YEDITEPE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

IMAGE RECONSTRUCTION FROM FMRI ACTIVITY IN VISUAL CORTEX

A Thesis Submitted

by

Handenur Genç

In Partial Fulfillment

of the Requirements for the Degree of

Master of Science

in

Biomedical Engineering

Supervisor

Assist. Prof. Dr. Andaç Hamamcı

Istanbul - 2024

IMAGE RECONSTRUCTION FROM FMRI ACTIVITY IN VISUAL CORTEX

by

Handenur Genç

Approved by:

Assist. Prof. Dr. Andaç Hamamcı
(Yeditepe University) ……
(Thesis Supervisor)

Assist. Prof. Dr. İpek Baz
(Yeditepe University) ……

Assist. Prof. Dr. Ali Bayram
(İstanbul University) ……

DATE OF APPROVAL: …. / …. / 20….

DECLARATION OF ORIGINALITY

I hereby declare that this thesis is my own work and that all information in this thesis has been obtained and presented following academic rules and ethical conduct. I have fully cited and referenced all material and results as required by these rules and conduct, and this thesis study does not contain any plagiarism. The necessary permissions have been obtained if any material used in the thesis requires copyright. No material from this thesis has been used to award another degree.

To the best of my knowledge and belief, it contains no material previously published or written by another person nor material accepted for the award of any other degree except where due acknowledgment has been made in the text.

I accept all kinds of legal liability that may arise in cases contrary to these situations.

Handenur Genç

……

ABSTRACT

IMAGE RECONSTRUCTION FROM FMRI ACTIVITY IN VISUAL CORTEX

Recently, the importance of using fMRI data to identify neural activity in the brain has increased significantly in neuroscience. The core principle of fMRI is blood-oxygen-level-dependent (BOLD) contrast, which occurs from a mismatch between blood flow and oxygen metabolism during localized brain activity. Studies of how the brain interprets complex visual objects and the perception of subjects are critical to understanding how the brain works. It's becoming more realistic to gain insights into how the brain perceives the external world quantitatively. Deep learning models and computer vision technologies have significantly contributed to this understanding with encouraging findings from studies. The convergence of quantitative cognitive process analysis and cutting-edge computational models in this context offers novel opportunities to understand the complex connections between brain function and perceptual decision-making. This study focuses on reconstructing visual objects from fMRI activation patterns that occur when visual stimuli are presented in arbitrary object categories. The BOLD5000 dataset that contains about 5,000 distinct slow event-related fMRI scans, was used to assess the applicability of statistical learning methodologies that involve neuroscience. A cortical surface model was extracted for each subject to investigate a relationship between fMRI activity and perceived stimuli. Specifically, occipital surface models were created to capture the complex geometry of the visual cortex, which plays a crucial role in processing visual information. These models were designed to address the spatial resolution challenges inherent in volumetric fMRI data by providing a more accurate representation of the cortical surface. This paradigm purposely hides the intrinsic spatial relationships present in the stimulus image. This study uses occipital surface models obtained from fMRI data to reconstruct visual images using latent diffusion model (LDM).The performance recorded AlexNet(2) and AlexNet(5) 2-way comparison scores of 93.6% as low-level metrics, while the high-level metric InceptionV3 recorded a score of 49.5% for the reconstructed images in surface-based with cross attention modality.

ÖZET

GÖRSEL KORTEKSTEKİ FMRI AKTİVİTİTESİNDEN GÖRÜNTÜ YENİDEN YAPILANDIRILMASI

Fonksiyonel Manyetik Rezonans Görüntüleme (fMRI) verilerine dayalı beyindeki nöronal aktivitenin tanımlanması, son yıllarda sinir bilim araştırmalarında önemli bir yer edinmiştir. fMRI temelde, lokal beyin aktivasyonu sırasında kan akışı ile oksijen metabolizması (BOLD) arasındaki değişikliklerden kaynaklanan beyindeki nöronal aktiviteyi saptamaya dayalı bir ölçüm yöntemdir. Karmaşık görsel nesneleri nasıl algılandığı ve beyin tarafından işlendiğini anlamak, beyin işleyişini kavramak için kritiktir. İnsanların dünyayı algılama süreçlerinin nöral aktiviteden niceliksel bir anlayışa sahip olmak giderek daha gerçekçi hale gelmektedir. Derin öğrenme modelleri ve bilgisayar görü, bu alandaki çalışmalara önemli katkılarda bulunmuş ve umut verici bulgular elde edilmiştir. Sinir bilim, niceliksel bilişsel işlem analizi ve hesaplamalı model gelişmelerinin bu birleşimi, beyin aktiviteleri ile algılar arasındaki karmaşık ilişkiyi anlama konusunda yeni fırsatlar sunmaktadır. Bu çalışma, rastgele nesne kategorilerinde görsel uyaranlar olarak sunulduğunda meydana gelen fMRI beyin aktivasyon desenlerinden görsel nesnelerin yeniden yapılandırılmasına odaklanmaktadır. Bu çerçevede istatistiksel öğrenme metodolojilerinin, nörobilimle ilgili değerlendirme bağlamında ne kadar etkili olduğunu incelemek için yaklaşık 5,000 ayrı yavaş olayla ilişkilendirilmiş fMRI taramalarını içeren BOLD5000 veri kümesi kullanılmıştır. Her bir katılımcı için bir kortikal yüzey modeli çıkarılarak fMRI beyin aktivitesi ile algılanan uyaranlar arasında bir ilişki kurulmuştur. Özellikle, uyaran görüntüsünde mevcut olan içsel uzamsal ilişkileri koruyarak görsel korteksin karmaşık geometrisini yakalayarak hacimsel fMRI verilerindeki mekansal çözünürlük zorluklarını çözmek oksipital yüzey modelleri önerilmiştir. fMRI verilerinden elde edilen oksipital yüzey modelleri kullanılarak görsel görüntülerin latent difüzyon modeli (LDM) ile yeniden yapılandırılmasını önermektedir. Modelin performansı, düşük seviye metriklerde AlexNet(2) ve AlexNet(5) ile yapılan iki yönlü karşılaştırmalarda %93,6 puan kaydederken, yüksek seviye metriklerde InceptionV3 ile yeniden yapılandırılan görüntüler için %49,5 puan elde etmiştir.

ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to my advisor, Assist. Prof. Dr. Andaç Hamamcı, for his invaluable guidance, support, and dedication throughout my academic journey. His insightful mentorship and unwavering commitment have been instrumental in shaping this thesis. I am truly grateful for the knowledge and encouragement he provided, which greatly contributed to completing this work.

I also want to express my heartfelt appreciation to my family and husband for their unconditional love, support, and sacrifices. Your emotional and financial support have been my pillars of strength, and your unwavering encouragement has motivated me throughout this journey. This achievement would not have been possible without your belief in me, and I am forever grateful for all you have done to help me reach this milestone.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY iii
ABSTRACT iv
ÖZET v
ACKNOWLEDGEMENTS vi
TABLE OF CONTENTS vii
LIST OF FIGURES x
LIST OF TABLES xvi
LIST OF ABBREVIATIONS xvii
1. INTRODUCTION 1
1.1. PROBLEM STATEMENT 3
1.2. RELATED WORKS 3
1.3. BACKGROUND 5
1.3.1. Retinotopic Mapping and Brodmann Areas 5
1.3.2. fMRI Setup: Visual Stimulation Experiment 10
1.3.3. Cortical Surface Model 12
1.3.4. Diffusion Models 16
1.4. CONTRIBUTION OF THE THESIS 20
2. METHODOLOGY 21
2.1. DATASET 21
2.1.1. Participants 21
2.1.2. Visual Stimulus and Experimental Design 21
2.1.3. fMRI Data Acquisition 23
2.2. DATA PRE-PROCESSING 23
2.2.1. DICOM Mosaics to Patient Coordinate System Mapping 23
2.2.2. FreeSurfer Single Subject Pipeline Analysis Overview 24
2.2.3. fMRI Analysis 28
2.2.4. Volume to Surface Mapping using Freesurfer Functions 32
2.2.5. Occipital Patch Removal and Flattening 33
2.2.6. Extracting Flatmap Surfaces 35
2.3. ARCHITECTURE OVERVIEW 36

2.3.1. Image to Image VQGAN First Stage Model 38
2.3.2. FMRI to Image VQGAN Network Model 39
2.3.2.1. Surface-Based Network Model 39
2.3.2.2. Volume-Based Network Model 39
2.3.3. Latent Diffusion Models 40
2.3.3.1. Cross-attention conditioning 40
2.3.3.2. Concatenation conditioning 41
2.4. IMPLEMENTATION DETAILS 42
2.4.1. EXPERIMENTS 43
2.4.1.1. Evaluation Metrics 43
2.4.1.2. Average BOLD Signal Analysis and Fourier Transform Analysis for Slow Event-Related Design 44
2.4.1.3. Evaluation Methodology 45
2.4.1.4. Evaluation of Image to Image VQGAN Model Evaluation 45
2.4.1.5. Evaluation of Cross-Attention and Concatenation Mechanisms in Surface-Based Models 45
2.4.1.6. Evaluation of Volume-Based Model and Surface-Based Model 45
2.4.1.7. Evaluation of fMRI Encoder: Frozen and Trainable Configurations 46
3. RESULTS 47
3.1. AVERAGE BOLD SIGNAL AND FOURIER ANALYSIS 47
3.2. QUALITATIVE RESULTS OF FIRST STAGE AND LATENT DIFFUSION MODELS WITH DIFFERENT MODALITIES 52
3.2.1. Image to Image VQGAN Model 52
3.2.2. fMRI to Image VQGAN Models with Different Modalities 53
3.2.2.1. Surface-Based Model 53
3.2.2.2. Volume-Based Model 55
3.2.3. Conditioning Mechanisms in LDM: Cross-Attention and Concatenation 56
3.2.3.1. Surface-Based Model with Cross Attention Mechanism 56
3.2.3.2. Surface-Based Model with Concatenation Mechanism 57

3.2.3.3. Surface-Based Model with Trainable Conditioning Concatenation Mechanism.....59
3.2.3.4. Volume-Based Model with Concatenation Mechanism 61
3.3. QUANTITATIVE RESULTS OF FIRST STAGE AND FMRI VQGAN MODEL WITH DIFFERENT MODALITIES 63
3.4. QUANTITATIVE RESULTS OF LATENT DIFFUSION MODEL WITH DIFFERENT MODALITIES 64
3.5. QUALITATIVE COMPARISON OF RESULTS WITH EXISTING STUDIES IN LITERATURE 66
3.6. QUANTITATIVE COMPARISON OF RESULTS WITH EXISTING STUDIES IN LITERATURE 71
4. DISCUSSION 73
5. CONCLUSIONS 75
REFERENCES.....76
APPENDIX A.....85
APPENDIX B 86

LIST OF FIGURES

Figure 1.1. Illustrates the vertical and horizontal signal travel after reaching the V1 layer. (left) Diagram of the visual fields and pathway. (right) 6

Figure 1.2. Stimuli: Rotating sectors (a) Ring contraction or expansion (b) (left) Maps of angle (polarity) and eccentricity (mid) Retinotopic visual areas (right) 7

Figure 1.3. TWR Measurements. High-contrast checkerboard patterns that move smoothly and periodically through polar angles (wedge) or eccentricity (ring) are used as traveling wave stimuli. An expanded view of this surface near the calcarine sulcus is overlaid with a color map showing the response phase at each location for polar angle (A) and eccentricity experiments (B). The solid white lines indicate the boundaries of V1 in the calcarine sulcus. 7

Figure 1.4. (A) The human visual cortex (orange overlay) occupies about 20% of the cerebral cortex and is located in the occipital lobe and posterior parts of the parietal and temporal lobes. A map of the visual field can be found in V1, which is located in and around the calcarine sulcus (dotted line). (B) VFM illustration in V1. This image is part of Sir Isaac Newton's 1989 portrait by Godfrey Kneller. The figure illustrates how the visual field (left) is transformed and represented on the V1 cortical surface (right) using a mathematical description proposed by Schwartz (1977). The LVF stimulates V1 in the RH; the image representation is inverted, and the center of the visual field, near the eye, is greatly expanded (cortical magnification). 8

Figure 1.5. Map of retinal eccentricity (a) Cortical magnification varies with eccentricity. (b) For visual regions V1, V2, and V4, the receptive field size (diameter) is a function of the receptive field center (eccentricity). A "hinged" line well describes the size-to-eccentricity relationship in each region. (c) A cartoon illustration of receptive fields with sizes based on physiological measurements. The center of each array is the fovea. The size of each circle is proportional to its eccentricity based on the corresponding scaling parameter (slope of the fitted line in c). A larger scaling parameter indicates larger receptive fields at a given eccentricity. The model used overlapping pooling regions (linear weighting functions) to tile the image uniformly. They are separable and of constant size when expressed in polar angle and log eccentricity. (d) [1] 9

Figure 1.6. Brodmann areas of the brain (left) lateral surface (right) medial surface 10

Figure 1.7. fMRI setup: Patient responds to the visual/auditory stimuli using the response box 11

Figure 1.8. The triangular tessellation's metric properties 13

Figure 1.9. Flattened left hemispheres 15

Figure 1.10. Three LHs in a lateral view after morphing. 16

Figure 1.11. The forward diffusion process's Markov chain generates a sample by gradually adding noise. 17

Figure 1.12. The reverse diffusion process Markov chain generates a sample by gradually removing noise. 18

Figure 2.1. Schematic representation of the task paradigm. "On" denotes the presentation of the stimuli, and "Off" denotes no presentation of the stimuli. A fixation cross centered on a blank shown for 6 at the beginning and 12 seconds at the end of each run. 23

Figure 2.2. FreeSurfer subject analysis recon-all pipeline overview 25

Figure 2.3. Orig and corrected orig surface (white surface) 26

Figure 2.4. Recon-all process outputs 28

Figure 2.5. The white and pial surfaces reconstructed with recon-all provided by FreeSurfer and positioned on the participant's T1W template. 29

Figure 2.6. Results of using SDC on the EPI 30

Figure 2.7. Brain mask calculated on the BOLD signal (red contour) and the masks used for a/tCompCor. The aCompCor mask (magenta contour) is a conservative CSF and WM mask for extracting physiological and movement confounds. The fCompCor mask (blue contour) contains the top 5% most variable voxels within a heavily eroded brain mask. 30

Figure 2.8. Bbregister was employed to generate transformations from EPI-space to T1W-space 31

Figure 2.9. Freeview graphical interface 32

Figure 2.10. Surface overlays with Brodmann areas V1 and V2 33

Figure 2.11. Cutting the occipital patch from the surface and flattening. (both left and right hemispheres) The same process performed for all participants. 34

Figure 2.12. Data acquisition by interpolation of BOLD Occipital surfaces with related a stimulus hemodynamic response 35

Figure 2.13. A schematic representation of Image to Image VQGAN model. 38

Figure 2.14. A schematic representation of surface-based VQGAN model. 39

Figure 2.15. A schematic representation of volume-based VQGAN model. 40

Figure 2.16. A schematic representation of cross attention mechanism. 41

Figure 3.1. Overall average signal with all stimulus onsets for runs 1-2 of Session-01 for participant CSI1 47

Figure 3.2. Average signal with first 4 stimulus onsets for Session-01 Runs 1-2 for participant CSI1 48

Figure 3.3. BOLD Signal with 10 secs windows of stimulus presentations and FCs (excluding initial and final FCs (left col: LH right col: RH)....49

Figure 3.4. Frequency Components of BOLD Signals in LH and RH for Session 01 - Run 1-2....50

Figure 3.5. (left) A normalized visualization to compare BOLD signals over different brain regions (V1, V2, MT) and hemispheres. For each run, the signals for both hemispheres of these 3 regions are combined into a single graph. (right) Design validation studies from original paper....51

Figure 3.6. Qualitative results of image to image first stage model on train set (left column is original stimulus) (Images in red boxes represent 4 reconstructed sample images corresponding to the original stimulus on the left) 52

Figure 3.7. Qualitative results of the image to the image first stage model on the test set (left column is original stimulus) (Images in red boxes represent 4 reconstructed sample images corresponding to the original stimulus on the left) 53

Figure 3.8. Qualitative results of the surface-based model in train dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left) 54

Figure 3.9. Qualitative results of the surface-based model in test dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left) 54

Figure 3.10. Qualitative results of the volume-based model in train dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left) 55

Figure 3.11. Qualitative results of the volume-based model in test dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left) 56

Figure 3.12. Qualitative results of the surface-based model in train dataset samples on LDM with cross attention conditioning (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)....57

Figure 3.13. Qualitative results of the surface-based model in test dataset samples on LDM with cross attention conditioning (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)....57

Figure 3.14. Qualitative results of the surface-based model in train dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)....58

Figure 3.15. Qualitative results of the surface-based model in test dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)....59

Figure 3.16. Qualitative results of the surface-based model in the train set samples on LDM with trainable concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left) 60

Figure 3.17. Qualitative results of the surface-based model in test set samples on LDM with trainable concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)....61

Figure 3.18. Qualitative results of the volume-based model in train dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 6 reconstructed sample images corresponding to the original stimulus on the left)....62

Figure 3.19. Qualitative results of the volume-based model in test dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 6 reconstructed sample images corresponding to the original stimulus on the left)....62

Figure 3.20. Referenced study results from Takagi et al. The ground truth images are the original stimuli. 'Z' columns are generated from single trial beta weights of fMRI activity alone, 'C' columns are generated using only text inputs, and 'Z_c' from both. 68

Figure 3.21. Referenced study results from Ozcelik et. al. The first column is the ground truth image (test set). The second column shows reconstructions from the full Brain-Diffuser model with all of its components. The third column is for reconstructions of the Only-VDVAE model. The remaining columns are for Brain-Diffuser with one of its components excluded, in order: without VDVAE, without CLIP-Text, and without CLIP-Vision. 69

Figure 3.22. Referenced study results from Usma et al. Generated images using fMRI encoder freeze. The original images are labeled in a red box, and the generated images are labeled in a blue box. All the blue box images were generated during the test using the z_{fmri} 70

Figure 3.23. Referenced study results from Usma et al. Generated images using an fMRI encoder trainable. The original images are labeled in a red box, and the generated images are labeled in a blue box. All the blue box images were generated during the test using the z_{fmri} 70

Figure A.1. Github project page of the occipital surface extraction process....85

LIST OF TABLES

Table 2.1. A general design of data per subject 22

Table 3.1. The maximum amplitude and the corresponding period values of BOLD signals from the LH and RH are determined through Fourier analysis. 50

Table 3.2. Quantitative results of the reconstructed image quality on the test set. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. 63

Table 3.3. Quantitative results of the reconstructed image quality. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. The best results for some metrics are shown in green cells. 65

Table 3.4. Decoding accuracy (Pearson's correlation coefficients) of latent representations and data quality metrics. Mean±s.e.m. across all features are shown for z and c. Ours are in red font.....71

Table 3.5. Quantitative results of the reconstructed image quality on the test set. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. Our results are in the red cells and indicated by the arrow pointing up or down, respectively.....72

LIST OF ABBREVIATIONS

I	Identity matrix
$q(h_t\|h_0)$	Forward diffusion process
$p_\theta(h_{t-1} \mid h_t)$	Reverse diffusion process
T	Sampling timestep
$h_1, \ldots, h_T$	Latent variables
$\sqrt{1-\beta_t}h_{t-1}$	Mean
$\beta_t I$	Covariance
$\mathcal{L}$	Loss Function
ϵ	Gaussian distribution
β_t	Diffusion rate
τ_θ	Conditioning mechanism
$\mathbf{e}_{kn}$	Unit vector pointing from vertex k to vertex n
J	Overall energy functional
J_d	Energy functional penalizing metric distortions
J_a	Energy functional penalizing area distortions
J_s	Energy functional for surface inflation
A_i^t	Area of the i-th triangle at time t
A_i^0	Area of the i-th triangle at $t = 0$
d_{ij}^t	Distance between vertices i and j at time t
d_{in}^0	Distance between vertices i and n on the original surface
$\frac{\partial J_a}{\partial \mathbf{x}_k}$	Gradient of J_a w.r.t vertex $\mathbf{x}_k$
$\frac{\partial J_d}{\partial \mathbf{x}_k}$	Gradient of J_d w.r.t vertex $\mathbf{x}_k$
AC/PC	Anterior commissure/Posterior commissure
AGI	Artificial general intelligence
AI	Artificial intelligence
APA	Anatomical pattern analysis
BIDS	Brain imaging data structure

BOLD	Blood oxygen level-dependent
CA	Correlation analysis
CLIP	Contrastive language-image pre-training
CMRR	Center for Magnetic resonance research
COCO	Common objects in context
CSF	Cerebrospinal fluid
DICOM	Digital imaging and communications in medicine
DDPM	Denoising diffusion probabilistic model
ECOC	Error-correcting output codes
ELBO	Evidence lower bound
EPI	Echo planar imaging
FC	Fixation cross
fMRI	Functional magnetic resonance imaging
GAN	Generative adversarial network
GLM	General linear model
GM	Gray matter
HPF	High-pass filter
Hz/Px	Hertz per pixel
ICA	Independent component analysis
INU	Intensity non-uniformity
LH	Left hemisphere
LPIPS	Learned perceptual image patch similarity
LDM	Latent diffusion model
LVF	Left visual field
MLP	Multi-layer perceptron
MNI	Montreal neurological institute
MRI	Magnetic resonance imaging
MSE	Mean squared error
MSBGM	Multi-subject bayesian generative model
MVPA	Multi variate pattern analysis
NSD	Natural scene dataset

PA	Posterior to anterior
PCC	Pearson correlation coefficient
PEPOLAR	Phase encoding POLARity
PET	Positron emission tomography
PixCorr	Pixel-wise correlation
PSNR	Peak signal-to-noise ratio
RAM	Random access memory
RH	Right hemisphere
RNN	Recurrent neural network
ROI	Region of interest
RVF	Right visual field
SEM	Standard error of the mean
SPM	Statistical parametric mapping
SRNN	Spiking recurrent neural network
SSIM	Structural similarity index metric
SUN	Scene understanding
SVM	Support vector machine
SwAV	Swapping assignments between views
T1W	T1-weighted
TE	Echo time
TR	Repetition time
TWR	Travelling wave retinotopy
WM	White matter
dof	Degree of freedom
SEM	Standard error of the mean
VDVAE	Versatile diffusion variational autoencoder
VFM	Visual field map
VGG	Visual geometry group
VQGAN	Vector quantized generative adversarial network
VQVAE	Vector quantized variational autoencoder

1. INTRODUCTION

The human brain is an extraordinary organ that challenges our current understanding with its dynamic structure and complexity. This biological structure, in which billions of neurons work in harmony by forming complex connections, shapes the essence of our existence, from the most basic reflexes to the complex interactions of thoughts and emotions. Neuroscience focuses on the idea that the activity of neuronal populations in the brain represents specific sensory inputs from the external world, investigating in-depth research better to understand the complexity of cognitive processes and emotional experiences. In this context, the human visual system emerges as the primary player in processing most external sensory information, challenging researchers to delve further into the intricate mechanics of visual information processing. Functional magnetic resonance imaging (fMRI) has significantly accelerated these studies with its ability to capture neural activity non-invasively. It has provided an invaluable tool for understanding the complex relationship between neuronal activity and the external sensory world, thereby strengthening the capability to unravel the cognitive basis of visual perception. Since fMRI provides a noninvasive window on brain activity with whole-brain coverage and spatial resolution, it is ideal for exploring representation-related challenges in the human brain. [2, 3]

Artificial intelligence (AI) has significantly impacted neuroscience, shaping how researchers study and interpret brain function. One of the primary objectives has been to develop machines capable of perceiving, learning, and reasoning like humans. This goal, known as the Turing test, aims to create Artificial general intelligence (AGI) systems that mimic human intelligence. Since beginning the development of sensors in the 1950s, this effort has led researchers to take inspiration from the mechanical and functional aspects of the brain. These early models, however simplistic, provided the framework for more complex neural networks, culminating with the introduction of backpropagation in the late 1980s.[4] Backpropagation revolutionized tasks like image and speech recognition by allowing neural networks to adjust their connections dynamically. Despite its origins in neuroscience, intense learning deviates significantly from biological plausibility in today's AI paradigm. Although backpropagation remains an effective training method, its biological significance remains unclear. However, brain-inspired AI systems promise to improve our understanding of neural

information processing.

The brain's evolutionary-honed architecture provides a blueprint for intelligence, and brain-inspired mechanisms can improve AI capabilities and efficiency. Among the various applications of AI in neuroscience, modeling working memory has become a significant focus. Working memory, essential for tasks that include temporary information storage and processing, has been effectively simulated using Recurrent Neural Networks (RNNs), which mimic the recurrent connections observed between neurons in the cerebral cortex. Notably, Spiking Recurrent Neural Networks (SRNNs) have further moved the development of this field by providing increasing energy efficiency and biological plausibility. [5]

AI has significantly advanced our understanding of the brain's visual system and its role in visual processing. Machine learning models trained to recognize neural activity patterns have achieved remarkable success in reconstructing both perceived and imagined visual images and have provided valuable information about the brain's visual processing pathways. [5–9] In particular, neural networks trained in visual tasks have demonstrated the potential to produce representations comparable to those of the brain's visual system, emphasizing the synergistic relationship between AI and neuroscience.

AI has transformed the analysis of neural data, especially in fMRI studies. Machine learning models are trained to decode brain activity models associated with specific tasks and provide in-depth information about attention processes and decision-making mechanisms. AI is critical to developing neural networks inspired by the brain to decode and interpret human visual processing. These networks have been performed with remarkable accuracy in object recognition and motion detection tasks by modeling the hierarchical processing processes in the brain's visual streams. Another significant contribution of AI to neuroscience is analyzing behavior and neural correlates. Researchers have used AI to automate the processing of large-scale brain imaging datasets, enabling them to detect and identify behavioral connections. Advanced deep learning systems are developed to categorize animal behavior automatically; in some cases, these systems outperform human accuracy. AI has contributed significantly to analyzing neuronal activity patterns and enabled individual neuron activity to be detected quantitatively.

1.1. PROBLEM STATEMENT

High-resolution image reconstruction from neuronal activity is a frontier challenge at the intersection of neuroscience and AI. Traditional visual reconstruction methods are based on direct imaging and basic pattern recognition techniques, often insufficient to capture the nuanced complexities of neural responses given to visual stimuli. The complex and noisy nature and the high dimension of neural data make it difficult for current techniques to effectively decode visual information in brain activity. In addition, traditional models cannot adequately explain the dynamic and complex neural activation patterns corresponding to specific visual experiences, leading to blurred or distorted reconstructions with insufficient resolution for practical use.

This study uses LDMs to explore more profound layers of neural encoding and accurately translate brain activity into complex visual representations. These models are designed to process and continuously improve brain activity data, increasing the fidelity of the reconstructed images. This study aims to improve the quality and resolution of images reconstructed from brain activity by combining today's deep learning algorithms with modern neuroimaging techniques and to approach the direct visualization of human cognition and perception.

