Integrating Deep Learning and Bayesian Inference in Neuroimaging Analysis

by

Yuan Zhong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2025

Doctoral Committee:

Professor Jian Kang, Chair
Assistant Professor Yang Chen
Professor Timothy Johnson
Associate Professor Zhenke Wu

Yuan Zhong
ylzhong@umich.edu
ORCID iD: 0009-0001-9927-3206

© Yuan Zhong 2025

ACKNOWLEDGEMENTS

I feel deeply fortunate to have completed both my master's and Ph.D. studies at the University of Michigan. The past several years in the Department of Biostatistics have been an incredibly rewarding journey, which is filled with warmth, inspiration, and growth, thanks to the many friendly, diligent, and intelligent students, professors, and staff who have surrounded me. This dissertation would not have been possible without the guidance, support, and encouragement of many people.

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jian Kang, for his continuing support, mentorship, and patience throughout my doctoral journey. I have been continuously inspired by Dr. Kang's insights, deep understanding of data, and the creative approach to tackling challenging research questions. His guidance not only shaped my dissertation but also taught me how to think critically and pursue scientific questions with curiosity and persistence.

I would also like to sincerely thank my committee members, Drs. Timothy Johnson, Zhenke Wu, and Yang Chen, for their invaluable feedback, insightful discussions, and continuous encouragement throughout the development of this dissertation. My heartfelt appreciation also goes to my collaborators, Drs. Lili Zhao, Weining Shen, Gang Chen, and Paul Taylor, for their generous sharing of expertise, thoughtful discussions, and meaningful contributions across our joint research projects. I am grateful to all past and current members of Dr. Kang's lab, whose collaboration, friendship, and countless discussions have made this journey both intellectually inspiring and enjoyable.

Finally, I owe my deepest gratitude to my family, who have given me freedom to explore my passions, supported every decision unconditionally and always believed in me. I would also like to thank my boyfriend, Qingyang Wang, for all the emotional support despite the distance between us. And to all my friends in Michigan, thank you for the laughter, memories, and encouragement that made these years so meaningful. I am grateful to have shared this journey with you and I look forward to all the paths that lie ahead for us.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES v

LIST OF TABLES vii

LIST OF APPENDICES viii

ABSTRACT ix

CHAPTER

1 Introduction 1

1.1 Neuroimaging 1
1.2 Image regression models 3
1.3 Bayesian methods and Gaussian process prior 4
1.4 Deep learning methods 4
1.5 Contribution 6

2 SIMBA: Scalable Image Modeling Using a Bayesian Approach, A Consistent Framework for Including Spatial Dependencies in fMRI Studies . 7

2.1 Introduction 7
2.2 Methods 10
2.2.1 Model Specification 10
2.2.2 Kernel approximation 11
2.3 Posterior Computation 13
2.3.1 Model reparameterization 13
2.3.2 Gibbs sampler 14
2.3.3 Variational Inference 15
2.4 Simulation and validation 16
2.4.1 Ex. 1: Simulated data 16
2.5 Real data applications 18
2.5.1 Ex.2: Neuroimaging Analysis Replication and Prediction Study . . 18
2.5.2 Ex. 3: Adolescent Brain Cognitive Development Study 22
2.6 Discussion and Conclusion 26

3 Deep Kernel Learning Based Gaussian Processes for Bayesian Image Regression Analysis 29

3.1 Introduction . . . 29
3.2 Methods . . . 32
3.2.1 Kernel learning via deep neural network . . . 32
3.2.2 Image regression models . . . 33
3.2.3 Prior specification and model representation . . . 34
3.3 Theory . . . 35
3.4 Posterior Computation . . . 39
3.5 Simulations . . . 40
3.5.1 Nonparametric regression . . . 41
3.5.2 Image-on-scalar regression . . . 42
3.5.3 Scalar-on-image regression . . . 43
3.6 Brain Imaging Applications . . . 45
3.6.1 Cortical surface task fMRI . . . 45
3.6.2 EEG brain-computer interfaces . . . 48
3.7 Conclusion . . . 53

4 Activation-guided Diffusion Models for fMRI Synthesis: Integrating Image Regression and Generation for Statistical Inference . . . 54

4.1 Introduction . . . 54
4.2 Methods . . . 56
4.3 Experiments . . . 60
4.3.1 Synthetic MNIST Data with Gaussian noise . . . 61
4.3.2 Cortical surface fMRI generation . . . 65
4.3.3 Hypothesis testing via synthetic null distributions . . . 66
4.4 Discussion . . . 70

5 Conclusion . . . 71

APPENDICES . . . 73

BIBLIOGRAPHY . . . 99

LIST OF FIGURES

FIGURE

2.1 Comparison of estimated spatial effects and activation region selection across methods under two noise levels. Each row displays a map of the true spatial effect of either $\beta_0(\cdot)$ or $\beta_1(\cdot)$, as well as the corresponding estimates from each method. The non-gray shapes show the distinct "significant" activation regions, defined to be where $|E_s| > 0.95$. 19

2.2 Comparison among voxel-level effect estimation in GLM, BML, and our proposed SIMBA on various axial slices of fMRI results. The overlay colors show the effect sizes, and statistical information is used to threshold the data in the "highlight" style [Allen et al., 2012, Taylor et al., 2023]: suprathreshold regions are opaque and outlined, with other regions fading with decreasing statistical value. 23

2.3 Posterior predictive checking curves for Bayesian models for real fMRI data (Ex. 2, NARPS). The yellow lines represent 150 posterior draws of predictive data distribution, and the black line shows the observed data density for all voxels. Having a larger overlap and centering of posterior draws over the observed data fit (black line) suggests better model fit. Here, the SIMBA models have much better posterior check results than the other two models. 24

2.4 Cross-site prediction mean squared errors (MSEs) for single-site analyses on the ABCD cortical surface fMRI data, shown separately for the left hemisphere (left) and right hemisphere (right). 25

2.5 Estimated associations and selected activation regions between fMRI data and cognitive scores in the ABCD study for the left and right hemispheres. The values represent the expected change in the 2-back versus 0-back contrast (expressed as z-statistics) for a one-unit increase in cognitive score, while holding all other demographic information. 26

3.1 Comparison of true spatially varying functions with estimated effects from DKLP and other competing methods in simulation studies across two image regression frameworks. 46

3.2 Cross-site prediction MSEs for cortical surface fMRI data collected at baseline on the left hemisphere in the ABCD study. 48

3.3 Selected brain regions using the ABCD cortical surface fMRI data in left and right hemispheres at baseline and follow-up time points. Blue regions are voxels selected using all-site data and yellow regions are reproducible voxels selected from 21 single-site analyses. 49

3.4 Posterior selection probabilities of EEG brain signals across sixteen brain channels at twelve equally spaced time points. Node sizes represent the magnitude of the selection probabilities on each channel. 52

4.1 Overview of the proposed BACNet framework. Stage I applies BNPR with a DKLP prior to estimate subject-specific activation maps from observed noisy fMRI contrast maps. Stage II implements a conditional diffusion model with a U-Net backbone, learning the predicted noises at each denoising step. Stage III introduces ControlNet to incorporate the estimated activation maps as spatial priors, which are injected into the U-Net through zero-initialized convolutional layers. 57
4.2 MNIST synthetic images added by Gaussian noise with $\sigma = 0.5$ and spatial digit noises. Each figure shows the observed noisy images (top), original clean images with sparse continuous patterns (middle), and estimated activation region from Bayesian nonparametric regression (bottom). 63
4.3 Comparison of reconstructed images from different generative models on the MNIST dataset with independent Gaussian noise and spatially digit noise. The top row shows the estimated signal maps, ground-truth images, followed by outputs from the BACNet and other comparative baseline generative models. . . . 64
4.4 Comparison of generated fMRI contrast maps with the true target contrast maps across different generative models. 67
4.5 Identified activation regions associated with cognitive score under different analysis frameworks, including (a) reference results from the all-site full ABCD data; (b) single-site analysis; (c) single-site analysis using a bootstrap-based null distribution (d) single-site analysis using the proposed generative null approach, where null distributions are generated from the BACNet model. 69

B.1 Comparison of true spatially varying functions with estimated effects from DKLP and other competitive methods in image-on-scalar regression simulations with $V = 64^2$. 88
B.2 Comparison of true spatially varying functions with estimated effects from DKLP and other competitive methods in scalar-on-image regression simulations with $V = 32^2$. 89
B.3 Cross-site prediction MSEs on cortical surface fMRI data in the ABCD study. . 91
B.4 Posterior selection probabilities on additional participants at twelve equally spaced time points on sixteen EEG brain channels. Node sizes represent the magnitude of the selection probabilities on each channel. 92
B.5 Sensitivity analysis of different DNN architectures in terms of the number of layers, the number of nodes per layer, and the activation function applied to ABCD cortical surface data for the left and right hemispheres at baseline and follow-up time points. The color scheme represents values within the range [-1, 1]. 94

C.1 BACNet generation results under different initial noise values. 98

LIST OF TABLES

TABLE

2.1 Evaluation results for simulation studies comparing mean squared error, true positive rate, false discovery rate, and coverage. Values are reported as mean percent, with standard deviations in parentheses, over 100 simulated datasets. The best-performing value(s) for each simulation is highlighted in bold. 20

3.1 Evaluation results on nonparametric regression, comparing the estimation accuracy of image functions and kernel functions. Mean values with standard deviations (in parentheses) are reported for each metric and scaled by factors of 0.01. The two best-performing metrics are highlighted in bold. 42
3.2 Evaluation results for image-on-scalar regression simulations. Mean values with standard deviations (in parentheses) are reported for each metric in unit 0.01. The best performing metric is highlighted in bold. 44
3.3 Evaluation results from scalar-on-image regression reported in mean and standard deviation (in parentheses) in scaled units. The two best-performing metrics are highlighted in bold. 47
3.4 Individual-level BCI character prediction accuracy percentage averaged across testing sessions for ALS and healthy control participants. 51

