Duala: Dual-level Alignment of Subjects and Stimuli for Cross-Subject fMRI Decoding

Supplementary Material

1. Implementation Details

1.1. Explanation of Metrics

Low-level metrics. Pixelwise Correlation (PixCorr) measures the linear correspondence between reconstructed and ground-truth images at the pixel level, while the Structural Similarity Index (SSIM) [8] evaluates perceptual similarity in terms of luminance, contrast, and structure. AlexNet-based metrics, Alex(2) and Alex(5) [3], compute the correlation between feature activations extracted from the second and fifth convolutional layers, providing a view of reconstruction quality at different representational depths.

High-level metrics. To quantify semantic alignment, we follow prior decoding studies and compare deep features from Inception-V3 [6], CLIP ViT-L/14 [4], EfficientNet-B1 [7], and SwAV-ResNet50 [1]. For Incep and CLIP, we adopt a two-way identification paradigm: for each test sample, the ground-truth image embedding is compared against embeddings from the reconstruction and a distractor, and accuracy is computed as the percentage of correct discriminations. Eff and SwAV assess feature-level correspondence by computing the average correlation distance between image and reconstruction embeddings, where lower values indicate better semantic preservation.

Retrieval metrics. Retrieval evaluates how precisely the latent representations preserve cross-modal associations. For image retrieval, each fMRI embedding is matched to the image embedding with which it shares the highest cosine similarity. A trial is counted as correct if the paired image is selected. Brain retrieval inverts this process by retrieving the correct fMRI embedding given an image embedding.

1.2. Inference Details

All additional components introduced in our method (*e.g.*, the semantic alignment loss, the relational consistency loss, and the subject-level distribution perturbation) are used only during training and are discarded at test time. During inference, Duala directly maps a new subject's fMRI response to the aligned latent space without any additional optimization. As a result, the number of inference parameters is essentially identical to that of existing methods [2, 5].

2. Method Details

2.1. Details of the Semantic Alignment Loss

Eq.4 in the main paper is implemented as a standard cosine triplet loss over the subject-specific embeddings $z_i^{S_N}$.

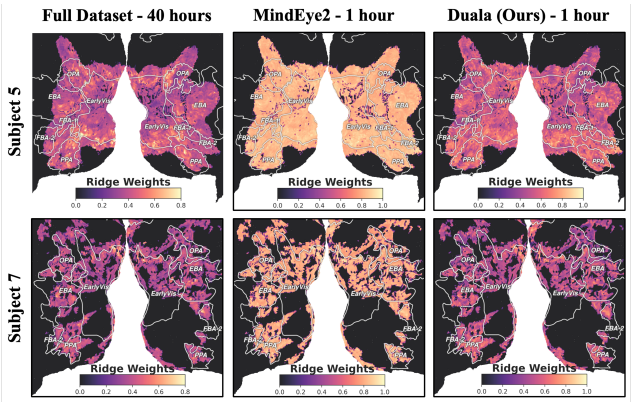

Figure 1. Visualization of transfer quantity in brain heatmaps with Subject 5 and Subject 7.

In practice, the number of stimulus categories is relatively large ($C = 80$), while a single mini-batch (batch size is 10) may contain very few positives and negatives for some classes. To obtain more reliable triplets, we maintain a small first-in-first-out memory bank of recent ℓ_2-normalized embeddings and their category labels. At each iteration, the current batch provides the anchor embeddings, and positives/negatives for the triplet loss are sampled from the union of the current batch and the memory bank. The loss still has exactly the form of Eq.4. The memory bank only enlarges the candidate pool of same-category and different-category examples across training steps and introduces no additional trainable parameters.

2.2. Details of the Relational Consistency Loss

For completeness, we provide clarifications regarding the relational consistency loss used during subject adaptation.

Prototype estimation. As defined in the main paper, the prototype of class c is the average of the normalized embeddings assigned to that class. Since mini-batches may contain few samples, we maintain a simple running average for each image i,

$$\mathbf{v}_i^{(t)} = m\,\mathbf{v}_i^{(t-1)} + (1-m)\,\mathbf{h}_i^{(t)},$$

where $\mathbf{h}_i^{(t)}$ is the current normalized embedding and m is the momentum. Class prototypes are then computed by averaging all available $\mathbf{v}_i^{(t)}$ for that class, together with the current batch features to ensure gradient flow. This mechanism introduces no additional learnable parameters and only stabilizes prototype estimation.

Table 1. Quantitative comparison with the Mindtuner model (reproduced). Results averaged across subjects 1, 2, 5, and 7 from the Natural Scenes Dataset with 1 hour of data. **Bold** indicates the best performance.

