A Technical Appendices and Supplementary Material

Sections

A.1 Model Architecture

Our BrainCoDec consists of three main components:

Voxel context token input projection. For each in-context voxel, we concatenate its response function parameter ω_k and measured neural activation β_k into a context token. We repeat this stage across voxels of interest across the brain for a single novel stimulus. A single-layer residual MLP blocks first projects this concatenated voxel context token. The residual MLP applies LayerNorm, LeakyReLU, dropout, and two linear layers with a skip connection.

Contextual decoder transformer. We employ a transformer encoder with 8 self-attention layers to perform aggregated encoder inversion across all voxel tokens and register tokens, allowing the model to infer the stimulus from encoder weights and voxel responses. Each block uses a pre-normalization architecture, we first apply LayerNorm to the inputs, scale the sequence by $logV$, where V is the number of in-context voxels, and then perform self-attention. The attention output is added back with dropout. Then we apply the second LayerNorm followed by a SwiGLU feed-forward network with residual connection.

Image embedding prediction head. After the transformer, we keep register tokens only, and apply an MLP to the concatenated register tokens. This yields a single predicted image embedding.

We primarily evaluate our model using CLIP, due to its excellent visual brain predictivity [Conwell et al., 2024], and additionally assess variants based on DINOv2 [Oquab et al., 2023] and SigLIP [Zhai et al., 2023]. The CLIP variant (encoding dimension $E = 512$) contains approximately 55.70M parameters, while the DINOv2 ($E = 768$) and SigLIP ($E = 1152$) variants comprise roughly 88.76M and 157.35M parameters, respectively. For all models we utilize the ViT-B variant.

A.2 Implementation Details

Training is implemented in PyTorch on two NVIDIA RTX 4090 GPUs (48GB each). At each training step, we sample a batch of in-context voxel tokens together with their target image-embedding vectors and feed them through BrainCoDec to obtain predicted embeddings. We train the model with a supervised objective that combines a cosine-similarity loss and an InfoNCE loss between predicted and ground-truth embeddings. Dropout is applied in all residual and attention blocks to regularize the model and mitigate overfitting. We optimize BrainCoDec using AdamW with an initial learning rate of 1×10^{-5} and a decoupled weight decay of 1×10^{-2}. In the first pretraining stage, each mini-batch samples a fixed set of 200 in-context voxels. In the second context-extension stage and the third finetuning stage, each mini-batch randomly samples between 200 and 4000 in-context voxels. The learning rate is scheduled with a cosine-annealing scheduler over the total number of training steps, gradually decaying to a minimum of 1×10^{-6}. We use the HuggingFace `Accelerate` library to jointly prepare the model, optimizer, data loaders, and scheduler for (potentially) distributed training. The same training protocol is applied to the CLIP, DINOv2, and SigLIP variants, differing only in the choice of backbone embedding dimension.

In the main paper, we focus on NSD S1/S2/S5/S7, as these are the four subjects that completed scanning from the dataset. We train 15 models total based on three backbones. For each backbone we train five variants (four where a single subject is held out, and one model where we train on all four subjects). Note, all of these models are effectively fine-tuned variants of the model that was trained with synthetic data only. The variants where a single subject is held out is used respectively for testing on S1/S2/S5/S7 from NSD to ensure there is no data contamination. For NSD S3/S4/S6/S8 and BOLD5000, we use the variant trained on all four NSD complete subject.

Our code will be open sourced once the review process is concluded. We thank the reviewers for your understanding.

For this supplemental, we first present the results for the subjects that completed NSD scanning (S1/S2/S5/S7), then we present the subjects that did not (S3/S4/S6/S8). Unless otherwise noted, in all cases the model has not seen data from a particular subject during training.

A.3 Quantitative table for S2-8

Table S.1: **Quantitative comparison on NSD Subjects 1, 2, 5, and 7.**

Model	S1	S2	S5	S7
		% **Top-1 Accuracy** (↑)		
MindEye2	4.11 ± 1.41	3.82 ± 1.10	2.87 ± 1.19	2.51 ± 1.64
TGBD	1.27 ± 0.16	0.56 ± 0.12	0.84 ± 0.16	0.39 ± 0.09
BrainCodec-200	$\mathbf{25.5 \pm 3.02}$	$\mathbf{22.9 \pm 2.98}$	$\mathbf{23.2 \pm 2.63}$	$\mathbf{19.2 \pm 2.42}$
		% **Top-5 Accuracy** (↑)		
MindEye2	12.9 ± 2.55	10.7 ± 3.14	9.58 ± 3.61	6.49 ± 2.87
TGBD	3.89 ± 1.25	2.33 ± 0.91	3.34 ± 0.99	1.41 ± 0.78
BrainCodec-200	$\mathbf{56.6 \pm 3.21}$	$\mathbf{52.4 \pm 4.08}$	$\mathbf{55.8 \pm 2.47}$	$\mathbf{51.2 \pm 3.50}$
		% **Mean Rank** (↓)		
MindEye2	24.70 ± 2.07	25.10 ± 2.40	26.03 ± 3.14	25.63 ± 2.67
TGBD	48.50 ± 2.87	50.87 ± 3.13	47.13 ± 3.20	49.47 ± 2.43
BrainCodec-200	$\mathbf{4.43 \pm 0.47}$	$\mathbf{4.23 \pm 0.33}$	$\mathbf{3.93 \pm 0.27}$	$\mathbf{3.73 \pm 0.30}$