4.1 Metric evaluation of generative performance on the MNIST simulation study, comparing the proposed model framework with baseline models. MSEs are evaluated on estimated sparse activation regions. Lower FID, MSE, and MSE_{A} and higher SSIM and PSNR indicate better performance. 65
4.2 Metric evaluation of generative performance on real fMRI data in the ABCD study, comparing the proposed model framework with baseline models. Lower FID, MSE, and MSE_A and higher SSIM, PSNR, and PCC indicate better performance. 68
4.3 Metric comparison of brain activation detection across inference methods, using all-site GLM results as the reference truth. 69

B.1 DNN configurations in each study section, including the number of hidden layers, the number of hidden units per layer, the activation functions, the kernel truncation numbers, and the learning rates. 93
B.2 Sensitivity analysis on BCI study for individual-level classifier for character-level prediction accuracy using different DNN configurations in terms of the number of layers, the number of nodes per layer, and the activation function. 95

LIST OF APPENDICES

A Supplementary Materials for Chapter II 73

B Supplementary Materials for Chapter III 77

Proof for Theoretical Results 77
Additional Results 86
DNN Specifications 93
Sensitivity Analysis 93

C Supplementary Materials for Chapter IV 96

Architecture details 96
Additional results 96

ABSTRACT

Neuroimaging has substantially advanced our understanding of brain structure, function, and disease. Techniques such as magnetic resonance imaging (MRI) and electroencephalography (EEG) have become essential tools in neuroscience and clinical research. However, statistical analysis of neuroimaging data presents several key challenges, including low signal-to-noise ratios, complex spatial and temporal dependence, and high dimensionality. Although Bayesian inference and deep learning methods have each demonstrated promise in addressing these challenges, there is a lack of unified frameworks that effectively combine their strengths. This dissertation develops novel methods that integrate Bayesian hierarchical modeling and deep learning methods to address challenges in regression analysis, spatial inference, and image generation in large-scale neuroimaging studies.

Chapter II develops a scalable image modeling framework under a Bayesian approach (SIMBA) using Gaussian process (GP) priors for multi-subject fMRI analysis. Although GP priors can capture smooth, global spatial correlations across voxels, their applications in whole-brain inference have been limited by the computational cost of Markov chain Monte Carlo (MCMC) sampling in high-dimensional settings. We address this limitation through low-rank kernel approximations and model reparameterizations, with an efficient Gibbs sampling algorithm that achieves whole-brain-level posterior inference within minutes for datasets exceeding 222,000 voxels. To further enhance computational scalability, we develop a variational inference (VI) algorithm that converges within seconds, achieving an 80% reduction in runtime, or approximately five-fold speed-up, compared to Gibbs sampling. The proposed method provides a practical and scalable solution for spatial Bayesian analysis of fMRI data in neuroimaging research.

Chapter III introduces a deep learning–based Bayesian nonparametric regression framework that enhances flexibility in modeling spatial and temporal correlations in neuroimaging data using GP priors. Traditional GP models rely on fixed, pre-specified kernel functions, which often limit their ability to adapt to complex spatial structures and generalize across imaging outcomes. To address this limitation, we propose the deep kernel learning process (DKLP), which adaptively learns the kernel functions from data using deep neural networks (DNNs). We establish theoretical results on posterior convergence rates for both kernel

and spatially varying functions. Through simulation studies, we show that DKLP achieves superior accuracy in parameter estimation and sparse region selection while maintaining a low false discovery rate in various regression settings. In task-based cortical surface fMRI analysis from the Adolescent Brain Cognitive Development (ABCD) study, DKLP reduces cross-site predictive mean squared errors by approximately 10% and yields the most stable single-site region selection. In an EEG-based brain–computer interface (BCI) study, DKLP improves individual-level character decoding accuracy for amyotrophic lateral sclerosis (ALS) participants from 75% to 80%, demonstrating practical improvements in both predictive performance and model generalization.

Chapter IV addresses the challenge of generating individualized task-based fMRI contrast maps using generative diffusion models. Unlike natural images, fMRI data exhibit sparse activation signals embedded in structured, spatially correlated noise with substantial inter-subject variability. Existing generative approaches often fail to preserve this structure or provide interpretable control over subject-level activations. We propose the Brain Activation Controlled Network (BACNet), which integrates Bayesian nonparametric regression with ControlNet-based diffusion modeling to incorporate spatially informative activation maps during image generation. This approach improves both biological interpretability and control over spatial features in synthetic images. Experiments on simulated MNIST data and real fMRI data from the ABCD study show that the proposed BACNet achieves the best trade-off between fidelity and structural accuracy compared to baseline generative models. Furthermore, we develop a synthetic null hypothesis testing framework that utilizes the generative model to construct empirical null distributions for small-sample studies. In the ABCD dataset, the conventional single-site analysis recovers only 3% of significant regions identified in the full multi-site data, whereas our generative inference recovers 44% of the voxels, demonstrating substantial gains in statistical power, reproducibility, and interpretability.

In summary, this dissertation establishes unified Bayesian modeling and deep learning frameworks for neuroimaging analysis. The proposed methods advance statistical inference, improve computational scalability, and enable biologically meaningful data generation, providing powerful tools for large-scale brain imaging research.

CHAPTER 1

Introduction

1.1 Neuroimaging

Over the past few decades, neuroimaging has become an essential technique in biomedical research for studying human brain structures and functions. Insights gained from neuroimaging studies have substantially advanced our understanding of neurological conditions and provide valuable information for disease diagnostics, prognostics, and drug development. With the rapid development of new data acquisition techniques, neuroimaging modalities, including magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and electroencephalography (EEG), provide unique insights in neuroscience research with distinct spatial and temporal resolutions.

Among these imaging techniques, functional MRI (fMRI) is a noninvasive technique that measures neural activity by capturing changes in blood oxygenation levels, known as the blood-oxygen-level-dependent (BOLD) signals. fMRI data are commonly represented in either volumetric or cortical surface data. Volumetric fMRI captures signals on a three-dimensional grid of voxels, where each voxel represents a small unit of brain tissue within the volume. In contrast, cortical surface fMRI measures the left and right hemispheres of the brain's outer gray matter layer on a two-dimensional manifold surface. During an fMRI session, BOLD signals are recorded over time, producing voxel-wise time series data that capture dynamic brain responses. There are two primary types of fMRI data: task-based and resting-state fMRI. In task-based fMRI sessions, participants are scanned while performing cognitive, sensory, or motor tasks, allowing researchers to map brain regions that respond to particular stimuli or actions. Massive univariate time-series regression models are typically applied independently at each voxel, producing coefficient or test statistic maps, which are known as contrast maps. These maps are often used as processed data for downstream group-level analysis. In resting-state fMRI, subjects are scanned at rest to investigate intrinsic functional connectivity based on spontaneous low-frequency fluctuations across brain regions.

In recent years, many large-scale fMRI datasets, such as the Adolescent Brain Cognitive Development (ABCD) Study [Jernigan et al., 2018], UK Biobank [Sudlow et al., 2015], the Human Connectome Project (HCP) [Glasser et al., 2016], and the Neuroimaging Analysis Replication and Prediction Study (NARPS) [Botvinik-Nezer et al., 2020], have substantially advanced neuroimaging research. These datasets offer diverse populations and experimental designs, providing valuable data sources for statistical modeling and neuroimaging studies.

Electroencephalography (EEG) is another noninvasive neuroimaging method that records the brain's electrical activity through electrodes placed on the scalp. A growing area of research involves EEG-based brain-computer interfaces (BCIs), which aim to translate neural signals into control commands for external devices without requiring muscular activity. These systems have demonstrated potential in assistive technologies for communication, rehabilitation, and motor control, particularly for individuals with severe physical impairments. One widely studied BCI application is the P300 speller, which enables users to select characters on a screen by detecting the P300 event-related potential (ERP) signals by visual stimuli [Farwell and Donchin, 1988]. The effectiveness of such systems relies heavily on accurate signal processing and classification techniques, including stepwise linear discriminant analysis (swLDA) [Farwell and Donchin, 1988], support vector machines (SVM) [Kaper et al., 2004], and deep learning methods such as EEGNet, a compact convolutional neural network architecture proposed by Lawhern et al. [2018].

Statistical methods in neuroimaging data face numerous challenges due to its complex data structure. First, neuroimaging datasets are inherently high-dimensional. For example, fMRI time-series scans or contrast maps typically contain over 100,000 voxels per subject, while EEG recordings involve dense temporal sampling from dozens or more electrodes across various brain channels. Moreover, neuroimaging signals are subject to substantial inherent noise and low signal-to-noise ratios due to measurement errors, body motions, scanner drift, and physiological artifacts. Additionally, there are large and intricate spatial and temporal dependencies, as well as a high degree of variability across subjects, which vary across modalities and tasks during data collection. These complex structures impose significant challenges for statistical models to capture these correlations and variabilities while retaining the ability to generalize across studies and populations. Finally, unlike domains such as computer vision, where large-scale labeled datasets are readily available, neuroimaging studies often involve relatively small sample sizes due to the cost, time, and ethical considerations involved in data collection. These challenges make neuroimaging studies a particularly challenging domain for statistical inference and predictions, while keeping the field active for more advanced statistical methods to deliver scientifically reliable results and interpretations.

