From Flat to Round: Redefining Brain Decoding with Surface-Based fMRI and Cortex Structure

Supplementary Material

A. Data Preprocessing

fMRI Data. This paper is based on the Natural Scenes Dataset (NSD)[1] [1]. Since we use spherical data as model's input, we employ the data preprocessed with FreeSurfer, which is provided by the official NSD dataset. Taking subj01 as an example, the original fMRI path we use is:

```
nsddata_betas/ppdata/subj01/fsaverage
/betas_fithrf_GLMdenoise_RR/
```

The data values here are single-trial beta weights estimated by applying a general linear model (GLM) to the raw fMRI time series, representing the voxel-wise response and its correlation with visual stimuli. The data provided by the official source are registered to the FreeSurfer standard fsaverage7 surface. We resample the data to a 40,962 sphere to match the compatibility with SphericalUNet [16]. Resampling is performed using the open-source SphericalUNet [16] package. Then, we independently apply zero-score normalization for each voxel within the `train` data of a subject. The `val` and `test` data are zero-centered using the mean and variance from the `train` data. The data split follows the standard setup used in previous works.

Cortex Structure Data. Still taking subj01 as an example, our cortical structural data comes from the path:

```
nsddata/freesurfer/subj01/surf
```

We use four types of structural information: cortical thickness, surface area, sulcal depth, and curvature. Taking the left hemisphere as an example, we use the following four files: `lh.thickness`, `lh.area`, `lh.sulc`, and `lh.curv`. We apply the same method as with the fMRI data to resample it to the fsaverage6 surface. Then, we performed zero-score normalization on each individual file (*i.e.*, each hemisphere of each subject).

Image-Text Pair. All images in the NSD dataset are derived from the MS-COCO [5] dataset. The text annotations we use are from the official MS-COCO[2] dataset.

[1] https://naturalscenesdataset.org
[2] http://images.cocodataset.org

B. Technical Details

Random Rotation of Sphere Positional Emb. The pseudocode is shown in Algorithm B-1. In the algorithm, we apply Rodrigues' rotation formula, a mathematical tool used for rotating 3D coordinates in 3D Euclidean space.

Algorithm B-1: Random Rotation Augmentation of Sphere Positional Embedding

---

input : Original spherical coordinates $\mathbf{x} \in \mathbb{R}^{n\times 3}$
Maximum rotation angle $\theta_{\max}$
output: Augmented coordinates $\mathbf{x}' \in \mathbb{R}^{n\times 3}$

get a random rotation axis $\mathbf{v}$
$\phi \sim \mathcal{U}(0, 2\pi)$
$\varphi \sim \mathcal{U}(0, \pi)$
$\mathbf{v} = (v_x, v_y, v_z) = (\sin\phi\cos\varphi, \sin\phi\sin\varphi, \cos\phi)$
get a random rotation angle θ
$\theta \sim \mathcal{U}(0, \theta_{\max})$
apply Rodrigues' rotation formula
$\mathbf{K} = \begin{pmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{pmatrix}, \mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
$\mathbf{R} = \mathbf{I} + \sin\theta \cdot \mathbf{K} + (1 - \cos\theta) \cdot \mathbf{K}^2$
$\mathbf{x}' = \mathbf{x}\mathbf{R}$
return $\mathbf{x}'$

Positive Sample Mixup. We present the pseudocode for positive sample mixup augmentation in Algorithm B-2. To uniformly sample a point $\mathbf{w}$ as the mixup weight in the convex polytope defined by equation (3), we sample a point from the 3-dimensional Dirichlet distribution $\mathcal{D}$ with parameters $\mathbf{a} = (1, 1, 1)$. The N-dimensional Dirichlet distribution is the N-dimensional generalization of the Beta distribution, and its probability density function is given by:

$$p_{\mathcal{D}}(\mathbf{w}|\mathbf{a}) = \frac{1}{B(\mathbf{a})} \prod_{n=1}^{N} w_n^{a_n - 1}. \tag{B-1}$$

Here, $B(n)$ is the normalization factor, ensuring that the probability density function integrates to 1 over the domain:

$$B(\mathbf{a}) = \frac{\prod_{n=1}^{N} \Gamma(a_n)}{\Gamma\left(\sum_{n=1}^{N} a_n\right)}. \tag{B-2}$$

Algorithm B-2: Positive Sample Mixup

input : Three fMRI scans $\mathbf{x} = (x_1, x_2, x_3)$
Mixup ratio λ
Mixup number K
output: Augmented samples $\tilde{x}_1, \tilde{x}_2, \cdots$

$t \sim \mathcal{U}(0, 1)$
if $t < \lambda$ **then**
  for $k \leftarrow 1$ **to** K **do**
    $\mathbf{a} = (1, 1, 1)$
    $\mathbf{w} \sim \mathcal{D}(\mathbf{a})$ # Dirichlet Dist., Eq. (B-1)
    $\tilde{x}_k = \mathbf{w} \cdot \mathbf{x}$
  return $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_K$
else
  return x_1, x_2, x_3

Multi Positive Samples InfoNCE. We modify the InfoNCE [7] to accommodate the scenario where multiple positive samples exist within a batch. The algorithm for multi positive samples InfoNCE is in Algorithm B-3.