The results of this study have the potential to improve the discipline of neural decoding significantly and have important implications for brain-computer interfaces, opening up new opportunities for communication and visualization for people unable to speak or perform physical tasks. Also, It may expand our understanding and improve the diagnosis of neurological disorders by providing a more detailed perspective on the visual processing abnormalities characteristic of neurological disorders.

1.2. RELATED WORKS

Studies focusing on understanding the brain activity generated by visual stimuli can be categorized as stimulus category classification [10–17], stimulus identification [18–20] and stimulus reconstruction [21–42] The stimulus category classification refers to the studies that

divide visual stimuli into predetermined categories according to the known activity patterns in the brain. These studies generally provide a general understanding of the species or classes presented. Stimulus identification is about defining individual stimuli from a known sequence, and researchers use this information by analyzing neural patterns and distinguishing the image-specific features of the presented images, such as the position, size, and angle of the images. Stimulus reconstruction includes decoding the neural signature associated with viewing experience and using it to reconstruct a visual representation corresponding to the experienced stimulus. A low-complexity level study sought to determine whether it's possible to identify previously unseen images using fMRI data collected while viewing a collection of natural images to decode or read out the mental content of visual stimuli resulting from brain activity. [10–13, 21] The researchers used a two-step methodology to do this. In the first stage, they estimated a quantitative receptive field model for each voxel in the brain. [21] The model was based on a Gabor wavelet pyramid to explain tuning along the space, orientation, and spatial frequency dimensions. This model was estimated by recording fMRI data while subjects looked at a collection of natural images. In the second stage, they used estimated receptive field patterns to determine which specific image was seen based on measured voxel activity patterns.

Another study introduced the APA framework to address instability and spatial localization concerns in task-based fMRI datasets for decoding human brain visual inputs. [14] The proposed method combines ECOC and binary classifiers. The process involves training multiple binary classifiers for each category of visual stimulus and using a coding matrix to make predictions based on the closest Hamming distance. Comparative analyses evaluate the APA framework, including alternative methodologies such as Multi-Class SVM, Multi-Layer Sensor, Selected ROI, and graph-based approaches. [14] Decoding neural activity faces difficulties in generalization due to individual variability. The cost and complexity of collecting fMRI data, limited sample sizes, and noisy datasets make accurate model training difficult. Distinguishing between volitional and spontaneous imagery and coping with a limited number of decoding outputs further compound the complexity of this task. Advanced methodologies are required to improve generalization, model accuracy, and negotiate the complexity of neural representation to overcome these challenges. The development of computational models in computer vision, particularly deep learning and GANs, has recently

opened up new opportunities to examine the visual perception mechanisms of the human brain. Early attempts using noise-based models, such as GANs, provided the framework for this study by proving the potential to reconstruct visual stimuli from neural activity. In addition to these achievements, researchers have increasingly included more complicated GAN architectures capable of generating high-resolution, photorealistic images. Integrating these advanced GAN models into frameworks for brain decoding has yielded encouraging outcomes, enabling more accurate and detailed reconstruction of perceived stimuli. A few studies have successfully implemented GAN-based algorithms to generate human faces from fMRI data and provided valuable information about the representation of the facial features of the brain. [29, 39] In many studies, existing approaches are focused on a novel approach to decoding visual stimuli by using bidirectional information flows within the human visual system to decode information from different brain areas. [43] Researchers have investigated the strategic selection of specific voxels from different regions of the visual cortex and integrated them into the decoding model under the assumption that these voxels correspond to the different responsibilities of various visual cortex regions in processing different aspects of visual information. Beyond improving decoding accuracy, the primary goal was to show the internal relationships between visual areas. This approach aimed to expand our understanding of the complex organizational dynamics that regulate the neural functioning of visual stimuli, providing valuable information about hierarchical representations within the visual cortex. [15, 16]

1.3. BACKGROUND

1.3.1. Retinotopic Mapping and Brodmann Areas

Retinotopic mapping helps to understand brain activity related to visual areas in modern neuroscience. This method maps the spatial organization of the visual cortex by focusing on how specific retinal locations correspond to stimulus activity. Through retinotopic mapping, researchers have contributed to a broader understanding of cognitive processes and emotional experiences by gathering information on how the brain interprets visual information topographically. The visual cortex is located in the occipital lobe behind the brain, which

processes, integrates, and analyzes visual information from the retina. This region is divided into five areas based on its function and structure (V1 to V5). Visual information from the retinas first passes through the thalamus, synapsing in the lateral geniculate nucleus. Afterward, this information leaves the lateral geniculate and is directed to V1, the first area of the visual cortex. V1 centered around the calcarine sulcus. It's important for receiving and processing visual information and is the best-understood part of the visual cortex. [44–46]

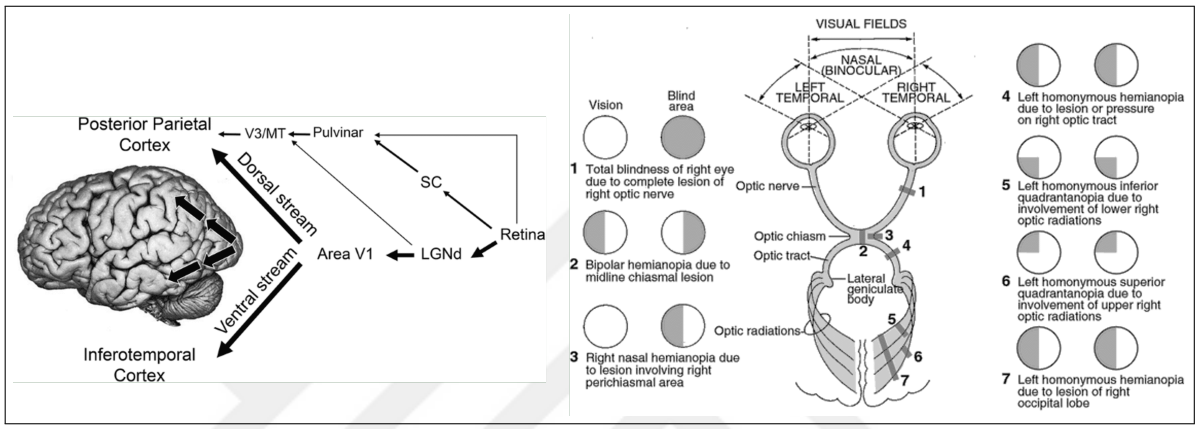

Figure 1.1. Illustrates the vertical and horizontal signal travel after reaching the V1 layer. (left) Diagram of the visual fields and pathway. (right)

V1 called the primary visual cortex, contains special cells with simple and complex cells, lines, edges, and orientations. The information these cells receive is transferred to the secondary visual cortex V2, where more complex visual features are processed. V2 is a structure that responds to increasing levels of complexity and visual stimuli, including color, spatial frequency, and object orientation, by processing data from V1. V2 receives strong forward connections from V1 and sends robust connections to V3, V4 and V5. It also sends strong feedback connections to V1. The information left from the second visual area is specialized for different processing pathways divided into dorsal and ventral flows. The dorsal stream is typically associated with object recognition, whereas the ventral stream focuses on spatial tasks and visual-motor abilities. The process includes specific groups of neurons in brain regions identified by retinotopic mapping that respond to different visual information. [3, 47–51]

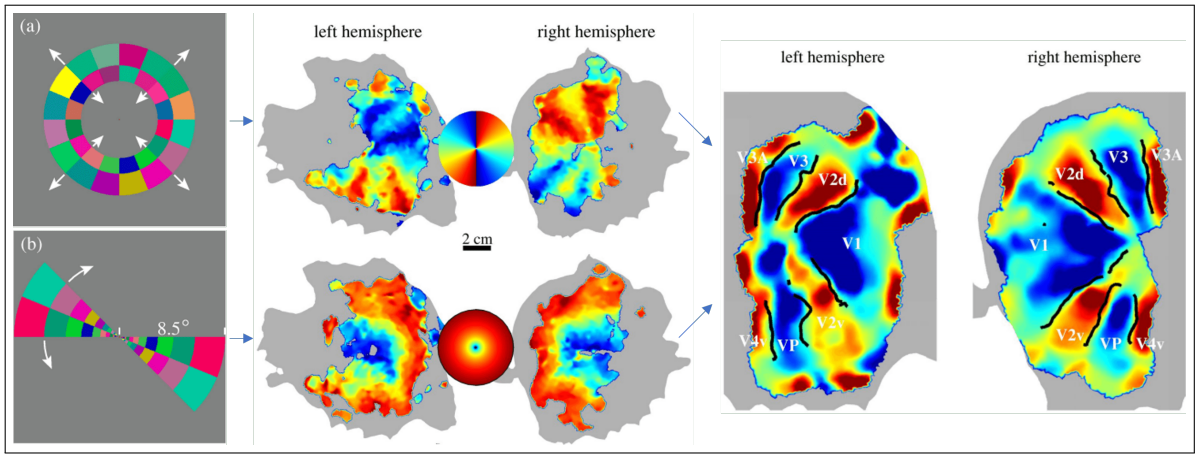

Figure 1.2. Stimuli: Rotating sectors (a) Ring contraction or expansion (b) (left) Maps of angle (polarity) and eccentricity (mid) Retinotopic visual areas (right)

Since the mid-1990s, the Travelling Wave Retinotopy (TWR) technique has been the gold standard for visual field mapping. In fMRI studies, visual stimuli with designs like expanding or contracting rings and rotating sectors are used to determine these visual field maps. This technique uses periodic stimuli that move smoothly across the visual field to measure orthogonal dimensions like polar angle and eccentricity.

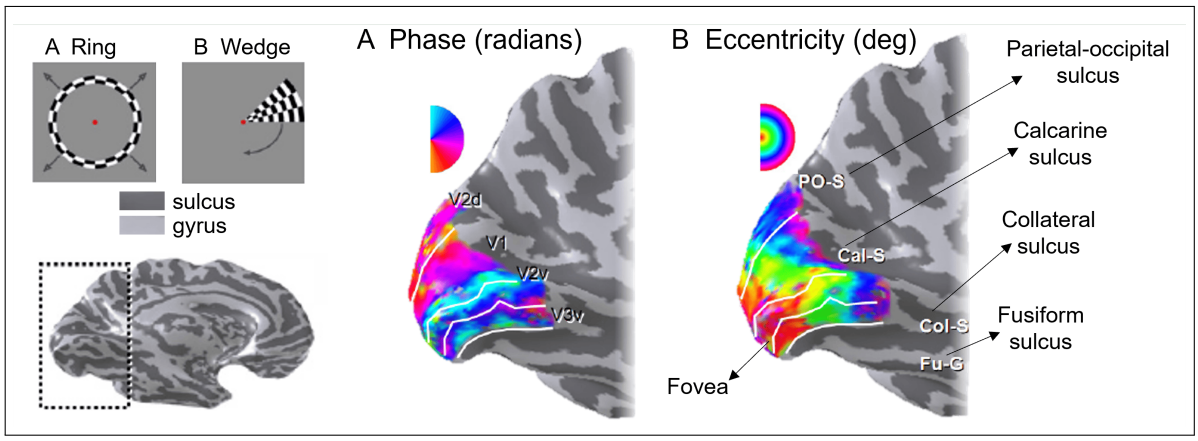

Figure 1.3. TWR Measurements. High-contrast checkerboard patterns that move smoothly and periodically through polar angles (wedge) or eccentricity (ring) are used as traveling wave stimuli. An expanded view of this surface near the calcarine sulcus is overlaid with a color map showing the response phase at each location for polar angle (A) and eccentricity experiments (B). The solid white lines indicate the boundaries of V1 in the calcarine sulcus.

For polar angle mapping, a stimulus shaped like a high-contrast, flickering checkerboard wedge covers a small range of specific polar angles from the fovea to the periphery, displaying

each voxel's preferred polar angle. The sector stimulus sequentially activates different polar angle representations of the visual field by rotating in discrete, even steps either clockwise or counterclockwise around a central fixation point. Eccentricity mapping includes the expansion or contraction of a ring-shaped stimulus in discrete steps between the fovea and the periphery. It's important to measure these orthogonal dimensions to define visual field maps (VFMs) because they allow the unique mapping of neuronal responses within a single voxel in the cortex to a location in visual space.

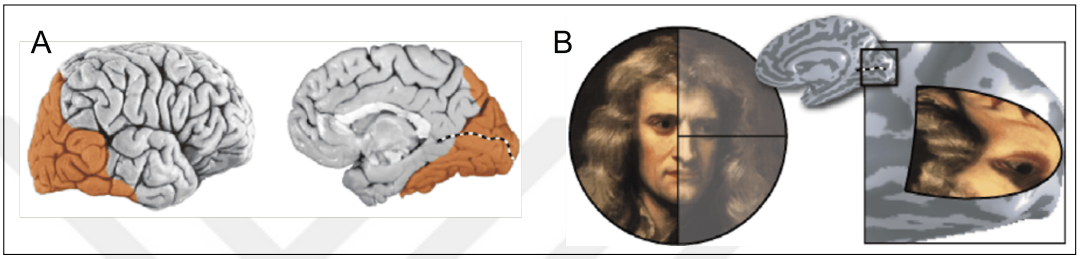

Figure 1.4. (A) The human visual cortex (orange overlay) occupies about 20% of the cerebral cortex and is located in the occipital lobe and posterior parts of the parietal and temporal lobes. A map of the visual field can be found in V1, which is located in and around the calcarine sulcus (dotted line). (B) VFM illustration in V1. This image is part of Sir Isaac Newton's 1989 portrait by Godfrey Kneller. The figure illustrates how the visual field (left) is transformed and represented on the V1 cortical surface (right) using a mathematical description proposed by Schwartz (1977). The LVF stimulates V1 in the RH; the image representation is inverted, and the center of the visual field, near the eye, is greatly expanded (cortical magnification).

In retinotopic mapping, peripheral visual stimuli are processed by a significantly smaller number of neurons than in the fovea, which is the center of the visual field and represents a region where a large number of neurons process information from a small area. It is known as cortical magnification and describes how neuron density affects how a stimulus is processed depending on where it's in the visual field. The high density of neurons in the fovea allows high-resolution visual processing, whereas the density decreases significantly in the periphery. It contributes to changes in the perceptual performance of different functions in the visual field, and cortical magnification plays an important role in understanding how this performance is organized. For instance, high visual acuity and detail perception are achieved in the fovea, but these properties are lost in the periphery. The variation in visual performance depending on the location of the visual stimulus is closely related to cortical magnification, and this process is of critical importance in understanding the functioning of

the visual cortex.

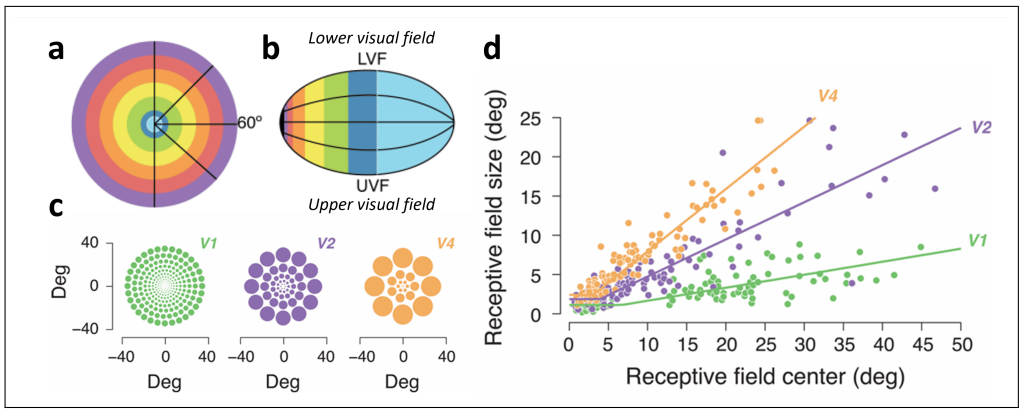

Figure 1.5. Map of retinal eccentricity (a) Cortical magnification varies with eccentricity. (b) For visual regions V1, V2, and V4, the receptive field size (diameter) is a function of the receptive field center (eccentricity). A "hinged" line well describes the size-to-eccentricity relationship in each region. (c) A cartoon illustration of receptive fields with sizes based on physiological measurements. The center of each array is the fovea. The size of each circle is proportional to its eccentricity based on the corresponding scaling parameter (slope of the fitted line in c). A larger scaling parameter indicates larger receptive fields at a given eccentricity. The model used overlapping pooling regions (linear weighting functions) to tile the image uniformly. They are separable and of constant size when expressed in polar angle and log eccentricity. (d) [1]

Named after the German neurologist Korbinian Brodmann, Brodmann areas represent a systematic approach to mapping the human cortex based on cytoarchitectural organization. He published his cortical Map for the first time in 1908, which provides a detailed description of the different cortical areas and their unique histological characteristics. Brodmann's mapping technique includes studying histological parts of the human brain to identify different cortical areas with varying cellular compositions. This system, which assigns specific numbers to different cortical regions in humans and other mammals, made it easier to identify and understand these specific brain regions at that time by clarifying terminology and concepts related to these brain regions.[52–54]

Before Brodmann's technique, some cortical areas, including Broca's and Wernicke's, had been identified as associated with language functions. It's named after the neurologists who first identified these areas based on their observations of language deficits in patients with brain lesions. Broddman expanded his approach to comprehensive cortex mapping by defining all functional areas according to their gross anatomical features and microscopic structure. [52–54]

Brodmann areas are regions of the cerebral cortex divided into distinct regions responsible for various cognitive functions. Brodmann regions 1, 2, and 3 handle sensory information from the body and coordinate motor movements. Brodmann area 4 helps the beginning and coordination of voluntary movements. Brodmann area 9 is related to higher cognitive functions, including working memory and abstract reasoning. Brodmann area 17, the primary visual cortex, is responsible for processing visual information. Brodmann areas 21 and 22 are essential for language processing and sound recognition. Brodmann areas 23, 24, 28, and 33, located in the cingulate gyrus, regulate emotions, pain perception, and memory. Finally, Brodmann regions 44 and 45, frequently referred to as Broca's area, are required for generating languages and syntactic processing. [52–54]

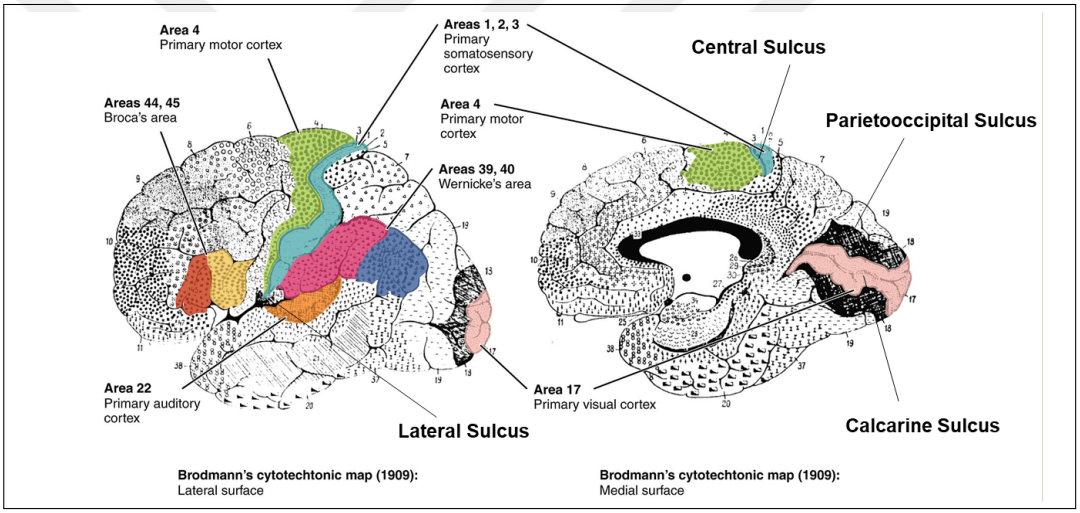

Figure 1.6. Brodmann areas of the brain (left) lateral surface (right) medial surface

1.3.2. fMRI Setup: Visual Stimulation Experiment

In fMRI visual stimulation experiments, the primary goal is to investigate the neural correlates of visual processing by monitoring brain activity in response to various visual stimuli. Participants are positioned within a high-field fMRI scanner, which typically operates at 3T or higher, providing enhanced spatial and temporal resolution. During the experiment, participants are presented with various visual stimuli, from simple geometric shapes to complex dynamic scenes, projected onto a screen within the scanner's field of view.

The participants' heads are carefully restrained using customized head coils or foam pads,

ensuring stability throughout the scanning session to minimize motion artifacts and enhance data quality. The visual stimuli are carefully controlled and can be manipulated in terms of contrast, brightness, and spatial frequency to trigger specific neural responses associated with visual perception and attention.

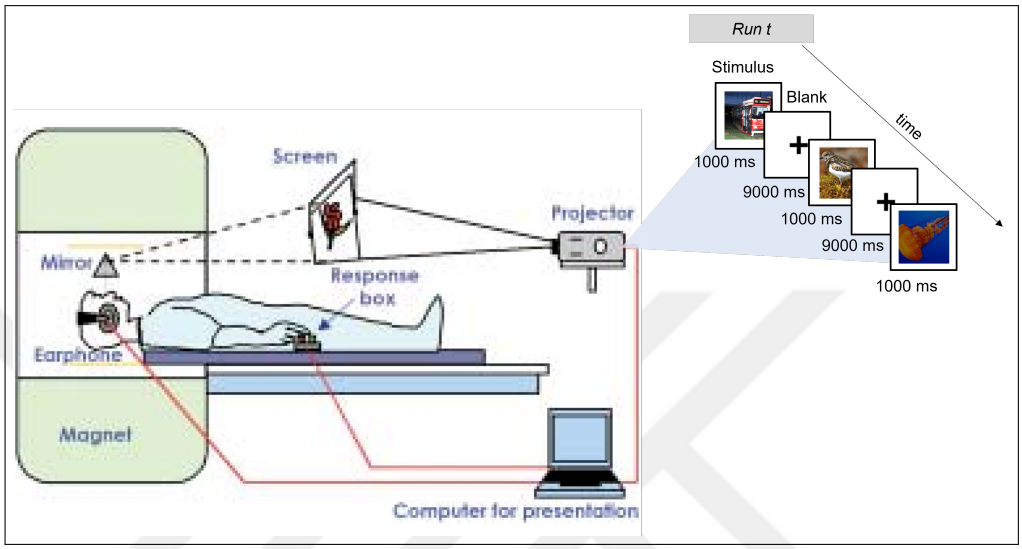

Figure 1.7. fMRI setup: Patient responds to the visual/auditory stimuli using the response box

The fMRI technique primarily measures changes in blood oxygenation level-dependent signals (BOLD), which reflect the hemodynamic response to neural activity. When neurons become active, they consume more oxygen, leading to localized blood flow and oxygenation changes. This BOLD signal is captured through a series of rapid imaging sequences, often using echo-planar imaging (EPI) techniques that allow for the collection of multiple slices of brain images in quick succession.

Stimulus presentation is often paired with task conditions, such as object recognition or passive viewing, to explore different cognitive processes. Behavioral data may also be collected alongside fMRI data, providing complementary insights into participants' perceptual and cognitive states during the experiment. The design of these studies is critical, as it influences the interpretation of the BOLD response and its relationship to underlying neural mechanisms. Through rigorous experimental design, including counterbalancing and randomization of stimulus presentation, researchers can draw robust conclusions about the neural networks involved in visual processing.

1.3.3. Cortical Surface Model

Cortical surface models represent the highly folded and convoluted sheet of the human cerebral cortex, where a significant portion of the surface area—up to 60-70%—is buried within the sulci. This intricate structure is challenging to analyze using traditional volumetric methods, as distances measured in 3D space often misrepresent the actual distances along the cortical sheet, with errors averaging around 45-50%. To overcome this, cortical surface models are constructed by segmenting brain tissues from high-resolution MRI scans, differentiating cortical gray matter (GM) from other tissue types. The segmented data is then used to generate a mesh-like representation of the brain's surface, accurately capturing its topological and geometric properties. These models allow for the study of critical cortical metrics, including thickness, surface area, and curvature, while supporting detailed analysis of the brain's functional and structural organization. Cortical surface models provide a significant advantage in analyzing the brain's topographic maps. The neocortex contains multiple functional areas organized spatially coherent, including retinotopic maps in the visual system, where cortical regions represent specific parts of the visual field; tonotopic maps in the auditory cortex, corresponding sound frequencies; and somatotopic maps in the somatosensory and motor cortices, organized according to the body's layout. These maps are inherently 2D, aligning closely with the cortical sheet's structure. Surface-based methods are particularly well-suited for studying these mappings, as they respect the natural organization of cortical functions, unlike volumetric methods that obscure such relationships. [55]

Mapping the highly folded cortical surface to parameterizable shapes, such as a sphere or flat plane, is crucial in cortical surface analysis. This process enables visualization, computational analysis, and cross-subject alignment while minimizing metric distortions. However, due to the cortex's intrinsic curvature, achieving a distortion-free mapping is mathematically impossible, as proven by Gauss's theorem on differing Gaussian curvatures. Consequently, these mappings introduce some degree of distortion in distances, areas, or angles. The primary objective is to minimize these distortions while preserving the surface's topological structure. Inflation, flattening, and spherical morphing address this challenge by constructing energy functionals that balance distortion minimization with topological preservation. [55]

The term used to minimize metric distortions is defined as follows. Consider a mesh of

V vertices irregularly distributed over a surface S embedded in 3D Cartesian space. The distance between the i-th and j-th vertices at time t, denoted as d_{ij}^t, is used to construct the mean-squared energy functional J_d.

$$J_d = \frac{1}{2V}\sum_{i=1}^{V}\sum_{n\in N(i)}\left(d_{in}^t - d_{in}^0\right)^2, d_{in}^t = \|x_i^t - x_n^t\| \tag{1.1}$$

The (x, y, z) position of vertex i at time t is given by x_i^t, and d_{in}^0 represents the distance between the i-th and n-th vertices on the original surface. $N(i)$ is the set of vertices in the neighborhood of vertex i. The gradient of J_d with respect to the k-th vertex is computed as:

$$\frac{\partial J_d}{\partial \mathbf{x}_k} = \sum_{n\in N(k)}(d_{kn}^t - d_{kn}^0)\mathbf{e}_{kn}, \tag{1.2}$$

where $\mathbf{e}_{kn}$ is the normalized unit vector that points from vertex k to vertex n.

