

# Supplementary Materials:

## Lite-Mind: Towards Efficient and Robust Brain Representation Learning

Anonymous Authors

### A ADDITIONAL METHOD DETAILS

#### A.1 Implementation details

**Data preprocessing.** We downloaded the NSD dataset from the official website and used Takagi’s code to extract *nsdgenal* fMRI voxels, while Takagi extracted the *stream* region of the NSD dataset. We noticed that MindEye scaled the fMRI voxel values in advance, while we did not. The difference in fMRI voxels input data is shown in the example in Figure 6. NSD image files come from *nsd-stimuli.hdf5* file and have a unified size of  $425 \times 425$ . We did not perform any data augmentation on the image and straightly extracted the hidden layer representation (size of  $257 \times 768$ ) of the image through CLIP ViT-L/14 for training.

**Hyper-parameters.** On the NSD dataset, during training DFT Backbone, the weight decay is set to 7,  $\tau$  is  $1/e^8$ ,  $\alpha = 1$ , and CLIP’s contrastive loss is unidirectional for image retrieval and fMRI retrieval. Owing to the lightweight nature of Lite-Mind, the batch size is set to 500, the learning rate is  $1e-3$ , patch size is set to 480, and Filter library size is 4. Filter block layers are the same for all subjects.

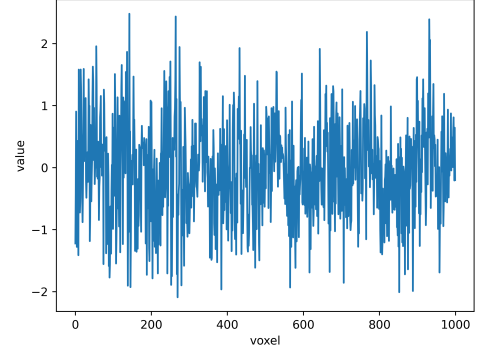
About LAION-5B retrieval, for DFT Backbone, the weight decay is set to 7,  $\tau$  is  $1/e^8$ , while the weight decay is  $6.02e-2$  for diffusion projector. CLIP’s contrastive loss is unidirectional for image retrieval,  $\alpha = 0.5$ , the batch size is set to 80 and the learning rate is  $1.16e-3$  for DFT Backbone while  $1.1e-4$  for diffusion projector. Note that hyper-parameters in Diffusion Projector are the value of the open-source DALLE-2 from <https://github.com/lucidrains/DALLE2-pytorch>.

On the GOD dataset, the hyperparameters are the same as those on the NSD dataset, except for the batch size and patch size which are set to 1200 and 8 for all 5 subjects on the GOD dataset, respectively.

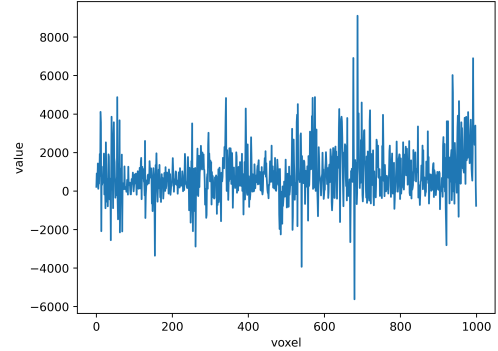
### B ADDITIONAL EXPERIMENT RESULTS

#### B.1 Additional Results of Retrieval

To demonstrate the applicability of our method, we also conducted experiments on three other Subjects, i.e., Subject 2, Subject 5, and Subject 7, on the NSD dataset, and the experimental results are shown in Table 7. In order to control model size similarity, we did not make more targeted model adjustments for subjects with shorter voxel lengths. It can be seen that the retrieval accuracies of Subjects 1 and 2 are greater than 90%, and the retrieval accuracies of Subjects 5 and 7 are also above 80%. The results prove that DFT Backbone can efficiently work on different subjects. Note that MindEye only includes the results of the overall model in Subjects 2, 5, and 7 in its Appendix, and it does not evaluate the effect of individual MLP Backbone. As a result, Table 7 does not present any evaluation results for MindEye.