Algorithm B-3: $\mathcal{L}_{\texttt{Info}}$ for Multi Positive Samples

input : Query $\mathbf{q} \in \mathbb{R}^n$
Positive key $\mathbf{k} \in \mathbb{R}^n$
Temperature coefficient τ
Image index $\mathbf{I} \in \mathbb{N}^n$
output: The InfoNCE loss l

$\mathbf{q} = \mathbf{q}/||\mathbf{q}||$
$\mathbf{k} = \mathbf{k}/||\mathbf{k}||$
$\mathbf{W} = \mathbf{q} \cdot \mathbf{k}^\top / \tau$ # $\mathbf{W} \in \mathbb{R}^{n \times n}$
Let sample mask $\theta_{ij} = \begin{cases} 1 & \text{if} \quad \mathbf{I}_i = \mathbf{I}_j \\ 0 & \text{if} \quad \mathbf{I}_i \neq \mathbf{I}_j \end{cases}$
$\mathbf{e}_i^{\texttt{all}} = \log \sum_{j=1}^{n} \exp\left(\mathbf{W}_{ij}\right)$ # $\mathbf{e}^{\texttt{all}} \in \mathbb{R}^n$
$\mathbf{e}_i^{\texttt{pos}} = \log \sum_{j=1}^{n} \theta_{ij} \exp\left(\mathbf{W}_{ij}\right)$ # $\mathbf{e}^{\texttt{pos}} \in \mathbb{R}^n$
$l = \sum_{i=1}^{n} \left(\mathbf{e}_i^{\texttt{all}} - \mathbf{e}_i^{\texttt{pos}}\right) / n$
return l

C. Training Details

Sphere Tokenizer. The training is conducted on a single NVIDIA A800 80GB GPU. The hyperparameters are shown in Tab. C-1. We use a cosine learning rate scheduler during training. To improve the robustness of the model, we introduce two data augmentation techniques: mixup with parameters `ratio=0.3` and `beta=0.3`, and the random rotation augmentation (maximum rotation angle: 5°) for position condition mentioned in our paper §3.2. To make the model focus more on visual brain regions, we only compute the loss for visual voxels.

fMRI Encoder. The training is conducted on a single NVIDIA A800 80GB GPU. The hyperparameters are shown in Tab. C-1. We use a cosine learning rate scheduler during training. To improve the model's generalization ability, we applied augmentation to the images. Each training image has a 50% chance of being randomly horizontally flipped. The color properties are perturbed as follows: brightness, contrast, and saturation with a maximum perturbation of 0.2, and hue with a maximum perturbation of 0.1. Then the image is randomly rotated by up to 30° and scaled within the range of [0.8, 1.0]. The temperature coefficient of $\mathcal{L}_{\texttt{biInfo}}$ is 0.1. To avoid the influence of outliers, the CLIP embeddings are clamped to the range [-1.5, 1.5].

D. More Results

Sphere Tokenizer. We present more reconstruction results of the sphere tokenizer autoencoder in Fig. D-1, which demonstrates its effectiveness.

Comparison to Previous Work. We report the quantitative evaluation metrics in Tab. 1. The results for Mind-Vis [2] and MindEye [10] are cited from the report in [9], while the result for UMBRAE [12] is from the report in [11]. The authors of NeuroPictor [4] fine-tune the model for each subject to achieve higher performance. However, to ensure a fair comparison with other methods, we cite and report the version w/o fine-tuning. The remaining methods are cited from their respective original papers.

Quantitative and Qualitative Results. We present the quantitative results on each subject of NSD [1] `test` in Tab. D-1. We also provide additional decoding results in Fig. D-2, D-3, and D-4. Readers can download all reconstruction images results from here: `https://huggingface.co/datasets/yusijin/sphere_tokenizer_results_NSD`.