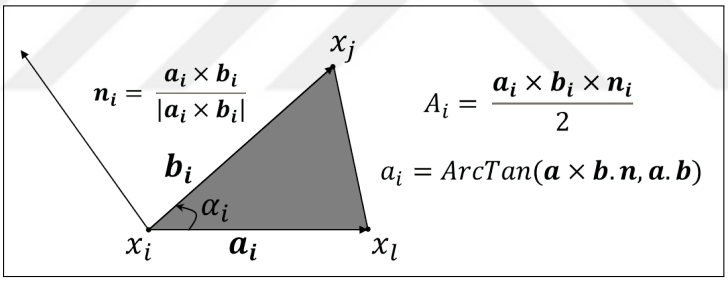

Figure 1.8. The triangular tessellation's metric properties

Using the metric properties of the surface defined by the triangular tessellation, an energy functional J_a is formulated to penalize negative areas in proportion to the difference between the current and the original areas of each triangle:

$$J_a = \frac{1}{2T}\sum_{i=1}^{T} P(A_i^t)(A_i^t - A_i^0)^2, \quad \text{where} \quad P(A_i^t) = \begin{cases} 1, & A_i^t \le 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.3}$$

where T is the total number of triangles, A_i^t represents the area of the i-th triangle at time t, and A_i^0 is its initial area at time $t = 0$. For simplicity, the functional dependence of A_i^t on the positions of the vertices and their neighbors is suppressed in this notation. To minimize J_a,

its gradient with respect to the vertex positions $\mathbf{x}_k$ is computed as:

$$\frac{\partial J_a}{\partial \mathbf{x}_k} = \frac{1}{T} \sum_{i=1}^{T} (A_i^t - A_i^0) \frac{\partial A_i^t}{\partial \mathbf{x}_k}. \tag{1.4}$$

Expanding the partial derivative of A_i^t concerning the vertex position $\mathbf{x}_k$ requires the chain rule, where the change in the area depends on the contributions from the triangle's edges $\mathbf{a}_i$ and $\mathbf{b}_i$. The gradient of the area is determined by the relative positions of the vertices within the triangle and the associated edge vectors. These contributions are determined by whether the vertex directly affects the edges, with adjustments depending on its role in forming the triangle. It ensures that the unfolding process penalizes distortions effectively while preserving the surface's geometric integrity.

The complete energy functional includes both the distance term (J_d) and the areal term (J_a), which together aim to minimize metric distortions while also promoting the unfolding of the surface. The energy functional is expressed as:

$$J = \lambda_d J_d + \lambda_a J_a \tag{1.5}$$

The coefficients λ_a and λ_d control the relative importance of two objectives, with λ_a initially taking much larger values than λ_d, gradually decreasing over time as the surface unfolds successfully. Additionally, the gradients are smoothed through iterative averaging during the numerical integration, allowing entire compressed or expanded regions to be moved coherently in the appropriate direction. It is similar to decimation, followed by upsampling with interpolation. Each scale, defined by the number of iterations in the averaging, can equilibrate before reducing the scale and continuing. The actual minimization of $J(x)$ is then achieved using gradient descent with line minimization. The complex folding of the cortical surface makes inflation crucial for visualizing functional activity within the sulci. Surface inflation aims to preserve the cortical surface's original shape and metric properties while making the sulcal activity more visible. The energy function for surface inflation includes a spring force term that smooths the surface, as well as the metric-preservation term to preserve

the local distances between vertices:

$$J_s = \frac{1}{2V} \sum_{i=1}^{V} \sum_{n \in N(i)} \|x_i - x_n\|^2 + \lambda_d J_d \tag{1.6}$$

where, $N(i)$ refers to the set of nearest neighbors of vertex i. Euler's method with momentum integrates this energy function, iterating until the surface reaches a smooth state, as measured by the goodness-of-fit of the polyhedral approximation.

Flattening a cortical hemisphere with minimal distortion includes making several cuts along the medial aspect of the surface. These cuts include one around the corpus callosum to remove subcortical structures, one along the fundus of the calcarine sulcus, and a set of equally spaced radial cuts. After the cuts, the surface is projected onto a plane, with the plane's normal determined by the cut surface's average normal. After the projection, the surface is unfolded by minimizing the energy functional, using randomly sampled distances in a 0.5 cm radius from each vertex as the neighborhood. This procedure reduces distortions and preserves the surface's original topological and metric properties as much as possible.

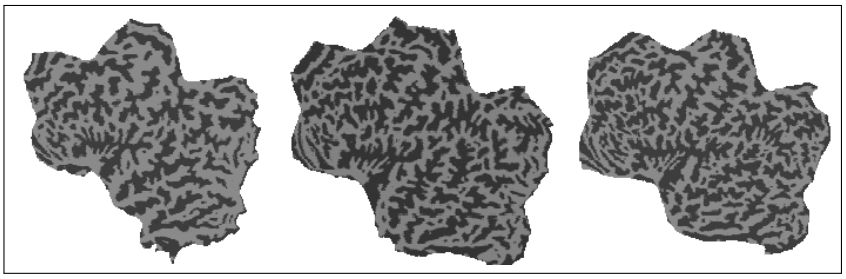

Figure 1.9. Flattened left hemispheres

A surface-based coordinate system is created by morphing the cortical surface into a parameterizable shape, typically a sphere. This transformation allows the surface to retain its original topological structure, preserving local connectivity while minimizing the distortions that result from flattening. In contrast to the flat surface, which involves cuts that alter its topology, the spherical surface preserves its topological integrity. Instead of Euclidean distances, geodesic distances are used in the sphere, providing a more accurate calculation of metric properties such as distances, areas, and angles. It allows for a uniform coordinate system that is useful for comparisons across subjects and further functional cortex analysis.

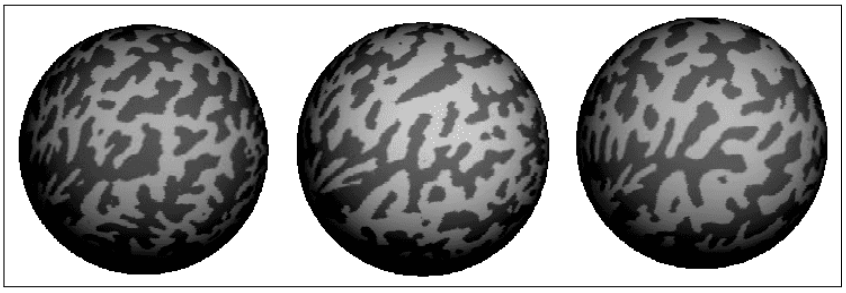

Figure 1.10. Three LHs in a lateral view after morphing.

1.3.4. Diffusion Models

Diffusion models are generative models inspired by the natural diffusion process observed in physics. In this process, the particles move to lower concentration regions from higher concentration regions and eventually reach equilibrium. These models formulated as complex latent variable models defined by the equation $p_\theta(h_0) = \int p_\theta(h_{0:T})\, dh_{1:T}$, where $h_1, \ldots, h_T$ are latent variables of the same dimensionality as the observed data $h_0 \sim q(h_0)$. Initially manifested as random noise $h_0 \sim \mathcal{N}(0, \sigma^2)$, the data is progressively transformed into complex, high-quality outputs through a structured process. [56, 57]

DDPM expands the basic concepts of diffusion models by focusing on the de-noising step in the reverse diffusion process. These models use a framework that simulates data corruption during the forward process and data reconstruction in the reverse process. It's based on a trained neural network to predict and eliminate the noise added at every step. This way, the model can gradually convert corrupted data into high-quality reconstructions. Diffusion models consist of forward diffusion, reverse diffusion, and sampling. The forward diffusion process adds noise into a simple Gaussian distribution, progressively increasing the data's complexity. [56, 57] The forward diffusion process is defined as follows:

$$q(h_{1:T} \mid h_0) := \prod_{t=1}^{T} q(h_t \mid h_{t-1}), \quad q(h_t \mid h_{t-1}) := \mathcal{N}\left(h_t; \sqrt{1-\beta_t} h_{t-1}, \beta_t I\right) \tag{1.7}$$

Each step in this sequence transforms the data using a Gaussian distribution, with the mean set as $\sqrt{1-\beta_t} h_{t-1}$ and the covariance as $\beta_t I$, where β_t called the diffusion rate is predetermined by a variance scheduler and I is the identity matrix, ensuring that each distribution is isotropic

Gaussian. The intermediate noisy images generated from timestep 1 to T are referred to as latent and maintain the same dimension as the original image.

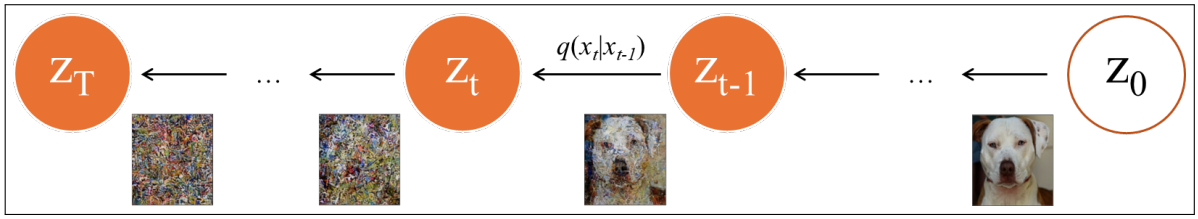

Figure 1.11. The forward diffusion process's Markov chain generates a sample by gradually adding noise.

Applying a well-calibrated schedule for β_t and using sufficiently large timesteps, repeated application of this forward diffusion progressively transforms the data distribution into an isotropic Gaussian distribution. The variance scheduler regulates the addition of the controlled Gaussian noise (ϵ) to the original image at each time. This controlled noise adds to the step-by-step increase in the complexity of the data. It converts the original clean image into a sequence of noisy latents, resulting in a distribution close to the target Gaussian model. Considering the forward diffusion equation, the h_t image is sampled from a normal distribution:

$$h_t = \sqrt{1-\beta_t}h_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1.8}$$

The variance parameters β_t in the forward diffusion process can be adapted by reparameterization techniques or fixed as hyperparameters. The effect of the reverse process is partially ensured by using Gaussian conditionals in $p_\theta(h_{t-1}|h_t)$, an approach that proves efficient when β_t values are kept small. A distinctive feature of the forward process is its capacity to facilitate the sampling of h_t at any selected timestep t in an exact manner. It is achieved by defining $\alpha_t = 1 - \beta_t$ and calculating $\overline{\alpha_t} = \prod_{s=1}^{t} \alpha_s$, which enables the explicit formulation of the conditional distribution of h_t given h_0 as:

$$q(h_t|h_0) = \mathcal{N}(h_t; \sqrt{\overline{\alpha_t}}h_0, (1-\overline{\alpha_t})I) \tag{1.9}$$

By substituting β with α and using the addition property of Gaussian distribution. The forward diffusion process can be defined in terms of α as:

$$h_t = \sqrt{\alpha_t} h_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1.10}$$

The joint distribution $p_\theta(h_{0:T})$, also known as the reverse process, is a Markov chain that begins with the distribution $p(h_T) = \mathcal{N}(h_T; 0, I)$. and seeks to reconstruct the original data by methodically removing the noise added during the forward process. It's mathematically expressed as:

$$p_\theta(h_{0:T}) := p(h_T) \prod_{t=1}^{T} p_\theta(h_{t-1} \mid h_t), \quad p_\theta(h_{t-1} \mid h_t) := \mathcal{N}(h_{t-1}; \mu_\theta(h_t, t), \Sigma_\theta(h_t, t)) \tag{1.11}$$

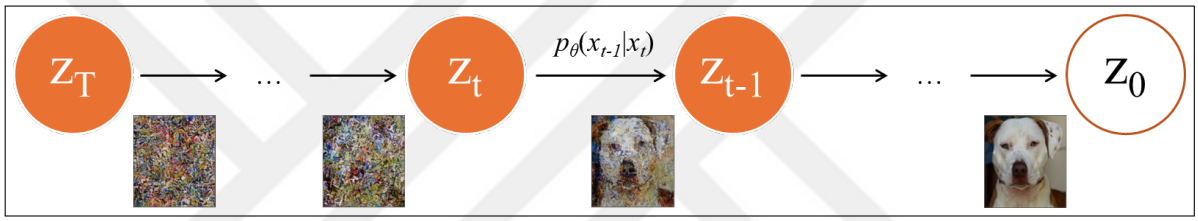

Figure 1.12. The reverse diffusion process Markov chain generates a sample by gradually removing noise.

After training, the model generates new samples by reversing the noise-added data ($h_0 \sim p_\theta(h_0 \mid h_1, h_2, ..., h_T)$. It transforms noise into meaningful patterns in this way and allows the generation of new data that closely match the statistical properties of the original data set. The training of diffusion models focuses on optimizing the variational bound on the negative log-likelihood of correctly reconstructing the original data from noisy states. The objective is expressed as follows:

$$E\left[-\log p_\theta(h_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(h_{0:T})}{q(h_{1:T} \mid h_0)}\right] = \mathbb{E}_q\left[-\log p(h_T) - \sum_{t\geq 1} \log \frac{p_\theta(h_{t-1} \mid h_t)}{q(h_t \mid h_{t-1})}\right] := \mathcal{L} \tag{1.12}$$

In this configuration, $p(h_T)$ is modeled as a Gaussian, reflecting the final noise state. In contrast, $p_\theta(h_{t-1}|h_t)$ and $q_(h_t|h_t - 1)$ monitor the model's capacity to effectively de-noise or reverse the diffusion process. DDPMs are trained with a simplified ELBO version that reduces computational needs while increasing model efficiency. The simplified training objective is

described as follows:

$$\mathcal{L}_{\text{simple}}(\theta) := \mathbb{E}_{t,h_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}h_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right] \tag{1.13}$$

In this formula, ϵ represents the Gaussian noise, ϵ_θ is the neural network estimate of this noise, and $\overline{\alpha}_t$ indicates the cumulative noise level up to step t. This loss function minimizes the MSE between the actual added noise in the forward diffusion process and the predicted noise from the model in the reverse diffusion process. Sampling t uniformly across the diffusion timestep enables the model to effectively handle various levels of data corruption, training the network to perform high-fidelity reconstructions across different stages of noise addition. This strategy effectively trains the model by focusing on individual noise removal steps, resulting in increased robustness and adaptability.

This study focuses on LDMs that optimize the diffusion process using latent representations. Instead of processing high-dimensional data directly, this model uses an encoder-based model like an autoencoder to encode the data into a lower-dimensional latent representation. This approach significantly reduces computational costs while improving the model's efficiency. Moreover, it retains the ability to capture and reconstruct complex data distributions, thus preserving the quality and fidelity of the generated samples.

1.4. CONTRIBUTION OF THE THESIS

This thesis aims to implement advanced deep learning and neuroimaging techniques to reconstruct high-resolution stimulus images of human brain activity. The study generated an overall cortical surface model from T1W MRI scans by preprocessing a series of images that overlapped with functional scan images combined with the surface model pipeline provided by the Freesurfer software. It was followed by mapping the cortical surface pattern produced by visual brain activity in the functional scans. The occipital surfaces that cover the V1 and V2 areas calculated in the cortical region and responsible for visual information processing were removed and flattened with the creation of surface models. As a result, occipital patches were obtained, and visual activity values were mapped. GLM, CA, ICA, MVPA, and other techniques are widely used to reconstruct visual experiences from brain activity. These techniques help to map brain functions and connections related to visual processing by analyzing complex neural signals obtained from fMRI data. No studies in the existing literature identified using cortical surface modeling and occipital patch extraction approaches for stimulus reconstruction from brain activity.

2. METHODOLOGY

This section describes the participants' demographic characteristics, task design, fMRI acquisition, network architecture, implementation details and evaluation metrics, and preprocessing steps.

2.1. DATASET

We used the publicly available large-scale 3T fMRI dataset, the BOLD5000 dataset. [58] The researchers selected 5,000 different images from 3 popular datasets: SUN, COCO, and ImageNet, due to the prominence of computer vision and the broad spectrum of image categories they represent. The scale and diversity of these datasets, paired with the slow event-related fMRI design, support a fine-grained analysis of the brain response to a wide range of visual categories, semantics, and features.

2.1.1. Participants

The BOLD5000 collected from 4 participants (CSI1, CSI2, CSI3, and CSI4). All participants were familiar with MRI procedures and could complete all scanning sessions with minimal effect on data quality. The demographics of participants were as follows: CSI1—male, age 27, right-handed; CSI2—female, age 26, right-handed; CSI3—female, age 24, right-handed; CSI4—female, age 25, right-handed. Each participant reported no history of psychiatric or neurological disorders and no current use of psychoactive medications.

2.1.2. Visual Stimulus and Experimental Design

This study used the diverse and large-scale BOLD5000 dataset to explore the neural representation of visual features, categories, and semantic content. A complete dataset acquired for 3 participants, including 16 fMRI scanning sessions. For CSI4, there were 9 functional sessions due to discomfort in the MRI. While 8 of these sessions consist of 9 runs, additional functional localizer runs were performed for the remaining 7 sessions

for the 10. sessions and only 9 runs, providing 8 localizer runs across 15 sessions. A complete dataset collected across 16 MRI scanning sessions, where 15 were functional sessions acquiring task-relevant data, and the remaining session collected high-resolution anatomical and diffusion data. In addition, physiological data, including respiration and heart rate, were recorded using wireless sensors. The experimental design also included gathering behavioral data using localizer tasks and judgments of scene valence. 112 out of 4,916 different images were randomly selected to be presented 4 times, with 1 image viewed 3 times by each participant To analyze the impact of image repetition. These 113 images selected to maintain proportional representation in the image dataset breakdown: 1/5 of the images were SUN images (7 SUN images), 2/5 were COCO images (14 COCO images), and the remaining 2/5 were ImageNet images (14 ImageNet images). Following the image repetitions, each participant presented with 5,254 images. At the initial and end of each

Table 2.1. A general design of data per subject

Subjects	4
fMRI Sessions	16
Total fMRI Scene Runs	142
Total Functional Localizer Runs	8
Total Scene Trials	5,254
Unique Scene Stimuli	4,916

run, participants viewed a fixation cross on a blank, black screen for durations of 6 seconds and 12 seconds, respectively. Following the initial fixation cross, the 37 stimuli displayed sequentially. Each image was presented for 1 second, followed by a 9-second gap with a fixation cross. With 37 stimuli per run, this resulted in a total presentation time of 370 seconds, plus an additional 18 seconds for pre- and post-stimulus fixation periods. In this respect, each run took 388 seconds (or [6 s + 37 blocks × [1 s + 9 s] + 12 s] = 6 minutes and 28 seconds) of data.

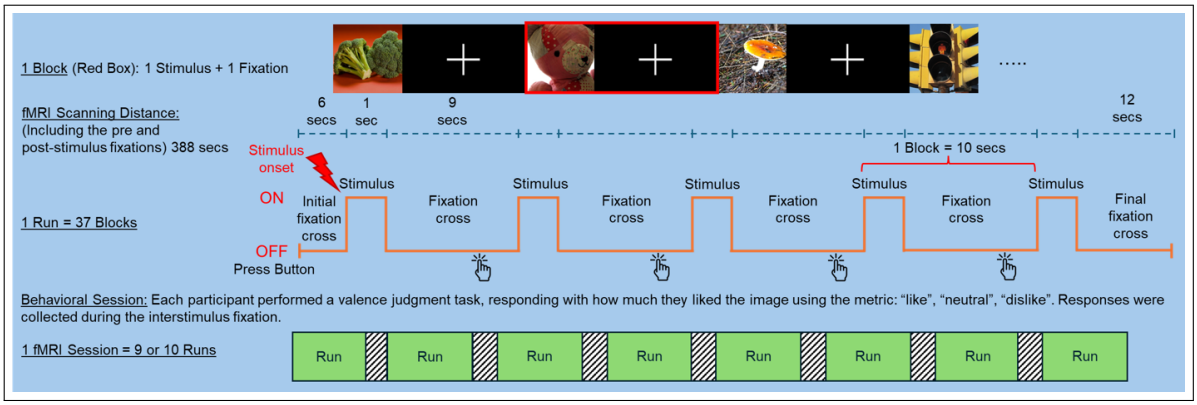

Figure 2.1. Schematic representation of the task paradigm. "On" denotes the presentation of the stimuli, and "Off" denotes no presentation of the stimuli. A fixation cross centered on a blank shown for 6 at the beginning and 12 seconds at the end of each run.

2.1.3. fMRI Data Acquisition

The MRI scan was gathered on the Carnegie Mellon University campus using a 3T Siemens Verio MR scanner with a 32-channel phased array head coil. The functional scans generated using a T2*-weighted gradient recalled EPI multi-band pulse sequence from the University of Minnesota's CMRR. The acquisition parameters included 69 slices co-planar with the AC/PC, 2 × 2 mm^2 in-plane resolution, a 106 × 106 matrix size, 2 mm slice thickness with no gap, interleaved acquisition, a field of view of 212 mm, phase partial Fourier scheme of 6/8, TR of 2000 ms, TE of 30 ms, flip angle of 79 degrees, bandwidth of 1814 Hz/Px, echo spacing of 0.72 ms, excite pulse duration of 8200 microseconds, multi-band factor of 3, phase encoding direction of PA, fat saturation on, and advanced shim mode on.

2.2. DATA PRE-PROCESSING

We performed the data preprocessing steps as outlined below.

2.2.1. DICOM Mosaics to Patient Coordinate System Mapping

DICOM mosaics converted into 3D volumes before further analysis. DICOM mosaics require preprocessing steps before being reconstructed into a 3D volume. It involves analyzing

the mosaic structure, extracting individual slices based on the total number of slices and block arrangement, and applying a 4x4 affine transformation matrix to account for DICOM orientation and map voxel coordinates to physical dimensions within the patient coordinate system. This transformation includes rotation, scaling based on pixel spacing and slice spacing, and translation for accurate alignment and reconstruction of the 3D volume.

Algorithm 2.1. The nii.gz format converted to volumes from all sessions of all participants. (CSI1, CSI2, CSI3, CSI4)

```
1: mri_convert T1w_MPRAGE_CSI1.nii T1w_MPRAGE_CSI1.nii.gz
```

2.2.2. FreeSurfer Single Subject Pipeline Analysis Overview

We used FreeSurfer version v6.1.32, a widely used neuroimaging software with a standardized and automated pipeline, to create cortical surface models from T1W anatomical scans. The pipeline consists of multiple steps, from acquiring high-resolution T1W scans that strongly contrast white matter (WM) and gray matter (GM). FreeSurfer uses these to segment cortical surfaces and subcortical structures into distinct regions. FreeSurfer transforms the cortical structure into a 2D surface representation to mitigate the partial voluming effect, where a single voxel may contain signals from multiple areas, complicating source determination. During this process, boundaries between different tissues identified, and the surfaces are inflated into spheres while automatically correcting defects. A schematic representation of the FreeSurfer single-subject analysis pipeline illustrated in Figure 2.2.

First, an automated method developed by the Montreal Neurological Institute (MNI) transforms individual high-resolution T1W scan data into standard Talairach coordinates. (Talairach and Tournoux) [59] This procedure calculates transformation parameters by maximizing the correlation between the individual volume and a large average volume composed of previously aligned brains, using a gradient descent method at multiple scales. As a result, a transformation matrix that converts image coordinates to Talairach coordinates is obtained and stored for use in subsequent processing stages. Following the transformation of images into Talairach coordinates, intensity normalization performed. High-resolution T1W MRI images are often affected by magnetic susceptibility artifacts and RF field

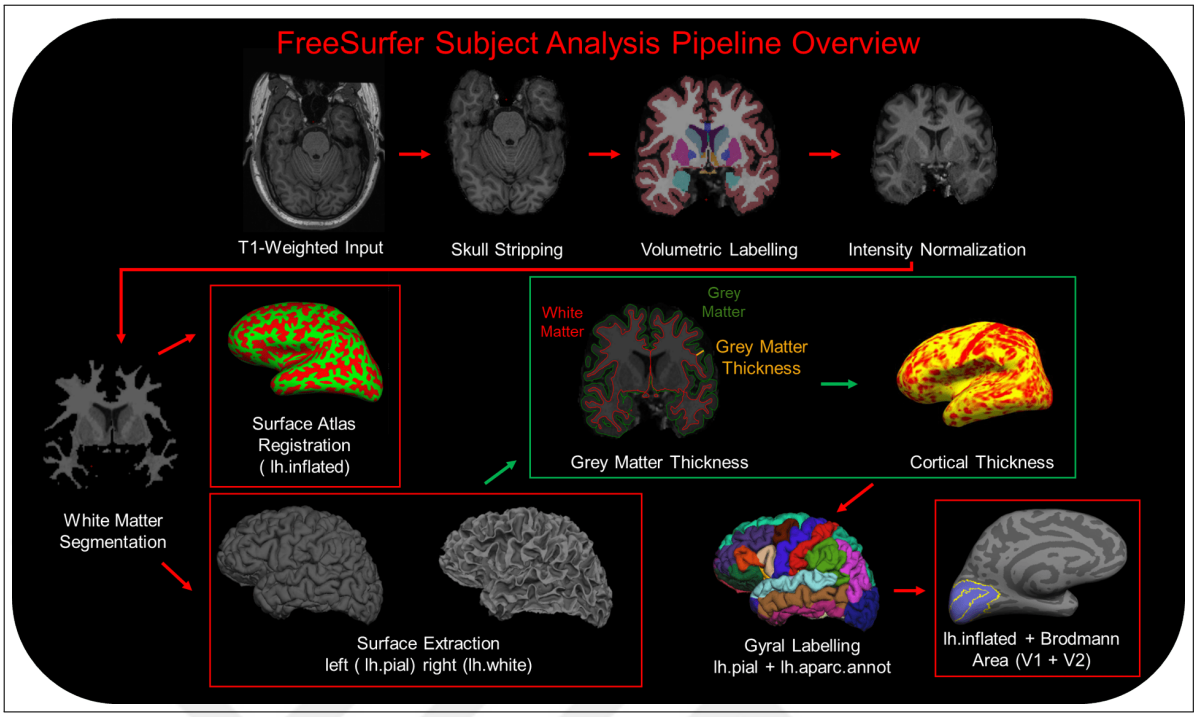

Figure 2.2. FreeSurfer subject analysis recon-all pipeline overview

inhomogeneities, resulting in tissue intensity and contrast variations. It leads to identical tissue types (e.g., WM and GM) appearing with varying intensities depending on their spatial location, negatively impacting the accuracy of segmentation processes. This step applies a fast and effective method to correct these intensity variations.

The procedure assumes that WM is the highest intensity tissue in slices parallel to the $x - y$ plane. This intensity centered on a desired reference value, and the remaining intensity variations are corrected using a Voronoi partitioning algorithm. The corrected intensity values provide a consistent foundation for subsequent analysis steps, such as segmentation and surface reconstruction. This process is applied after the Talairach registration, ensuring the image data are standardized and optimized for further analysis.

The next step in the reconstruction process includes automated skull stripping of the intensity-normalized image. This step deforms a tessellated ellipsoidal template to match the shape of the skull's inner surface. The deformation driven by two forces: (1) an MRI-based force that pushes the template outward to separate it from the brain and (2) a curvature-reducing force that enforces smoothness, encoding prior knowledge about the smoothness of the inner skull surface. The MRI-based force relies on non-local information

by sampling MRI data along the surface normal at each vertex of the tessellated template. It ensures that the surface pushed toward regions consistent with CSF, characterized by low MRI intensity values while repelling it outward from the contiguous areas consistent with brain tissue with higher MRI intensities. This approach based on the "shrink wrapping" method [60] and combines elements of active contour methods.