(a) MindEye



(b) Lite-Mind

**Figure 6: The fMRI averaged activities of MindEye and Lite-Mind responding to the same image, respectively. The figures only visualize the activities of the first 1000 voxels for illustration.**

Method	Voxel Length	Parameters	Retrieval	
			Image $\uparrow$	Brain $\uparrow$
Lite-Mind(Subj 1)	15724	12.51M	94.6%	97.4%
Lite-Mind(Subj 2)	14278	12.49M	94.1%	98.2%
Lite-Mind(Subj 5)	13039	12.47M	80.5%	86.3%
Lite-Mind(Subj 7)	12682	12.47M	81.7%	82.3%

**Table 7: Additional retrieval performance for individual subjects on 982 test images of the NSD dataset.**

Method	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
Lite-Mind(Subject 1)	.134	.332	78.8%	88.9%	88.5%	88.8%	.730	.451
Lite-Mind(Subject 2)	.120	.328	78.0%	89.4%	86.3%	87.4%	.730	.446
Lite-Mind(Subject 5)	.123	.332	79.4%	90.0%	88.8%	89.9%	.712	.440
Lite-Mind(Subject 7)	.121	.331	78.7%	88.8%	87.8%	88.5%	.723	.448

**Table 8: LAION-5B retrieval alternative reconstruction performance for the specific subject.**

## B.2 Additional Results of LAION-5B Retrieval

In order to better reflect the retrieval performance of Lite-Mind on LAION-5B, we presented the performance indicators of LAION-5B retrieval substitution reconstruction for other subjects, i.e., Subject 2, Subject 5, and Subject 7, in Table 8, corresponding to the average performance of subjects in Table 1. Similarly, the visualization results of the other subjects in Figure 7 correspond to the image samples of Subject 1 in Figure 4. Based on the comprehensive table and graph, it can be found that Lite-Mind has good generalization on all four subjects, verifying the LAION-5B retrieval ability of Lite-Mind on different subjects. Meanwhile, as shown in Figure 7, the retrieval performance of LAION-5B completely depends on the retrieval accuracy of the CLS model in the test set, such that images retrieved incorrectly in the test set may also have retrieval bias on LAION-5B, for example treating a teddy bear as an image of a cat or dog as shown in Figure 7. However, both MindEye and Lite-Mind exhibit relatively low retrieval accuracy with aligned CLS embedding models. In the future, it would be beneficial to explore models that improve the alignment of CLS embeddings or employ more efficient methods to directly perform retrieval through hidden layers in LAION-5B.

## B.3 Additional Results of Zero-shot Classification

In the main body of the paper, we only demonstrated the zero-shot classification effect of Lite-Mind on the GOD dataset. The corresponding retrieval results of each Subject are shown in the Table 9 below.

Method	Voxel Length	Parameters	Image Retrieval↑	
			top1	top5
Lite-Mind(Subj 1)	4466	15.50M	30.0%	60.0%
Lite-Mind(Subj 2)	4404	15.46M	38.0%	58.0%
Lite-Mind(Subj 3)	4643	15.64M	38.0%	72.0%
Lite-Mind(Subj 4)	4133	15.24M	42.0%	62.0%
Lite-Mind(Subj 5)	4370	15.43M	26.0%	54.0%

**Table 9: Retrieval performance on the GOD dataset.**

## B.4 Hyper-parameters Experiment

In this section, we explore the influence of some hyperparameters on the model, including patch size and number of filters  $M$ , to verify the parameter sensitivity of the model. All the experimental results on the NSD dataset are from Subject 1, with a retrieving pool size of 300.

*Patch size.* We conducted experiments by varying the patch size as presented in Table 10. The results exhibit stable retrieval accuracy above 90%, which indicates that our DFT Backbone is not sensitive to part size. Since we found patch size has less effect on retrieval accuracy, we chose a relatively large patch size (i.e., 480) to ensure fewer parameters and faster convergence.

Patch size	Parameters	Retrieval	
		Image↑	Brain↑
50	14.0M	91.6%	96.4%
200	12.7M	92.9%	96.9%
480	12.5M	94.6%	97.1%
600	12.3M	94.2%	97.3%
900	11.9M	92.1%	95.5%

**Table 10: Retrieval performance for different patch size on the NSD dataset.**

*Filter library size.* We conducted experiments by varying the Filter library size as presented in Table 11. Specifically, there is a significant performance improvement from  $M = 1$  to  $M = 4$ , while a slight fluctuation is observed for  $M = 4/8$ . By setting  $M = 4$ , the model has the ability to acquire diverse and distinct feature patterns from various dimensions of the frequency response while still maintaining an appropriate computational cost. Therefore, we determine that  $M = 4$  is the optimal choice on the NSD dataset.

