A DATA PREPROCESSING

Preprocessing before input: The original BOLD volumes in the HCP dataset have dimensions of $113 \times 136 \times 113$. The 3-D BOLD volumes are patchified by 3D-CNN as shown in Fig. 3(a) during experiments. For the NSD dataset, we extract ROIs in the visual cortex based on brain parcellation, including V1-V3, V3ab, hV4, ventral occipital (VO), intraparietal sulcus (IPS), lateral occipital (LO), middle temporal (MT), and parahippocampal cortex (PHC), covering both primary and high-level areas of the visual cortex. The number of voxels in each ROI varies across subjects, and the number of voxels for the same subject also varies across different ROIs. To enable model weight sharing across subjects, we align the dimensions of the BOLD signals in the same ROI across subjects. Additionally, to adapt the data to the Transformer model, we align the dimensions of the BOLD signals in the same subject across different ROIs. To achieve this, we employ principal component analysis to reduce the dimensionality of the BOLD signals to the minimum dimensionality value of 268 across all BOLD signals from all ROIs of all subjects. During the testing process, we can apply zero-padding to neural signals with dimensions lower than this value. In the methods using MLP as the backbone network, the patches of all ROIs are concatenated into a vector to input into the model. In the methods using ViT as the backbone network, one patch is one token.

Data split: The HCP dataset is split in accordance with the method described in Khosla et al. (2020). The first three movies are used for training and validation, while the fourth movie is used for test. For the single-subject decoding task, the training, validation, and test sets consist of 2000, 265, and 699 samples, respectively. For the multi-subject decoding task, the training sets of nine subjects are combined and randomly shuffled for model training, while the validation sets of the same nine subjects are combined for model validation. To account for the randomness, we report the average results of three random runs. A hemodynamic delay of 4 seconds estimated in Khosla et al. (2020) is used in this paper. For the NSD dataset, the stimuli viewed by all subjects and their corresponding neural responses are used for test in the single-subject decoding task. A validation set consisting of 1,000 randomly sampled examples is used for hyper-parameter tuning and convergence monitoring during training. The remaining data is used for training. For the multi-subject decoding task, the training sets of eight subjects are combined and randomly shuffled, and a validation set of 1,000 examples is randomly sampled. The remaining data is used for training. Due to the randomness of data split, we report the average results of five random splits.

Multimodal feature extraction: We concatenate the WordNet annotations of each movie frame to form the textual information for stimuli in the HCP dataset. In the NSD dataset, each stimulus image is associated with five captions. Following the approach in Lin et al. (2022), we compute the similarity between each image and its five captions in the multimodal feature space of CLIP. We use half of the maximum similarity score as the threshold and randomly select one caption from the selected candidates to serve as the textual information for the stimulus. We extract the text features by inputting the textual information into the CLIP text encoder. For both datasets, we truncate the image and text features with a threshold of 1.5 and normalize their L2 norms to 1. We then compute the average of the image and text features to obtain the multimodal feature $\mathbf{f}_{hlv}$, which guides the learning of $\mathbf{z}_{hlv}$.

B THE CHOICE OF AN RSA-BASED LOSS

Using RSA to guide the representation learning of fMRI is superior to directly mapping fMRI representations to the CLIP embeddings. There are three main reasons:

1. Mapping-based methods tend to focus on learning local information when embedding fMRI representations into the CLIP space (e.g., minimizing the L2 norm between true and predicted values). However, they may overlook the global topological structure of the CLIP space, which can affect the interpretability and generalization capabilities of the embedding space. In contrast, using RSA loss ensures the global topological structure similarity between the embedding space and the CLIP space.
2. There is a gap between the representation spaces of CLIP and fMRI. It is more challenging to directly learn the mapping from fMRI representations to CLIP representations than to learn the topological relationship. Forcing alignment may lead to overfitting and poor generalization.

Table B4: Results of the RSA-based method (CLIP-MUSED) and mapping-based methods on the NSD dataset.

Method	mAP ↑	AUC ↑	Hamming ↓
Mapping-Based	$.247 \pm .026$	$.844 \pm .033$	$.035 \pm .002$
RSA-Based	$\mathbf{.258 \pm .017}$	$\mathbf{.877 \pm .021}$	$\mathbf{.030 \pm .002}$

3. Mapping-based methods introduce additional trainable parameters in the mapping network, and the architecture of the mapping network also needs to be delicately optimized.