Following skull stripping, the next step is WM segmentation. It assigns tissue types to individual voxels using a combination of intensity-based thresholds and geometric information to segment WM accurately. Traditional segmentation methods rely on global grayscale thresholds to classify tissue types. Still, these approaches are highly sensitive to image artifacts such as partial voluming, magnetic susceptibility effects, and RF field inhomogeneities. It uses intensity-based and geometric criteria to refine the classification and address these limitations. By exploiting the laminar structure of the cortex, it ensures that the cortical surface remains smooth, with finite curvature, and minimizes topological defects. Local geometric information, such as the plane of least intensity variance, is used to classify ambiguous voxels at the boundaries between tissue types. Recon-all computes the interface between WM and GM for both hemispheres, storing surface estimates as **lh.orig** and **rh.orig**. The estimations are improved for accuracy and stored as **lh.white** and **rh.white**.

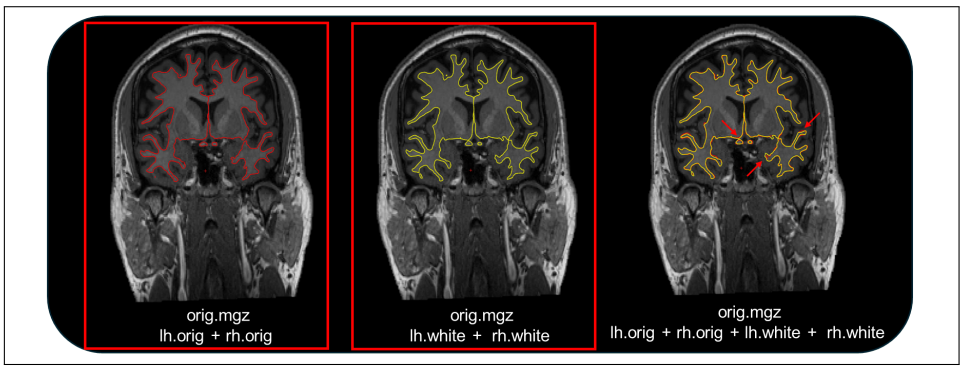

Figure 2.3. Orig and corrected orig surface (white surface)

After WM segmentation, the process of reconstructing the white matter surface begins. This step models the boundary between WM and GM using a triangle-based mesh (tessellation). This surface accurately represents the complex folds and sulci of the cortical structure. The reconstruction process aims to preserve the topological accuracy of the surface while minimizing connectivity issues that may arise from segmentation errors. Once the surface

created, smoothing algorithms and deformation techniques are applied to align the surface with MRI intensity data. During this process, an energy function minimized to ensure that the surface adapts to the cortical folds. In the deformation process, each point on the surface is updated according to the positions of its neighboring points, allowing the surface to adapt to the complex anatomical structure of the cortex. It ensures that features like sulci and gyri accurately represented while maintaining the topological integrity of the surface for a robust analysis foundation.

Algorithm 2.2. Reconstructing the cortical surface for single subject command. All participants subjected to the same process. (CSI1, CSI2, CSI3, CSI4)

```
1: recon-all -i T1w_MPRAGE_CSI1.nii.gz -s CSI1 -all
```

Recon-all extends the sensors from the WM surface (WM) to detect the edge of GM, resulting in the generation of **lh.pial** and **rh.pial** surfaces. After surface reconstruction, Freesurfer automatically corrects defects on the inflated surface. These corrections may include softening or interpolation techniques to fill missing or incorrect data points. The pial surface then inflated to form the inflated surface, which can further expanded and transformed into a sphere. These surfaces stored as **lh.inflated** and **rh.inflated**.

Finally, recon-all uses cortical and subcortical atlases (such as the Desikan-Killian or Destrieux atlas) to parcellate the brain into ROI. The parcellated regions stored as part of the output. These atlases and surfaces enable detailed analyses, such as measuring cortical thickness or BOLD signals across sulci and gyri.

Recon-all provides various outputs, including different surfaces (original, white, pial, inflated), parcellated areas, and quality control metrics. These are useful for advanced analyses, such as volumetric studies, cortical thickness measurements, and group comparisons.

The Figure 2.4 shows the outputs of the recon-all analysis process. The recon-all uses T1W scans to create cortical surface models, organizing outputs into folders. The **mri** folder contains anatomical scans, **label** stores brain region parcellation files, **scripts** holds command scripts, **stats** includes morphometric measurements, **surf** stores surface models, **tmp** contains temporary files, and **touch** holds tactile data for visualization.

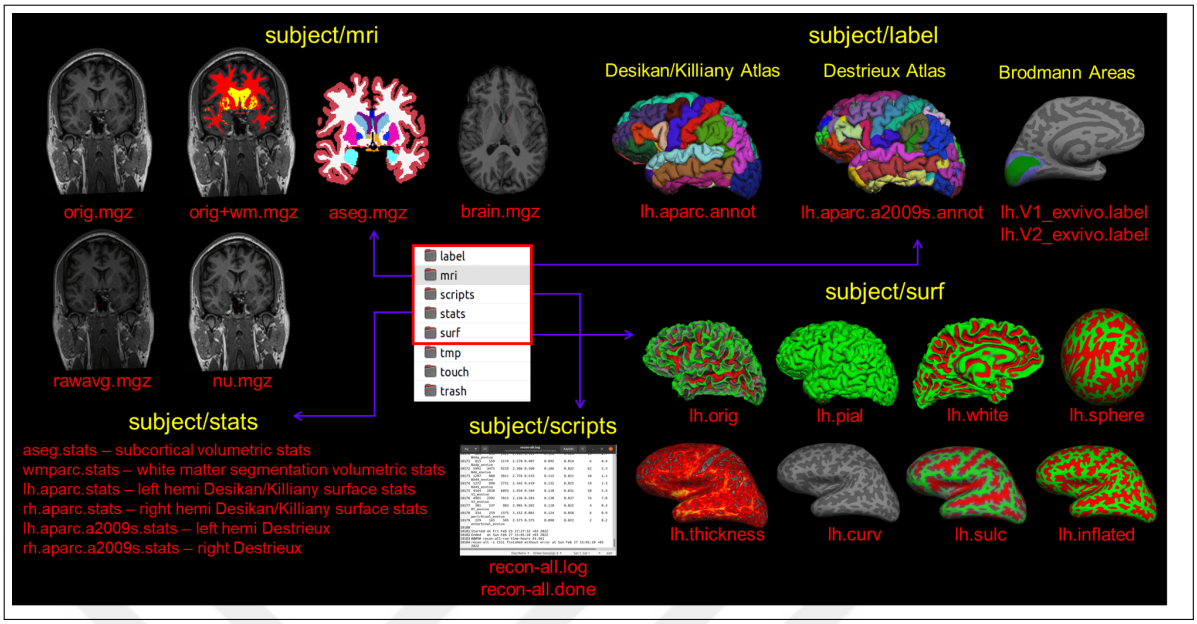

Figure 2.4. Recon-all process outputs

2.2.3. fMRI Analysis

This study used fMRI data, and a series of preprocessing steps were applied, described in detail below, and made available as part of the open source dataset. [61] All fMRI scans initially transformed from DICOM format to standardized BIDS [62] using a modified version of Dcm2bids. [63] All analyses performed with data with a pre-scan normalization filter. The default MRIQC pipeline used to extract image quality metrics and evaluate data quality. [64] The preprocessing included corrections performed with fMRIPrep 1.1.4 [64, 65], a Nipype 0.13.1 based on tool. [66] INU in each T1W volume was corrected using N4BiasFieldCorrection v2.1.0 [67], and then non-brain tissues (skull-stripped) were removed using antsBrainExtraction.sh v2.1.0, using the OASIS template. The brain surfaces created with FreeSurfer v6.0.1 recon-all [68], and the previously estimated brain mask refined using a custom variation designed to reconcile ANT-derived and FreeSurfer-derived segmentations of the cortical GM of Mindboggle. [69] Spatial normalization performed for the ICBM 152 Nonlinear Asymmetrical template version 2009c [70] through nonlinear registration using the antsRegistration tool of ANTs v2.1.0 using brain-extracted versions of both the T1W volume and the template. [71] Segmentation of brain tissues into the CSF, WM, and GM was analyzed in brain-extracted T1W volumes using fast FSL v5.0.9. [72] Functional data

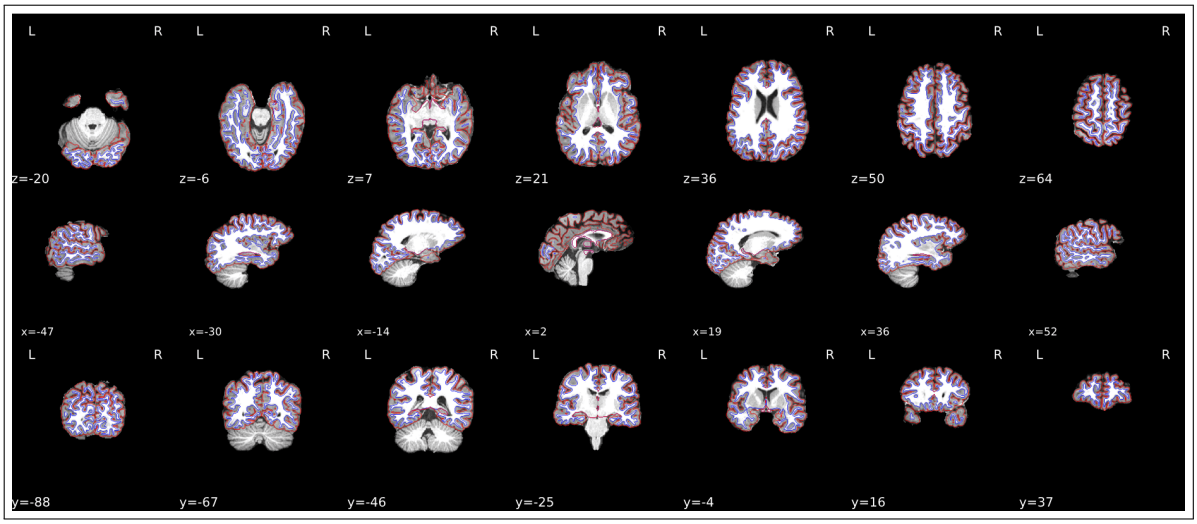

Figure 2.5. The white and pial surfaces reconstructed with recon-all provided by FreeSurfer and positioned on the participant's T1W template.

performed motion correction using mcflirt from FSL v5.0.9, and distortion corrected using the PEPOLAR technique implementation with 3dQwarp from AFNI v16.2.07. [73] This was followed by co-registration for the corresponding T1W scans using boundary-based registration [74] with 9 dof, employed bbregister from FreeSurfer v6.0.1. The process included combining motion-correcting transformations, field distortion-correcting warp, and BOLD to T1W transformation into a single step.

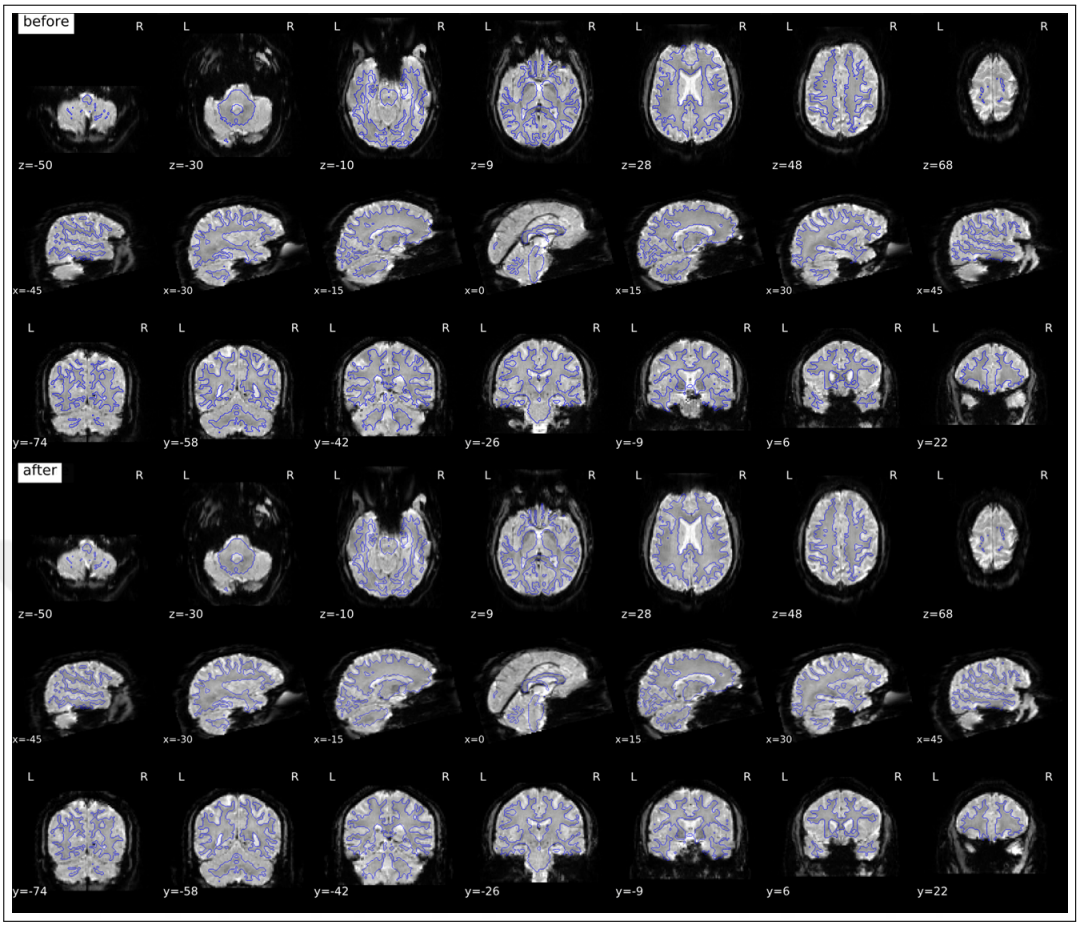

Figure 2.6. Results of using SDC on the EPI

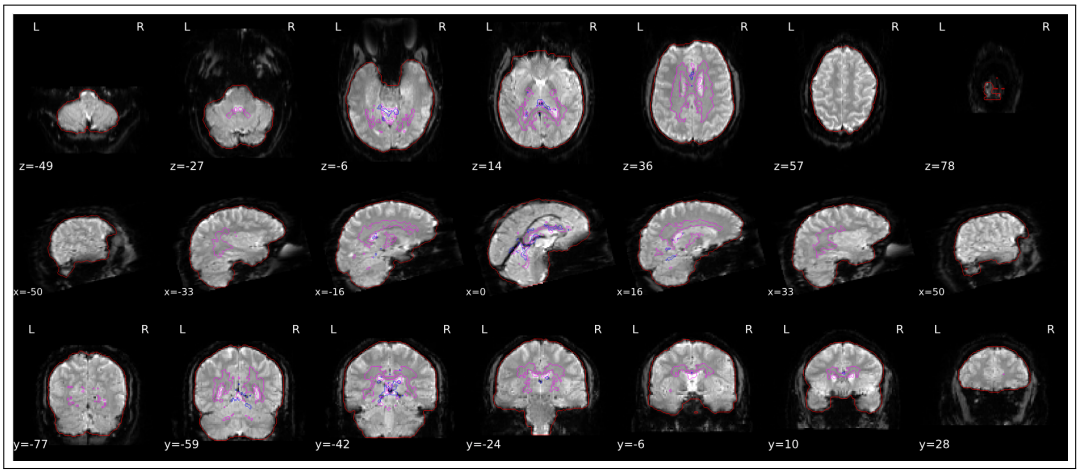

Figure 2.7. Brain mask calculated on the BOLD signal (red contour) and the masks used for a/tCompCor. The aCompCor mask (magenta contour) is a conservative CSF and WM mask for extracting physiological and movement confounds. The fCompCor mask (blue contour) contains the top 5% most variable voxels within a heavily eroded brain mask.

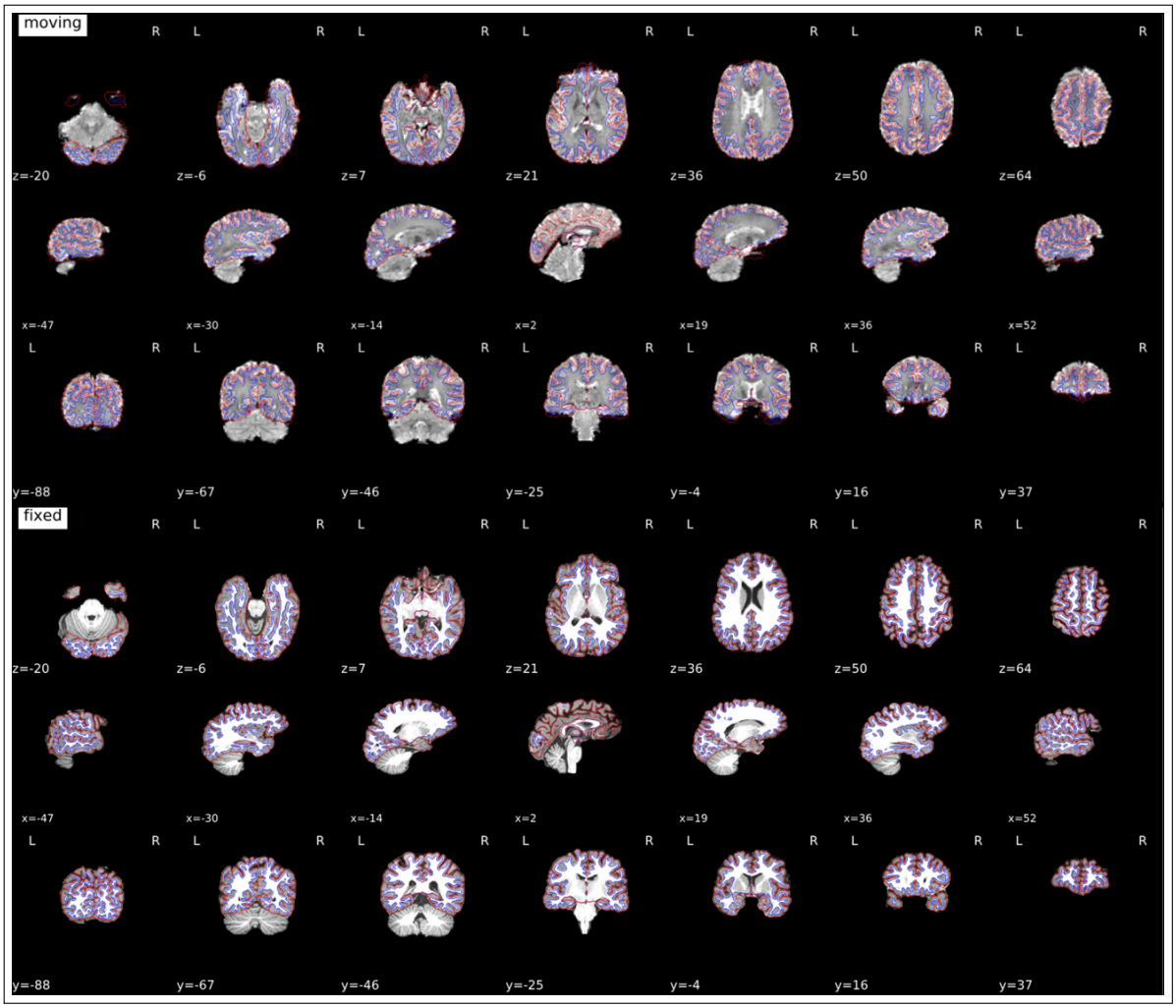

Figure 2.8. Bbregister was employed to generate transformations from EPI-space to T1W-space

It performed using antsApplyTransforms from ANTs v2.1.0 with Lanczos interpolation. The detailed pipeline and reports of these preprocessing steps are publicly available in the fMRIPrep documentation and repository. After preprocessing, data for 5,254 images analyzed. Each session's data, which included 9 or 10 runs, were processed through a GLM, where nuisance variables, including six motion parameter estimates, the average signals from CSF and WM masks, and the global signal from the whole-brain mask, were regressed. A regressor for each run in the GLM and an HPF of 128 seconds applied to the data. Residual time series from each voxel within an ROI were extracted and demeaned in all image presentations.

2.2.4. Volume to Surface Mapping using Freesurfer Functions

Volume-to-surface mapping is critical in transitioning neuroimaging data from a 3D voxel-based representation to a 2D cortical surface model. This process includes defining two essential surfaces: (1) the WM boundary surface, delineating the interface between WM and GM, and (2) the pial surface, marking the boundary between GM and cerebrospinal fluid (CSF). These surfaces establish the spatial boundaries necessary for the projection of data. Functional and other MRI modalities are recorded in structural MRI scans to ensure accurate spatial alignment. The mapping of volumetric data onto these surfaces accomplished using interpolation techniques. These techniques assign values from the 3D voxel space to the vertices of the cortical mesh constructed from the brain's surface. Common interpolation methods, such as nearest neighbor and linear interpolation, are used to optimize data placement. The **mri_vol2surf** function, available in FreeSurfer v6.1.32, facilitates this transformation by effectively projecting voxel intensities from volumetric data onto the cortical surfaces. The Freeview interface in FreeSurfer v6.1.32 used to visualize the results of volume-to-surface mapping. This interface provides an intuitive platform for examining cortical surface projections, further supporting structural and functional neuroimaging data integration.

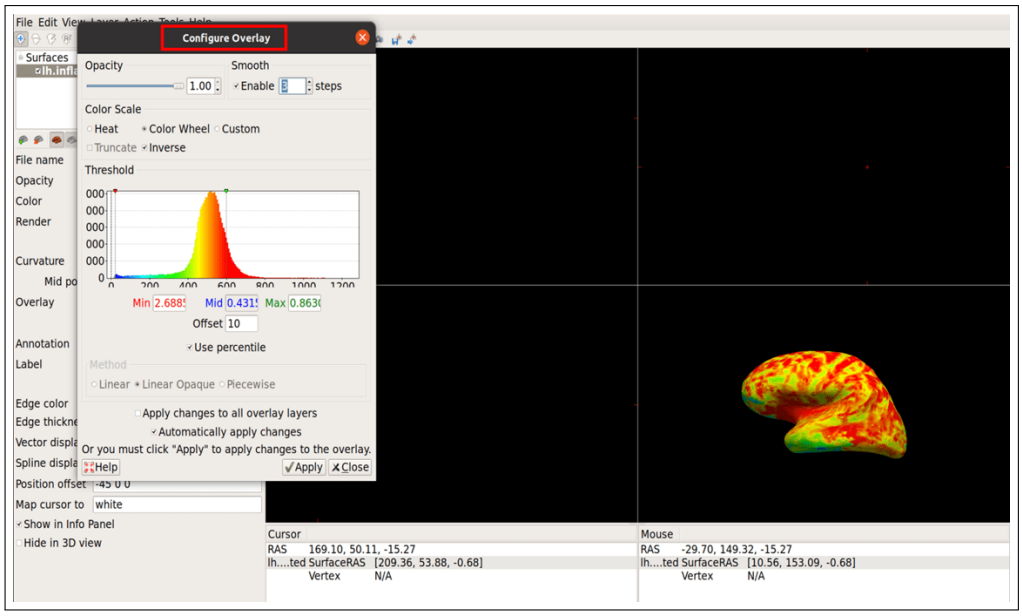

Figure 2.9. Freeview graphical interface

2.2.5. Occipital Patch Removal and Flattening

The Tksurfer tool from FreeSurfer v6.1.32 created occipital patches on cortical surfaces. Tksurfer is a software package for analyzing and visualizing surfaces provided by FreeSurfer v6.1.32. It enables interactive viewing of 3D models of created cortical surface reconstructions and analysis of different cortex regions. Removal of the occipital patch

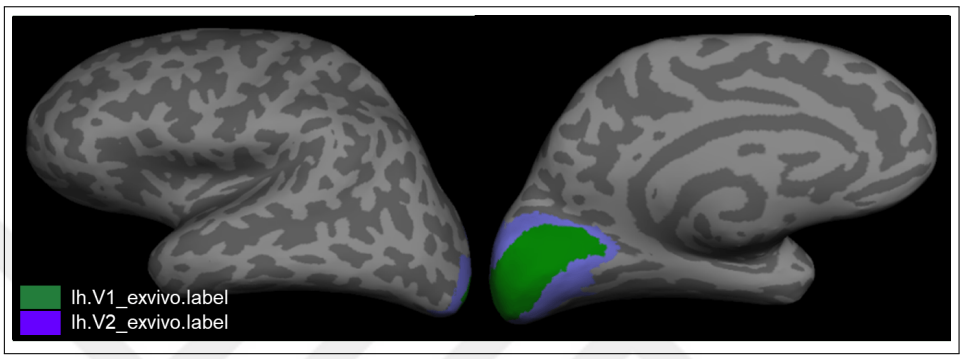

Figure 2.10. Surface overlays with Brodmann areas V1 and V2

involves several steps. First, the tksurfer users access GUI using the command-line interface. The occipital patches that cover the entire V1 and V2 areas extracted using Brodmann regions created from a single-subject analysis pipeline. Within the GUI, users select points along the calcarine fissure, a landmark in the brain. The "CUT LINE" button then pressed to define the cutting plane obtained by selecting three points: two on the medial side and one on the lateral side of the brain. These points help to describe the specific region of the brain's surface removed. 4. point is selected to specify the area of the surface to be kept. This step is important to remove the rest of the surface while leaving the targeted occipital patch intact. After that, the "CUT PLANE" button pressed to create the occipital patch. The resulting occipital patch from this process can saved as "?h.occip.patch.3d" for further analysis or visualization. Following occipital patch extraction, the **mris_flatten** function

Algorithm 2.3. Launches the Tksurfer tool, which allows visualization and navigation of cortical surface data for participant CS1. The lh.inflated argument specifies the inflated surface of the left hemisphere. The -gray flag load curvature file (?h.curv)

```
1: tksurfer CSI1 lh.inflated -gray
2: tksurfer CSI1 rh.inflated -gray
```

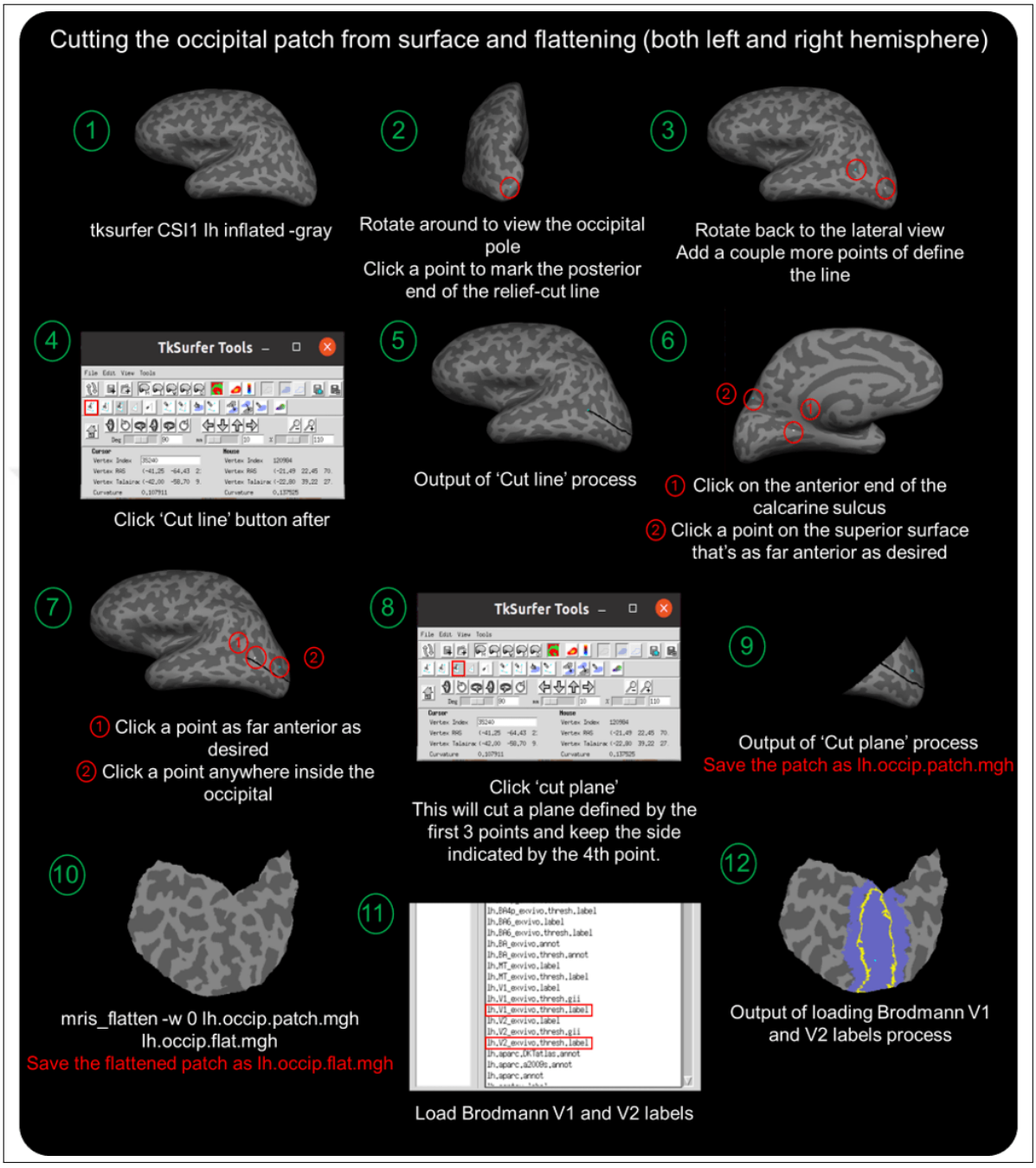

Figure 2.11. Cutting the occipital patch from the surface and flattening. (both left and right hemispheres) The same process performed for all participants.

flatten the occipital patch surface. The 3D surface is transformed into a 2D surface using a flattening algorithm, keeping spatial relationships among various brain regions. In our case, this function takes the surface of the extracted occipital patch as input.