Method	Venue	Subject	Low-Level				High-Level				Retrieval	
			PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff.↓	SwAV↓	Image↑	Brain↑
MindTuner [2]	AAAI'25	Avg	0.218	0.410	86.7%	92.7%	83.8%	81.4%	0.804	**0.456**	83.8%	80.5%
Duala (Ours)	-	Avg	**0.230**	**0.416**	**87.9%**	**93.5%**	**85.4%**	**83.5%**	**0.781**	**0.445**	**84.5%**	**81.1%**
MindTuner [2]	AAAI'25	Subj1	0.241	0.415	89.56%	94.46%	84.83%	83.12%	0.788	0.444	94.77%	**92.29%**
Duala (Ours)	-	Subj1	**0.253**	**0.418**	**90.81%**	**95.29%**	**86.62%**	**85.11%**	**0.768**	**0.433**	**94.77%**	91.22%
MindTuner [2]	AAAI'25	Subj2	0.231	0.423	88.78%	**94.56%**	84.73%	81.98%	0.794	0.447	92.91%	90.76%
Duala (Ours)	-	Subj2	**0.236**	**0.428**	**89.21%**	94.35%	**85.35%**	**83.74%**	**0.778**	**0.443**	**94.64%**	**91.47%**
MindTuner [2]	AAAI'25	Subj5	0.194	0.408	84.72%	91.56%	83.89%	81.21%	0.808	0.459	75.63%	70.40%
Duala (Ours)	-	Subj5	**0.214**	**0.413**	**86.50%**	**93.07%**	**86.36%**	**84.46%**	**0.772**	**0.439**	**76.59%**	**71.20%**
MindTuner [2]	AAAI'25	Subj7	0.204	0.394	83.75%	90.29%	81.57%	79.16%	0.828	0.475	71.92%	68.69%
Duala (Ours)	-	Subj7	**0.215**	**0.405**	**85.17%**	**91.16%**	**83.34%**	**80.72%**	**0.805**	**0.463**	**71.94%**	**70.53%**

Table 2. Ablation studies on Duala. Results of subject 1 from NSD Dataset with 1 hour of data. **Bold** indicates the best performance.

Method	Low-Level				High-Level				Retrieval	
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff.↓	SwAV↓	Image↑	Brain↑
baseline	0.243	0.418	89.33%	94.28%	84.24%	83.35%	0.791	0.446	93.31%	89.92%
Duala w/o SSP	0.244	0.414	89.70%	94.34%	84.45%	83.68%	0.779	0.441	93.84%	90.59%
Duala w/o SDP	0.253	0.415	90.66%	94.98%	85.78%	83.31%	0.775	0.437	92.34%	**91.41%**
Duala (Ours)	**0.253**	**0.418**	**90.81%**	**95.29%**	**86.62%**	**85.11%**	**0.768**	**0.433**	**94.77%**	91.22%

Class and pair selection. For each pair of classes (c_1, c_2), we precompute a reference similarity

$$S^{\text{ref}}_{c_1,c_2} = \frac{1}{|\mathcal{K}_{c_1,c_2}|} \sum_{k \in \mathcal{K}_{c_1,c_2}} S^{s_k}_{c_1,c_2},$$

where $\mathcal{K}_{c_1,c_2}$ denotes the set of pre-trained subjects in which both classes appear. During adaptation, only pairs with available reference values are included in the valid set Ω, and the loss in Eq. (6) is computed.

3. Additional Results

3.1. Functional Alignment Analysis

We provide the full Transfer Quantity (TQ) visualizations for Subject 5 and Subject 7 in addition to the Subject 1 and 2 results presented in the main paper. As shown in Figure 1, the overall alignment pattern remains consistent with our earlier observations. Across both subjects, high-TQ activations remain distinctly concentrated in canonical visual areas, mirroring the structure observed in the 40-hour full-dataset model. This demonstrates that our cross-subject mapping preserves region-specific functional organization even under more challenging individual differences. In contrast, the MindEye2 [5] produces more dispersed TQ distributions, with elevated values appearing broadly across the cortex rather than localized to visual pathways.

3.2. Reproducing MindTuner for Fair Comparison

In the main paper, we report the performance of MindTuner [2] using the results provided by the original publication since the official implementation of MindTuner was not publicly released.