1.2 Image regression models

A wide range of research questions in neuroimaging can be addressed using regression models. In task-based fMRI studies, a common objective is to identify brain regions activated by specific tasks or stimuli. Traditionally, this has been achieved through massive univariate analysis (MUA) [Friston, 2003], where univariate statistical tests such as t-tests, ANOVA, or linear regression are independently applied to each voxel. To address multiple comparisons, procedures such as the Bonferroni correction or false discovery rate (FDR) control are typically employed. MUA is widely adopted and implemented in standard neuroimaging software packages. However, this approach fails to account for the inherent spatial or temporal correlations across voxels and cannot adequately capture the complexity of neural activation patterns. As a result, MUA often shows limited statistical power and interpretability, especially when analyzing spatially varying effects.

To address these limitations, more flexible regression models have been developed within the framework of functional data analysis, where imaging data and regression coefficients are treated as smooth functions defined over space or time. For instance, nonparametric regression approaches facilitate the estimation of complex, nonlinear relationships without requiring a parametric functional form. Additionally, several regression frameworks have emerged in neuroimaging analysis, including image-on-scalar regression (ISR), scalar-on-image regression (SIR), and image-to-image regression (IIR). Each of these frameworks provides different perspectives for relating imaging data to covariates under various research questions.

ISR treats image data as response variables to learn their associations with scalar covariates, such as age, gender, or clinical measurements. These models are particularly useful in characterizing regional effects and detecting activations in multi-subject fMRI studies [Li et al., 2021, Zhang et al., 2023a, Lin et al., 2024, Whiteman et al., 2024]. SIR utilizes image data as high-dimensional predictors to predict scalar outcomes, such as behavioral scores or clinical diagnostics [Goldsmith et al., 2014, Kang et al., 2018, Xu and Kang, 2023]. Regularized regression, basis function expansions, and spatial priors are often employed in SIR to address the high dimensionality of the image domain and to improve interpretability. IIR models one image as a function of another and is useful for modeling associations between different modalities [Guo et al., 2022, Ma et al., 2023] while accounting for high dimensionality and various spatial relationships in image data.

1.3 Bayesian methods and Gaussian process prior

Bayesian methods provide a powerful and flexible framework for modeling complex data structures. They can effectively handle hierarchical dependencies, incorporate prior knowledge, and quantify estimation uncertainties. In recent years, a wide range of Bayesian models has been developed for neuroimaging applications, demonstrating strong performance and improved interpretability by explicitly modeling the spatial and temporal dependencies inherent in neuroimaging data.

Among Bayesian nonparametric priors, the Gaussian process (GP) prior is one of the most extensively studied and widely applied in both theory and practice. A GP defines a prior over functions, fully specified by a mean function and a covariance kernel function that encodes the correlations between spatial locations. In neuroimaging studies, GP priors are commonly used to model latent spatial or temporal functions, such as activation patterns, spatially varying regression coefficients, or structured noise processes. Many Bayesian regression models with GP or variations of GP priors have been proposed in neuroimaging analysis [Kang et al., 2018, Menacher et al., 2024, Goldsmith et al., 2014, Lin et al., 2024, Whiteman et al., 2024] for various regression tasks, showing their capability to capture complex spatial dependencies and improve the estimation of activation patterns.

However, two major challenges limit the scalability and flexibility of standard GP models in high-dimensional neuroimaging datasets. First, exact posterior inference requires the inversion or Cholesky decomposition of the kernel matrix, which requires a computational time complexity in $\mathcal{O}(V^3)$ and memory cost in $\mathcal{O}(V^2)$, where V is the number of voxels. This makes whole-brain level applications impractical because V usually exceeds hundreds of thousands. Second, fixed parametric kernels are often too simple to adequately capture the complex, heterogeneous spatial structures observed in neuroimaging data. These patterns may vary across study populations and image outcomes, leading to poor generalization and requiring extensive manual tuning to adapt the spatial properties of each dataset.

1.4 Deep learning methods

Deep learning has become an increasingly influential tool in imaging research, offering strong representational power for modeling complex, high-dimensional, and nonlinear patterns in imaging data. In practice, deep neural networks (DNNs) have demonstrated success in a variety of neuroimaging tasks, including disease classification, image segmentation, functional connectivity analysis, and spatiotemporal pattern recognition [Shen et al., 2017, Lee et al., 2017, Zhou et al., 2021]. Moreover, DNNs are theoretically supported by the univer-

sal approximation theorem, which guarantees their ability to approximate a broad class of continuous functions with arbitrary precision, given sufficient network depth and complexity [Scarselli and Tsoi, 1998, Hornik et al., 1989].

Additionally, DNNs have been employed to model spatially varying functions within various regression frameworks. For instance, Zhang et al. [2023a] uses DNN architectures to estimate spatially varying regression functions in ISR under a frequentist approach. Under the context of IIR, Ma et al. [2023] proposes a multi-stage framework to predict task-based fMRI contrast maps from resting-state data, providing a biologically meaningful interpretation of features between image modalities.

Beyond regression tasks, generative models have gained increasing attention for synthesizing neuroimaging data. These models aim to learn the data-generating distribution and produce realistic synthetic brain images that preserve both anatomical and functional plausibility. Among recent developments, denoising diffusion probabilistic models (DDPMs) [Ho et al., 2020] have emerged as a state-of-the-art class of generative models, surpassing traditional frameworks such as generative adversarial networks (GANs) [Goodfellow et al., 2014] and variational autoencoders (VAEs) [Kingma et al., 2013] in terms of fidelity and stability [Dhariwal and Nichol, 2021]. DDPMs learn to reverse a forward diffusion process that incrementally adds noise to the data, thereby reconstructing samples from noise through a series of learned denoising steps.

A key feature of diffusion models is controllable generation, where the generated images can be guided by additional information such as class labels or clinical covariates. There have been numerous successful applications in natural image generation and various conditional generation tasks, including controllable image generation [Cao et al., 2024], inpainting [Lugmayr et al., 2022], super-resolution [Moser et al., 2024], and image segmentation [Amit et al., 2021]. One common approach is classifier guidance, in which a separately trained classifier is used during sampling to adjust the denoising steps toward a target class label [Dhariwal and Nichol, 2021]. To avoid additional training costs, classifier-free guidance (CFG) becomes a more flexible alternative [Ho and Salimans, 2022], which is widely adopted in recent works. In CFG, the diffusion model is trained jointly on both conditional and unconditional samples by randomly dropping conditioning labels during training. This is especially useful in neuroimaging, as datasets often contain missing labels. By treating the unlabeled samples as unconditional inputs, CFG allows the model to retain them in the training data, improving data efficiency and reducing information loss.

Zhao et al. [2025] provides a comprehensive review of recent advances in diffusion-based generative modeling in neuroimaging, highlighting successful applications across diverse modalities, including MRI, EEG, and PET. However, the application of diffusion models

to task-based fMRI data remains underexplored. Beyond the general challenges of fMRI analysis, an important difficulty lies in simultaneously modeling the generative distributions of both full fMRI contrast maps and the task-relevant activation regions, which are the primary targets of scientific interest. Activation maps tend to be sparse and embedded within complex noise structures, making it difficult to accurately recover them when using contrast maps as the direct generative target. Moreover, direct generation targeting observed fMRI data offers limited control over subject-specific generation, hindering the ability to synthesize brain images that reflect individual activation patterns or covariate profiles. As a result, ensuring biological and clinical plausibility in generated fMRI images remains an open problem.

1.5 Contribution

In this dissertation, we develop novel methodologies that integrate Bayesian inference with deep learning to address key challenges in neuroimaging analysis, focusing on regression models and image generation. In Chapter 2, we introduce a scalable image model using a Bayesian approach (SIMBA) for whole-brain voxel-wise fMRI analysis using GP priors. By low-rank kernel approximations and efficient Gibbs sampling and variational inference algorithms, we enable full posterior inference within minutes or seconds. This framework provides spatially interpretable activation estimates and uncertainty quantification, making Bayesian whole-brain analysis computationally feasible for large-scale neuroimaging studies. In Chapter 3, we propose a deep kernel learning process (DKLP) model that adaptively learns kernel basis functions in GP priors through DNNs under various regression tasks. This method enhances the flexibility of traditional GP models by enabling data-driven learning of spatial covariance structures under diverse regression settings. The proposed DKLP prior model substantially improves spatially varying functions estimation and region detection in both simulated and real neuroimaging data applications. In Chapter 4, we develop a diffusion-based generative modeling framework that combines Bayesian regression and DDPMs for individualized synthesis of task-based fMRI contrast maps. Beyond conditional generation based on clinical or demographic covariates, our model introduces spatial conditioning using features learned from nonparametric regression, allowing spatially guided and anatomically plausible image generation. This multi-stage design separates the learning processes for marginal image distribution and the distribution conditional on activation regions, thereby providing greater interpretability and control over the generative process. In Chapter 5, we conclude with a summary of contributions, limitations, and future research directions.

CHAPTER 2

SIMBA: Scalable Image Modeling Using a Bayesian Approach, A Consistent Framework for Including Spatial Dependencies in fMRI Studies

2.1 Introduction

Functional magnetic resonance imaging (fMRI) has emerged as one of the most widely used neuroimaging techniques for noninvasive mapping of brain activity. In typical task-based fMRI studies, participants are scanned while performing specific cognitive or motor tasks, with the primary goal of identifying brain regions that are activated in response to the experimental conditions and often response contrasts. By measuring blood-oxygen-level-dependent (BOLD) signals, fMRI provides rich spatial and temporal information on brain function. These data are indexed by spatial locations, which can be represented either in a volumetric format, where responses are measured at voxels across the three-dimensional brain volume, or in a cortical surface format, where responses are mapped onto two-dimensional surface meshes of the cortex. In both cases, these spatial locations can be denoted using three-dimensional coordinates to describe their relative positions within the brain volume.