Filters	Parameters	Retrieval	
		Image↑	Brain↑
1	9.0M	89.8%	96.4%
2	10.2M	93.5%	97.4%
4	12.5M	94.6%	97.1%
8	17.2M	94.4%	96.7%

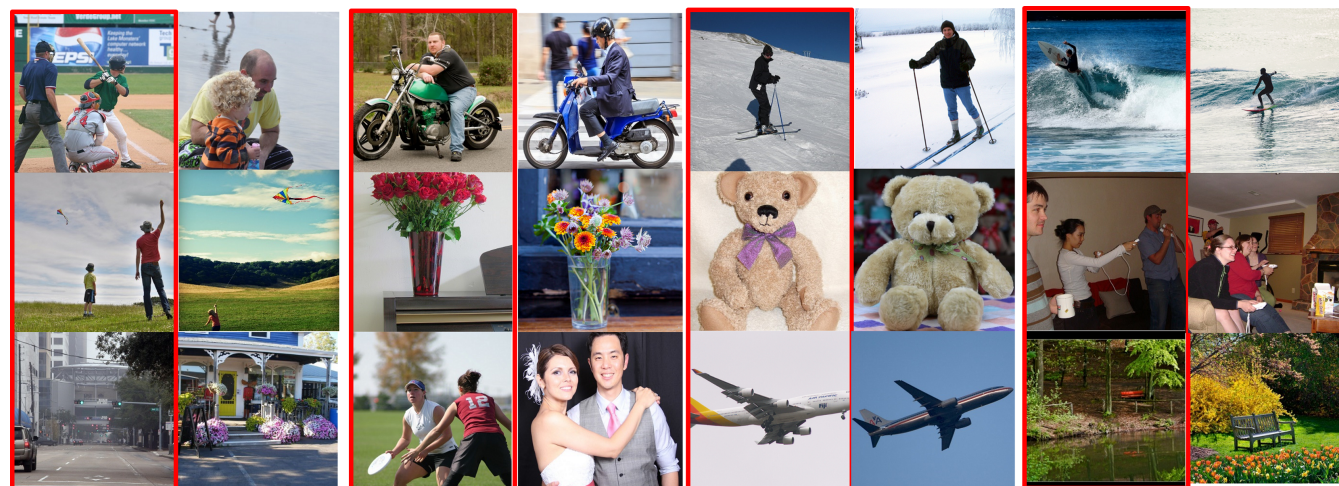
**Table 11: Retrieval performance for different Filter library size on the NSD dataset.**

## C ADDITIONAL VISUALIZATION

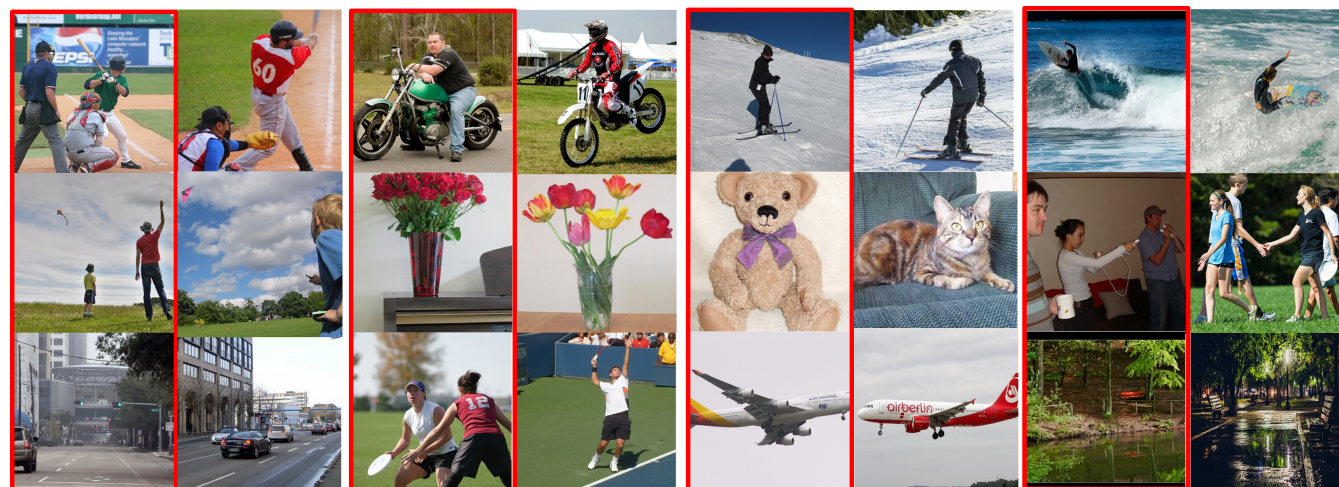
### C.1 Training Curve

We visualized the training process of all four subjects on the NSD dataset in Figure 8. As the training epochs increased, the loss of the training set rapidly decreased, while the accuracy of the test set