We also conduct a comparative experiment between RSA-based and mapping-based methods on the NSD dataset. The results in Table B4 demonstrate the superiority of the RSA-based loss. The coefficient ($\lambda = 0.0001$) of the mapping-based loss item has been optimized.

C PEFORMANCE COMPARISON ON EACH SUBJECT

We compare the performance of SS-ViT, which is a single-subject decoding method with CLIP-MUSED, as they both utilize ViT as the backbone network. Fig. C5 compares the mAP of our method and SS-ViT on each subject. As shown in the figure, our method outperforms SS-ViT on all subjects. Fig. C6 show the comparison results on the NSD dataset. It is evident that our method consistently outperforms SS-ViT on the majority of subjects. These findings suggest that the neural representations shared among subjects learned by our method are superior to those learned by the single-subject method.

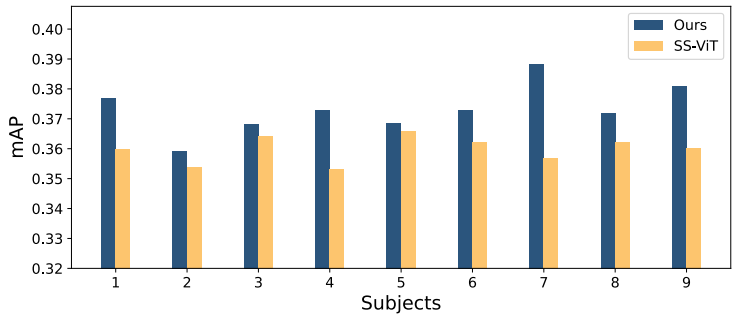

Figure C5: Performance comparison between SS-ViT and our method on each subject of the HCP dataset.

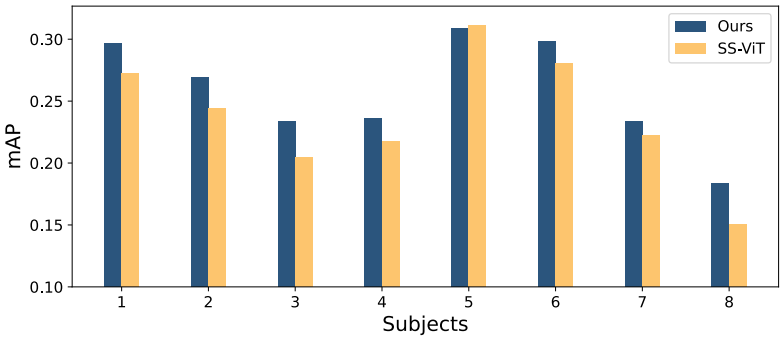

Figure C6: Performance comparison between SS-ViT and our method on each subject of the NSD dataset.

D GUIDANCE EFFECT ON THE SINGLE-SUBJECT METHOD

In the experiments, we only employ CLIP guidance in our proposed multi-subject model. To investigate the impact of CLIP guidance on the single-subject model, we modified the SS-ViT to

Table D5: Results of two single-subject methods CLIP-SS-ViT and SS-ViT, with and without CLIP guidance, and CLIP-MUSED on the NSD dataset. All the improvement of CLIP-MUSED compared to other cases is significant (t-test, $p < 0.05$), where the p-values have been corrected with the Holm-Bonferroni method for multiple comparisons.

Method	mAP ↑	AUC ↑	Hamming ↓
SS-ViT	$.238 \pm .005$	$.815 \pm .008$	$.032 \pm .000$
CLIP-SS-ViT	$.234 \pm .002$	$.822 \pm .006$	$.032 \pm .001$
CLIP-MUSED	$\mathbf{.258 \pm .017}$	$\mathbf{.877 \pm .021}$	$\mathbf{.030 \pm .002}$

enable the learning of neural representations under the guidance of CLIP, i.e., CLIP-SS-ViT. The model structure and composition of the loss function for CLIP-SS-ViT are the same as our method, with the only difference being that CLIP-SS-ViT is trained on single-subject data. We carefully tuned the trade-off parameters for CLIP-SS-ViT, and the optimal parameters are $\lambda_{\perp} = 0.0001$, $\lambda_{hlv} = 0.0001$, and $\lambda_{llv} = 0.0001$. The results are shown in Table D5.