Traditional fMRI analysis studies operate in a massive univariate analysis (MUA) framework. First, a time series regression model is independently fitted to each spatial location on single-participant data to obtain estimated coefficients associated with task effects or contrasts between different conditions. These coefficients comprise the individual-specific effect estimates or contrast maps, which are then used for subsequent or group-level analyses to map brain activations. Then, MUA is applied at the group level, where each spatial location is treated independently, and univariate hypothesis tests such as t-tests or general linear models (GLMs) are applied at each location. This approach is often followed by multiple

testing adjustment procedures of statistical values, such as family-wise error (FWE) rate adjustment or false discovery rate (FDR) control. Although MUA is computationally efficient and widely used in popular neuroimaging software such as FSL (FMRIB Software Library; Jenkinson et al. [2012]), SPM (Statistical Parametric Mapping; Ashburner [2012]), and AFNI (Analysis of Functional NeuroImages; Cox [1996]), it has critical limitations. MUA results in information waste by ignoring spatial correlations and common information shared across voxels, as it assumes independence at each spatial location and lacks model flexibility. They also operate solely on statistical values, ignoring the importance of the effect estimates from modeling [Chen et al., 2017]. Furthermore, multiple testing adjustment methods commonly applied in MUA pipelines often cause information loss by producing binary decisions rather than providing full inference on the magnitude and uncertainty of activation effects [Marchini and Presanis, 2004, Chen et al., 2019].

To address these limitations, it is crucial to account for the spatial dependence inherent in fMRI data within a consistent and appropriate statistical modeling framework. Bayesian models with spatial priors provide a principled approach to encode this dependence, allowing models to borrow strength across neighboring locations and improving both model performance and interpretability. Popular methods include conditional autoregressive (CAR) models and Gaussian Markov random fields (GMRFs), which impose local spatial smoothness through sparse precision matrices [Descombes et al., 1998, Woolrich et al., 2004, Mejia et al., 2020, Zhang et al., 2016, Sidén et al., 2017]. Other extensions incorporate mixture models with Dirichlet process clustering to account for latent spatial heterogeneity [Jbabdi et al., 2009, Zhang et al., 2016]. More recently, Gaussian process (GP) priors have gained increased attention in neuroimaging as a flexible method for modeling globally smooth spatial functions, with various applications in modeling activation regions, individual-specific effects, and spatially structured residual noise [Whiteman et al., 2024, Lin et al., 2024].

In general, Bayesian spatial models face significant computational challenges in whole-brain fMRI analyses, primarily due to the high dimensionality of the data, with the number of spatial locations often exceeding 10^5. Fully Bayesian inference via Markov Chain Monte Carlo (MCMC) sampling becomes computationally infeasible, particularly when accounting for spatial dependence across the entire brain. To mitigate this issue, some studies have focused on region-based analyses, either restricting inference to small regions of interest (ROIs) [Hartvig and Jensen, 2000] or averaging voxel-level effects within predefined anatomical regions [Chen et al., 2019], both of which sacrifice the fine spatial resolution inherent in voxel-wise fMRI data. To improve scalability, approximate inference methods have been widely explored. Many studies adopt variational inference (VI), which offers a computationally efficient alternative to MCMC by approximating the posterior through opti-

mization [Zhang et al., 2016, Penny et al., 2005, Sidén et al., 2017]. Another useful direction is based on the integrated nested Laplace approximation (INLA). Mejia et al. [2020] introduced spatial Bayesian GLMs that combine INLA with GMRF priors, taking advantage of sparse precision matrices to achieve substantial computational gains. This approach has been shown to deliver accurate posterior estimates and can scale to tens of thousands of voxels or surface vertices with runtimes on the order of hours. However, because GMRF priors are defined through the sparsity structure of the precision matrix, the induced spatial dependence is limited to local neighborhoods. This restriction makes them well-suited for local smoothing but less flexible in capturing global or long-range spatial correlations.

GP priors offer a powerful alternative for modeling globally smooth and spatially varying effects. However, exact inference under GP priors involves dense covariance matrix inversions with computational complexity $\mathcal{O}(V^3)$ and memory cost $\mathcal{O}(V^2)$, where V is the number of spatial locations. This makes traditional MCMC approaches infeasible for whole-brain fMRI analyses, particularly for Hamiltonian Monte Carlo (HMC) methods commonly used in software like Stan [Carpenter et al., 2017] and PyMC [Salvatier et al., 2016]. Whiteman et al. [2024] proposed a two-stage MCMC-based framework using Vecchia approximations for cortical surface fMRI data. Lin et al. [2024] developed a population-weighted image-on-scalar regression method using low-rank GP kernel approximations, with a maximum likelihood estimation method for parameter estimation instead of Bayesian inference. Despite these advances, there remains a lack of a unified, fully Bayesian framework capable of performing computationally efficient inference across the whole brain volume using GP priors.

To address these challenges, we propose a Scalable Image Model using a Bayesian Approach (SIMBA) for modeling spatial effects in group-level fMRI analysis. This is a unified Bayesian modeling framework with GP priors that can efficiently perform whole-brain level analysis on fMRI data. The SIMBA model explicitly targets effect estimation at each spatial location across the brain volume, while simultaneously incorporating participant-specific effects and spatially correlated residuals. This framework allows for flexible and statistically consistent modeling of inter-participant variability with spatial dependencies, while also enabling posterior predictive checks as a principled method for validating model fit and assessing the adequacy of spatial inferences.

To reduce computational complexity in the implementation of this method, we adopt a low-rank approximation of the GP covariance kernel, projecting the high-dimensional spatial functions to a lower-dimensional subspace. Our approach shares a similar kernel decomposition and model reparameterization strategy with Lin et al. [2024], extending it to full Bayesian inference and allowing for more flexible kernel choices. We implement efficient posterior inference through both Gibbs sampling and variational inference, enabling whole-brain

Bayesian spatial inference at voxel-level resolution within minutes or seconds. This makes our method feasible for high-dimensional fMRI data, which has previously been unattainable in standard neuroimaging practice. We demonstrate the performance of our approach through three example cases: first, with simulation studies where results can be compared with known results; and then with real fMRI volumetric data from the Neuroimaging Analysis Replication and Prediction Study (NARPS) [Botvinik-Nezer et al., 2020] in a whole brain analysis and finally with a gray matter-focused analysis of real data from the Adolescent Brain Cognitive Development (ABCD) study [Casey et al., 2018] to identify cortical activation patterns.

The remainder of this paper is organized as follows. In Section 2.2, we introduce the SIMBA framework and the corresponding prior specifications. In Section 2.3, we demonstrate the kernel approximation, model reparameterization, and Bayesian inference algorithms, including the Gibbs sampler and variational inference. In Section 2.4, we present simulation studies that evaluate model performance, followed by a real data application in Section 2.5 on task-based fMRI volumetric data from the NARPS dataset and cortical gray matter surface data from the ABCD study. We conclude with a discussion in Section 2.6.

2.2 Methods

2.2.1 Model Specification

Suppose fMRI data are collected from n individuals, each measured on V spatial locations across the brain. Let $\mathcal{S} = \{s_v\}_{v=1}^{V} \subset \mathbb{R}^d$ denote the d-dimensional coordinates of the spatial locations of the volumetric voxels or surface vertices. For simplicity of notation, we refer to all locations as voxels indexed by v throughout. Additionally, while we use a single index to label voxels, the original geometric spatial structure is maintained and used within the modeling process.

We denote $y_i(s_v)$ as the observed image response at the spatial location s_v and let x_{ir} represent the r-th covariate for participant i, such as demographic or clinical information, with $i = 1, \ldots, n$ and $r = 0, \ldots, q$. The proposed SIMBA model is specified as:

$$y_i(s_v) = \sum_{r=0}^{q} x_{ir}\alpha_r + \sum_{r=0}^{q} x_{ir}\beta_r(s_v) + \eta_i(s_v) + \epsilon_i(s_v), \quad \epsilon_i(s_v) \sim \mathcal{GP}(0, \sigma_\epsilon^2 K). \tag{2.1}$$

We include $x_{i0} = 1$ for all $i = 1, \ldots, n$ to model the global intercept α_0 and the spatially varying intercept $\beta_0(s_v)$ for baseline brain activity patterns. Here, α_r represents the global effect of the r-th covariate shared across all locations, and $\beta_r(s_v)$ captures the spatially

varying population-level effect of the same covariate at location s_v. Participant-specific deviations from the voxel-level population effect are represented by $\eta_i(s_v)$, and $\epsilon_i(s_v)$ denotes spatial noise, which is assumed to follow a zero-mean Gaussian process with a covariance kernel K and unknown kernel variance σ_ϵ^2. To ensure model identifiability, we impose the constraints: $\sum_{i=1}^{n} \eta_i(s_v) = 0$ for all $v = 1, \ldots, V$, and $\sum_{v=1}^{V} \beta_r(s_v) = 0$ for $r = 0, \ldots, q$. Our primary parameter of interest is the voxel-level effect $\alpha_r + \beta_r(s_v)$ for $r = 0, \ldots, q$, which combines both the global and spatial varying association between the intercept or the r-th covariate and brain activities.