Subject 2



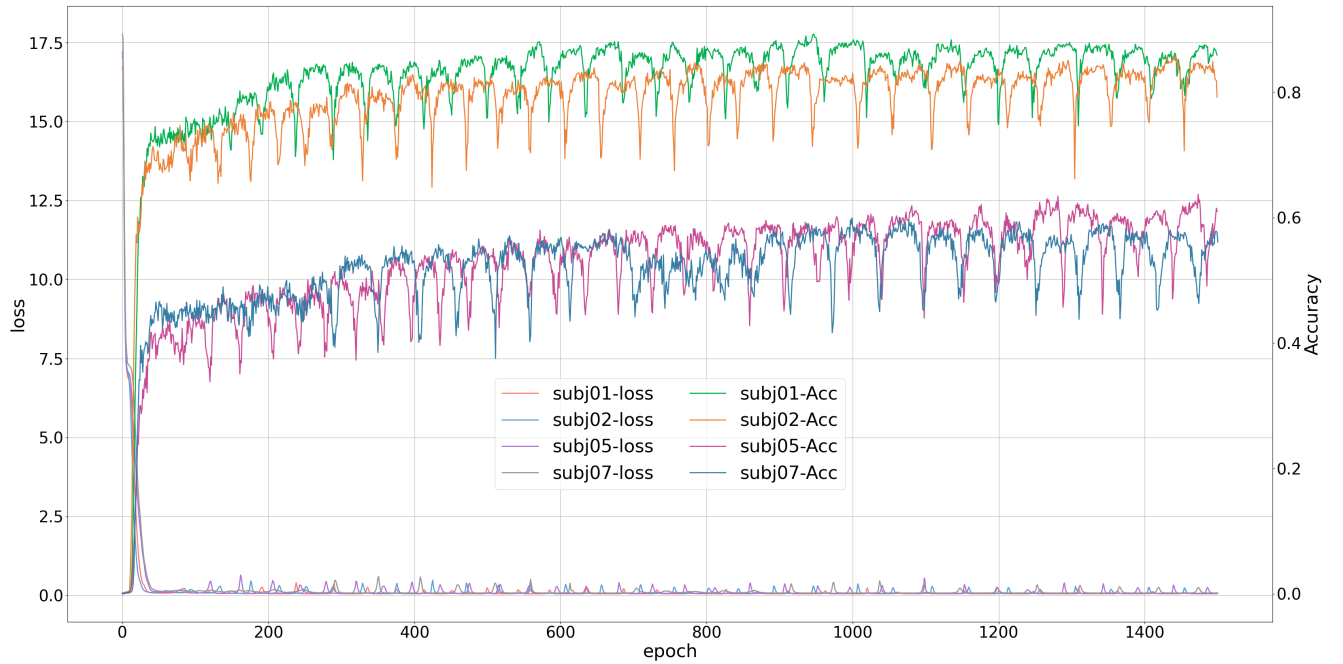
Subject 5



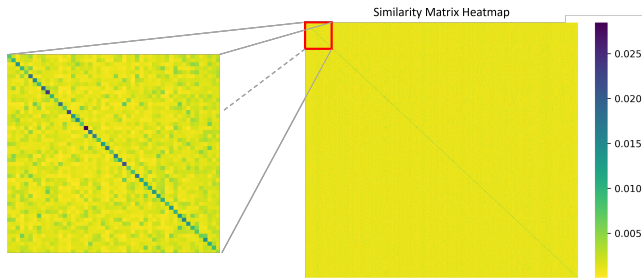
Subject 7

Figure 7: Additional retrieval results corresponding to Figure 4. The left column marked by a red box in every two columns represents the original image seen by the subject, and the right column represents the image retrieved on LAION-5B.





**Figure 8: Lite-Mind's training loss curve and testing accuracy curve for Subject 1, 2, 5, 7 on the NSD dataset. The testing accuracy is calculated based on a retrieval pool of size 982.**



**Figure 9: Lite-Mind's retrieval heatmap on the NSD dataset for Subject 1. The larger figure on the right represents  $982 \times 982$ 's heatmap, and the smaller figure on the left represents the  $50 \times 50$  subgraph.**

rapidly increased and then showed a slow upward trend. Accuracy refers to the hit rate of correct retrieval from 982 test set images.

## C.2 Retrieval Heatmap

We visualized the retrieval heatmap for Subject 1 on all 982 test images of the NSD dataset in Figure 9. It can be observed that the similarity is highest on the diagonal, and the color of the retrieved heat map is darker. It shows that Lite-Mind has effectively retrieved corresponding images, even if there are many similar images in the test set, which verifies the fine-grained ability of Lite-Mind.