Algorithm 2.4. FreeSurfer function for flattening occipital patches

```
1: mris_flatten -w 0 lh.occip.patch.mgh lh.occip.flat.mgh
```

2.2.6. Extracting Flatmap Surfaces

We performed multidimensional interpolation on irregularly spaced data using the `LinearNDInterpolator` function of the `Scipy` library. This function is necessary to create a continuous coherent surface from isolated data points corresponding to cortical surface measurements obtained from fMRI scans. This process coordinates these points into a 2D grid by specifying points on the 3D cortical mesh as input and scalar values corresponding to various neurological activity measurements. After interpolation, the methodology uses the convex hull method to further refine the cortical surface data's transformation. In this transformation, the convex hull defines the outer boundary of the set of points on the 2D grid, providing that all the projected points enclosed within the minimal possible perimeter. As a result, a flat representation of the occipital surface generated, keeping the data's spatial integrity.

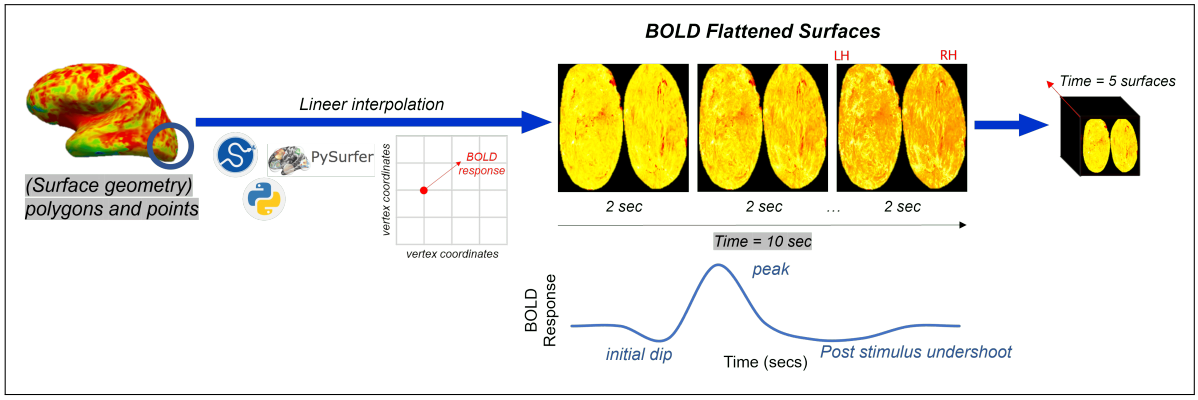

Figure 2.12. Data acquisition by interpolation of BOLD Occipital surfaces with related a stimulus hemodynamic response

2.3. ARCHITECTURE OVERVIEW

The study design and methodology have several key architectural components. This section provides an overview of the primary architectures used in the study, detailing their core components, loss functions, and the stages included in the training pipeline. The VQVAE model's loss function consists of three main components, each guiding the model towards generating high-quality reconstructions.

The ***reconstruction loss*** measures how close the reconstructed output image is to the original input image. It typically calculated using the mean squared error (MSE). Essentially, it ensures that the decoder generates outputs that are as similar as possible to the input images, focusing on the accuracy of the reconstruction.

The ***codebook update loss*** helps the model gradually improve codebook entries throughout training. It allows the selected codebook vector to better align with the encoder's output, making it more representative of the latent space as the model learns. This continuous refinement helps produce more precise latent representations, improving overall output quality.

The ***commitment loss*** is essential to ensure that the encoder adheres to a specific codebook entry, preventing latent representations from deviating excessively from the discrete codebook. This loss penalizes the difference between the encoder's output and the quantized codebook vector selected. A "stop-gradient" operation blocks gradients from propagating backward through this component during backpropagation. As a result, the encoder is encouraged to remain closely aligned with the chosen codebook entry.

$$\mathcal{L}_{VQVAE}(E,G,Z) = \underbrace{\|x-\hat{x}\|_2^2}_{\text{Reconstruction Loss}} + \underbrace{\|\text{sg}[E(x)]-z_q\|_2^2}_{\text{Codebook Loss}} + \underbrace{\|\text{sg}[z_q]-E(x)\|_2^2}_{\text{Commitment Loss}} \tag{2.1}$$

The VQGAN model adds additional layers of complexity to improve image quality. The ***perceptual loss*** ensures that the generated images capture high-level features similar to those observed in actual images. It is achieved by comparing the generated image features with those of a pre-trained neural network, such as VGG, providing the model with a 'realistic'

measure based on high-level feature alignment. An adaptive weighting term dynamically adjusts, allowing the model to simultaneously optimize both perceptual quality and fidelity and balance perceptual and adversarial losses during training.

Lastly, the ***adversarial loss*** component of the GAN framework pushes the model to generate images that are indistinguishable from real ones, improving the visual coherence and realism of the reconstructions.

$$Q^* = \arg \min_{E,G,Z} \max_{D} \mathbb{E}_{x \sim p(x)} \left[\mathcal{L}_{VQVAE}(E, G, Z) + \lambda \mathcal{L}_{\text{GAN}}(\{E, G, Z\}, D) \right] \tag{2.2}$$

where $\mathcal{L}_{\text{GAN}}(\{E, G, Z\}, D)$ is adversarial training with patch-based discriminator D, formulated as:

$$\mathcal{L}_{\text{GAN}}(\{E, G, Z\}, D) = [\log D(x) + \log(1 - D(\hat{x}))] \tag{2.3}$$

The adaptive weight λ is defined to balance the perceptual reconstruction loss and adversarial loss as:

$$\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta} \tag{2.4}$$

where $\delta = 10^{-6}$ is a small constant added for numerical stability, and G_L represents the gradient of its input concerning the last layer L of the decoder.

The LDM loss function designed to evaluate the model's ability to reconstruct noise in the latent representation of images. This loss function consists of several key elements, beginning with the latent representation, which contains the noisy latent code at a given time step. This noisy latent, derived through the latent diffusion process, forms the basis for the denoising process. The noise prediction component of the loss function allows the model to iteratively predict the noise introduced into the latent representation at each step of the diffusion process. This iterative approach enables the model to progressively reverse the added noise, accurately reconstructing the original latent representation.

The conditioning mechanism plays a crucial role in conditional generation tasks, such as text-to-image synthesis. Conditioning provides context that guides the denoising process, allowing LDMs to adjust the generated output according to specific conditions.

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\varepsilon(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{2.5}$$

where, ϵ is the actual noise sampled from a normal distribution, added to the latent representation during the diffusion process. ϵ_θ is the predicted noise added to the latent representation. $\tau_\theta(y)$ represents conditioning information that guides the model in generating output based on specific conditions. z_t is the noisy latent code at timestep t, derived from the encoder output of the latent diffusion process.

The objective is to minimize the difference between the actual noise ϵ and the predicted noise ϵ_θ, indicating how well the model can reconstruct the latent representation of the image from compressed latent space.

2.3.1. Image to Image VQGAN First Stage Model

In the first stage of the Image-to-Image VQGAN model, training performed in the pixel space to reconstruct images. The input images resized to `256×256` and normalized to `[1,1]`.

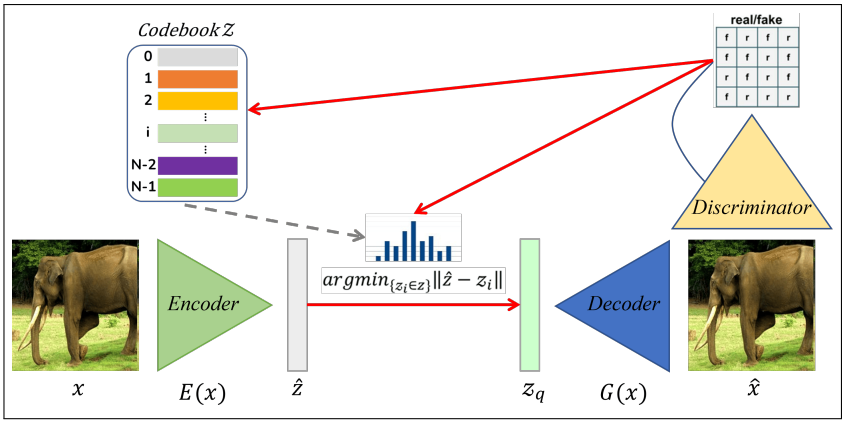

Figure 2.13. A schematic representation of Image to Image VQGAN model.

2.3.2. FMRI to Image VQGAN Network Model

The fMRI-to-Image VQGAN model trained on both surface-based and volume-based data modalities.

2.3.2.1. Surface-Based Network Model

A Surface-based model is used for decoding tasks. Each specific visual stimulus is paired with corresponding cortical surface representations, capturing neural activity during a 1 second stimulus presentation followed by a 9 second fixation period. This process is generates 5 surface-based volumes per stimulus. The scans scaled by their maximum value to normalize all pixel intensities, mathematically explained as $x'_{i,j} = \frac{x_{i,j}}{\max(x)}$ represents the original pixel intensity, whereas $x'_{i,j}$ is the normalized intensity.

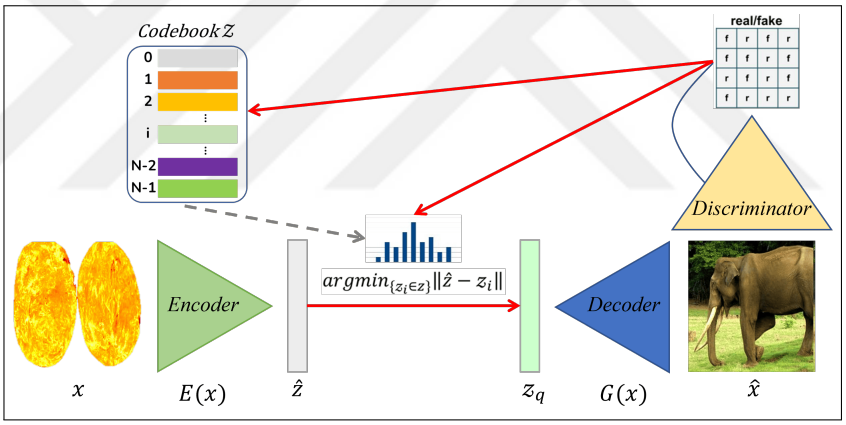

Figure 2.14. A schematic representation of surface-based VQGAN model.

The pre-trained decoder model of the *Image-to-Image VQGAN* framework used for decoding. The decoder weights were frozen in this configuration, allowing for effective decoding of neural activity patterns.

2.3.2.2. Volume-Based Network Model

In this stage, a 3D ResNet-101 model was used as the encoder to generate latent representations from volume-based fMRI data. The decoder used the pre-trained weights of an *Image to image VQGAN model*, enabling the decoding process to use prior information for reconstructing stimuli. This architecture facilitated training latent representations from the volume-based model, aligning them with the corresponding stimulus for accurate reconstruction. Data for

each run consisted of a series of fMRI volumes with dimensions `106×106×68` along the coronal, sagittal, and axial axes, respectively. These volumes were acquired every 2000 ms, producing 194 volumes per run. For model training, the fMRI volumes resized to `72×88×72×5`, where the dimension of 5 captured the slow event-related design comprising a 1-second stimulus presentation followed by a 9-second fixation period.

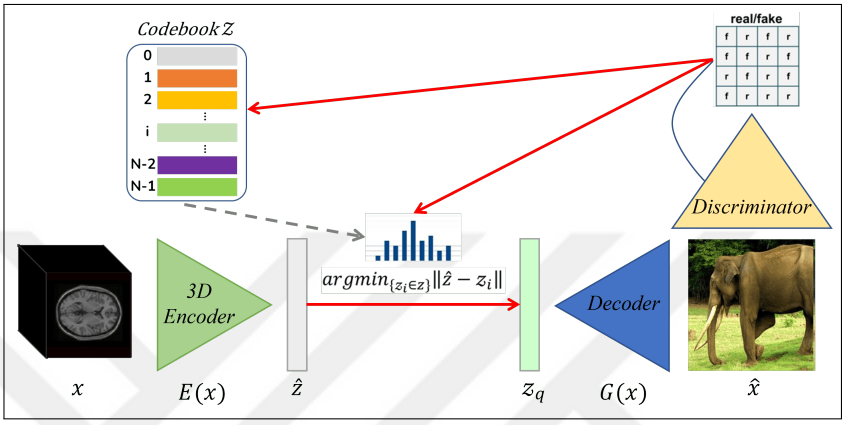

Figure 2.15. A schematic representation of volume-based VQGAN model.

2.3.3. Latent Diffusion Models

The conditioning mechanism trained using two different modalities, integrating fMRI data and visual stimuli.

2.3.3.1. Cross-attention conditioning

The cross-attention mechanism aligns two sequences —fMRI and stimulus representations— by allowing one sequence to focus on relevant details in the other. This mechanism effectively allows contextual information from one data type to guide or enhance another.

The query (Q) derived from the target sequence, which, in this case, corresponds to the latent representations of the generated images. The query aims to extract relevant guidance from the fMRI activity conditioning sequence. The key (K) and value (V) calculated from the conditioning, especially the stimulus latent representation, which encodes the brain activity patterns that guide the decoding process.

The attention mechanism calculates the relevance of each query to the keys through a dot

product operation, followed by a softmax function to normalize the attention scores:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.6}$$

where Q, K, and V are the query key, and value. These components derived from the intermediate representations using learnable projection matrices:

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\theta(y) \tag{2.7}$$

where $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ denotes a (flattened) intermediate representation of the U-Net implementing ϵ_θ and $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ are the learnable projection matrices. A domain-specific encoder τ_θ is introduced to preprocessing y from fMRI data. This encoder projects y into a latent representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$. It is then mapped to the intermediate layers of the U-Net through a cross-attention mechanism, enabling the alignment of neural activity patterns with the target visual stimuli.

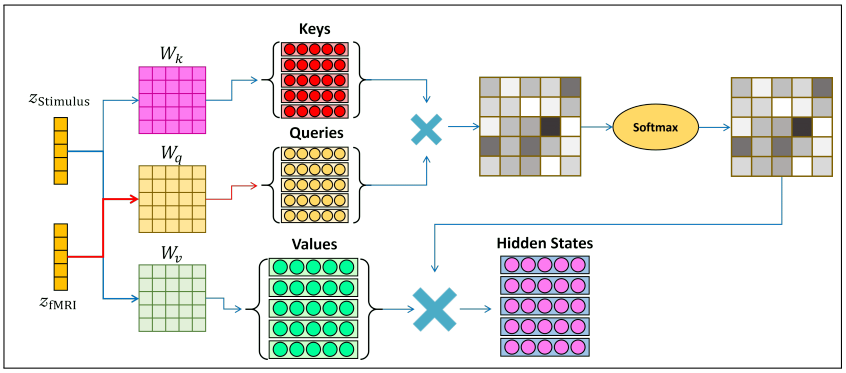

Figure 2.16. A schematic representation of cross attention mechanism.

2.3.3.2. Concatenation conditioning

The concatenation mechanism integrates the fMRI latent representation (z_{fMRI}). The stimulus latent representation (z_{stimulus}) by directly combining them along the channel dimension, resulting in a simple latent representation of shape $z_{\text{concat}} \in \mathbb{R}^{B \times 2C \times H \times W}$, where B represents the batch size, C denotes the number of channels in the latent representation, and H and W correspond to spatial dimensions.

2.4. IMPLEMENTATION DETAILS

The model trained on a Windows-based operating system with PyTorch 2.3.0 framework. The training performed on a 12 GB NVIDIA Geforce RTX 4070 GPU with 64 GB of RAM. In the first stage, *Image-to-Image VQGAN model*, is trained on the entire dataset, including all participants, to fine-tune the encoder and decoder. The data includes 5865 training samples, 1467 validation samples, and 30 test samples. This stage aimed to generate visual feature representations with a base `learning rate` of 4.5×10^{-6}, an embedding dimension of 1, and 8192 embeddings. The model configuration included an input resolution of 256, 3 input channels, and 3 output channels, with 128 channels fixed for processing and a channel multiplier of $[-1, -2, -4]$. Additional parameters included two residual blocks, no attention resolutions, a dropout rate 0.0, and a batch size set to 1. The loss parameters were configured in this stage with `disc_conditional` set to false, `disc_in_channels` as 3, `disc_start` of 0, `disc_weight` of 0.8, and `codebook_weight` of 1.0.

The second stage, *fMRI-to-Image VQGAN model*, was set up by including the decoder weights from the first stage model. This method facilitated the adaptation of the learned characteristics to address the complexity of the fMRI data. This stage used only single-subject data (CSI1) and included only the stimulus set from the ImageNet dataset. The input is a five-channel occipital surface model. The same model parameters from the first stage, including the base learning rate, the embedding size, and the channel configuration, were set with a batch size of 1. The training process took about 3 days, focusing on tasks that used the *fMRI-to-ImageNet VQGAN model* in pixel space. The fMRI data from Session 15 – Run 10 is used in the test stage, including 14 test samples, while the other sessions contain 1629 training samples and 408 validation samples. In the LDM model, a different configuration was applied, with a base learning rate of 1.0×10^{-6}, a linear start of 0.0015, and a linear end of 0.0205 over 1000 timesteps. The LDM trained using the `LambdaLinearScheduler` with parameters set to 10,000 warm-up steps and a long cycle length. The denoising U-Net configuration consisted of an image size of 64, 1 input and output channel, 192 model channels, attention resolutions of $[32, 16, 8, 4]$, 2 residual blocks, a channel multiplier sequence of $[1, 2, 2, 4]$, 8 heads and `resblock_updown` enabled. The loss function was $L1$, and the model conditioned by cross-attention to specific input data.

2.4.1. EXPERIMENTS

2.4.1.1. Evaluation Metrics

The model performance evaluated using 7 different image quality metrics, which evaluate high-level and low-level characteristics and provide an overview of image quality. These metrics provide a quantitative comparison between the proposed model and other existing literature.

Pixel-wise Correlation Metric (PixCorr), measures the linear correlation between the reconstructed and the original image in pixel space. The correlation coefficient calculated as follows:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.8}$$

where σ_X and σ_Y are the standard deviations of images X and Y, respectively, and $\text{cov}(X, Y)$ represents the covariance between the two images. This metric directly evaluates how closely the pixel intensities in the reconstructed image align with those in the original image, making it a useful measure for assessing low-level image quality.

Structural Similarity Index Metric (SSIM), is used to evaluate the structural similarity between two images based on luminance, contrast, and structure. The following formula gives the SSIM:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.9}$$

where μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance of x and y. C_1 and C_2 are constants that help stabilize the division in cases of weak denominators.

AlexNet n-way Comparison is performed by extracting features from the 2. and 5. layers of AlexNet (denoted as AlexNet(2) and AlexNet(5)) and comparing them using a 2-way comparison. The features of the last pooling layer of ***InceptionV3 n-way Comparison*** are

used for 2-way comparison, providing a better understanding of the image content.

SwAV-ResNet50 is used as a feature extractor to calculate the distance between images in the feature space. The distance metric defined as:

$$d = 1 - \frac{(\mu - \overline{\mu}) \cdot (\nu - \overline{\nu})}{\|(\mu - \overline{\mu})\|_2 \|(\nu - \overline{\nu})\|_2} \tag{2.10}$$

where μ and ν represent one-dimensional feature vectors, and $\overline{\mu}$ and $\overline{\nu}$ are their respective means. This metric helps quantify the similarity between reconstructed and original images based on high-level features.

EffNet-B represents distance metrics gathered from EfficientNet-B1 and is used for high-level comparison for evaluating how well the model preserves object recognition and semantic details. The first four metrics can considered low-level metrics, while the last three reflect high-level features. Together, these metrics provide a holistic view of the provided model's performance and allow for meaningful comparisons with other existing approaches.

2.4.1.2. Average BOLD Signal Analysis and Fourier Transform Analysis for Slow Event-Related Design

The average BOLD signal in V1 area analysis used for Session 01 - run 01-02. The method aims to evaluate the consistency and reliability of brain activation patterns across runs while also evaluating the temporal components of the hemodynamic response to subsequent stimuli. The temporal evolution of the BOLD signal examined to identify peak response times and evaluate the extent to which the signal returned to baseline between stimulus presentations. A Fourier transform was applied, using the `np.fft.fft` function from `Numpy`, to decompose the BOLD signals into their frequency components. Dominant frequency patterns and periodicities identified to learn more about the temporal structure of the BOLD response. The initial and final fixation periods excluded for frequency analysis to focus on stimulus-related activity and minimize contamination from basal fluctuations.

2.4.1.3. Evaluation Methodology

The evaluation process applied across all models' stages and configurations, encompassing training and test data sets to provide a comprehensive performance analysis. For quantitative evaluation, six sample images generated for each input image. Performance metrics were calculated for each sample individually by comparing them with the corresponding input image. The average metric value then calculated for each image, and the dataset's overall performance was obtained by averaging these per-image results across the entire set. The qualitative evaluation includes visual inspection of reconstructed outputs to assess structural integrity, texture, and alignment with input images or neural activity patterns. This dual evaluation allows for a detailed assessment of quantitative accuracy and perceptual coherence.

2.4.1.4. Evaluation of Image to Image VQGAN Model Evaluation

The first-stage model was assessed to evaluate its ability to reconstruct images in pixel space. The model's performance on training and test datasets analyzed to determine its effectiveness in capturing and reproducing visual features.

2.4.1.5. Evaluation of Cross-Attention and Concatenation Mechanisms in Surface-Based Models

The comparison between surface-based LDM with cross-attention and concatenation mechanisms focused on their ability to decode and reconstruct neural activity patterns. The evaluation performed for training and test datasets to provide a balanced perspective on performance.

2.4.1.6. Evaluation of Volume-Based Model and Surface-Based Model

The performance of volume-based and surface-based models compared to determine their relative effectiveness in decoding neural activity patterns. Both models were evaluated in training and test datasets, emphasizing their strengths and limitations in different fMRI modalities.

2.4.1.7. Evaluation of fMRI Encoder: Frozen and Trainable Configurations

The fMRI encoder analyzed in two configurations: frozen (using pre-trained features) and trainable. Both configurations were evaluated on training and test data sets to assess their impact on decoding performance within the LDM's conditional mechanism.

3. RESULTS

3.1. AVERAGE BOLD SIGNAL AND FOURIER ANALYSIS

The analysis of the averaged BOLD signal in the V1 region of the visual cortex revealed distinct patterns across runs, providing insights into the dynamics of neuronal activity in response to visual stimuli. In the first run, the BOLD signal consistently peaked approximately 6-8 seconds after each stimulus onset, followed by a partial return to baseline. However, the signal did not fully return to baseline before subsequent stimulus presentation, resulting in a cumulative overlap effect. This pattern became more complex as the session progressed, reflecting the continuous accumulation of hemodynamic responses over successive stimuli.