For each global effect, we assign a normal prior with mean zero and an unknown variance σ_α^2, denoted as $\alpha_r \sim \mathcal{N}(0, \sigma_\alpha^2)$. We incorporate spatial information by assigning GP priors with zero mean and the same predefined covariance function $K(\cdot, \cdot)$ as for the residuals $\epsilon_i(s_v)$ to the latent spatial functions $\beta_r(\cdot) \sim \mathcal{GP}(0, \sigma_\beta^2 K)$ and $\eta_i(\cdot) \sim \mathcal{GP}(0, \sigma_\eta^2 K)$. For computational simplicity, we assume a shared covariance kernel K across all spatial components, while allowing for different kernel variances $\sigma_\beta^2, \sigma_\eta^2$ and σ_ϵ^2 to capture varying degrees of spatial variation across the respective model components. In practice, different kernels could be specified for the individual-level effects if desired. The choice of kernel function determines the smoothness of spatially varying functions and spatial correlations across voxels. In this work, we use a Matérn kernel, which is widely adopted in spatial statistical modeling in fMRI studies [Mejia et al., 2020, Sidén et al., 2021] due to its flexibility in controlling both correlation range and smoothness. Details on the kernel construction and approximation strategies are provided in Section 2.2.2. We place half-Cauchy priors on the unknown standard deviations as $\sigma_\alpha, \sigma_\beta, \sigma_\eta, \sigma_\epsilon \sim \mathcal{C}^+(A)$, where $A = 100$ induces weakly informative priors. To enable efficient posterior sampling with conjugate priors, we reparameterize the half-Cauchy prior using the inverse gamma (IG) distribution as $\sigma_\alpha^2 \sim \text{IG}(\frac{1}{2}, \frac{1}{a_\alpha}), \sigma_\beta^2 \sim \text{IG}(\frac{1}{2}, \frac{1}{a_\beta}), \sigma_\eta^2 \sim \text{IG}(\frac{1}{2}, \frac{1}{a_\eta}), \sigma_\epsilon^2 \sim \text{IG}(\frac{1}{2}, \frac{1}{a_\epsilon})$, where $a_\alpha, a_\beta, a_\eta, a_\epsilon \sim \text{IG}(\frac{1}{2}, \frac{1}{A^2})$. This hierarchical formulation preserves conjugacy in the posterior computations while retaining the desirable heavy-tailed shrinkage property of the half-Cauchy prior.

2.2.2 Kernel approximation

A key challenge in applying GP priors to whole-brain fMRI data lies in the computational cost. Inference using GP priors requires the inversion of dense covariance matrices, which incurs $\mathcal{O}(V^3)$ in complexity. This makes standard GP inference infeasible for voxel-level analysis, where V often exceeds 100,000. To address this, we adopt a low-rank approximation strategy based on the Nyström method [Nyström, 1930]. This approach not only signif-

icantly reduces computational complexity but also preserves the flexibility to incorporate any parametric covariance kernel.

Let $K_V = K(\mathcal{S}, \mathcal{S}) \in \mathbb{R}^{V \times V}$ denote the Gram matrix over all spatial locations $\mathcal{S}$. We select $J \ll V$ inducing points $\mathcal{S}_J \subset \mathcal{S}$ (a process described below in Sec. 2.3.1), and define the reduced matrices $K_J = K(\mathcal{S}_J, \mathcal{S}_J) \in \mathbb{R}^{J \times J}$, $\quad K_{V,J} = K(\mathcal{S}, \mathcal{S}_J) \in \mathbb{R}^{V \times J}$. Following the Nyström method approximation strategy, the full kernel can be approximated as $K_V \approx K_{V,J} K_J^{-1} K_{V,J}^\top$, where the inducing points serve as a representative set that captures the dominant spatial structure, enabling a computationally efficient representation of the kernel.

In the next step, we represent this approximated matrix through a spectral decomposition for subsequent model inference. We first compute the Cholesky decomposition $K_J = R_J R_J^\top$, where $R_J \in \mathbb{R}^{J \times J}$ is a lower triangular matrix. Plugging this into the Nyström expression yields: $K_V \approx K_{V,J} R_J^{-\top} R_J^{-1} K_{V,J}^\top = \widetilde{K} \widetilde{K}^\top$, where $\widetilde{K} = K_{V,J} R_J^{-\top} \in \mathbb{R}^{V \times J}$. We then apply singular value decomposition (SVD) on $\widetilde{K}$ to get $\widetilde{K} = \Psi D U^\top$ where $\Psi \in \mathbb{R}^{V \times J}$ is an orthonormal matrix, i.e. $\Psi^\top \Psi = I_J$, where $I_J \in \mathbb{R}^{J \times J}$ is the J-dimensional identity matrix. $D \in \mathbb{R}^{J \times J}$ is a diagonal matrix containing the singular values of $\widetilde{K}$, and $U \in \mathbb{R}^{J \times J}$ is another orthonormal matrix. Then the approximated kernel can be written as $K_V \approx \Psi \Lambda \Psi^\top$ with $\Lambda = D^2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_J)$, which aligns with the kernel decomposition in Mercer's theorem. The obtained matrices Ψ and Λ, which contain the approximated basis functions and their corresponding eigenvalues from the spectral decomposition, are then used to reparameterize the regression model formulation, as described in Section 2.3.1. For the kernel approximation of the individual-specific deviations $\eta_i(\cdot)$, we consider a smaller number $J_\eta \ll J$ of basis functions to prevent overfitting and reduce computational cost, resulting in a lower-rank approximation characterized by $\Psi_\eta \in \mathbb{R}^{V \times J_\eta}$ and $\Lambda_\eta \in \mathbb{R}^{J_\eta \times J_\eta}$ derived in the same manner as above, but with J_η inducing points instead. In practice, setting $J_\eta = \lfloor 0.1J \rfloor$ is typically sufficient to capture individual spatial variation while maintaining computational efficiency and preventing overfitting.

We apply this kernel decomposition to represent the spatially varying coefficients via the Karhunen–Loève (KL) expansion. The spatially varying components in the model can be expressed as:

$$\beta_r(\cdot) = \sum_{j=1}^{J} \psi_j(\cdot) \sqrt{\lambda_j} \theta_{\beta_{rj}}, \quad \eta_i(\cdot) = \sum_{l=1}^{J_\eta} \psi_{\eta_j}(\cdot) \sqrt{\lambda_{\eta_j}} \theta_{\eta_{ij}}, \quad \epsilon_i(\cdot) = \sum_{j=1}^{J} \psi_j(\cdot) \sqrt{\lambda_j} \theta_{\epsilon_{ij}}. \tag{2.2}$$

Here, $\psi_j(\cdot)$ and $\psi_{\eta_j}(\cdot)$ denote the j-th basis from the orthonormal matrices Ψ and Ψ_η, while $\{\theta_{\beta_{rj}}\}_{r=0}^{q}$, $\{\theta_{\eta_{ij}}\}_{i=1}^{n}$, and $\{\theta_{\epsilon_{ij}}\}_{i=1}^{n}$ represent the j-th basis coefficients for the population-level, individual-level, and residual spatial effects, respectively. To allow for flexible spatially

varying coefficients, we assign independent normal priors with mean zero and unknown variances for each spatial component, denoted by $\theta_{\beta_{rj}} \sim \mathcal{N}(0, \sigma_\beta^2)$, $\theta_{\eta_{ij}} \sim \mathcal{N}(0, \sigma_\eta^2)$, and $\theta_{\epsilon_{ij}} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

2.3 Posterior Computation

2.3.1 Model reparameterization

To achieve scalable Bayesian inference, we reformulate the proposed model in a low-dimensional subspace induced by the kernel approximation described in Section 2.2.2. Let $\mathbf{y}_i \in \mathbb{R}^V$ denote the vector of observed image responses for participant i, and let $\mathbf{x}_i \in \mathbb{R}^{q+1}$ be the vector of intercept and q covariates. Then, Model (2.1) can be expressed in vectorized form as:

$$\mathbf{y}_i = \mathbf{x}_i^\top \boldsymbol{\alpha} \mathbf{1}_V + \mathbf{x}_i^\top \boldsymbol{\theta}_\beta \Lambda^{1/2} \Psi^\top + \boldsymbol{\theta}_{\eta_i} \Lambda_\eta^{1/2} \Psi_\eta^\top + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma_\epsilon^2 K), \quad K \approx \Psi \Lambda \Psi^\top, \tag{2.3}$$

where $\mathbf{1}_V \in \mathbb{R}^V$ is a vector of all ones, $\boldsymbol{\alpha} \in \mathbb{R}^{q+1}$ is the vector for common covariate effects shared by all locations, $\boldsymbol{\theta}_\beta \in \mathbb{R}^{(q+1)\times J}$ contains the basis coefficients for the population-level effects after KL expansion, and $\boldsymbol{\theta}_{\eta_i} \in \mathbb{R}^{J_\eta}$ contains coefficients for the participant-specific effects. Here, $\Psi \in \mathbb{R}^{V\times J}$ and $\Lambda \in \mathbb{R}^{J\times J}$ represent the approximated orthonormal basis functions and the associated eigenvalues from the low-rank kernel approximation described in Section 2.2.2.

Direct inference in the original space remains computationally prohibitive due to the large dimension V. To avoid this bottleneck, we reparameterize the model by projecting the data onto the low-dimensional subspace. Specifically, we define the projection matrices: $\Phi = \Psi \Lambda^{-1/2} \in \mathbb{R}^{V\times J}$, $\Phi_\eta = \Lambda_\eta^{1/2} \Psi_\eta^\top \Psi \Lambda^{-1/2} \in \mathbb{R}^{J_\eta \times J}$, and compute the transformed response vector for participant i as $\widetilde{\mathbf{y}}_i = \mathbf{y}_i^\top \Phi \in \mathbb{R}^J$. Under this reparameterization, the model becomes:

$$\widetilde{\mathbf{y}}_i = \mathbf{x}_i^\top \boldsymbol{\alpha} \mathbf{1}_V \Phi + \mathbf{x}_i^\top \boldsymbol{\theta}_\beta + \boldsymbol{\theta}_{\eta_i} \Phi_\eta + \widetilde{\boldsymbol{\epsilon}}_i, \quad \widetilde{\boldsymbol{\epsilon}}_i \sim \mathcal{N}(0, \sigma_\epsilon^2 I_J). \tag{2.4}$$

This transformation reduces the original high-dimensional regression problem into a much lower-dimensional multivariate regression with only J outcomes per participant. As a result, posterior inference requires only operations involving matrices of size J or J_η, rather than V. This reparameterization effectively circumvents costly matrix inversions and enables efficient, vectorized posterior computations within the Bayesian framework.