Table S.2: **Quantitative comparison on NSD Subjects 3, 4, 6, and 8.**

Model	S3	S4	S6	S8
		% **Top-1 Accuracy** (↑)		
MindEye2	3.50 ± 1.13	3.19 ± 1.16	2.69 ± 1.42	2.33 ± 1.86
TGBD	0.65 ± 0.13	0.75 ± 0.15	0.61 ± 0.12	0.17 ± 0.05
BrainCodec-200	$\mathbf{19.0 \pm 1.86}$	$\mathbf{16.1 \pm 1.75}$	$\mathbf{20.1 \pm 2.52}$	$\mathbf{14.4 \pm 1.56}$
		% **Top-5 Accuracy** (↑)		
MindEye2	10.33 ± 3.30	9.95 ± 3.45	8.04 ± 3.24	4.95 ± 2.50
TGBD	2.67 ± 0.94	3.00 ± 0.96	2.38 ± 0.89	0.44 ± 0.68
BrainCodec-200	$\mathbf{48.3 \pm 2.34}$	$\mathbf{42.3 \pm 3.01}$	$\mathbf{48.7 \pm 3.00}$	$\mathbf{53.3 \pm 4.02}$
		% **Mean Rank** (↓)		
MindEye2	25.40 ± 2.63	25.73 ± 2.90	25.83 ± 2.90	25.43 ± 2.43
TGBD	49.63 ± 3.17	48.37 ± 3.17	48.30 ± 2.80	50.63 ± 2.07
BrainCodec-200	$\mathbf{4.97 \pm 0.30}$	$\mathbf{5.97 \pm 0.30}$	$\mathbf{4.53 \pm 0.30}$	$\mathbf{3.03 \pm 0.27}$

A.4 Retrieval visualizations for NSD

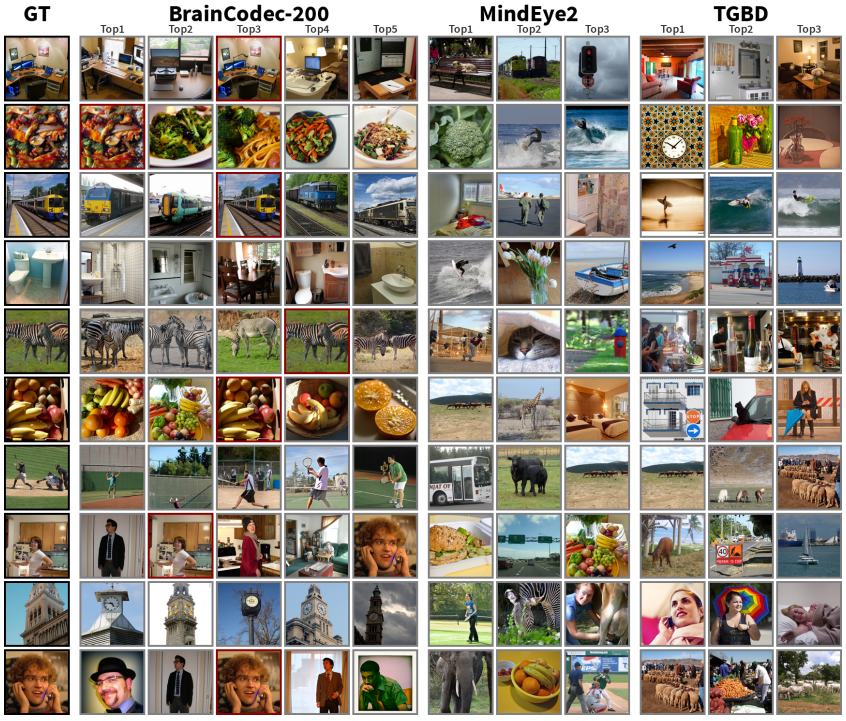

Figure S.1: Image retrieval comparison on an unseen subject (S1).