## C.3 Information Alignment

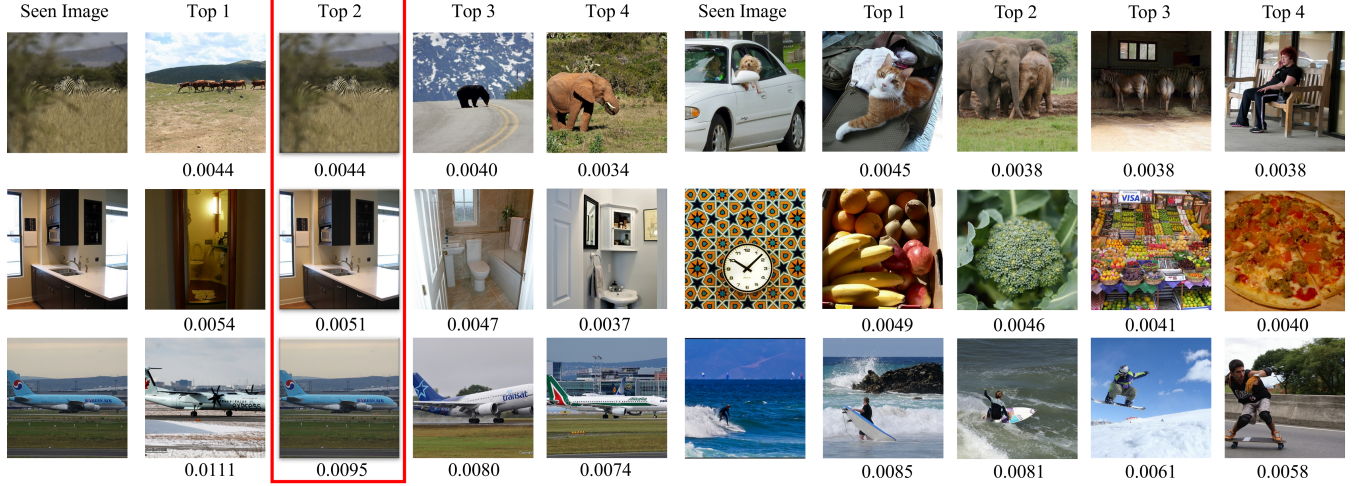
We visualized the T-SNE plot between the voxel embeddings output by DFT Backbone and the image embeddings of frozen CLIP as the accuracy of the test set improved, as shown in Figure 13. We can observe that as the training progresses, the retrieval accuracy of the test set improves, and the shape of voxel embeddings tends to be closer to image embeddings, indicating the success of contrastive learning.

## C.4 More Retrieval Cases

In this section, we visualize retrieval failure cases in Figure 10, although only dozens of images are not in the Top 1. From the left half of the Figure, it can be seen that though not in Top 1, Lite-Mind still retrieved ground-truth in Top 2, and images in Top 4 are similar (either semantically similar or structurally similar, eg. animals in the wild or an airplane on the runway). On the contrary, a smaller proportion of ground-truth did not appear within the Top 4, as shown in the right half of the Figure. From the perspective of the images themselves, most of the scenes are too complex, even abstract (as shown in the second image), which may be the reason for the retrieval failure.

## C.5 Visualization in Frequency Domain

We visualized the weights of Filter library of Filter Blocks, as shown in Figure 11 and 12. Visualization is divided into the real part (left) and imaginary part (right) of filter weights. It can be observed that different filters have varying degrees of attention to different tokens, and the frequency domain better captures this characteristic.



**Figure 10: Partial failure retrieval results of Lite-Mind on all 982 test images for Subject 1. The number below each image represents the similarity score.**

Interestingly, the weight of the imaginary part for the first and last tokens is almost always 0, indicating that the noise is distributed in these two tokens.

## D THEORETICAL ANALYSIS.

### D.1 Complex Multiplication.

For two complex number values  $Z_1 = (a + jb)$  and  $Z_2 = (c + jd)$ , where  $a$  and  $c$  is the real part of  $Z_1$  and  $Z_2$  respectively,  $b$  and  $d$  is the imaginary part of  $Z_1$  and  $Z_2$  respectively. Then the multiplication of  $Z_1$  and  $Z_2$  is calculated by:

$$\begin{aligned} Z_1 Z_2 &= (a + jb)(c + jd) \\ &= ac + j^2 bd + jad + jbc \\ &= (ac - bd) + j(ad + bc) \end{aligned} \quad (1)$$