The number of basis functions J used in the low-rank kernel approximation is a critical hyperparameter in our model. It controls both the approximation accuracy and computational

efficiency. A larger value of J allows the model to capture more detailed spatial variability, which is necessary for recovering complex activation patterns in fMRI data. However, including more basis functions increases computational costs and may lead to overfitting, especially when the data contains substantial noise. Conversely, selecting too few basis functions may result in underfitting, failing to adequately represent the true underlying spatial structure and the important activation signals. As an additional practical consideration in the mechanics of identifying a reasonable basis size, a different value of J defines a different approximated kernel and thereby a different likelihood function, particularly on the transformed data space. As a result, common Bayesian model selection criteria such as the widely applicable information criterion (WAIC) or the deviance information criterion (DIC) are not well-suited in our case, because the likelihood functions are not directly comparable across models with different numbers of basis functions.

Therefore, we have developed the following approach to select the number of basis functions J. We first define an upper bound $J_{\max} \leq V$, which is determined by the available computational capacity. We then compute the covariance matrix $K_{J_{\max}} \in \mathbb{R}^{J_{\max} \times J_{\max}}$ using the selected kernel function with hyperparameters empirically estimated from the data. From this truncated representation, we obtain the leading eigenvalues $(\lambda_j)_{j=1}^{J_{\max}}$ and choose a candidate range of J where the cumulative variance $\sum_{j=1}^{J} \lambda_j$ explains between 80% and 98% of the total variance $\sum_{j=1}^{J_{\max}} \lambda_j$. Within this range of J, we perform a grid search based on predictive performance evaluated via leave-one-out cross-validation (LOOCV). Specifically, we compute the voxel-wise predictive mean squared error (PMSE) by leaving out one participant at a time and using the fitted model from the remaining participants to predict the response. This approach allows us to select the value of J that yields the best predictive accuracy without relying on likelihood comparisons.

2.3.2 Gibbs sampler

With the reparameterized model in (2.4) and prior specifications, all model parameters are assigned independent conjugate priors, enabling efficient posterior inference via Gibbs sampling. Specifically, all model parameters can be sampled directly from their full conditional distributions. The full set of parameters includes:

$$\boldsymbol{\Theta} = \{(\alpha_r)_{r=0}^{q}, (\boldsymbol{\theta}_{\beta_r})_{r=0}^{q}, (\boldsymbol{\theta}_{\eta_i})_{i=1}^{N}, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\eta^2, \sigma_\epsilon^2, a_\alpha, a_\beta, a_\eta, a_\epsilon\}.$$

Detailed derivations of these conditional posteriors are provided in the Supplementary Materials.

A key advantage of our formulation is that it avoids high-dimensional matrix inversions

typically required in GP-based models. For models that include only an intercept term, the posterior covariance matrix of $\boldsymbol{\theta}_\beta$ is diagonal, allowing for vectorized independent sampling of each coefficient in the J-dimensional vector. This structure significantly simplifies posterior computation. When additional covariates are included, the size of the posterior covariance of $\boldsymbol{\theta}_{\beta_r}$ grows with the number of covariates J, but remains low-dimensional and computationally tractable, since in real applications J is typically small. For the individual-specific basis coefficients $\boldsymbol{\theta}_{\eta_i}$, we can consider Φ_η as the design matrix, and then the posterior is multivariate normal distribution with a covariance matrix in $\mathbb{R}^{J_\eta \times J_\eta}$. Since we set $J_\eta \ll J \ll V$, the dimensionality remains low, and posterior sampling remains efficient by using Cholesky decomposition on the posterior precision matrix.

The total per-iteration computational complexity of our Gibbs sampler is $\mathcal{O}(NJJ_\eta + J_\eta^3)$, with a memory cost of $\mathcal{O}(VJ)$. This demonstrates a significant improvement from the original complexity of $\mathcal{O}(V^3)$ in time and $\mathcal{O}(V^2)$ in memory, which is achieved by low-rank kernel approximation and model reparameterization. The full Gibbs sampling algorithm is implemented in the Python module PyTorch [Ansel et al., 2024] to leverage fast tensor operations and potential GPU acceleration. In practice, we run multiple MCMC chains with different initial values and assess convergence using the Gelman–Rubin diagnostic [Gelman and Rubin, 1992] on conditional log-likelihoods.

2.3.3 Variational Inference

While Gibbs sampling offers exact posterior inference under our model formulation, it can still be memory-consuming to store thousands of posterior samples with high-dimensional voxels or basis functions. To further improve computational efficiency, we in addition develop a variational inference (VI) approach as an alternative to MCMC-based methods.

Variational inference approximates the true posterior distribution by a simpler, tractable family of distributions, transforming the problem of sampling into an optimization task. Specifically, we adopt a mean-field variational family that assumes mutual independence among model parameters. The variational distribution factorizes as: $q(\boldsymbol{\Theta}) = \prod_{m=1}^{M} q(\theta_m)$ where θ_m represents a parameter in $\boldsymbol{\Theta}$. Each variational factor has a closed-form full conditional distribution derived from the conjugate prior, as discussed in Section 2.3.2.

We optimize the evidence lower bound (ELBO) by performing coordinate ascent variational inference (CAVI). In each iteration, we update each variational factor in closed form while holding others fixed, using the analytical updates derived from the conjugacy of the model. This ensures that each step is computationally efficient and scalable, requiring only matrix operations involving low-dimensional quantities of size J or J_η. The computational

cost per iteration of the VI algorithm is comparable to that of the Gibbs sampler. However, VI typically converges in fewer iterations than MCMC, making it advantageous for large-scale applications requiring faster posterior approximations. Full derivations of variational posterior distributions are included in the Supplementary Materials.

2.4 Simulation and validation

2.4.1 Ex. 1: Simulated data

To test the proposed model, we created synthetic datasets based on a single brain slice from a binary brain mask. These contain 4,999 brain voxels and represent a realistic fMRI image spatial structure. We evaluate the performance of our proposed model under both Gibbs sampling and variational inference algorithms, compared to alternative competing methods.

The simulated data follow Model (2.1) with $q = 1$ covariate. The true spatially varying intercept $\beta_0(\cdot)$ and covariate effect $\beta_1(\cdot)$ are each defined to be sparse and consist of two localized geometric regions with smoothly decaying effect sizes, as illustrated in Figure 2.1. In this study, we set $\alpha_0 = \alpha_1 = 0$ to ensure comparability with competing methods. The individual-level effects are independently drawn from a standard normal distribution. To assess model performance under different signal-to-noise ratio (SNR) conditions, we simulate datasets under two noise levels $\sigma_\epsilon = \{2, 5\}$, corresponding to SNR=$\{0.3, 0.05\}$. Under these settings, we also examine different sample sizes by considering two scenarios with number of individuals $n = \{50, 200\}$. The setting with low SNR and small sample size roughly mirrors our real fMRI data applications. Each simulation scenario is replicated with 100 datasets, each with the same fixed model parameters and different randomization seeds.

We evaluate the proposed SIMBA model using each of the Gibbs and variational inference (VI) implementations, and we also compare these to two existing approaches within the neuroimaging field. First is the classical general linear model (GLM), which performs massive univariate regression independently at each spatial location as a commonly used baseline in traditional fMRI analyses. This method does not incorporate spatial correlation or participant-level effects $\eta_i(s_v)$, and we use the classical Benjamini-Hochberg procedure for multiple-testing correction [Benjamini and Hochberg, 1995] at FDR = 0.05. The second comparison method is a Bayesian multi-level model (BML), which is adapted from Model (16) in Chen et al. [2019]. Although the original formulation was designed for region-level analysis, we extend the modeling implementation to voxel-level resolution to facilitate a more direct comparison. In contrast to GLM, BML borrows strength across spatial locations through hierarchical priors, enabling partial pooling of information across both individuals

and spatial locations. However, the BML method does not incorporate spatial correlations in the prior structure and models participant-level variability as constant across voxels. While the original BML model in AFNI is implemented using Hamiltonian Monte Carlo (HMC) sampling, in this voxelwise implementation we have adopted a Gibbs sampling formulation to improve computational efficiency and ensure a consistent comparison with our proposed method.

To evaluate modeling performance, we estimate the accuracy of the true spatial effects using the mean squared error (MSE) between the estimated and true main effects. In addition, we examine the activation region selection performance, which represents the ability to correctly identify the spatial locations with true non-zero effect by true positive rate (TPR) and suppress false detections in regions with no signal by FDR. For a unified assessment of significance across both frequentist and Bayesian methods, we follow the approach proposed in Taylor et al. [2023], which converts the frequentist adjusted p-values and Bayesian posterior probability of positive effect $P^+ = P(\beta_r(\cdot) > 0)$ to a "statistical evidence" measure E_s with a range of $[-1, 1]$. Specifically, for frequentist modeling, $E_s = (1 - p)$ for positive effects and $E_s = -(1 - p)$ for negative effects; for Bayesian modeling, $E_s = 2(P^+ - 0.5)$ for all posterior probabilities P^+. For the purpose of counting of true and false positives in these simulations, a voxel is referred to as "statistically significant" if $|E_s| > 0.95$, corresponding to strong evidence for either a positive or negative effect. Finally, we evaluate the uncertainty quantification of each method by computing the empirical coverage rates of the true parameters, based on either confidence intervals or credible intervals, respectively.