Figure S.2: Image retrieval comparison on an unseen subject (S2).

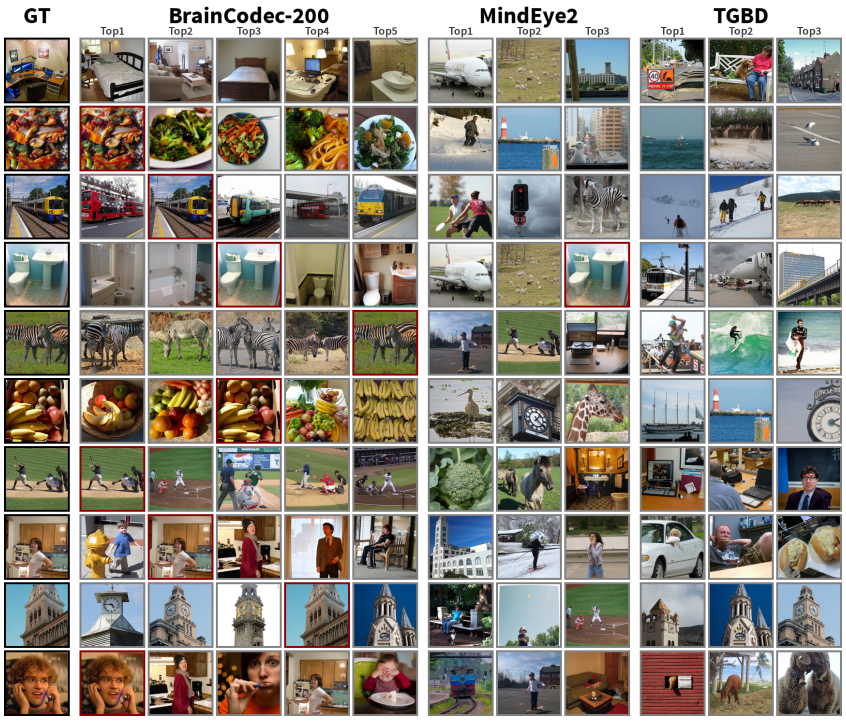

Figure S.3: Image retrieval comparison on an unseen subject (S5).

Figure S.4: Image retrieval comparison on an unseen subject (S7).

Figure S.5: Image retrieval comparison on an unseen subject (S3).

Figure S.6: Image retrieval comparison on an unseen subject (S4).

Figure S.7: Image retrieval comparison on an unseen subject (S6).

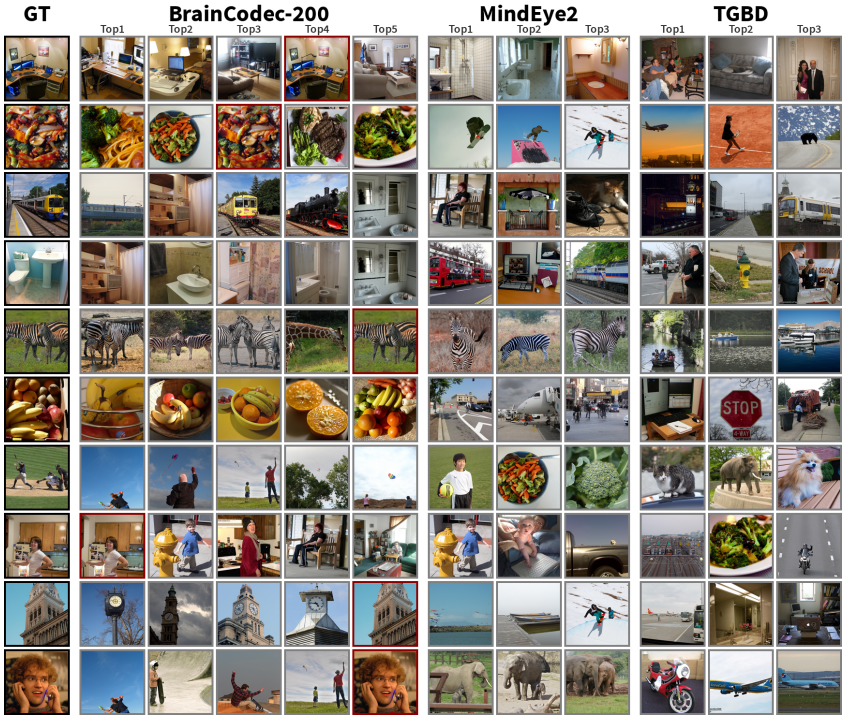

Figure S.8: Image retrieval comparison on an unseen subject (S8).