To ensure completeness and fairness, we additionally re-implemented MindTuner following the description in the paper. The reproduced model is trained under the same data splits and evaluation protocol as Duala. Despite following the same data splits and evaluation protocol as used for Duala, the reproduced model yields lower accuracy than the reported results. We attribute this discrepancy to unreported implementation choices and tuning details that cannot be precisely recovered from the paper alone. Therefore, we retain the original reported numbers for the main comparison, while the reproduced results are included in Table 1 as a reference to illustrate the range of performance achievable under a faithful re-implementation. Under the same controlled experimental setup, our Duala method consistently outperforms the reproduced MindTuner across all low-level, high-level, and retrieval metrics. This confirms the effectiveness and stability of our dual-level alignment decoding design.

3.3. Effectiveness of Key Components

To further assess the role of the Subject-level Distribution Perturbation (SDP) module, we include an additional variant that retains only the stimulus-level components ($\mathcal{L}_{\text{sa}}$ + $\mathcal{L}_{\text{rc}}$) while removing SDP entirely. As shown in Table 2,

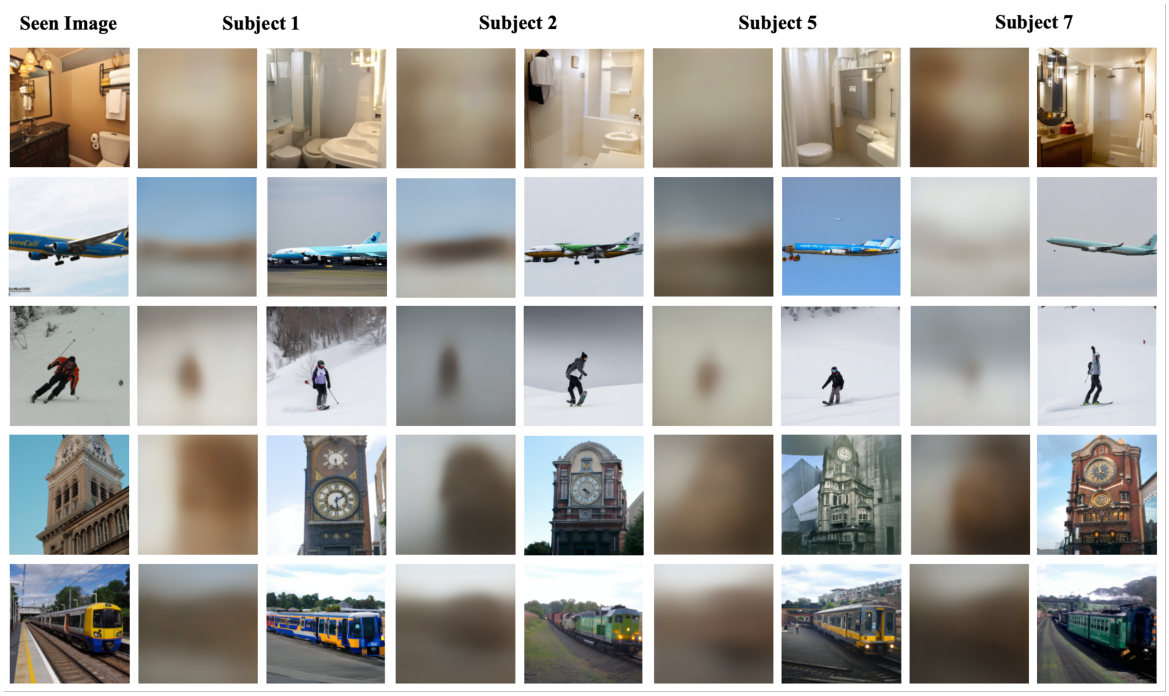

Figure 2. Visualization of reconstruction results with only 1 hour of data.

incorporating SDP yields clear gains in high-level semantic alignment. This validates that SDP improves subject-specific alignment beyond what stimulus-level regularization can achieve alone.

3.4. Subject-Specific Visualizations

We provide in Figure 2 additional reconstruction examples for individual subjects using the 1-hour setting. Across subjects, Duala reliably recovers core semantic content from fMRI signals, though the level of fine-grained detail naturally varies with subject-specific noise patterns and cortical response differences. These observations suggest that, under limited data such as the 1-hour setting, semantic alignment can be made robust across subjects. Nevertheless, the color appearance remains the most challenging aspect. Some reconstructed images exhibit noticeable shifts in hue or saturation even when the semantic content is correct.

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1

[2] Zixuan Gong, Qi Zhang, Guangyin Bao, Lei Zhu, Rongtao Xu, Ke Liu, Liang Hu, and Duoqian Miao. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14247–14255, 2025. 1, 2

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 1

[5] Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. In *International Conference on Machine Learning*, 2024. 1, 2

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 1

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[8] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1