For BML and SIMBA-Gibbs methods, we independently run three MCMC chains, each with 4,000 burn-in iterations and 1,000 posterior samples for inference. For the proposed SIMBA model, we used a Matérn kernel with a length scale of 0.3, which was empirically estimated from the data. The number of basis functions J used in the low-rank kernel approximation was selected for each scenario based on the lowest PMSE from LOOCV, as described in Section 2.3.1. Here, the selected values of J ranged from 90 to 160 across different simulation settings, and we set $J_\eta = \lfloor 0.1J \rfloor$ as mentioned previously to avoid overfitting and save computation costs. Empirically, we found that the estimation accuracy of the population-level main effects was not sensitive to the choice of J_η. For the simulation scenario where $n = 200$ and $J = 160$, the proposed SIMBA takes approximately 10 seconds per $1,000$ iterations using the Gibbs sampler and less than 1 second using the VI algorithm, demonstrating substantial computational efficiency.

Table 2.1 summarizes the evaluation metrics for our proposed SIMBA model in comparison to alternative methods. Across all simulation scenarios, the SIMBA model consistently outperforms the competing approaches. Specifically, SIMBA achieves the lowest mean

squared error (MSE) and near-nominal coverage rates close to 95%, illustrating accurate effect estimation. Both SIMBA implementations using Gibbs sampling and VI yield higher true positive rates (TPR) compared to GLM and BML methods, while maintaining low FDR below 0.05, indicating enhanced power for detecting true activations. In one scenario ($n = 200, \sigma_\epsilon = 2$), BML achieves a slightly higher TPR, but it also leads to substantially higher false signals.

Figure 2.1 illustrates the estimated spatial effects and region selection results for all four methods, compared against the true "significant" activation patterns (first column); each subfigure shows one representative simulation scenario from the full dataset, with given noise level. While the GLM and BML methods recover part of the activation regions, their estimates are noisy and spatially scattered, with notable false detections. In contrast, SIMBA more accurately reconstructs activation regions with smooth spatially varying effects and effectively suppresses noise outside the true activation areas. The SIMBA estimates closely match the ground truth in magnitude of the effect sizes and regions with non-zero effects, demonstrating its superior ability to recover smooth spatial signals. These results demonstrate the advantages of incorporating principled spatial priors via Gaussian process modeling for voxel-wise inference, particularly in low signal-to-noise regimes.

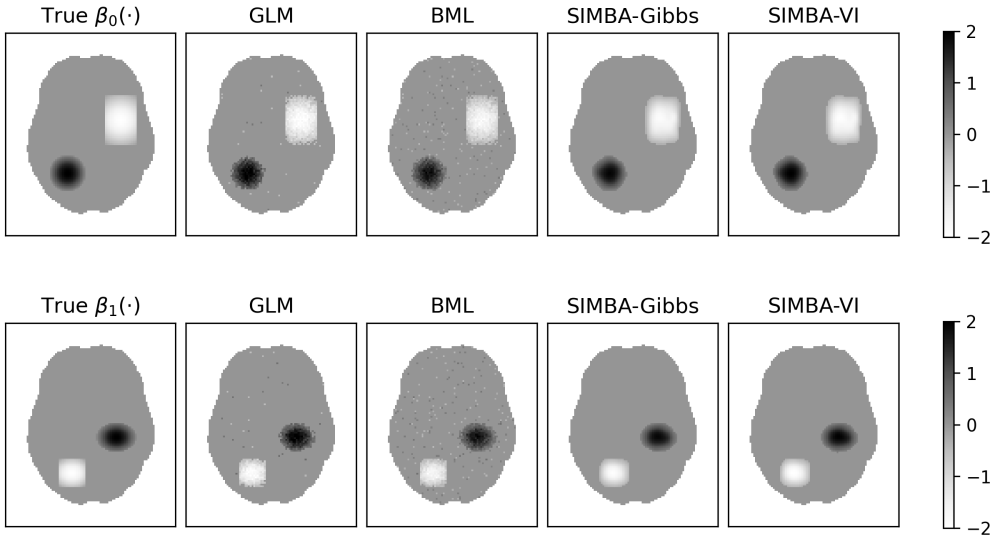
2.5 Real data applications

2.5.1 Ex.2: Neuroimaging Analysis Replication and Prediction Study

We evaluate the proposed SIMBA model in a group-level analysis using publicly available task-based fMRI data from the Neuroimaging Analysis Replication and Prediction Study (NARPS) [Botvinik-Nezer et al., 2020]. The NARPS project was a large-scale initiative designed to assess variability across neuroimaging analysis pipelines and evaluate reproducibility among research teams, which was also later re-examined separately in Taylor et al. [2023]. Both the unprocessed fMRI data and the results of the participating analysis teams have conveniently been made publicly available.

The dataset includes volumetric fMRI data collected during a mixed-gambles task, in which participants were asked to accept or reject monetary gambles with varying potential gains and losses. Each stimulus trial consists of a gain–loss pair, and participants would respond on a four-point scale (strongly/weakly accept/reject). In this study, we focus on the "equal range" group ($n = 54$ scanned participants), where the possible gain and loss values in the trials have the same value range.

(a) $n = 200, \sigma_\epsilon = 2$



(b) $n = 200, \sigma_\epsilon = 5$

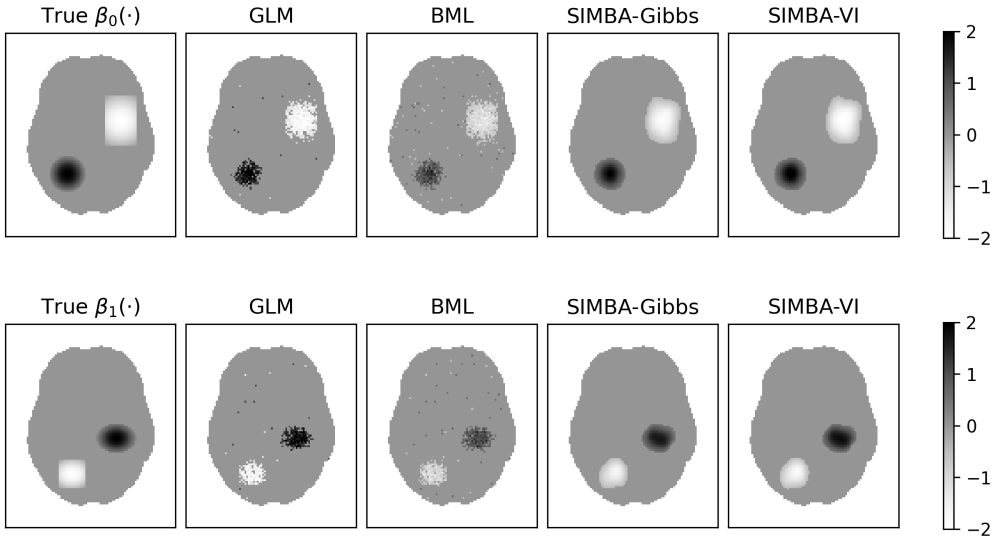


Figure 2.1: Comparison of estimated spatial effects and activation region selection across methods under two noise levels. Each row displays a map of the true spatial effect of either $\beta_0(\cdot)$ or $\beta_1(\cdot)$, as well as the corresponding estimates from each method. The non-gray shapes show the distinct "significant" activation regions, defined to be where $|E_{\text{s}}| > 0.95$.

Table 2.1: Evaluation results for simulation studies comparing mean squared error, true positive rate, false discovery rate, and coverage. Values are reported as mean percent, with standard deviations in parentheses, over 100 simulated datasets. The best-performing value(s) for each simulation is highlighted in bold.

n	σ_ϵ	GLM	BML	SIMBA-Gibbs	SIMBA-VI
			Mean Squared Error (MSE)		
50	2	9.3 (0.2)	6.8 (0.1)	1.2 (0.0)	**1.1** (0.0)
	5	48.5 (0.8)	14.7 (0.1)	2.6 (0.2)	**1.9** (0.1)
200	2	2.6 (0.0)	2.3 (0.0)	0.6 (0.0)	**0.5** (0.0)
	5	13.5 (0.2)	8.0 (0.3)	1.4 (0.2)	**1.2** (0.1)
			True Positive Rate (TPR)		
50	2	61.0 (1.0)	66.2 (1.0)	89.0 (0.9)	**89.5** (0.9)
	5	4.3 (1.0)	1.5 (0.5)	54.3 (2.0)	**58.2** (1.7)
200	2	87.6 (0.5)	**92.3** (0.4)	91.9 (0.4)	91.9 (0.5)
	5	51.3 (1.1)	58.7 (1.1)	82.9 (0.9)	**84.1** (0.9)
			False Discovery Rate (FDR)		
50	2	5.1 (0.8)	13.0 (1.0)	**6.4** (0.6)	6.5 (0.5)
	5	5.5 (3.2)	9.9 (6.5)	**0.0** (0.0)	**0.0** (0.1)
200	2	4.5 (0.6)	20.5 (0.8)	**4.0** (0.3)	**4.0** (0.3)
	5	4.5 (0.8)	10.9 (1.0)	**2.9** (0.4)	3.3 (0.5)
			Coverage		
50	2	94.6 (0.2)	93.6 (0.2)	97.8 (0.1)	**98.4** (0.1)
	5	94.5 (0.2)	92.0 (0.3)	99.7 (0.2)	**100.0** (0.0)
200	2	94.8 (0.2)	94.8 (0.2)	98.9 (0.1)	**99.1** (0.0)
	5	94.9 (0.2)	94.1 (0.3)	98.6 (0.2)	**99.2** (0.1)

Preprocessing was conducted using standardized pipelines incorporating FreeSurfer's recon-all [Fischl and Dale, 2000] for anatomical parcellation and AFNI's afni_proc.py [Cox, 1996, Reynolds et al., 2024] for full subject-level processing and voxel-level contrast generation, as described in Taylor et al. [2023]. Participant-level contrast regression coefficients were estimated using an independent time series regression model with a temporal structure of ARMA(1, 1) for the residuals at each voxel, with the gain and loss values treated as continuous amplitude modulators. The corresponding regression coefficients quantify the percentage of BOLD signal change per dollar. Specifically, we focus on the marginal effect of gain in this study. After subject-level processing and quality control using afni_proc.py's QC HTML and gen_ss_review_table.py (Taylor et al. [2024]; see Taylor et al. [2023] for details on criteria applied here) the final dataset included voxel-level contrasts from 47 individuals, each with $V = 222{,}230$ voxels across the brain. In this example, we consider an intercept at the group level without including additional covariates.