A.5 Context scaling of other unseen NSD subjects

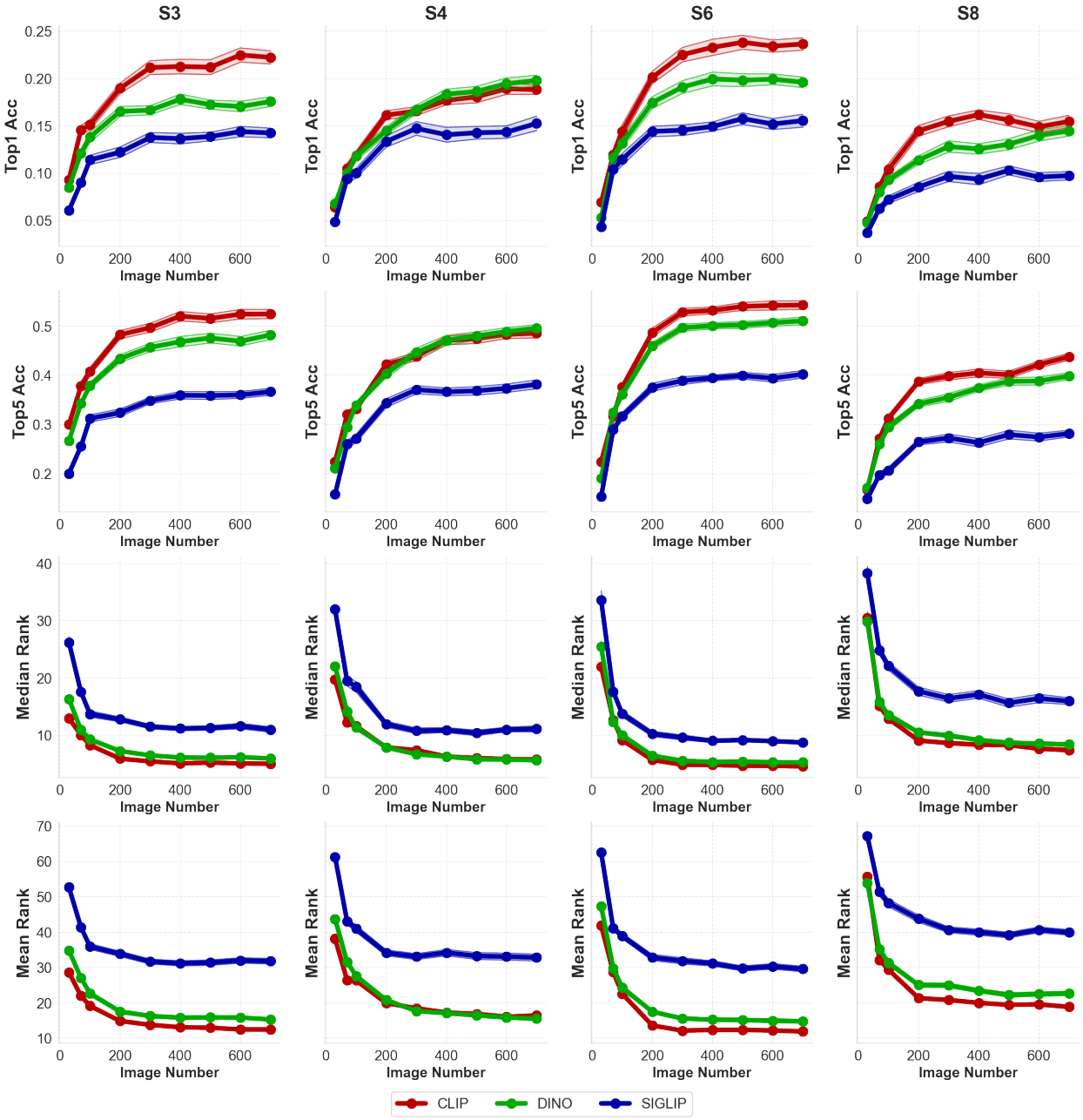

Figure S.9: Image-context scaling of BrainCoDec on NSD subjects 3, 4, 6, and 8.