It can be observed that, both guided by CLIP, our method outperforms the CLIP-SS-ViT. This superiority can be attributed to the multi-subject aggregation strategy employed in our method. On the single-subject model, the relatively weak effect of using CLIP as guidance may be due to the fact that the single-subject model does not need to handle individual differences. In other words, the original classification loss can implicitly learn the relationship between different fMRI representations of a single subject, and the guidance from CLIP is redundant for the model.

E GUIDANCE EFFECT OF DIFFERENT DNNS

Table E6 presents the model performance when the neural representation learning is guided by the topological relationship of visual stimuli across different DNN representation spaces. CLIP-Img/CLIP-Text refer to the utilization of high-level image/text features extracted by the image/text encoder of CLIP, instead of the multimodal features, during the learning of high-level tokens. In summary, the results depicted in Table E6 suggest that features extracted from CLIP exhibit a superior guidance effect when compared to the baseline models (ViT and AlexNet). Notably, textual information provides a richer semantic understanding of the visual stimuli and integrating it with the image features can serve to augment the performance of the model in guiding neural representation learning.

Table E6: Comparison of the guidance effect of different DNNs. All the improvement of CLIP compared to other cases is significant (t-test, $p < 0.05$) except for those underlined, where the p-values have been corrected with the Holm-Bonferroni method for multiple comparisons.

Methods	HCP			NSD		
	mAP ↑	AUC ↑	Hamming ↓	mAP ↑	AUC ↑	Hamming ↓
ViT	$.362 \pm .003$	$.562 \pm .004$	$.383 \pm .040$	$.229 \pm .009$	$.849 \pm .014$	$.035 \pm .001$
AlexNet	$.362 \pm .002$	$.558 \pm .005$	$.333 \pm .013$	$.225 \pm .012$	$.855 \pm .013$	$.034 \pm .002$
CLIP-Img	$.370 \pm .001$	$\underline{.577 \pm .001}$	$.291 \pm .002$	$.238 \pm .006$	$.863 \pm .008$	$.033 \pm .001$
CLIP-Text	$.370 \pm .002$	$\underline{.579 \pm .006}$	$.286 \pm .003$	$.223 \pm .006$	$.841 \pm .008$	$.036 \pm .001$
CLIP (OURS)	$\mathbf{.373 \pm .002}$	$\mathbf{.581 \pm .008}$	$\mathbf{.283 \pm .004}$	$\mathbf{.258 \pm .017}$	$\mathbf{.877 \pm .021}$	$\mathbf{.030 \pm .002}$

F VISUALIZATION

We present the visualization of between-subject representational similarity matrices (RSMs) of low-level and high-level tokens on two datasets, as depicted in Fig. F7. Notably, the similarity of tokens between subjects is observed to be higher on the HCP dataset compared to the NSD dataset. This can be attributed to the fact that in the HCP dataset, all subjects viewed the same stimuli, and the distribution of stimuli across subjects is uniform, whereas in the training set of the NSD dataset, the stimuli viewed by different subjects are mutually exclusive. Evidently, different subjects process

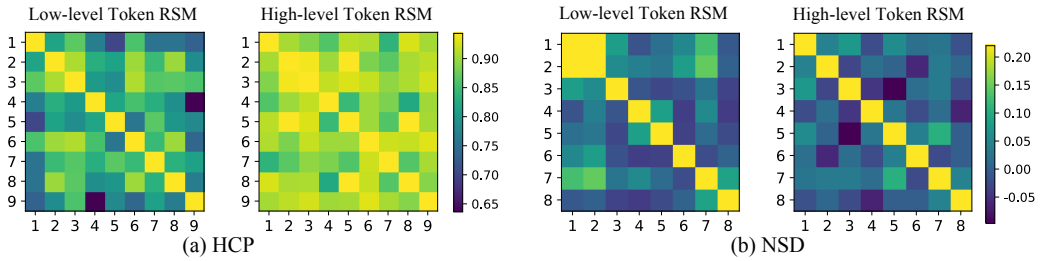

Figure F7: RSM between low-level and high-level tokens across subjects of our method on (a) the HCP dataset and (b) the NSD dataset.