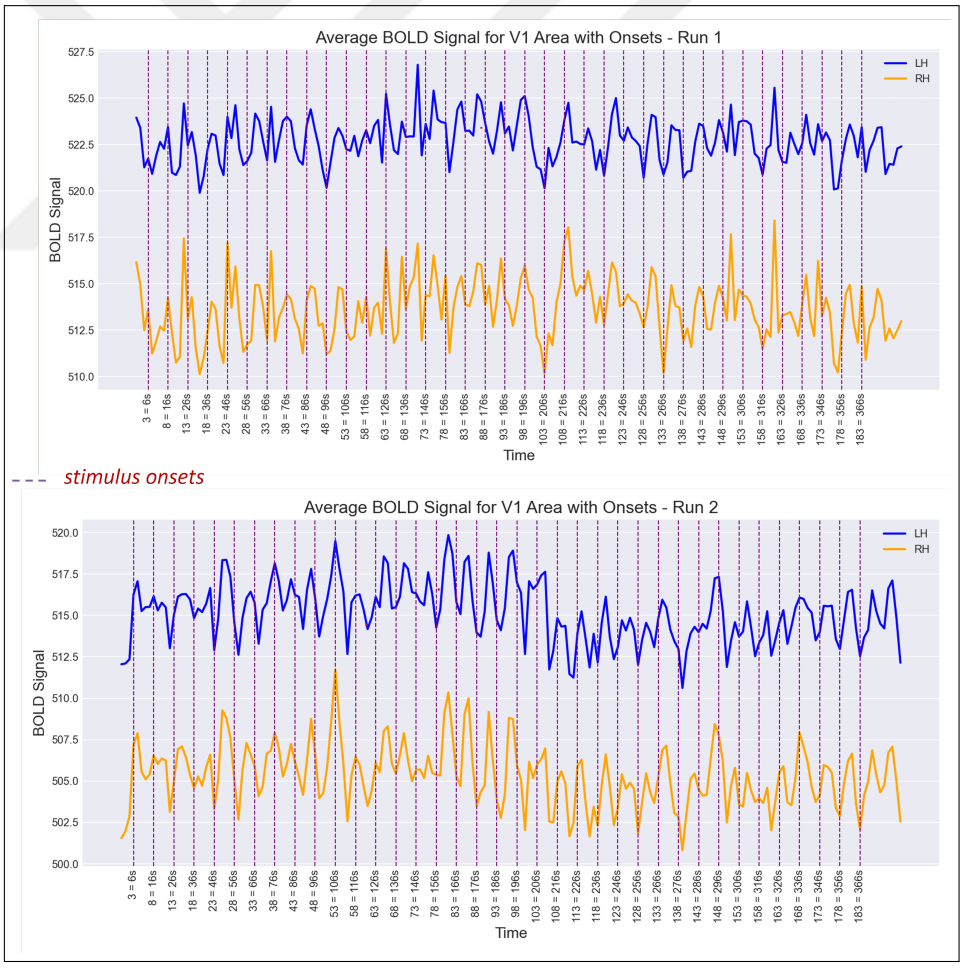

Figure 3.1. Overall average signal with all stimulus onsets for runs 1-2 of Session-01 for participant CSI1

In the second run, an unexpected early increase in the BOLD signal was observed prior to

the presentation of the first stimulus. This could be attributed to baseline activation or the participant's task anticipation. As the run progressed, the overlapping effect became more pronounced. These findings suggest that residual hemodynamic responses from previous stimuli influenced subsequent responses, posing challenges for interpreting distinct neural activation patterns.

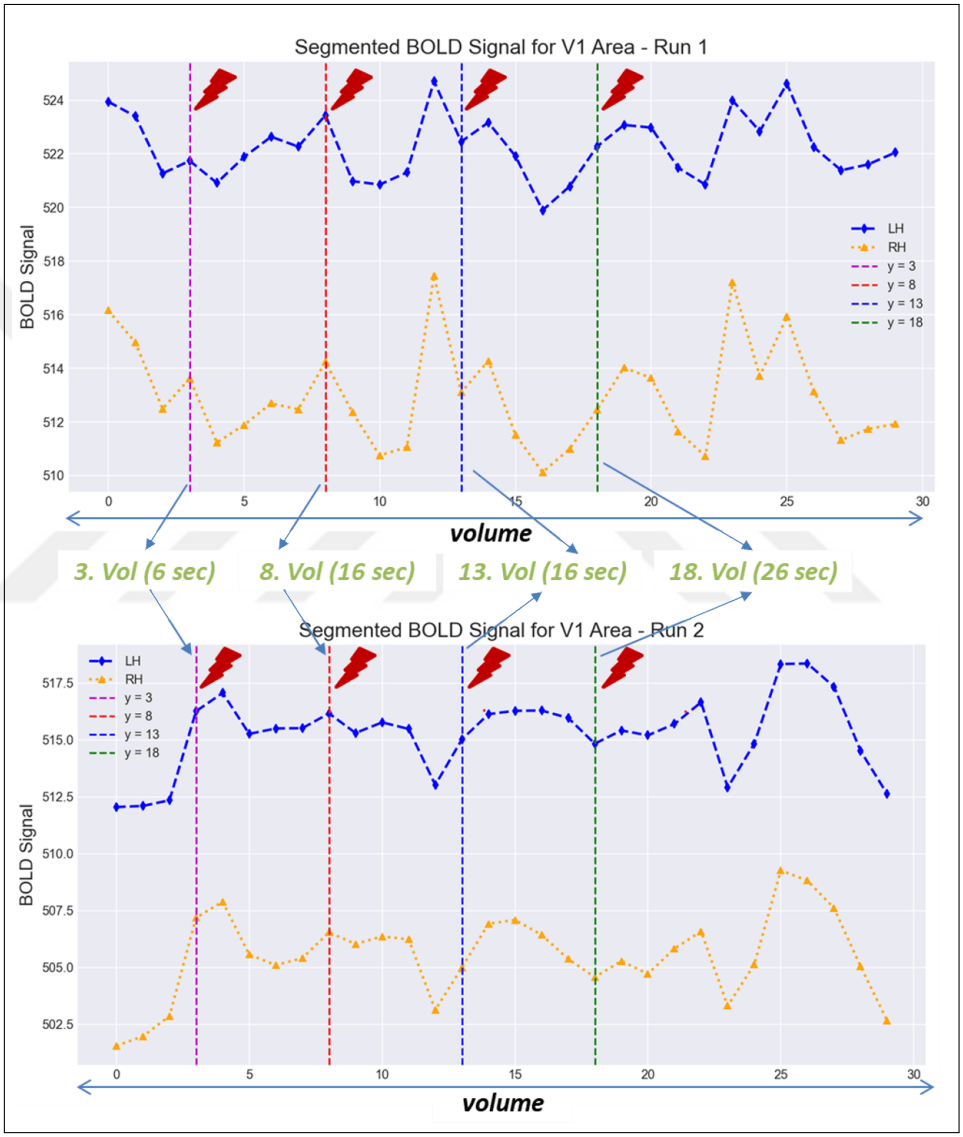

Figure 3.2. Average signal with first 4 stimulus onsets for Session-01 Runs 1-2 for participant CSI1

In this analysis, a Fourier transform was applied to extract the frequency components of the BOLD signals, focusing on identifying dominant frequency patterns and periodicities. The analysis excluded the initial and final fixation periods to ensure that only stimulus-related activity was captured. The resulting frequency plots, as shown in Figure 3.4, illustrate the

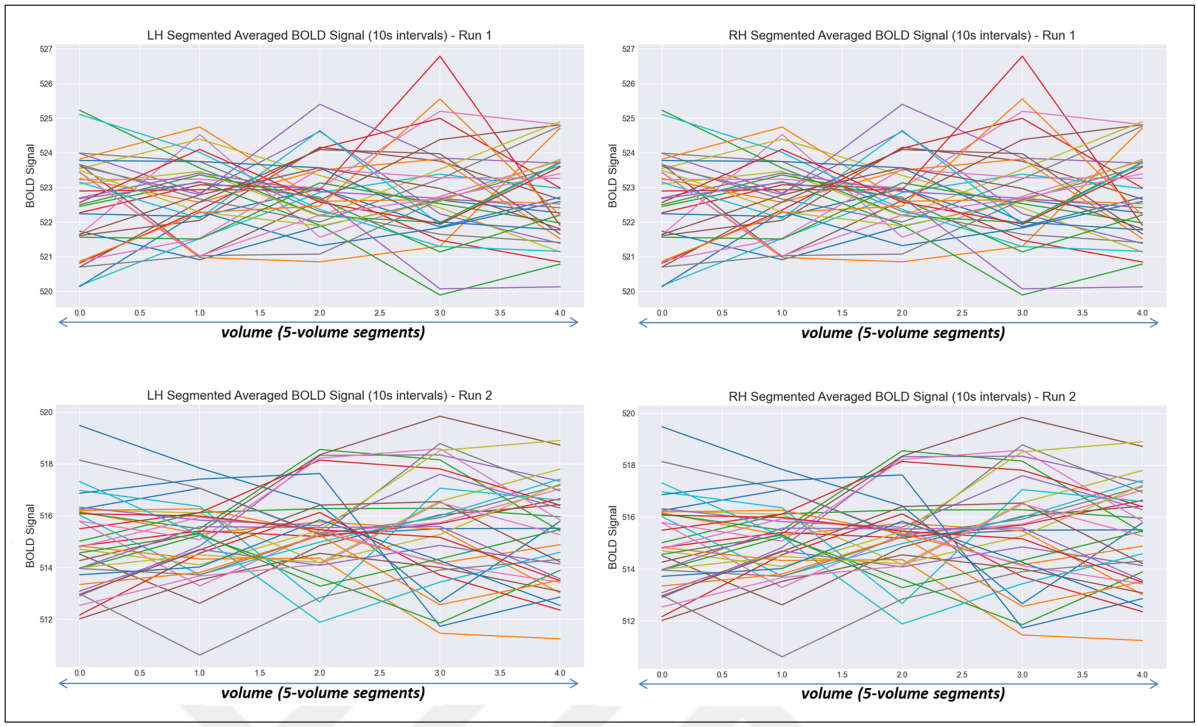

Figure 3.3. BOLD Signal with 10 secs windows of stimulus presentations and FCs (excluding initial and final FCs (left col: LH right col: RH)

dominant frequency components of the BOLD signal for the first two runs of the first session.

The Table 3.1 presents the maximum amplitude values and their corresponding periods for both hemispheres in all runs in the first session. These results offer quantitative insights into the periodic nature of the BOLD signal and its alignment with the experimental design. According to the original study, the 10-second intervals between stimuli were designed to allow the hemodynamic response to return to baseline before the subsequent stimulus onset. However, upon evaluating the data, it was observed that in specific runs, the BOLD signal required more than 10 seconds to return to baseline.

This cumulative effect may occur, where residual hemodynamic responses from prior stimuli overlap with responses to subsequent stimuli. This effect can lead to challenges in isolating stimulus-specific activity, reducing clarity, and signal resolution. Furthermore, these delays may contribute to overfitting during model training, as the model could interpret the cumulative effects of repeated stimuli as activity specific to individual stimulus. This could reduce the model's ability to generalize effectively to specific stimuli.

Additionally, the BOLD5000 used in this study was derived from 3 different datasets, each

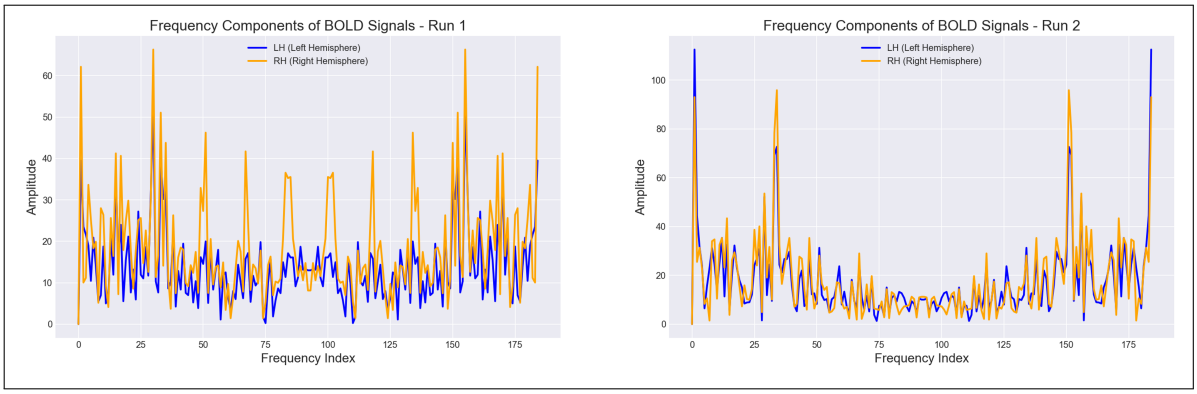

Figure 3.4. Frequency Components of BOLD Signals in LH and RH for Session 01 - Run 1-2

containing various scene categories that were presented to participants in a mixed manner in all runs and sessions. This may have introduced additional complexity to the BOLD signal, as cumulative effects could partially originate from responses to stimuli not included in the model's training data. This adds another layer of difficulty in interpreting neural activity, as the model may struggle to process and decode information it has not previously encountered.

Table 3.1. The maximum amplitude and the corresponding period values of BOLD signals from the LH and RH are determined through Fourier analysis.

Run	LH Max Amplitude	LH Max Period	RH Max Amplitude	RH Max Period
1	52.613	12.333	66.305	12.333
2	72.658	10.882	95.850	10.882
3	75.706	11.212	109.077	11.212
4	117.088	11.212	131.549	11.212
5	91.535	10.882	124.222	10.882
6	106.703	11.212	141.697	11.212
7	79.393	13.704	92.925	12.333
8	69.584	10.000	79.716	10.278
9	82.624	10.882	98.172	10.882
10	84.564	11.212	109.056	11.212

The graphs on the left in Figure 3.5 show the normalized BOLD signal on the y-axis and time in 10-second intervals on the x-axis, illustrating the average time course across all stimulus trials for each region of interest. This analysis examines whether the slow event-related design minimizes bleed-over from neighboring trials and compares the findings with the original study. Following each stimulus presentation, the BOLD signal in these regions exhibits a

steady increase, peaking approximately 6-8 seconds after stimulus onset before gradually returning to baseline. This pattern reflects the inherent delay in the hemodynamic response and its gradual decay following the stimulus.

The graphs on the right in Figure 3.5 , taken from the original study, depict the average time course of the BOLD signal across all stimulus presentations for each region of interest (ROI). These values represent the BOLD signal averaged across all voxels within each ROI and across all stimulus trials. Comparing these original study graphs with the current analysis highlights several key similarities. The results confirm that the BOLD signal consistently peaks 6-8 seconds after stimulus presentation in both analyses, followed by a gradual return to baseline. This observation emphasizes the reproducibility of the hemodynamic response in the V1, V2, and V5 regions. The consistency of these patterns demonstrates that the visual cortex responds similarly to stimuli in both studies, showcasing comparable results in terms of the reliability of the experimental design and analysis.

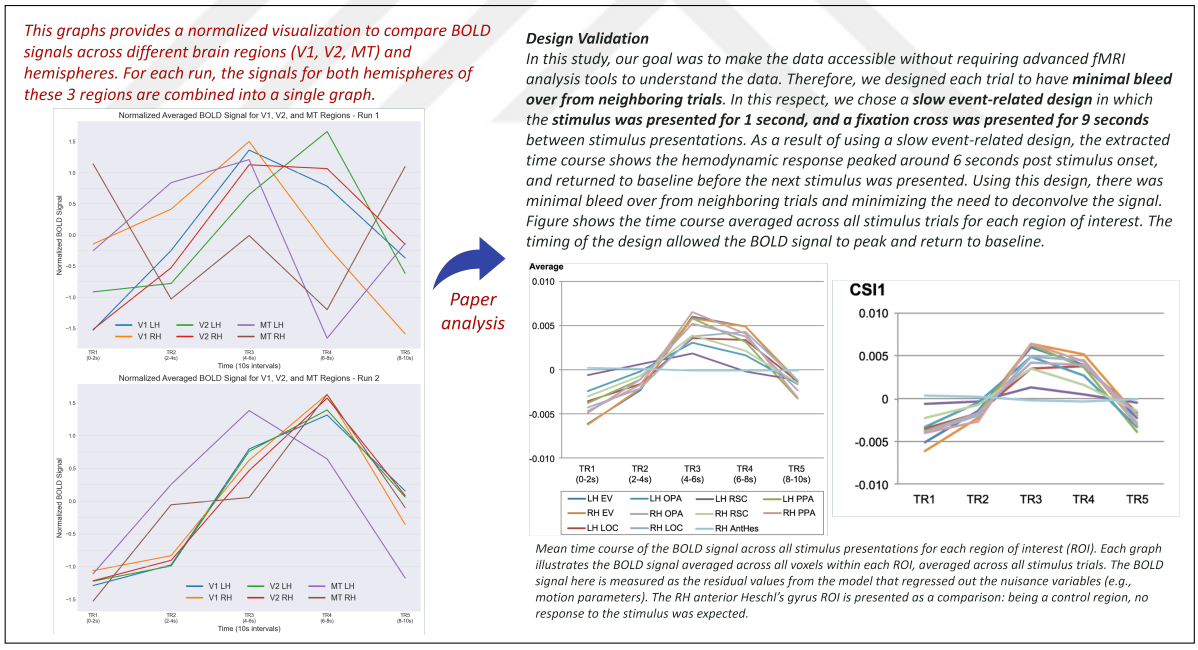

Figure 3.5. (left) A normalized visualization to compare BOLD signals over different brain regions (V1, V2, MT) and hemispheres. For each run, the signals for both hemispheres of these 3 regions are combined into a single graph. (right) Design validation studies from original paper

In addition to these analyses, the examination of the average BOLD signal and the frequency components derived from the Fourier transform reveals bleed-over between neighboring trials at specific time intervals. In particular, it was observed that, in some cases, the BOLD signal

did not fully return to baseline before the onset of the next stimulus. This situation resulted in the cumulative effect of hemodynamic responses between stimuli, making it more challenging to isolate signals specific to individual stimuli.

3.2. QUALITATIVE RESULTS OF FIRST STAGE AND LATENT DIFFUSION MODELS WITH DIFFERENT MODALITIES

3.2.1. Image to Image VQGAN Model

The results of the first-stage model were evaluated on both the training and test datasets. The model successfully aligned with the stimulus images, demonstrating significant convergence, as evidenced by the qualitative results shown in Figure 3.6 and Figure 3.7. The outputs generated indicated the model's ability to encode and reconstruct significant visual features by capturing not only the basic colors and forms of the input images but also their overall contextual structure. These findings demonstrate that the model effectively learned an accurate representation of the input images, resulting in high-quality reconstructions.

Figure 3.6. Qualitative results of image to image first stage model on train set (left column is original stimulus) (Images in red boxes represent 4 reconstructed sample images corresponding to the original stimulus on the left)

Figure 3.7. Qualitative results of the image to the image first stage model on the test set (left column is original stimulus) (Images in red boxes represent 4 reconstructed sample images corresponding to the original stimulus on the left)

3.2.2. fMRI to Image VQGAN Models with Different Modalities

3.2.2.1. Surface-Based Model

The surface-based model of the fMRI-VQGAN framework demonstrated a limited ability to reconstruct high-quality images from the training set, as shown in Figure 3.8. Although the overall quality of the images was low, the shapes, colors, and positions of objects within the images were sufficiently distinct to allow the general silhouettes to be recognized. These results suggest that the model captured some level of structural and spatial information during training.

For the test set, illustrated in Figure 3.9, the model's performance was limited. The generated images displayed vague forms, some images showing partial resemblance to the original colors. However, this similarity was inconsistent and varied between the data sets, making the observations somewhat subjective. Although specific images retained similar color distributions, this similarity was not consistently apparent in all cases.

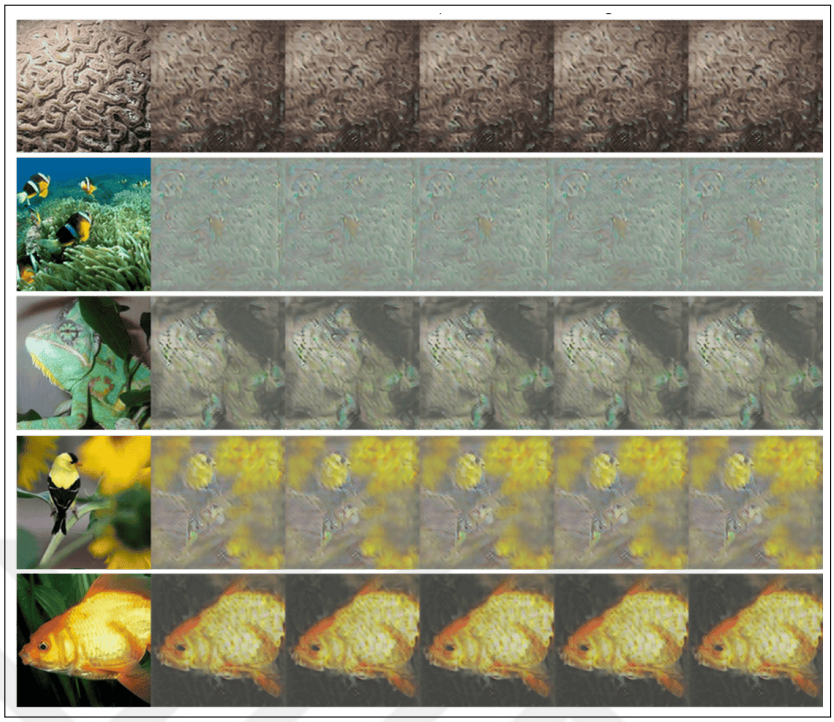

Figure 3.8. Qualitative results of the surface-based model in train dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

Figure 3.9. Qualitative results of the surface-based model in test dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

3.2.2.2. Volume-Based Model

The results of the volume-based model demonstrated noticeably different performance compared to the surface-based model. The generated images were often confusing, characterized by uniform and hazy forms in the data set as seen in Figure 3.10. The model failed to capture the unique features of each individual stimulus, resulting in repeated and generic outputs that lacked distinction between stimuli. This observation indicates that the volume-based model struggled to generate representations that were specific to the input data.

Similar patterns were observed during the test phase Figure 3.11, where the model consistently produced ambiguous reconstructions. This uniformity in the generated images suggests a potential underfit, as the model failed to learn the specific patterns associated with individual stimuli. Despite periodic evaluations of the reconstructions throughout training, the results remained largely unchanged, reinforcing the idea that the model may not have effectively captured the complexities of the volume-based input data.

Figure 3.10. Qualitative results of the volume-based model in train dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

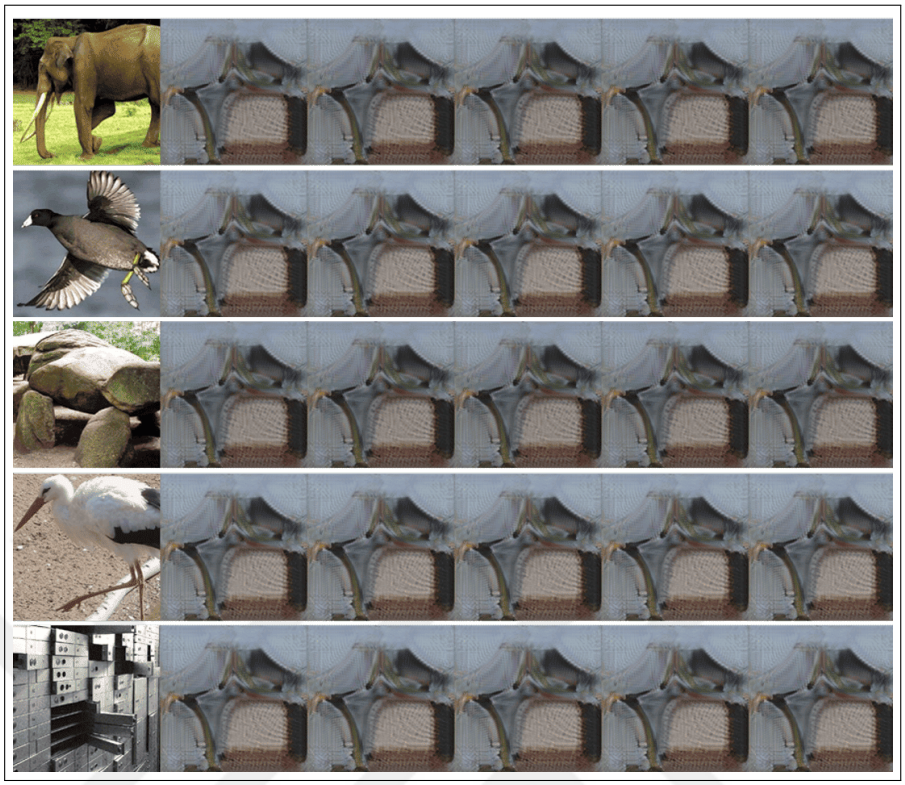

Figure 3.11. Qualitative results of the volume-based model in test dataset samples (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

3.2.3. Conditioning Mechanisms in LDM: Cross-Attention and Concatenation

3.2.3.1. Surface-Based Model with Cross Attention Mechanism

The performance of the surface-based LDM guided by the cross-attention conditioning mechanism was evaluated on both the training and test sets, as seen in Figure 3.12 and Figure 3.13, respectively. The analysis revealed that while the model occasionally generated images with colors and textures that were somewhat similar to the original data, the reconstructions lacked consistency.

When examining the 4 reconstruction samples for each stimulus, the model frequently produced outputs that appeared irrelevant or random, failing to capture the defining features of the input data. Although certain samples exhibited partial similarities, such as general color gradients or basic textures, these instances were infrequent and did not represent the majority of the reconstructions.

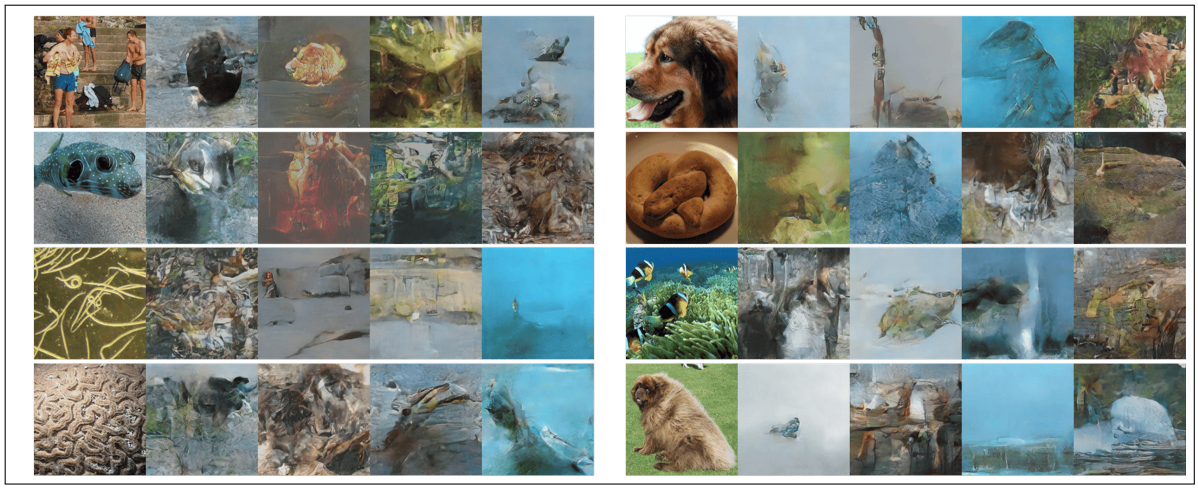

Figure 3.12. Qualitative results of the surface-based model in train dataset samples on LDM with cross attention conditioning (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

Figure 3.13. Qualitative results of the surface-based model in test dataset samples on LDM with cross attention conditioning (The first column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

3.2.3.2. Surface-Based Model with Concatenation Mechanism

This section presents the qualitative results of the LDM with a concatenation-based conditioning mechanism, focusing on the surface-based data in both the training and test sets.

For the training set, the reconstructions generated by the surface-based model generally exhibit preserved structure and finer details that are more consistent with the original stimuli,

as shown in Figure 3.14. The concatenation mechanism appears to enable the model to capture recognizable shapes and textures, resulting in reconstructions that maintain a certain level of visual coherence. The consistent alignment of key features across multiple reconstructions suggests that the model effectively uses the conditioning information provided by the surface-based data.