### D.2 Theorem Proof.

**Theorem 1.** Suppose that  $\mathbf{H}$  is the representation of raw fMRI voxel tokens and  $\mathcal{H}$  is the corresponding frequency components of the spectrum, then the energy of voxel tokens in the spatial domain is equal to the energy of its representation in the frequency domain. Formally, we can express this with the above notations:

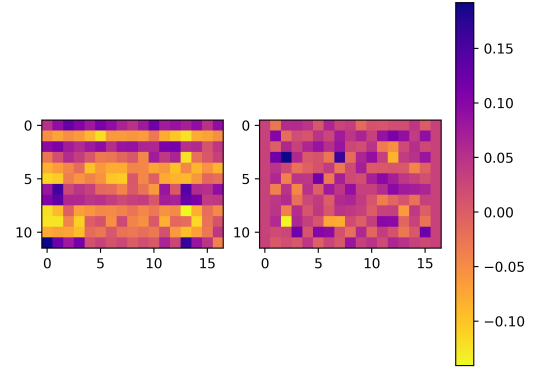
$$\int_{-\infty}^{\infty} |\mathbf{H}(v)|^2 dv = \int_{-\infty}^{\infty} |\mathcal{H}(f)|^2 df \quad (2)$$

Where  $\mathcal{H}(f) = \int_{-\infty}^{\infty} |\mathbf{H}(v)| e^{-j2\pi f v} dv$ ,  $v$  is the token dimension,  $f$  is the frequency dimension.

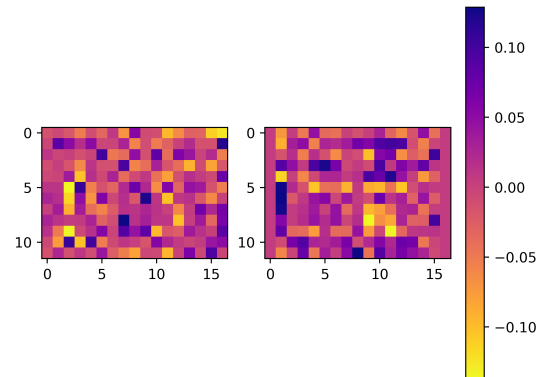
**PROOF.** Given the representation of raw voxel token series  $H \in \mathbb{R}^{C \times N}$ , let us consider performing integration in the  $N$  dimension (spatial dimension), denoted as the integral over  $v$ , then

$$\int_{-\infty}^{\infty} |\mathbf{H}(v)|^2 dv = \int_{-\infty}^{\infty} \mathbf{H}(v) \mathbf{H}^*(v) dv \quad (3)$$

where  $\mathbf{H}^*(v)$  is the conjugate of  $\mathbf{H}(v)$ . According to IDFT,  $\mathbf{H}^*(v) = \int_{-\infty}^{\infty} \mathcal{H}^*(f) e^{-j2\pi f v} df$ , we can obtain

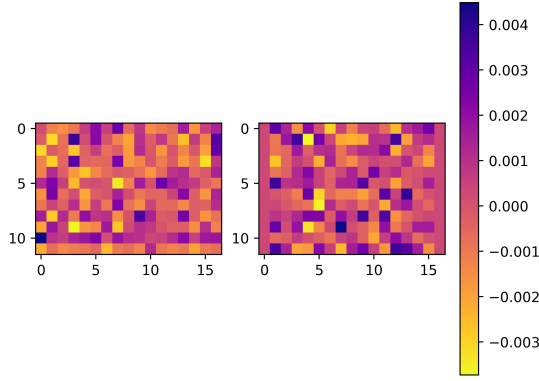


**(a) The real part and imaginary of a filter weights.**

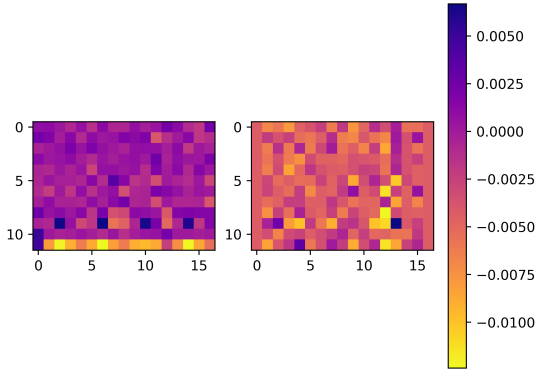


**(b) The real part and imaginary of another filter weights from the same Filter Library.**

**Figure 11: Weights visualization for DFT Backbone of 257×768 embedding length.**



(a) The real part and imaginary of a filter weights.



(b) The real part and imaginary of another filter weights from the same Filter Library.