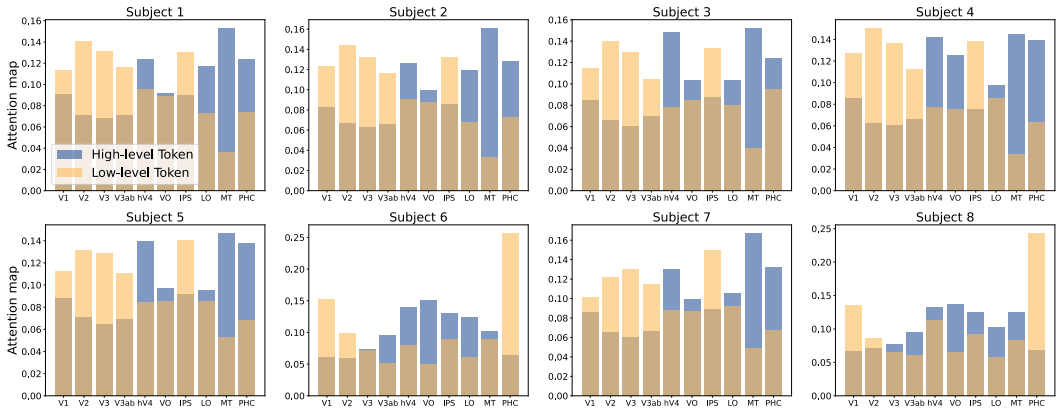

Figure F8: Attention maps between the low-level and high-level tokens of our method and different brain region tokens at the last Transformer self-attention layer for 8 subjects on the NSD dataset.

stimulus information in slightly distinct patterns even when viewing the same stimuli, and these differences are further amplified when presented with different stimuli. The tokens learned for each subject in CLIP-MUSED can encode these inter-subject variabilities, resulting in lower similarity of tokens between subjects on the NSD dataset with different stimuli for different subjects.

Fig. F8 shows the attention maps of the low-level and high-level tokens on different ROIs of the NSD dataset. The low-level tokens exhibit a higher attention allocation towards the primary and intermediate visual cortex regions on the left side of the bar chart, including V1-V4, while the high-level tokens exhibit a more pronounced attention towards the higher visual cortex regions such as LO, MT, and PHC. Prior research has established that MT plays a crucial role in depth perception (Born & Bradley, 2005), LO is involved in object recognition tasks (Grill-Spector et al., 2001), and PHC contributes to visual perception related to memory and spatial scenes (Aminoff et al., 2013). In contrast, Fig. F9 demonstrates the attention maps of the subject embeddings in the MS-EMB method, revealing a more focused attention on the V3, V4, and VO brain regions, but less on the higher visual cortex regions such as LO, MT, and PHC. In contrast, our method leverages both low-level and high-level tokens to allocate attention towards both low-level and intermediate visual cortex regions, as well as higher visual cortex regions. The difference in token attention maps between our method and the MS-EMB method partially elucidates the superiority of our method in classification performance.

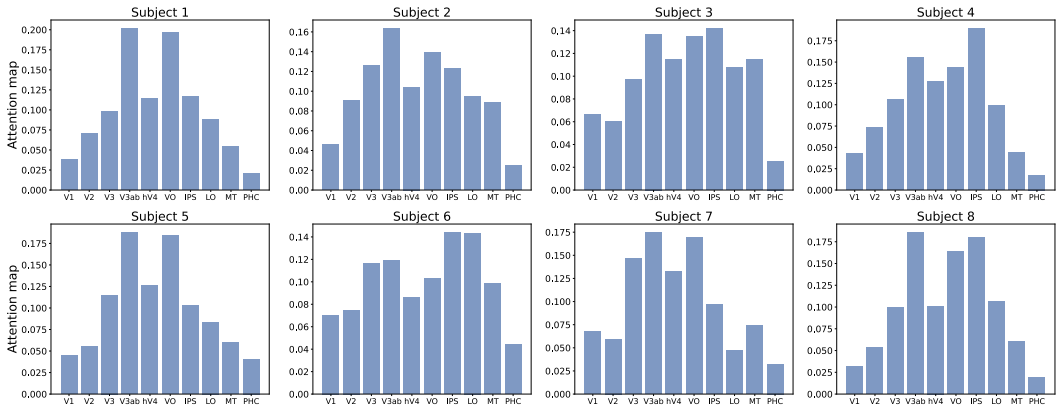

Figure F9: Attention maps between the subject embedding token of the MS-EMB method and different brain region tokens at the last Transformer self-attention layer for 8 subjects on the NSD dataset.