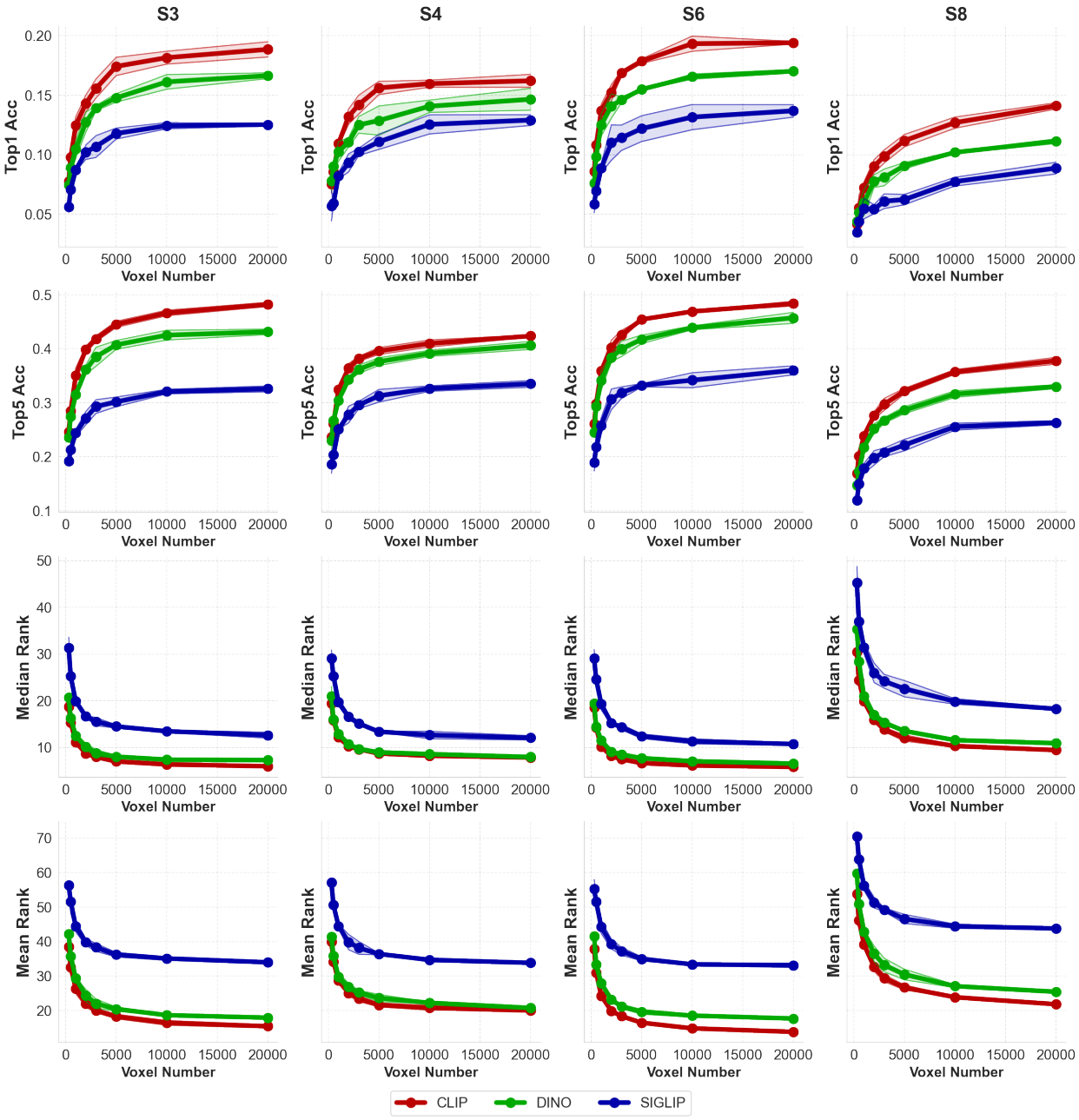

Figure S.10: Voxel-context scaling of BrainCoDec on NSD subjects 3, 4, 6, and 8.