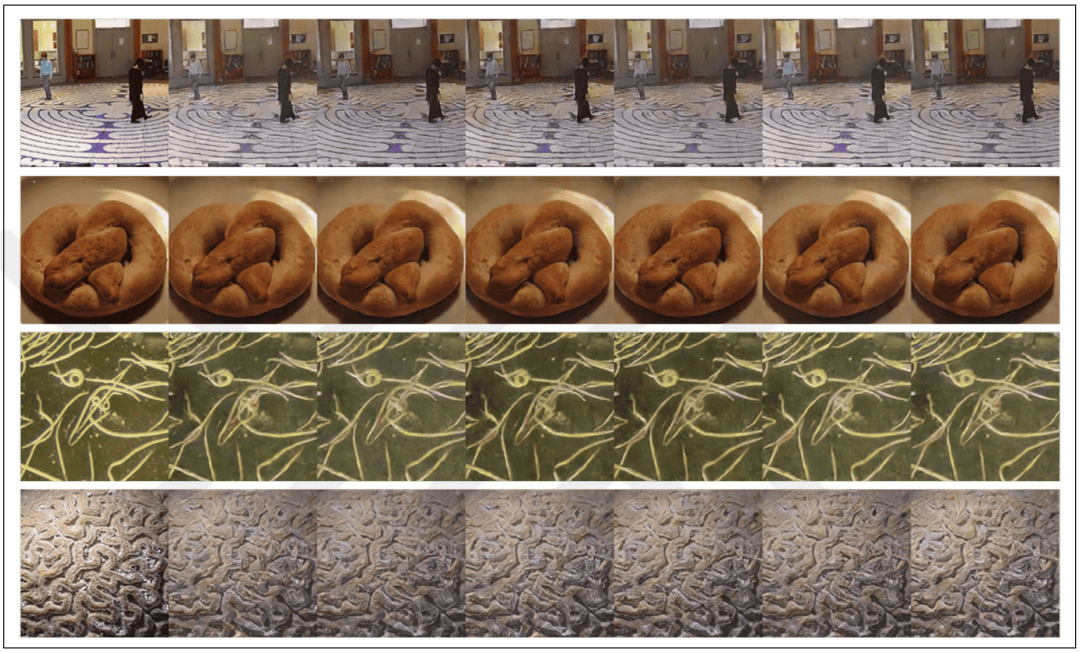

Figure 3.14. Qualitative results of the surface-based model in train dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

In the test set, the results indicate a decline in reconstruction quality, as shown in Figure 3.15. While the surface-based model still produces consistent textures, the generated images often lack clarity and specific details, leading to more abstract and less distinct representations. These outputs highlight the limitations of the concatenation conditioning mechanism when applied to unseen data, as the reconstructions fail to fully capture the nuanced features of the original stimuli.

Figure 3.15. Qualitative results of the surface-based model in test dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

3.2.3.3. Surface-Based Model with Trainable Conditioning Concatenation Mechanism

In the training set, the model showed the ability to generate reconstructions that were contextually aligned with the stimuli, as shown in Figure 3.16. The reconstructions partially exhibited diverse textures and colors, displaying distinct yet similar patterns. This indicates that the model effectively captured the general features of the training data while encoding stimulus-specific characteristics.

For the test set in Figure 3.17, the model showed a decrease in reconstruction quality, producing images lacking clarity and distinct features. The generated images often exhibited ambiguous characteristics that did not align closely with specific attributes of the original stimuli. The outputs were characterized by repetitive patterns, suggesting that the fMRI latent representations did not effectively generalize to unseen data.

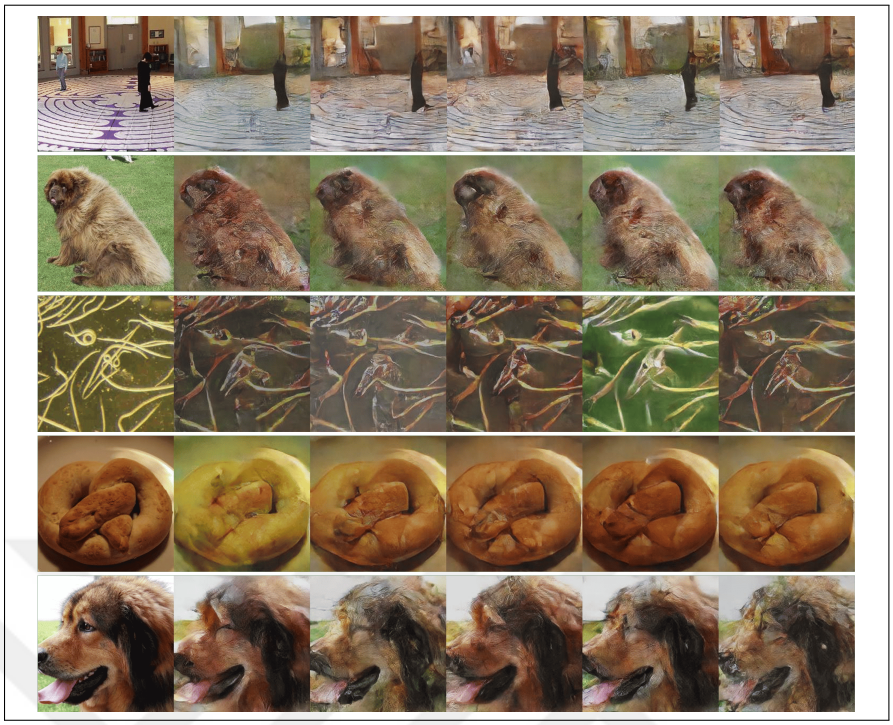

Figure 3.16. Qualitative results of the surface-based model in the train set samples on LDM with trainable concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

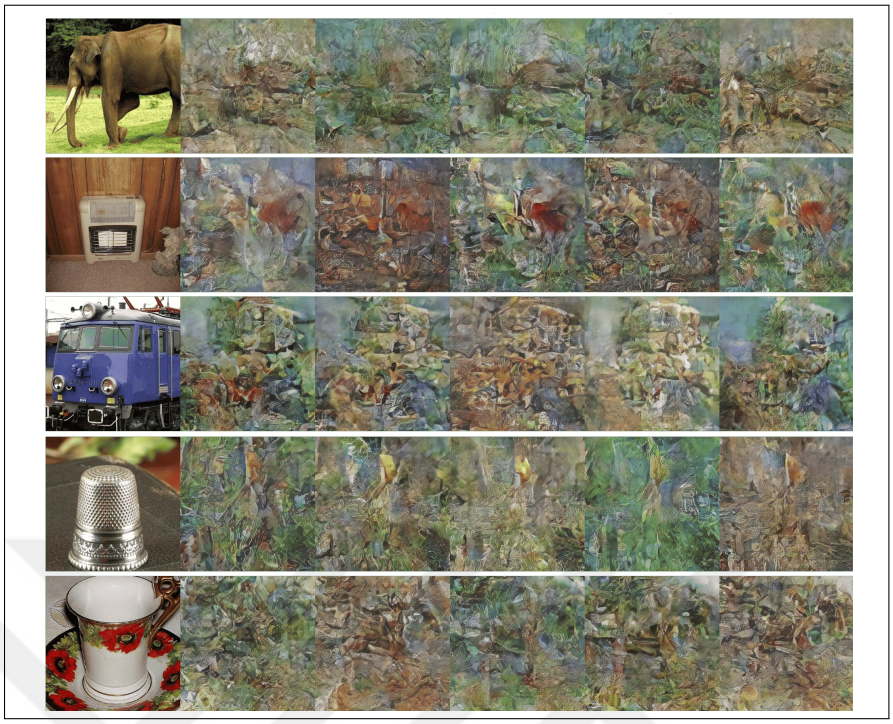

Figure 3.17. Qualitative results of the surface-based model in test set samples on LDM with trainable concatenation conditioning (First column is original stimulus) (Images on the right represent 5 reconstructed sample images corresponding to the original stimulus on the left)

3.2.3.4. Volume-Based Model with Concatenation Mechanism

The qualitative results of the LDM with a concatenation-based conditioning mechanism focus on the volume-based data in both the training and test sets.

For the training set, the volume-based model generates reconstructions that are more abstract and less structured compared to the surface-based approach, as shown in Figure 3.18. Although some textures and color distributions in the outputs vaguely resemble the original stimuli, the model struggles to replicate clear and recognizable forms. The generated images exhibit considerable variation, suggesting instability in capturing specific details of the stimuli. This lack of stability indicates challenges in using volume-based data to produce coherent representations during training.

In the test set, the reconstructions show increased abstraction and divergence from the original stimuli, as illustrated in Figure 3.19. Although there is greater diversity in texture and color, the generated images often do not provide meaningful resemblance to the visual characteristics of the input data. The outputs lack the clarity and structural fidelity required for an accurate

Figure 3.18. Qualitative results of the volume-based model in train dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 6 reconstructed sample images corresponding to the original stimulus on the left)

depiction, highlighting the difficulty of the model in generalizing to unseen data.

Figure 3.19. Qualitative results of the volume-based model in test dataset samples on LDM with concatenation conditioning (First column is original stimulus) (Images on the right represent 6 reconstructed sample images corresponding to the original stimulus on the left)

3.3. QUANTITATIVE RESULTS OF FIRST STAGE AND FMRI VQGAN MODEL WITH DIFFERENT MODALITIES

The first-stage model performed strongly in both train and test datasets. The metrics decreased slightly from train to test set, showing that the model aligns well at the pixel level, preserving structural similarity and efficiently capturing low-level information. Additionally, high-level metrics confirmed that the model successfully captured higher-level semantics, aligning with qualitative observations and validating the model's overall performance.

For the surface-based fMRI-VQGAN model, training set metrics outperformed those of the test set, but a significant drop in test set performance indicated potential overfitting and limited generalizability. The high-level metrics for the test set further underlined this limitation, as the scores were even lower, supporting the conclusion of overfitting. In contrast,

Table 3.2. Quantitative results of the reconstructed image quality on the test set. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better.

in LDM Model	Method	Dataset	Quantitative Measures						
			Low-Level Metrics				High-Level Metrics		
			PixCorr	SSIM	AlexNet(2)	AlexNet(5)	Inception	EffNet-B	SwAV
First Stage (pixel space)	Stimulus VQModel	train	.911	.586	92.6%	92.6%	92.6%	.146	.103
		test	.898	.536	92.1%	92.1%	94.1%	.255	.124
Conditioning	fMRI VQModel (surf)	train	.694	.336	94.1%	94.1%	92.7%	.638	.348
		test	.169	.231	94.1%	94.1%	52.4%	.972	.509
	fMRI VQModel (vol)	train	.131	.202	94.1%	94.1%	47.0%	.939	.494
		test	.176	.234	93.6%	93.6%	46.3%	.951	.473

the volume-based model showed lower overall performance in both datasets. However, its results were more stable in training and test sets, with some metrics even performing better in the test set than in the training set. This stability and lack of variation suggest potential underfitting, as the model appeared unable to learn the specific patterns required to improve its performance meaningfully.

When comparing the test set performance of surface-based and volume-based models, the volume-based model achieved slightly better scores on most metrics, except for AlexNet n-way comparisons and Inception metrics. Despite this, qualitative observations revealed that the volume-based model frequently generated the same image throughout the data set, indicating a lack of diversity and robustness in its reconstructions. This discrepancy highlights that

the slightly improved test set metrics for the volume-based model may not accurately reflect its qualitative performance, pointing to inconsistencies in its ability to generate meaningful output.

3.4. QUANTITATIVE RESULTS OF LATENT DIFFUSION MODEL WITH DIFFERENT MODALITIES

Based on the quantitative results provided in Table 3.3, a detailed analysis of the performance of LDM is performed. This includes comparisons across various conditioning mechanisms and dataset configurations, emphasizing the metrics that indicate poor or strong performance.

The surface-based cross-attention model demonstrates balanced but limited performance. While structural similarity metrics such as PixCorr and SSIM suggest consistent alignment between the training and test sets, the overall scores remain low. This indicates moderate success in capturing the structural layout of the data but limited refinement. Despite its steady low-level metrics, the high-level scores, such as EffNet-B and SwAV, remain high, reflecting poor semantic representation. This is consistent with qualitative results, where the model generated abstract but structurally aligned outputs. While it generalizes better than other configurations, it still struggles to capture meaningful high-level features.

Switching to the concatenation mechanism results in strong training set performance but significant overfitting. Structural metrics are notably high on the training set, reflecting the model's ability to learn patterns in seen data. However, these metrics drop sharply on the test set, signaling the model's failure to generalize. While low-level features are captured effectively in the training phase, the high-level metrics remain poor, aligning with qualitative observations of overfitted, less generalizable outputs. Although this configuration captures fine-grained details during training, its overall performance is undermined by its inability to extend this knowledge to unseen data.

Making the fMRI encoder trainable introduces slight improvements in balancing structural similarity and high-level semantic metrics. While the training set scores remain reasonable, test set results show moderate improvements compared to the frozen encoder concatenation model. The structural fidelity remains limited, but the semantic preservation shows

Table 3.3. Quantitative results of the reconstructed image quality. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. The best results for some metrics are shown in green cells.

Method	Dataset	Quantitative Measures							
		Low-Level				High-Level			
		PixCorr	SSIM	AlexNet(2)	AlexNet(5)	Inception	CLIP	EffNet-B	SwAV
Ours (surf-crossatt)	train	.149	.186	94.1%	94.1%	43.2%	-	.943	.484
	test	.142	.228	93.6%	93.6%	49.5%	-	.980	.505
Ours (surf-concat)	train	.890	.530	94.1%	94.1%	93.8%	-	.235	.169
	test	.092	.032	93.6%	93.6%	49.2%	-	.964	.540
Ours (surf-concat, *fMRI encoder trainable*)	train	.749	.369	94.1%	94.1%	85.6%	-	.653	.374
	test	.083	.142	93.6%	93.6%	43.5%	-	.993	.482
Ours (vol-concat)	train	.152	.182	94.1%	94.1%	49.6%	-	.945	.492
	test	.143	.249	93.6%	93.6%	44.7%	-	.947	.487

slight gains. This configuration highlights the potential of tuning the encoder to improve generalization and achieve a better balance between the datasets. However, it remains limited in its ability to achieve substantial improvements in retaining high-level features.

The volume-based model with the concatenation mechanism exhibits consistently low performance across both the training and test sets, a clear indication of underfitting. The structural similarity metrics, such as PixCorr and SSIM, remain low for both datasets, demonstrating the model's inability to learn detailed structural patterns effectively. Additionally, while AlexNet scores suggest recognition of basic shapes and patterns, they are insufficient to indicate a meaningful representation of finer details or higher-level semantics.

When comparing the cross-attention mechanism with the concatenation mechanism in surface-based models, the cross-attention mechanism demonstrates better generalization across both training and test datasets. Its relatively consistent performance suggests that the model is less prone to overfitting. However, despite its stability, the cross-attention mechanism struggles to preserve high-level semantics, as indicated by weaker performance in metrics that measure semantic fidelity.

In contrast, the concatenation mechanism shows a stronger ability to capture finer details and structural patterns in the training set. However, this comes at the cost of severe overfitting, as evidenced by a significant drop in performance when evaluated on the test set. While the concatenation mechanism excels at capturing low-level details during training, it fails to generalize effectively, resulting in poor performance on unseen data.

Switching from a frozen to a trainable fMRI encoder within the concatenation-based surface

model results in slight improvements in generalization. The trainable encoder balances structural similarity and semantic preservation better than its frozen counterpart. While the trainable setup reduces overfitting to some extent, the overall improvements are moderate and do not fully resolve the limitations in semantic representation. The trainable encoder slightly improves the model's ability to learn higher-level features, but overall performance remains limited.

Surface-based models generally outperform volume-based models in their ability to capture and preserve structural details, as seen in their relatively higher PixCorr and SSIM scores. They are better suited for capturing fine-grained patterns and low-level information, particularly in the training set. However, surface-based models are more prone to overfitting, especially when the concatenation mechanism is used. However, volume-based models exhibit consistent but insufficient performance in both datasets. This consistency, while avoiding overfitting, highlights the model's inability to learn meaningful patterns effectively, leading to underfitting. Volume-based models fail to capture both low-level structural details and high-level semantic features, making them less effective overall compared to surface-based configurations.

3.5. QUALITATIVE COMPARISON OF RESULTS WITH EXISTING STUDIES IN LITERATURE

The models developed by Takagi et al. shown in Figure 3.20 are based on single-trial beta weights derived from fMRI data. The reconstructed images produced by these models can be easily recognized in terms of basic shapes or silhouettes; however, fine low-level details and overall naturalness are not captured. In the 'Z' columns, generated from fMRI activity alone, general color gradients and basic spatial characteristics are often retained, but the content and semantic meaning of the original stimuli are not accurately conveyed. For example, an image of an airplane reconstructed in the 'Z' columns appears as abstract forms resembling buildings, and at times, the aircraft is reconstructed as a blurry mixture of shapes that could be interpreted as multiple objects or scenes. Similarly, a clock tower is transformed into an unrelated shape, resembling a human or animal figure, rather than preserving its original architectural form. These inconsistencies underline the challenges faced in reconstructing

detailed and semantically rich images from fMRI data alone.

In contrast, our model's performance is qualitatively weaker, with reconstructed images that exhibit significant deficiencies in both low-level details and semantic accuracy. Although some basic structures are recognizable, fine-grained elements and semantic richness are absent, and the reconstructed images do not convey the same level of meaning as the original stimuli. These results show that further improvement is needed to improve the model's ability to accurately capture both low and high level features of visual stimuli.

Ozcelik et al. in Figure 3.21 used a different method, using single-trial beta weights derived from GLMs that were optimized using GLMDenoise and ridge regression. The CLIP Vision model was used to transform the beta weights into visual features, while the CLIP text-to-image model was used to convert the textual descriptions into textual features. These two feature sets were used as conditions for the image reconstruction task. Ozcelik's model showed high performance, especially in complex scenes where structural integrity was preserved in various stimuli. Minor qualitative differences in pixel-level details and contrast were observed, but the model was successful in maintaining consistency and capturing high-level semantic information. Using both visual and textual features as conditions improved their model's ability to capture stimulus complexity, resulting in more coherent and detailed reconstructions.

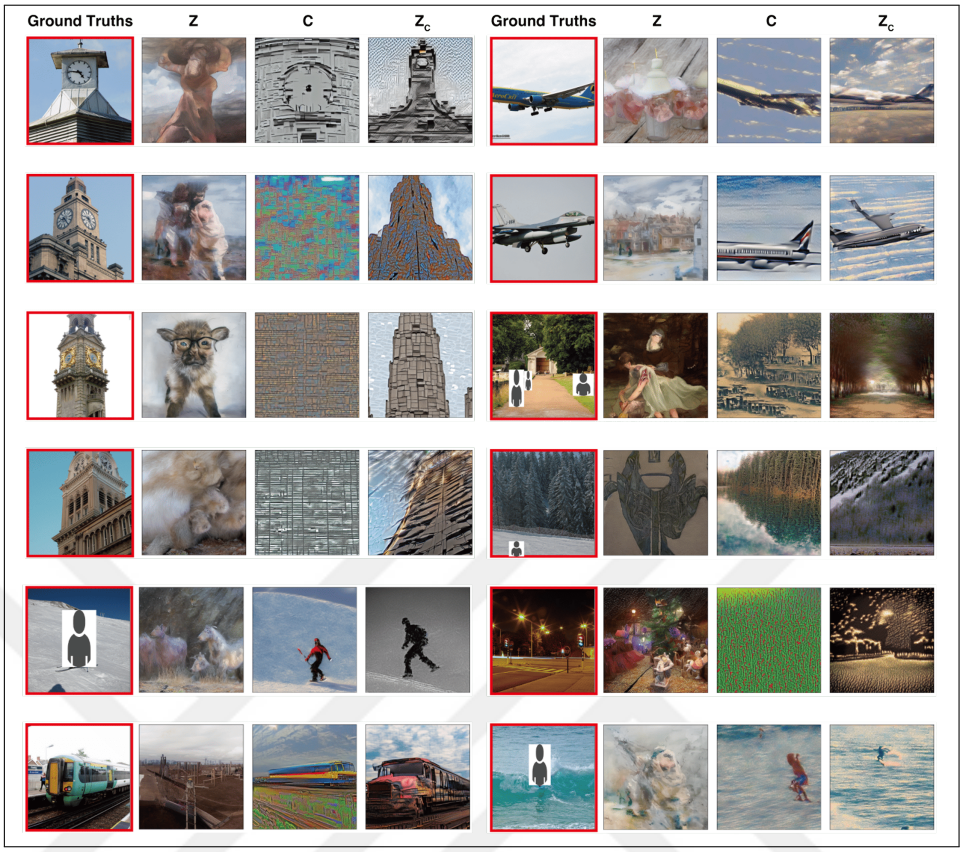

Figure 3.20. Referenced study results from Takagi et al. The ground truth images are the original stimuli. 'Z' columns are generated from single trial beta weights of fMRI activity alone, 'C' columns are generated using only text inputs, and 'Z_c' from both.

In the study by Usma et al. in Figure 3.22, a diffusion model was used, with fMRI volumes integrated as conditions for the reconstruction task. A 3D fMRI encoder was proposed, which encoded brain activity in a latent space using the temporal dimension as an additional channel. The reconstructed images generated by this model exhibited greater structural consistency and less randomness than those of other models. However, despite these improvements, fine-grained details were still missing, and the images exhibited artificial textures. For example, while the model could capture the general colors in the background of images, such as the tower in the 'G' image or the tennis court in the 'E' image, the identification of the objects remained uncertain. The tower, for instance, was reconstructed in a shape that was partially recognizable, but the architectural details were lost, resulting in a vague and imprecise representation.

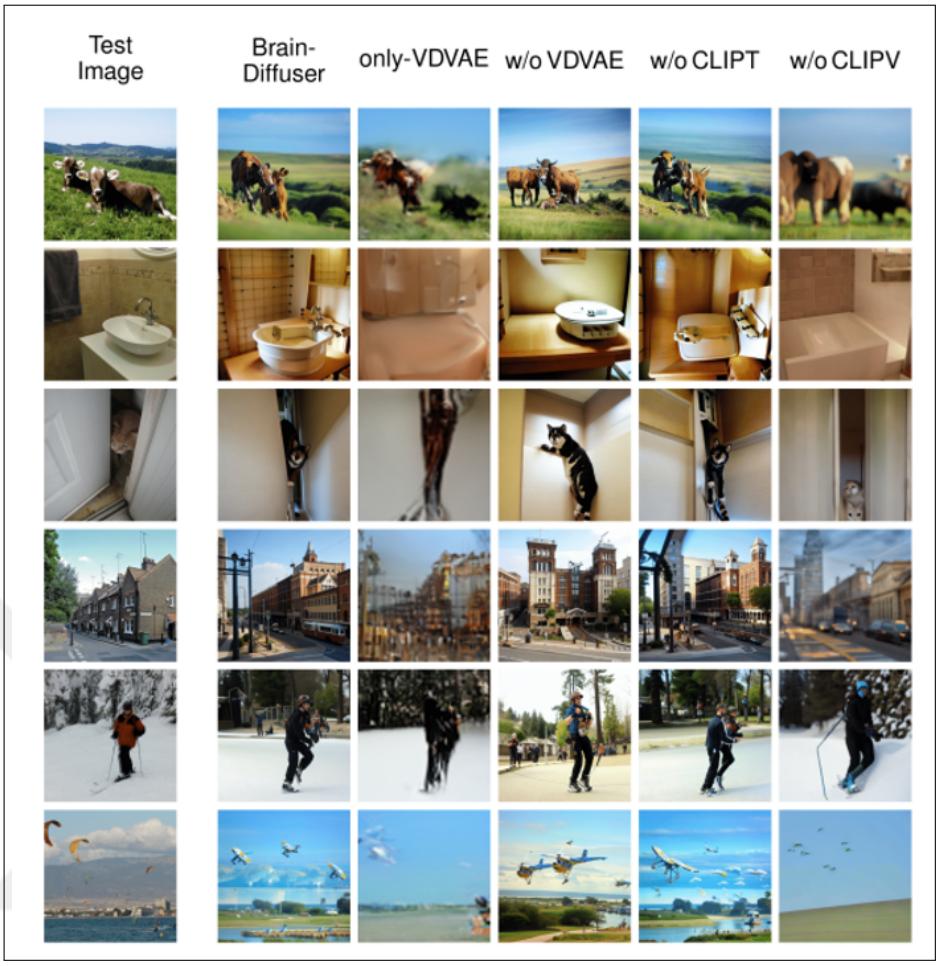

Figure 3.21. Referenced study results from Ozcelik et. al. The first column is the ground truth image (test set). The second column shows reconstructions from the full Brain-Diffuser model with all of its components. The third column is for reconstructions of the Only-VDVAE model. The remaining columns are for Brain-Diffuser with one of its components excluded, in order: without VDVAE, without CLIP-Text, and without CLIP-Vision.

In Figure 3.22 and Figure 3.23, the first-row image in the left column (depicting a tower) demonstrates that although the content is not accurately reconstructed, the position of the object and the color gradients it contains are better represented in the encoder trainable configuration. Similarly, the second-row image in the right column, which initially depicts a blue tennis court, is reconstructed as a body of water across all generated examples. However, the textures in the encoder trainable configuration appear more realistic, and the color gradients of the original image are better captured.

Figure 3.22. Referenced study results from Usma et al. Generated images using fMRI encoder freeze. The original images are labeled in a red box, and the generated images are labeled in a blue box. All the blue box images were generated during the test using the z_{fmri}

A similar result was observed in the SSIM of the quantitative metrics of our model, which measures image properties like brightness and contrast. Additionally, the shapes of objects such as the tower and the airplane in the first and third rows of the right column are better preserved in the encoder trainable configuration. This indicates that the trainable encoder configuration enables the model to use fMRI representations more effectively, resulting in improved alignment with the original image content and structure.