Inference	Low-Level				High-Level			
	`PixCorr` ↑	`SSIM` ↑	`AlexNet(2)` ↑	`AlexNet(5)` ↑	`Inception` ↑	`CLIP` ↑	`EffNet-B` ↓	`SwAV` ↓
subj01	0.172	0.314	78.6%	88.7%	84.8%	88.9%	0.736	0.396
subj02	0.167	0.302	77.7%	89.0%	85.9%	88.2%	0.733	0.394
subj05	0.163	0.305	78.6%	90.1%	86.4%	89.6%	0.717	0.393
subj07	0.157	0.298	78.0%	88.3%	83.2%	86.7%	0.746	0.409
Average	0.165	0.305	78.2%	89.0%	85.1%	88.3%	0.733	0.398

Table D-1. Quantitative results for each subject on NSD [1] `test`.

Hyperparameters	**Value**
encoder channels	[64, 128, 256]
dncoder channels	[32, 64, 128]
hidden channels	32
encoder ResNet blocks per down layer	4
decoder ResNet blocks per down layer	2
batch size	32
learning rate	4.0e-5
weight decay	0.05
max gradient norm	0.1
epoch	80

Table C-1. Hyperparameters for training sphere tokenizer.

Hyperparameters	**Value**
embedding dim	1024
MLP ratio	4
depth	24
num heads	16
projection dropout	0.5
batch size	64
learning rate	5.0e-4
weight decay	0.05
max gradient norm	0.5
epoch	30

Table C-2. Hyperparameters for training fMRI encoder.

E. Limitations and Future Work

Although we propose a novel approach for vision brain decoding, there are limitations that should be acknowledged.

Low fMRI Resolution. Our model is trained on low fMRI resolution, constrained by the spherical resolution supported by the standard SphericalUNet. Experimental results demonstrate that fMRI resolution has a significant impact on performance. In the future, we plan to (1) use higher-resolution spherical convolutions and (2) adjust the architecture of the sphere tokenizer to more efficiently handle low-resolution fMRI data.

Constrained Dataset Size. Similar to most previous work, we have only validated our results on the NSD dataset. Although current techniques can handle cross-subject decoding, they are far from achieving cross-dataset generalization. To achieve more robust vision brain decoding, larger-scale datasets are required.

Challenges in Practical Application. Our model is based on fMRI, which requires stringent conditions and high costs for data collection. This limits the widespread adoption and application of brain decoding technology. Some technologies that are more suitable for real-time brain activity recording, such as EEG [6] and fNIRS [3, 15], fall far short of fMRI decoding performance due to their low signal-to-noise ratio. This highlights the need for more efficient methods to enable models to capture representations of brain activity.

Challenges in Neuroscience. The scientific community still has no definitive understanding of the detailed mechanisms behind the functioning of the human brain. The development of brain decoding can provide novel perspectives on this, highlighting the significance of biological interpretability in brain decoding models. We will conduct a deeper exploration of this.

Privacy and Security Considerations. Personal brain activity data is highly sensitive private information, so protecting data security is crucial. For example, membership inference attacks [14] and model inversion attacks [8, 13, 17] can cause serious privacy leaks during the inference stage of the model. We will incorporate considerations of data security in the future.

Figure D-1. More fMRI vision voxels reconstruction results using the sphere tokenizer on the NSD [1] `test`.

Figure D-2. More fMRI-image reconstruction results on the NSD [1] `test`.

Figure D-3. More fMRI-image reconstruction results on the NSD [1] `test`.