A.6 Context scaling of unseen BOLD500 subjects

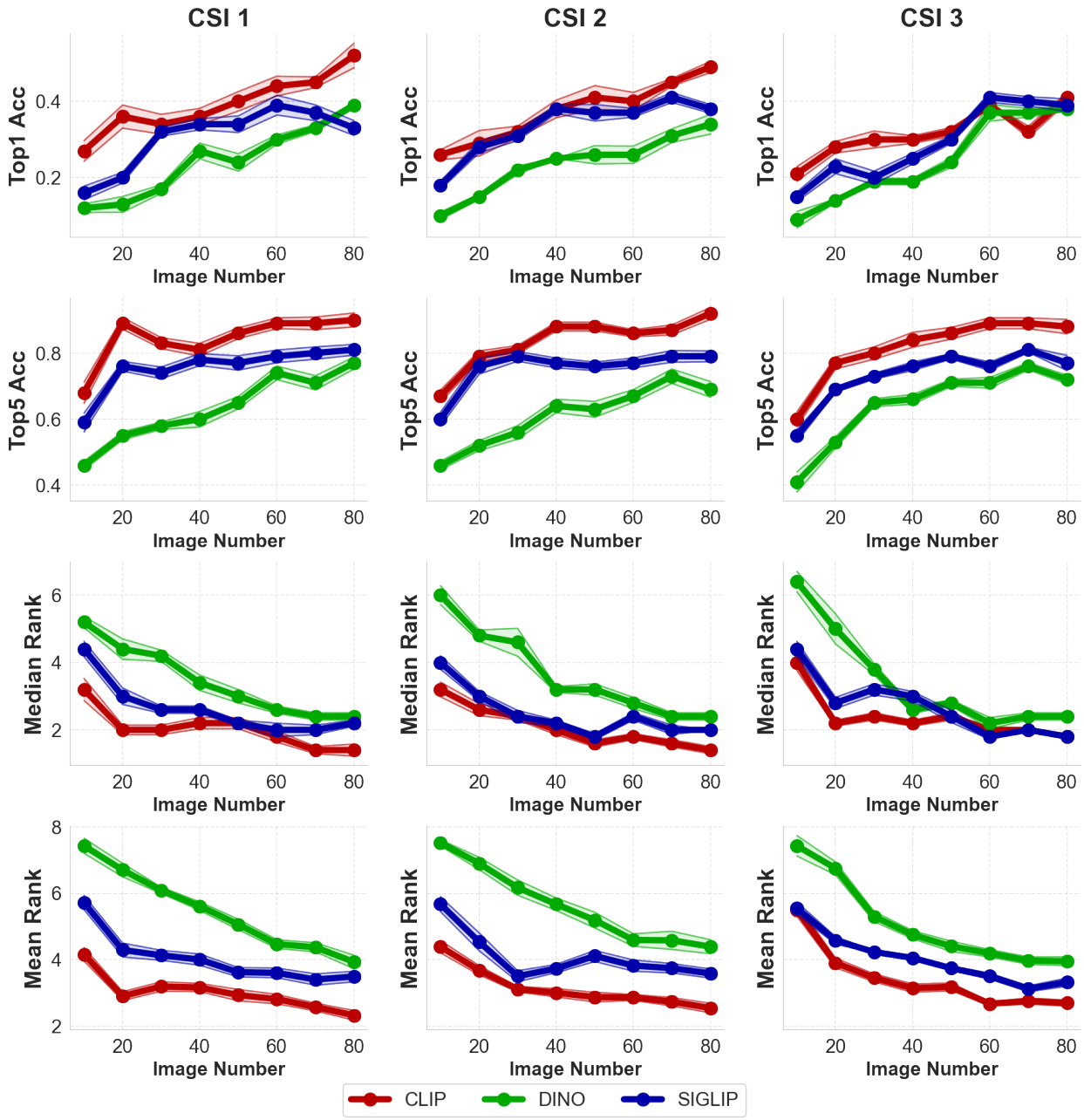

Figure S.11: Image-context scaling of BrainCoDec on BOLD5000 subjects.

