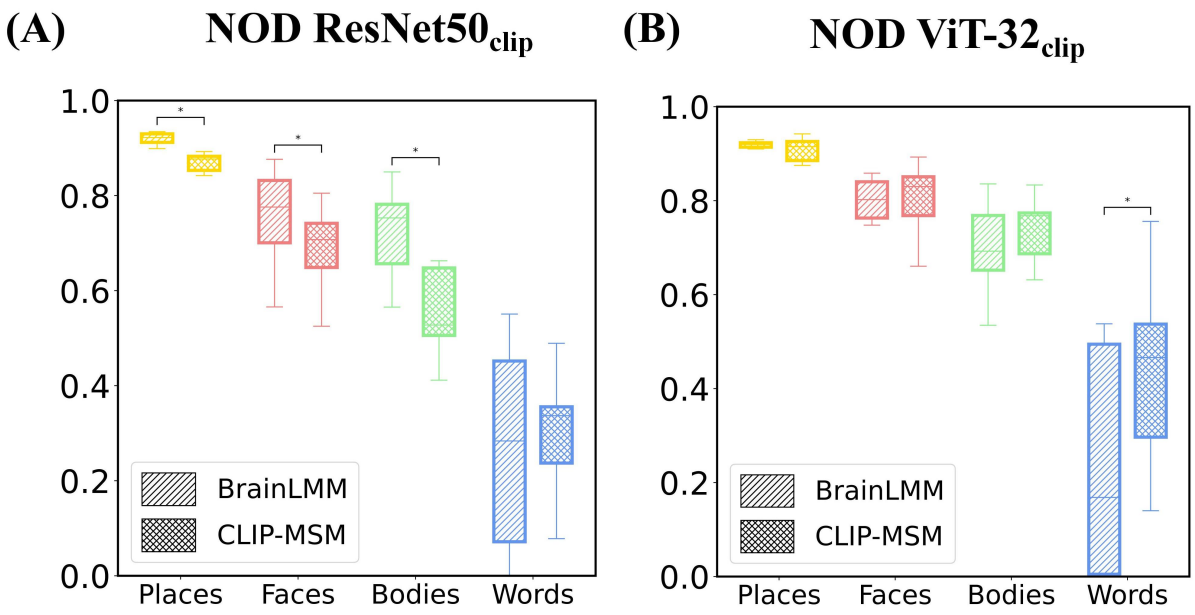

Figure 17: **Comparison of multi-semantic mapping between BrainLMM and CLIP-MSM.** Pearson correlation coefficients quantifying the alignment between multi-semantic mapping derived from both BrainLMM and CLIP-MSM and the ground-truth voxel responses across selective cortical regions. **(A)** Using $\text{ResNet50}_{\text{CLIP}}$ as the visual encoding backbone, BrainLMM achieves significantly higher prediction accuracy than CLIP-MSM for visual responses associated with places, food, faces, bodies, and words. **(B)** Using $\text{ViT-32}_{\text{CLIP}}$ as the visual encoding backbone, BrainLMM also outperforms CLIP-MSM in predicting responses for places, food, bodies, and words ($^{*}P < 0.05$, $^{**}P < 0.001$, paired *t*-test).