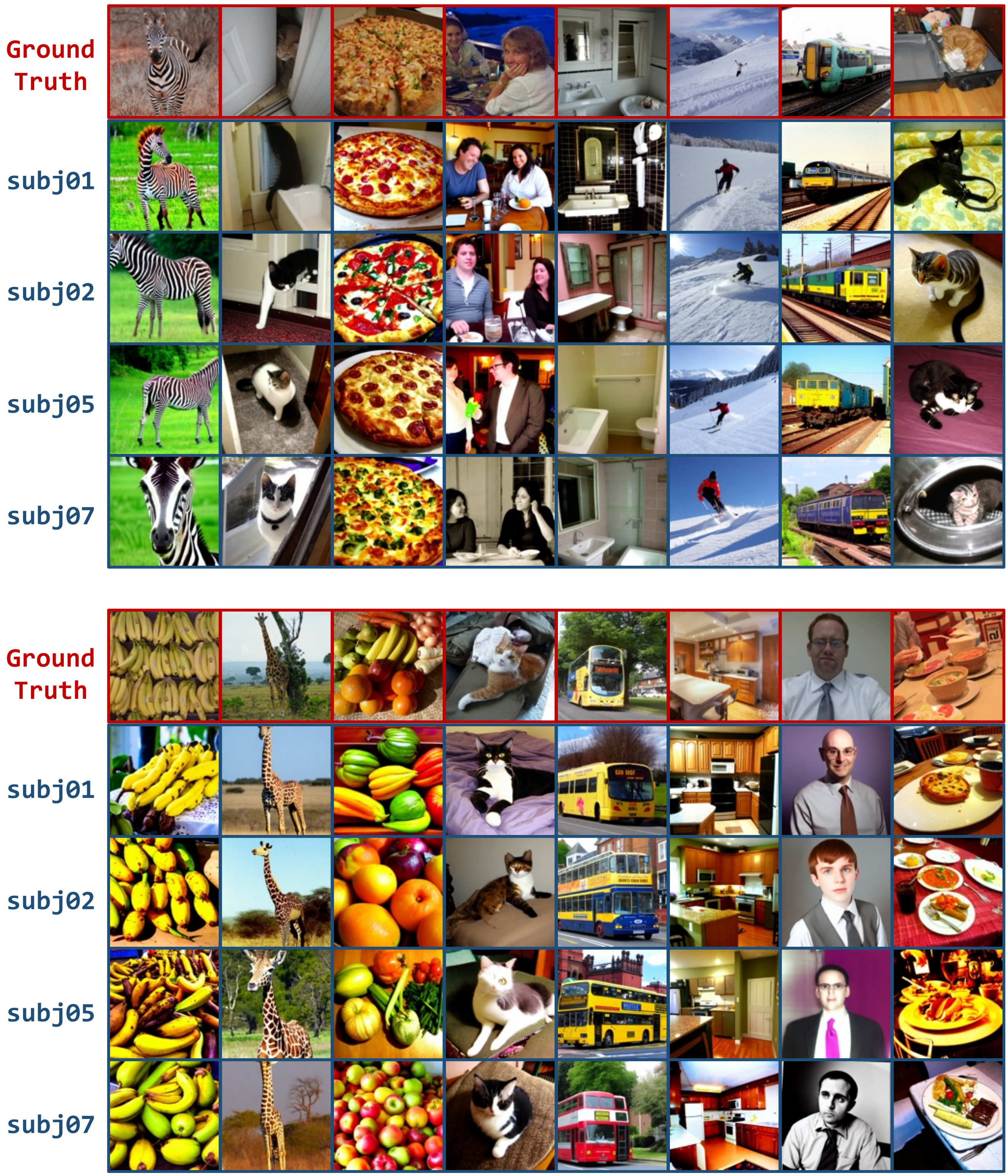

Figure D-4. More fMRI-image reconstruction results on the NSD [1] `test`.

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 1, 2, 3, 4, 5, 6, 7

[2] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023. 2

[3] Zhaojin Chen, Sijin Yu, Xuejiao Li, Huirong Lei, Jiyu Qian, Yingxue Yao, Zicong Zheng, Guodong Liang, Xiaofen Xing, Xin Zhang, et al. Sss: Signature-sequence-statistical model for exploring the impact of prenatal depression on newborns' brain using fnirs. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025. 3

[4] Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pages 56–73. Springer, 2025. 2

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[6] Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024. 3

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[8] Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *European Conference on Computer Vision*, pages 109–126. Springer, 2024. 3

[9] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 233–243, 2024. 2

[10] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[11] Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2025. 2

[12] Weihao Xia, Raoul de Charette, Cengiz Oztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *European Conference on Computer Vision*, pages 242–259. Springer, 2025. 2

[13] Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, Sijin Yu, Bin Wang, Shu-Tao Xia, and Ke Xu. Calor: Towards comprehensive model inversion defense. *arXiv preprint arXiv:2410.05814*, 2024. 3

[14] Hongyao Yu, Yixiang Qiu, Yiheng Yang, Hao Fang, Tianqu Zhuang, Jiaxin Hong, Bin Chen, Hao Wu, and Shu-Tao Xia. Icas: Detecting training data from autoregressive image generative models. *arXiv preprint arXiv:2507.05068*, 2025. 3

[15] Sijin Yu, Xuejiao Li, Huirong Lei, Yingxue Yao, Zhaojin Chen, Zicong Zheng, Guodong Liang, Xiaofen Xing, Xin Zhang, and Chengfang Xu. Attention-based-features-fusion emotion-guided fnirs classification network for prenatal depression recognition. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 12–23. Springer, 2024. 3

[16] Fenqiang Zhao, Shunren Xia, Zhengwang Wu, Dingna Duan, Li Wang, Weili Lin, John H Gilmore, Dinggang Shen, and Gang Li. Spherical u-net on cortical surfaces: methods and applications. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 855–866. Springer, 2019. 1

[17] Tianqu Zhuang, Hongyao Yu, Yixiang Qiu, Hao Fang, Bin Chen, and Shu-Tao Xia. Stealthy shield defense: A conditional mutual information-based approach against black-box model inversion attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. 3