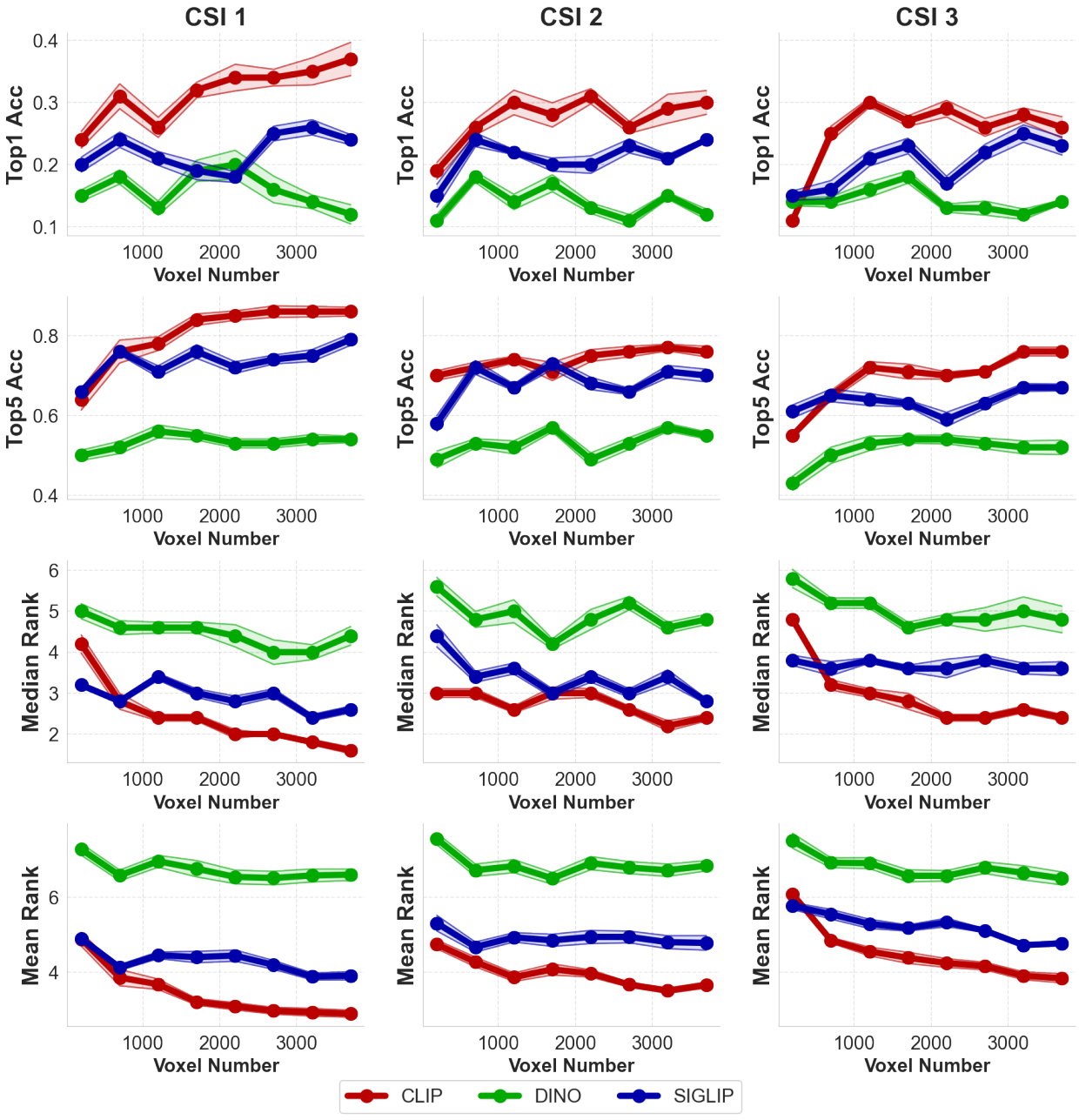

Figure S.12: Voxel-context scaling of BrainCoDec on BOLD5000 subjects.