**Figure 12: Weights visualization for DFT Backbone of 768 CLS embedding length.**

$$\begin{aligned}
 \int_{-\infty}^{\infty} |\mathbf{H}(v)|^2 dv &= \int_{-\infty}^{\infty} \mathbf{H}(v) \left[ \int_{-\infty}^{\infty} \mathcal{H}^*(f) e^{-j2\pi f v} df \right] dv \\
 &= \int_{-\infty}^{\infty} \mathcal{H}^*(f) \left[ \int_{-\infty}^{\infty} |\mathbf{H}(v)| e^{-j2\pi f v} dv \right] df \\
 &= \int_{-\infty}^{\infty} \mathcal{H}^*(f) \mathcal{H}(f) df \\
 &= \int_{-\infty}^{\infty} |\mathcal{H}(f)|^2 df
 \end{aligned} \tag{4}$$

Proved.  $\square$

Therefore, the energy of a voxel token series in the spatial domain is equal to the energy of its representation in the frequency domain.

**Theorem 2.** Given the voxel token series input  $\mathbf{H}$  and its corresponding frequency domain conversion  $\mathbf{H}$ , the operations of frequency-domain MLP on  $\mathbf{H}$  can be represented as global convolutions on  $\mathbf{H}$  in the spatial domain. This can be given by:

$$\mathcal{H}\mathcal{W} + \mathcal{B} = \mathcal{F}(\mathbf{H} * \mathcal{W} + \mathcal{B}) \tag{5}$$

Where  $\mathcal{F}$  is DFT,  $*$  is a circular convolution,  $\mathcal{W}$  and  $\mathcal{B}$  are the complex number weight and bias, and  $W$  and  $B$  are the weight and bias in the spatial domain.

**PROOF.** Suppose that we conduct operations in the  $N$  (i.e., token dimension), then

$$\mathcal{F}(\mathbf{H}(v) * W(v)) = \int_{-\infty}^{\infty} (\mathbf{H}(v) * W(v)) e^{-j2\pi f v} dv \tag{6}$$

According to convolution theorem,  $\mathbf{H}(v) * W(v) = \int_{-\infty}^{\infty} (\mathbf{H}(\tau) W(v - \tau)) d\tau$ , then

$$\begin{aligned}
 \mathcal{F}(\mathbf{H}(v) * W(v)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathbf{H}(\tau) W(v - \tau)) e^{-j2\pi f v} d\tau dv \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(v - \tau) e^{-j2\pi f v} dv \mathbf{H}(\tau) d\tau
 \end{aligned} \tag{7}$$

Let  $x = v - \tau$ , then

$$\begin{aligned}
 \mathcal{F}(\mathbf{H}(v) * W(v)) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(x) e^{-j2\pi f(x+\tau)} dx \mathbf{H}(\tau) d\tau \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(x) e^{-j2\pi f x} e^{-j2\pi f \tau} dx \mathbf{H}(\tau) d\tau \\
 &= \int_{-\infty}^{\infty} (\mathbf{H}(\tau) *) e^{-j2\pi f \tau} d\tau \int_{-\infty}^{\infty} (W(x) *) e^{-j2\pi f x} dx \\
 &= \mathcal{H}(f) \mathcal{W}(f)
 \end{aligned} \tag{8}$$