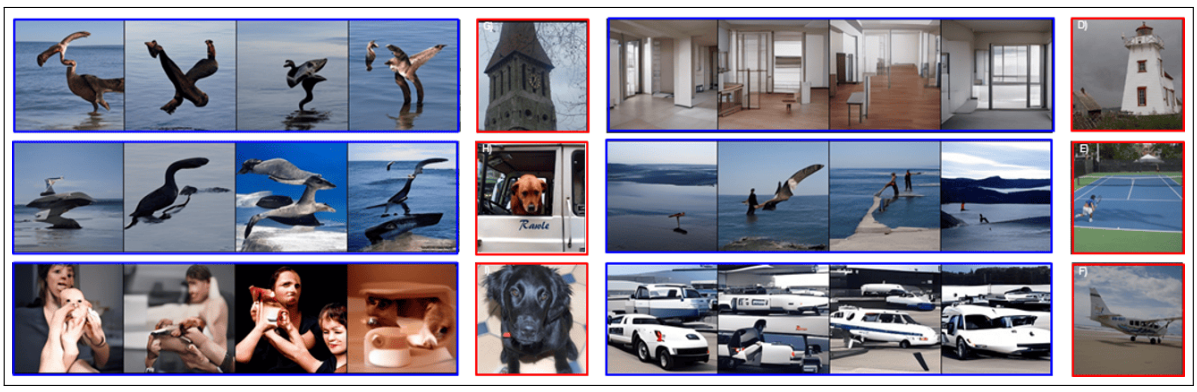

Figure 3.23. Referenced study results from Usma et al. Generated images using an fMRI encoder trainable. The original images are labeled in a red box, and the generated images are labeled in a blue box. All the blue box images were generated during the test using the z_{fmri}

3.6. QUANTITATIVE COMPARISON OF RESULTS WITH EXISTING STUDIES IN LITERATURE

The results of our quantitative evaluations are presented in the tables below. Although these results may not be directly comparable to those from other studies because of differences in experimental design, they provide a basis for understanding how our model performs relative to the existing literature. For example, while some studies have used the NSD dataset, which includes a diverse range of stimuli, such as COCO images, our study used stimuli from the ImageNet data set. While both COCO and ImageNet contain rich visual information, they differ in their focus: COCO provides a wide array of objects in complex scenes, whereas ImageNet is more centered on specific, singular objects designed for classification purposes. These differences in stimuli structure significantly affect the outcomes of our comparisons.

Additionally, many previous studies have used fMRI data from multiple subjects, while our study focuses solely on data from a single subject. The neural activity data corresponding to a stimulus image is limited in this context. Furthermore, single-subject data may introduce subject-specific variability, which, in turn, can reduce the generalizability of our findings. Therefore, any conclusions drawn from this study must be evaluated within the specific context of these methodological differences.

Table 3.4. Decoding accuracy (Pearson's correlation coefficients) of latent representations and data quality metrics. Mean±s.e.m. across all features are shown for z and c. Ours are in red font.

		Takagi et al.				**Ours**
		subj01	subj02	subj05	subj07	CSI1
Decoding accuracy	z	0.239 ± 0.137	0.213 ± 0.132	0.177 ± 0.134	0.145 ± 0.136	0.076 ± 0.028
	c	0.304 ± 0.108	0.295 ± 0.107	0.341 ± 0.109	0.296 ± 0.118	-

For instance, Table 3.5 presents a comparison of 8 different samples of a stimulus image, similar to methods used in other studies. In Table 3.4, we present results for the PCC metric, which has been used in another study using LDMs for fMRI-based image reconstruction. The averages of 5 different latent representations and the corresponding standard error of the mean (SEM) were calculated to ensure robust evaluation. While our model's performance, as shown in Table 3.5, is generally lower than that of other approaches, particularly in the evaluation

of low-level features, this finding aligns with similar observations in different studies. In terms of structural similarity and 2-way comparisons, our model underperforms significantly compared to existing methods. However, in the two-way comparison using AlexNet, our model shows improvement, performing better than the results reported in Takagi's study. The images reconstructed by our model appear clearer, whereas Takagi's reconstructions tend to be more blurred. This discrepancy may be due to the variations in datasets used, further emphasizing the importance of dataset choice in such comparative analyses.

Similar trends were observed in high-level feature analysis, where our model consistently underperformed when compared to other approaches. These results suggest that our model has not yet achieved the desired performance in both low-level and high-level quantitative measures. As a result, significant improvements are required to enhance the model's performance in other methods of comparison.

Table 3.5. Quantitative results of the reconstructed image quality on the test set. For PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and CLIP metrics, higher is better. For EffNet-B and SwAV distances, lower is better. Our results are in the red cells and indicated by the arrow pointing up or down, respectively.

Method	**Quantitative Measures**							
	Low-Level				**High-Level**			
	PixCorr$^{\downarrow}$	**SSIM**$^{\downarrow}$	**AlexNet(2)**	**AlexNet(5)**	**Inception**	**CLIP**	**EffNet-B**$^{\uparrow}$	**SwAV**$^{\uparrow}$
Takagi et al.	-	-	83.0%	83.0%	76.0%	77.0%	-	-
Gu et al.	.150	.325	-	-	-	-	.862	.465
Ozcelik et al.	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
Ours	.142	.228	93.6%	93.6%	49.5%	-	.980	.505

4. DISCUSSION

This study proposes that LDMs reconstruct visual experiences from fMRI activity. However, the results highlight several critical challenges and limitations, demonstrating that the proposed approach struggled to achieve the desired accuracy and generalizability.

The complexity of brain processes and limitations in fMRI data processing plays a significant role in the model's performance. fMRI technology relies on measuring the BOLD signal, which indirectly represents neuronal activity. However, the BOLD signal has intrinsic limitations, such as low temporal and spatial resolution and non-linear relationships with neuronal activity. These factors lead to discrepancies between the neuronal activity elicited by the stimuli and the captured BOLD signal, especially for complex or dynamic visual stimuli. It hinders the model's ability to reconstruct stimuli from latent representations accurately. Additionally, the activities of neuron groups in different brain regions can vary significantly between individuals, making it challenging to create a universal response with a standard model. Isolating neuronal activity associated with a specific stimulus is difficult due to complex interactions with other mental processes.

The cortical surface modeling procedure using fMRI volumes is a complex and sequential process that includes motion correction, intensity normalization, skull stripping, WM segmentation, and cortical parcellation. However, each step is prone to errors that can propagate through the pipeline, distorting anatomical accuracy, reducing data reliability, and leading to misinterpretation of anatomical structures.

Occipital surface data from a single participant were used in our study, and the stimuli derived from the ImageNet data set containing individual objects. This approach inherently restricted the amount of fMRI data available. Analyses like average BOLD signal across runs and Fourier transform analysis were applied. The average BOLD signal revealed the cumulative effects of the hemodynamic responses of specific stimuli during certain trials. However, Fourier transform analysis indicated that, in some runs, the periodicity of responses lasted longer than expected. This made it more challenging to isolate the fMRI data directly associated with a specific stimulus.

Moreover, the BOLD5000 dataset presented participants with a random mix of three popular

computer vision datasets. However, this study used volumes related to only the Imagenet stimulus dataset. The cumulative effects observed across runs might have limited the model's ability to generalize activations for stimuli from unseen datasets and relate them with specific activations.

However, reconstructing visual experiences from the fMRI activity task used a combination of pixel reconstruction loss, perceptual loss, and adversarial loss. However, after addressing the aforementioned data-related challenges, the generalization capability of the training process could be improved by considering factors related to the loss configuration and the training dynamics.

Firstly, one of the main contributors to overfitting might be the weight assigned to the adversarial loss in the total loss. An overemphasis on adversarial loss might cause the generator to prefer misleading the discriminator over learning generalizable features, resulting in memorization of the training data. Reducing the weight of the adversarial loss may help the model focus more on reconstruction fidelity while being realistic. In addition, introducing adversarial training at a very early stage may have destabilized the generator's learning process. Delaying the start of adversarial training could allow the generator to develop meaningful reconstruction capabilities before competing with the discriminator, thus improving stability and reducing overfitting. The reconstruction loss, which heavily relies on codebook regularization, may also contribute to overfitting. An overemphasis on compact latent representations may restrict the model's generalization ability. By reducing the weight of the codebook loss, the model's adaptability might be improved, and the contributions of the different loss components might be balanced. Finally, the balance between perceptual loss and pixel reconstruction loss is another critical factor. While perceptual loss encourages the model to capture fine-grained details, overreliance on it can lead to overfitting on specific textures and structures. Shifting this balance toward pixel loss could improve the model's ability to generalize across diverse inputs while preserving essential structural information.

5. CONCLUSIONS

This study focuses on using a conditional latent diffusion model to reconstruct visual experiences from fMRI activity. The findings highlight significant challenges arising from the limitations of fMRI data and the complexities of the reconstruction task. The low temporal and spatial resolution of fMRI, coupled with the non-linear relationship between BOLD signals and neural activity, made it difficult to reconstruct stimuli accurately.

In this study, only a single participant's occipital surface data were used, further narrowing the scope of the data set. Using stimuli from only the ImageNet dataset, consisting primarily of isolated objects, imposed additional challenges. The limited relation of fMRI activity with specific stimuli reduced data availability, and cumulative effects observed in average BOLD signals across trials diluted stimulus-specific information. Fourier transform analysis has also shown longer periodicity in some trials, complicating the extraction of fMRI signals directly related to individual stimuli.

Furthermore, the BOLD5000 dataset, which combines stimuli from three popular computer vision datasets, introduces its challenges. However, only a single stimulus data set was used for analysis in this study. The cumulative effects between runs may have limited the model's ability to generalize activations related to specific stimuli and associate them with specific neural responses.

REFERENCES

1. Freeman J, Simoncelli EP. Metamers of the ventral stream. *Nature neuroscience*. 2011;14(9):1195-201.

2. Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*. 2003 June;19(2 Pt 1):261-70.

3. Huff T, Mahabadi N, Tadi P. 5. *Neuroanatomy, Visual Cortex*. StatPearls Publishing. 2021:101-10. http://www.ncbi.nlm.nih.gov/books/NBK482504/.

4. Goertzel B. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*. 2014;5(1):1.

5. Surianarayanan C, Lawrence JJ, Chelliah PR, Prakash E, Hewage C. Convergence of Artificial Intelligence and Neuroscience towards the Diagnosis of Neurological Disorders—A Scoping Review. *Sensors*. 2023;23(6):3062.

6. Chen M, Han J, Hu X, Jiang X, Guo L, Liu T. Survey of encoding and decoding of visual stimulus via FMRI: an image analysis perspective. *Brain imaging and behavior*. 2014;8:7-23.

7. Cox D, Dean T. Neural Networks and Neuroscience-Inspired Computer Vision. *Current Biology*. 2014;24(18):R921-9. Available from: https://www.sciencedirect.com/science/article/pii/S0960982214010392.

8. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. *Nature neuroscience*. 2019;22(11):1761-70.

9. Marblestone AH, Wayne G, Kording KP. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*. 2016;10:94.

10. Behroozi M, Daliri MR. Predicting brain states associated with object categories from fMRI data. *Journal of integrative neuroscience*. 2014;13(04):645-67.

11. Cabral C, Silveira M, Figueiredo P. Decoding visual brain states from fMRI using an

ensemble of classifiers. *Pattern Recognition*. 2012;45(6):2064-74.

12. Song S, Zhan Z, Long Z, Zhang J, Yao L. Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. *PloS one*. 2011;6(2):e17191.

13. Tithi ID, Shuchi USK, Tasneem NA, Mobin MI, Alam MA. Brain fMRI image classification and statistical representation of visual objects. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE. 2019:1-6.

14. Yousefnezhad M, Zhang D. Anatomical pattern analysis for decoding visual stimuli in human brains. *Cognitive Computation*. 2018;10:284-95.

15. Qiao K, Chen J, Wang L, Zhang C, Zeng L, Tong L, et al. Category decoding of visual stimuli from human brain activity using a bidirectional recurrent neural network to simulate bidirectional information flows in human visual cortices. *Frontiers in Neuroscience*. 2019;13:692.

16. Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*. 2017;8(1):15037.

17. Battleday RM, Peterson JC, Griffiths TL. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*. 2020;11(1):5418.

18. Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, Lebihan D, et al. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*. 2006;33(4):1104-16.

19. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. *Nature*. 2008;452(7185):352-5.

20. Miyawaki Y, Uchida H, Yamashita O, Sato MA, Morito Y, Tanabe HC, et al. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*. 2008;60(5):915-29.

21. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*. 2011;21(19):1641-6.

22. Akamatsu ea Yusuke. Perceived image decoding from brain activity using shared information of multi-subject fMRI data. *IEEE Access*. 2021;9:26593-606.

23. Nishida S, Nishimoto S. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*. 2018;180:232-42.

24. Seeliger K, Güçlü U, Ambrogioni L, Güçlütürk Y, van Gerven MA. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*. 2018;181:775-85.

25. Qiao K, Chen J, Wang L, Zhang C, Tong L, Yan B. BigGAN-based Bayesian reconstruction of natural images from human brain activity. *Neuroscience*. 2020;444:92-105.

26. Han K, Wen H, Shi J, Lu KH, Zhang Y, Fu D, et al. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*. 2019;198:125-36.

27. Qiao K, Zhang C, Wang L, Yan B, Chen J, Zeng L, et al. Accurate reconstruction of image stimuli from human fMRI based on the decoding model with capsule network architecture. *arXiv preprint arXiv:180100602*. 2018.

28. Gaziv G, Beliy R, Granot N, Hoogi A, Strappini F, Golan T, et al. Self-supervised natural image reconstruction and rich semantic classification from brain activity. *bioRxiv*. 2020;6(9).

29. VanRullen R, Reddy L. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology*. 2019;2(1):193.

30. Meng L, Yang C. Dual-Guided Brain Diffusion Model: Natural Image Reconstruction from Human Visual Stimulus fMRI. *Bioengineering*. 2023;10(10):1117.

31. Shen G, Dwivedi K, Majima K, Horikawa T, Kamitani Y. End-to-end deep image

reconstruction from human brain activity. *Frontiers in computational neuroscience*. 2019;13:21.

32. Zhang C, Qiao K, Wang L, Tong L, Zeng Y, Yan B. Constraint-free natural image reconstruction from fMRI signals based on convolutional neural network. *Frontiers in human neuroscience*. 2018;12:242.

33. Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*. 2022.

34. Halac M, Isik M, Ayaz H, Das A. Multiscale Voxel Based Decoding For Enhanced Natural Image Reconstruction From Brain Activity. *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022:1-7.

35. Fang T, Qi Y, Pan G. Reconstructing perceptive images from brain activity by shape-semantic GAN. *Advances in Neural Information Processing Systems*. vol. 33. 2020:13038-48.

36. Lin S, Sprague T, Singh AK. Mind reader: Reconstructing complex images from brain activities. *Advances in Neural Information Processing Systems*. vol. 35. 2022:29624-36.

37. Kneeland R, Ojeda J, St-Yves G, Naselaris T. Reconstructing seen images from human brain activity via guided stochastic search. *arXiv preprint arXiv:230500556*. 2023.

38. Shen G, Horikawa T, Majima K, Kamitani Y. Deep image reconstruction from human brain activity. *PLoS computational biology*. 2019;15(1):e1006633.

39. Dado T, Güçlütürk Y, Ambrogioni L, Ras G, Bosch S, van Gerven M, et al. Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Scientific reports*. 2022;12(1):141.

40. Huang W, Yan H, Wang C, Yang X, Li J, Zuo Z, et al. Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks. *Neuroscience bulletin*. 2021;37:369-79.

41. Takagi Y, Nishimoto S. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs. *arXiv preprint*

arXiv:230611536. 2023.

42. Yu Z, Qiao K, Zhang C, Wang L, Yan B. End-to-End Image Reconstruction of Image from Human Functional Magnetic Resonance Imaging Based on the" Language" of Visual Cortex. *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*. 2020:176-81.

43. Du C, Li J, Huang L, He H. Brain Encoding and Decoding in fMRI with Bidirectional Deep Generative Models. *Engineering*. 2019;5(5):948-53. Available from: https://www.sciencedirect.com/science/article/pii/S2095809917305647.

44. Aine CJ, Supek S, George JS, Ranken D, Lewine J, Sanders J, et al. Retinotopic Organization of Human Visual Cortex: Departures from the Classical Model. *Cerebral Cortex*. 1996 05;6(3):354-61. Available from: https://doi.org/10.1093/cercor/6.3.354.

45. BrainVoyager. Retinotopic Mapping of Early Visual Areas. Accessed: 2024-04-18. https://www.brainvoyager.com/bv/doc/UsersGuide/AdditionalDocu/RetinotopicMapping/RetinotopicMappingOfEarlyVisualAreas.html.

46. Tyler CW, Likova LT, Chen CC, Kontsevich LL, Schira MM, Wade AR. Extended concepts of occipital retinotopy. *Current Medical Imaging*. 2005;1(3):319-29.

47. Vividly. Physiology of Vision: Visual System - Visual Cortex. Vividly. n.d. https://www.seevividly.com/info/Physiology_of_Vision/The_Brain/Visual_System/Visual_Cortex.

48. Tootell RB, Hadjikhani NK, Vanduffel W, Liu AK, Mendola JD, Sereno MI, et al. Functional analysis of primary visual cortex (V1) in humans. *Proceedings of the National Academy of Sciences*. 1998;95(3):811-7.

49. Wandell BA, Dumoulin SO, Brewer AA. Visual field maps in human cortex. *Neuron*. 2007;56(2):366-83.

50. Busse L. The Mouse Visual System and Visual Perception. *Handbook of Behavioral Neuroscience*. vol. 27. Elsevier. 2018:53-68.

51. Queensland Brain Institute. Visual Perception. Accessed: 2024-04-18. https://qbi.uq.edu.au/brain/brain-functions/visual-perception.

52. Judaš M, Cepanec M, Sedmak G. Brodmann's map of the human cerebral cortex — or Brodmann's maps? *Translational Neuroscience*. 2012;3(1):67-74. Available from: https://doi.org/10.2478/s13380-012-0009-x [cited 2024-04-19].

53. Simply Psychology. Brodmann Areas. Simply Psychology. 2024. Accessed: 2024-04-18. https://www.simplypsychology.org/brodmann-areas.html.

54. Kenhub. Brodmann Areas. Kenhub. 2024. Accessed: 2024-04-18. https://www.kenhub.com/en/library/anatomy/brodmann-areas.

55. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*. 1999;9:195-207.

56. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*. 2020;33:6840-51.

57. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. 2021.

58. Chang N, Pyles JA, Marcus A, et al. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*. 2019;6:49.

59. Talairach PJ. Co-planar stereotaxic atlas of the human brain. *(No Title)*. 1988.

60. AM D. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of Cognitive Neuroscience*. 1993;5:162-76.

61. Chang N, et al. BOLD5000. *OpenNeuro*. 2019. Available from: https://doi.org/10.18112/openneuro.ds001499.v1.3.0.

62. Brain Imaging Data Structure Consortium. Brain Imaging Data Structure (BIDS). Brain Imaging Data Structure Consortium. 2024. Accessed: 2024-04-30. https://bids.neuroimaging.io/.

63. Lee J, Contributors. Dcm2Bids: A tool to convert DICOM data to BIDS format. GitHub. 2024. Accessed: 2024-04-30. Available from: https://github.com/jooh/Dcm2Bids.

64. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one*. 2017;12(9):e0184661.

65. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: A Robust Preprocessing Pipeline for fMRI Data. fMRIPrep. 2024. Accessed: 2024-04-30. Available at: https://github.com/nipreps/fmriprep.

66. Gorgolewski KJ, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, et al. Nipype: Neuroimaging in Python Pipelines and Interfaces. Nipype. 2024. Accessed: 2024-04-30. Available at: https://github.com/nipy/nipype.

67. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*. 2010;29(6):1310-20.

68. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*. 1999;9:179-94.

69. Klein A, Ghosh SS, Bao FS, Giard J, Häme Y, Stavsky E, et al. Mindboggling morphometry of human brains. *PLoS computational biology*. 2017;13(2):e1005350.

70. Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*. 2009;47:S102.

71. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*. 2008;12(1):26-41.

72. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*. 2001;20(1):45-57.

73. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*. 1996;29(3):162-73.

74. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based

registration. *Neuroimage*. 2009;48(1):63-72.

75. Gattass R, Gross CG, Sandell JH. Visual topography of V2 in the macaque. *Journal of Comparative Neurology*. 1981;201(4):519-39.

76. Brewer AA, Barton B. Visual Field Map Organization in Human Visual Cortex. *Visual Cortex*. Rijeka: IntechOpen. 2012:Available from: https://doi.org/10.5772/51914.

77. Van Essen DC, Maunsell J. Two-dimensional maps of the cerebral cortex. *Journal of Comparative Neurology*. 1980;191(2):255-81.

78. Braddick O. Occipital Lobe (Visual Cortex): Functional Aspects. *International Encyclopedia of the Social Behavioral Sciences*. Oxford: Pergamon. 2001:10826-8. Available from: https://www.sciencedirect.com/science/article/pii/B0080430767034707.

79. Goebel R, Khorram-Sefat D, Muckli L, Hacker H, Singer W. The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience*. 1998;10(5):1563-73.

80. Wandell BA, Winawer J. Imaging retinotopic maps in the human brain. *Vision Research*. 2011;51(7):718-37. Vision Research 50th Anniversary Issue: Part 1. Available from: https://www.sciencedirect.com/science/article/pii/S0042698910003780.

81. DeYoe EA, Bandettini P, Neitz J, Miller D, Winans P. Functional magnetic resonance imaging (FMRI) of the human brain. *Journal of Neuroscience Methods*. 1994;54(2):171-87. Imaging Techniques in Neurobiology. Available from: https://www.sciencedirect.com/science/article/pii/0165027094901910.

82. Warnking J, Dojat M, Guérin-Dugué A, Delon-Martin C, Olympieff S, Richard N, et al. fMRI Retinotopic Mapping—Step by Step. *NeuroImage*. 2002;17(4):1665-83. Available from: https://www.sciencedirect.com/science/article/pii/S1053811902913042.

83. Leeds DD, Seibert DA, Pyles JA, Tarr MJ. Comparing visual representations across human fMRI and computational vision. *Journal of vision*. 2013;13(13):25-5.

84. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI.

Neuroimage. 2011;56(2):400-10.

85. Huang S, Shao W, Wang ML, Zhang DQ. fMRI-based decoding of visual information from human brain activity: A brief review. *International journal of automation and computing*. 2021;18(2):170-84.

86. Yarkoni T, Poldrack RA, Van Essen DC, Wager TD. Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends in cognitive sciences*. 2010;14(11):489-96.

87. Fischl B, Salat DH, Van Der Kouwe AJ, Makris N, Ségonne F, Quinn BT, et al. Sequence-independent segmentation of magnetic resonance images. *Neuroimage*. 2004;23:S69-84.

88. Prince JS, Charest I, Kurzawski JW, Pyles JA, Tarr MJ, Kay KN. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*. 2022 nov;11:e77599. Available from: https://doi.org/10.7554/eLife.77599.

89. Book AB. FreeSurfer Short Course: Recon-All. Online; accessed April 2024. https://andysbrainbook.readthedocs.io/en/latest/FreeSurfer/FS_ShortCourse/FS_03_ReconAll.html.

90. Jahn A. Introduction to Freesurfer. 2017. Accessed on Month Day, Year. Available from: https://www.youtube.com/playlist?list=PLIQIswOrUH6_DWy5mJlSfj6AWY0y9iUce.

91. Wu J, Ngo GH, Greve D, Li J, He T, Fischl B, et al. Accurate nonlinear mapping between MNI volumetric and FreeSurfer surface coordinate systems. *Human Brain Mapping*. 2018;39(9):3793-808. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24213.

92. LearnOpenCV. Denoising with Diffusion Probabilistic Models. [Online; accessed April 2024]. https://learnopencv.com/denoising-diffusion-probabilistic-models/.

APPENDIX A: FMRI PREPROCESSING: OCCIPITAL SURFACE EXTRACTION FROM VOLUME DATA

This appendix provides an overview of the preprocessing steps to extract the occipital surface from fMRI volume data. The methods and tools used in this process are detailed in the associated GitHub repository.

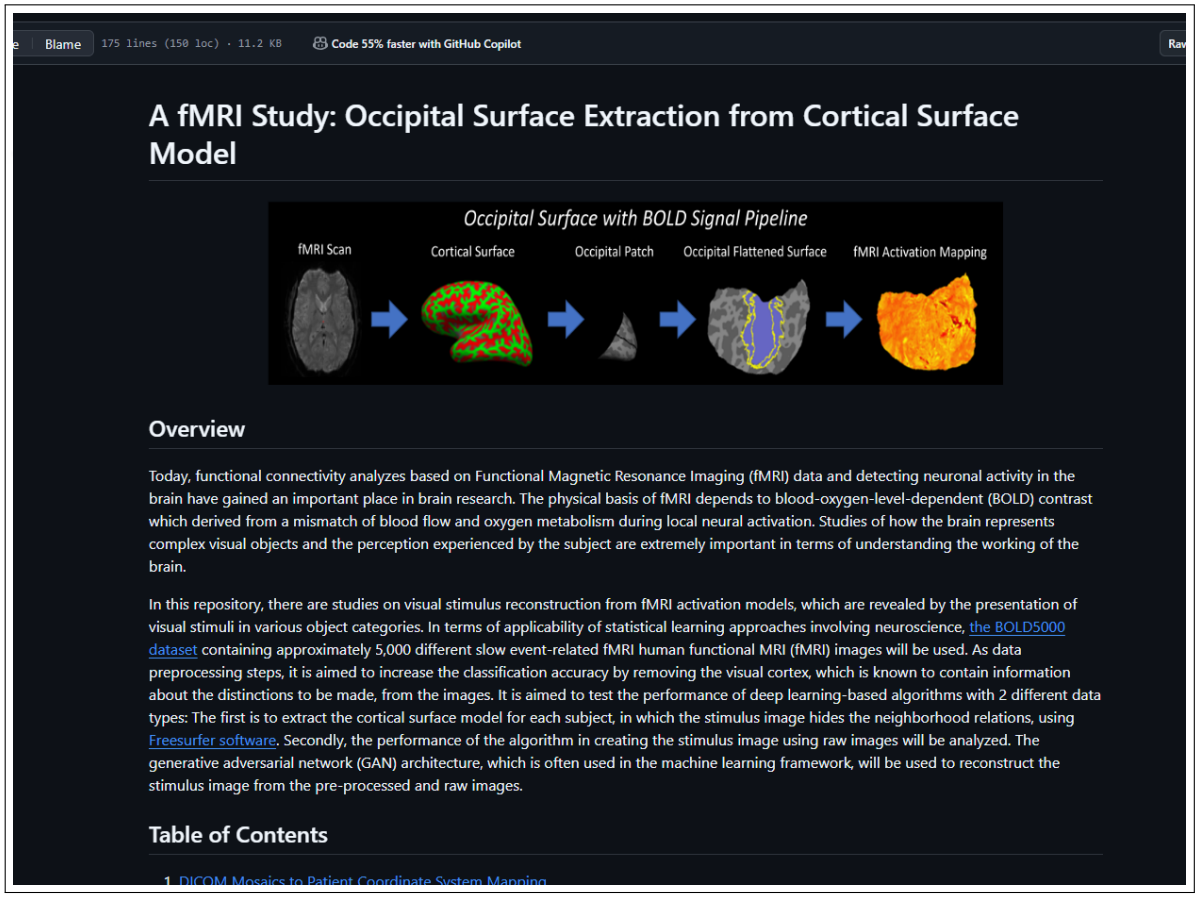

Figure A.1. Github project page of the occipital surface extraction process.

Code Availability: The complete codebase for the preprocessing pipeline is hosted on GitHub. You can find the repository at the following link:

https://github.com/genchandenur/Occipital_Surface_Extraction

APPENDIX B: LATENT DIFFUSION MODELS FOR HUMAN BRAIN ACTIVITY DECODING

Code Availability: The complete codebase for proposed models is hosted on GitHub. You can find the repository at the following link:

https://github.com/genchandenur/latent-diffusion