A.7 Attention UMAP for other subjects

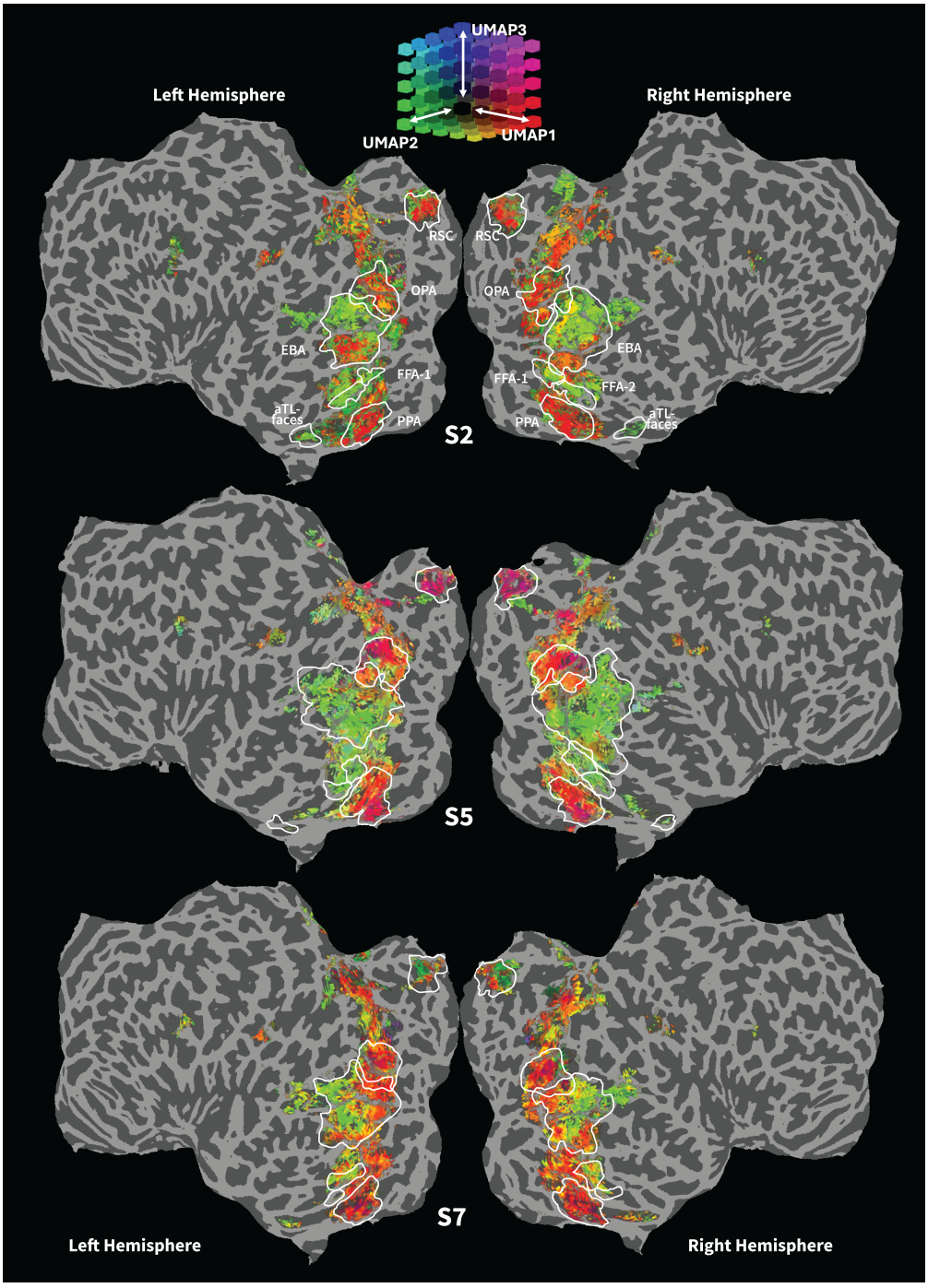

Figure S.13: Semantic attention patterns in BrainCoDec.

A.8 More Retrieval Results on unseen BOLD5000 Subjects

(a) Subject CSI 1 (b) Subject CSI 2 (c) Subject CSI 3

Figure S.14: Image retrieval results on BOLD5000 unseen subjects from fold 1 using 80 images as context. To note, since BOLD5000 provides only 20 test images, we visualize retrieval results from a pool of 500 images for rigorous evaluation.

A.9 Ablations

In this section, we compare different models on a variety of metrics. **PT only** indicates our model where it was only trained on synthetic data. **Inversion** is the model where we try to solve for the image embedding using gradient based optimization to recover the voxelwise activations using the stage-1 estimated voxelwise weights. For all models listed here we utilize 200 images and brain activation patterns from the novel subject as context.

Table S.3: **Quantitative comparison on model variants and ablations.**

Model	**S1**	**S2**	**S5**	**S7**
		$\%$ **Top-1 Accuracy (↑)**		
PT only	3.67 ± 0.69	3.24 ± 0.71	2.96 ± 0.64	2.68 ± 0.63
Inversion	1.61 ± 0.73	1.39 ± 0.86	2.04 ± 0.81	1.90 ± 0.77
BrainCoDec-200	25.5 ± 3.02	22.9 ± 2.98	23.2 ± 2.63	19.2 ± 2.42
BrainCoDec-200 no HO	28.3 ± 3.40	27.1 ± 3.21	29.4 ± 3.40	24.0 ± 3.36
		$\%$ **Top-5 Accuracy (↑)**		
PT only	14.0 ± 1.23	11.6 ± 1.42	9.70 ± 1.08	8.23 ± 0.94
Inversion	2.01 ± 0.53	1.98 ± 0.65	2.79 ± 0.63	2.21 ± 0.42
BrainCoDec-200	56.6 ± 3.21	52.4 ± 4.08	55.8 ± 2.47	51.2 ± 3.50
BrainCoDec-200 no HO	61.1 ± 2.19	61.1 ± 2.98	64.6 ± 2.71	56.8 ± 2.84
		$\%$ **Mean Rank (↓)**		
PT only	26.63 ± 0.93	27.70 ± 0.67	29.63 ± 0.87	30.93 ± 1.07
Inversion	45.87 ± 0.87	46.47 ± 0.90	43.97 ± 0.77	46.20 ± 1.27
BrainCoDec-200	4.43 ± 0.47	4.23 ± 0.33	3.93 ± 0.27	3.73 ± 0.30
BrainCoDec-200 no HO	2.67 ± 0.27	3.13 ± 0.30	2.50 ± 0.13	3.30 ± 0.23
		Cosine Similarity (↑)		
PT only	0.23 ± 0.05	0.20 ± 0.04	0.19 ± 0.05	0.20 ± 0.05
Inversion	0.32 ± 0.02	0.30 ± 0.02	0.31 ± 0.02	0.31 ± 0.07
BriancoDec-200	0.81 ± 0.01	0.80 ± 0.02	0.79 ± 0.03	0.79 ± 0.04
BrainCoDec-200 no HO	0.82 ± 0.01	0.81 ± 0.03	0.82 ± 0.03	0.80 ± 0.03

References

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1): 9383, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.