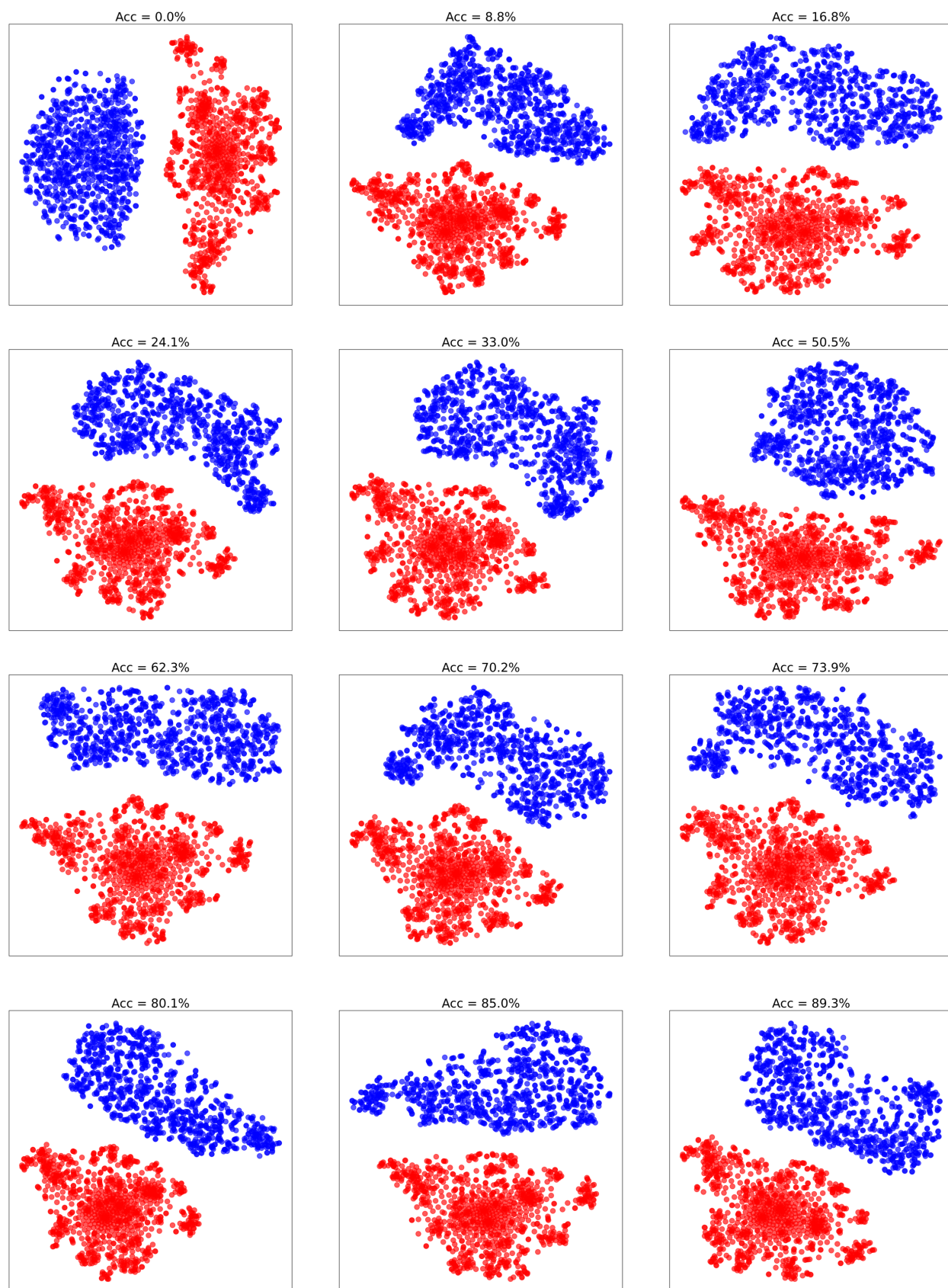
Accordingly,  $\mathbf{H}(v) * W(v)$  in the spatial domain is equal to  $\mathcal{H}(f) \mathcal{W}(f)$  in the frequency domain. Therefore, the operations of FreMLP ( $\mathcal{H}\mathcal{W} + \mathcal{B}$ ) in the token dimension (i.e.,  $v = N$ ) are equal to the operations ( $\mathcal{F}(\mathbf{H} * W + B)$ ) in the spatial domain. This implies that frequency-domain MLPs can be viewed as global convolutions in the spatial domain. Proved.  $\square$

### D.3 Complexity Analysis

For a fMRI voxel with a length of  $l$ , we divide it into  $n$  patches. Assuming  $L_1$  and  $L_2$  are the layer depths of MLP Backbone and DFT Backbone respectively, the middle layer dimension of MLP Backbone is  $D$ , and the alignment embedding dimension is  $n' \times D'$ , where  $n'$  is the the number of tokens of CLIP. The time complexity of MLP Backbone is  $O(ID + L_1 D^2 + n' D D')$ . For DFT Backbone, the time complexity of patchify and tokenization is  $O(ID')$ , and the time complexity of DFT, IDFT, and filtering for each layer is  $O(2nD' \log n + nD')$ . The time complexity of FreMLP is  $O(2nD' \log n + 2nn'D' + 2n'D')$ . Thus the time complexity of the entire DFT Backbone is:

$$\begin{aligned}
 &O(4nD' \log n + (n + 2nn' + 2n')D') \\
 &= O((n \log n + nn' + n')D')
 \end{aligned} \tag{9}$$

Quantitative analysis algorithm complexity for DFT Backbone has been shown in Table 3.



**Figure 13: T-SNE visualization between the voxel embeddings output by DFT Backbone and the image embedding of frozen CLIP. Accuracy in the title refers to the hit rate of correct retrieval from 982 test set images and the blue dots represent voxel embeddings.**