We compare our SIMBA model using both the Gibbs sampler and VI algorithm to classical GLM and the BML models. We evaluate voxel-level effect estimates, activation detection of suprathreshold regions, and model fitting to the observed data. The activation threhsold is defined using the same statistical evidence measure E_s introduced in Section 2.4. In this analysis of real fMRI data, we define the Bayesian posterior probability of positive intercept effect as $P^+ = P\{\alpha_0 + \beta_0(\cdot) > 0\}$, which includes the global intercept effect. For the GLM, voxel-wise activation is determined by t-tests with multiple testing adjustment at a false discovery rate of 0.05. The quality of model fit is further assessed using posterior predictive checks (PPCs) [Gelman et al., 1996, Gabry et al., 2019], by comparing predictive and observed data distributions across voxels and participants.

For the Gaussian process prior in our proposed model, we adopt a Matérn kernel with a length scale of 0.1, which was empirically estimated from the pairwise correlations of a subset of spatial locations. To implement the low-rank kernel approximation, we select $J = 1{,}200$ basis functions, chosen based on both the PMSE from LOOCV evaluation and the optimal PPC coverage of the observed data distribution. For the Bayesian models, we run three MCMC chains with different initializations, each for 5,000 iterations with 4,000 burn-in samples.

All computations are performed on a single CPU using an Apple Silicon M3 chip with 18 GB RAM. The SIMBA model using the Gibbs sampler achieves a runtime of approximately 25 seconds per 1,000 iterations, while SIMBA's VI algorithm converges in approximately 5 seconds. This represents a substantial reduction in computational cost compared to the previous BML method (Model (16) in Chen et al. [2019]), which used an HMC-based sampling algorithm and required approximately 30 minutes per 1,000 iterations for four MCMC

chains, using 48 CPUs to analyze 360 brain regions (as opposed to >200,000 voxels, in the SIMBA cases) and 47 participants.

Figure 2.2 compares voxel-level effect estimates and activation region estimation results from the three models. Overlay colors show the effect estimates in units of BOLD percent signal change per dollar. Following the "highlight, don't hide" visualization strategy proposed by Allen et al. [2012] and used in Taylor et al. [2023], we outline suprathreshold regions and display them with full opacity, while subthreshold areas are faded to preserve spatial context rather than being hard-thresholded. Both the GLM and BML models yield scattered selected voxels without spatial regularity, while BML shows decreased effect magnitudes from the global shrinkage through shared prior variance. As shown in Figure 2.2, the GLM approach only detects a few voxels with strong statistical evidence after FDR-based multiple testing adjustment, indicating limited statistical power. In contrast, the proposed SIMBA model obtains structured and spatially continuous patterns, suggesting improved power in detecting task-related effects and better alignment with known brain structures. Importantly, SIMBA does not artificially inflate statistical evidence, since regions with large effect sizes under GLM analysis are also captured by SIMBA, suggesting that the improvement comes from more efficient use of spatial information. This observation further highlights that FDR-based corrections in MUA may be overly conservative, discarding biologically meaningful effects.

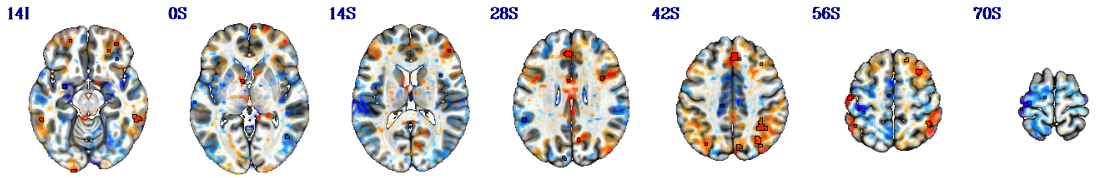
Finally, the PPC curves provide an important quantitative validation of our proposed framework. Figure 2.3 shows 150 posterior draws of predictive data distributions (yellow lines) along with the observed data density (black line) for each method. The closely matching distributions in Figure 2.3c and 2.3d indicate that SIMBA better captures the distribution features of the data than the competing approaches. This shows the ability of the SIMBA model to learn the complex spatial and distributional characteristics of individual-level regression coefficients on fMRI data via the Gaussian process prior.

2.5.2 Ex. 3: Adolescent Brain Cognitive Development Study

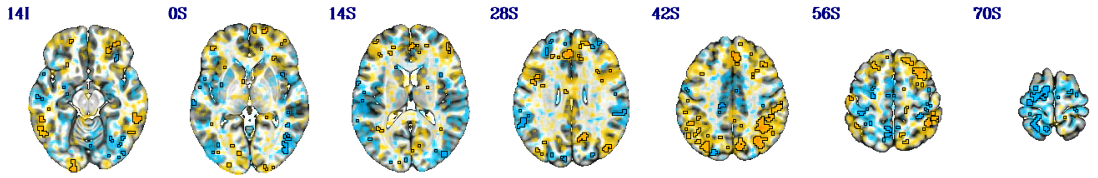
To further evaluate the efficacy of our proposed model, we additionally apply SIMBA to analyze the relationships between cortical gray matter surface fMRI data and clinical covariates from the Adolescent Brain Cognitive Development (ABCD) study in Release 4.0. The ABCD study is a large, multi-site longitudinal initiative aimed at understanding brain development from childhood through adolescence and identifying its biological and environmental determinants [Casey et al., 2018]. In this case, we provide an example of investigating the relationship between cognitive function and brain activity during working memory tasks

Comparison of Model Results in Ex. 2 (NARPS)

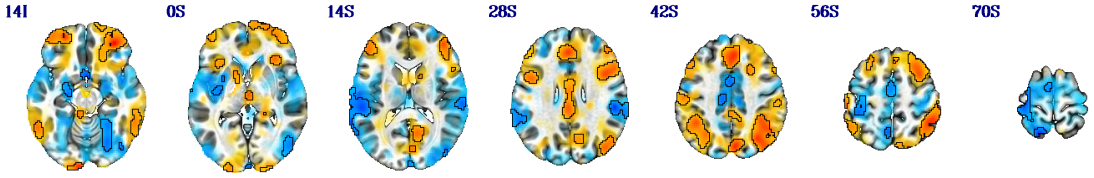
(a) GLM



(b) BML



(c) SIMBA-Gibbs



(d) SIMBA-VI

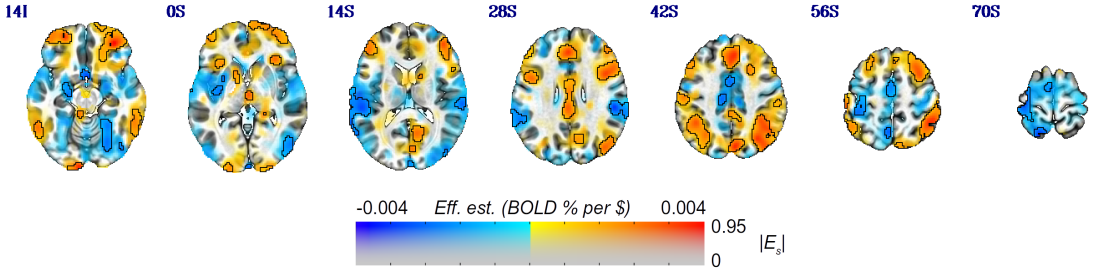
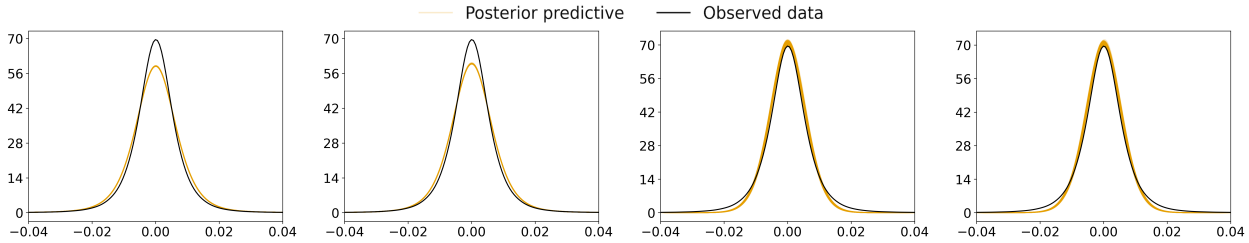


Figure 2.2: Comparison among voxel-level effect estimation in GLM, BML, and our proposed SIMBA on various axial slices of fMRI results. The overlay colors show the effect sizes, and statistical information is used to threshold the data in the "highlight" style [Allen et al., 2012, Taylor et al., 2023]: suprathreshold regions are opaque and outlined, with other regions fading with decreasing statistical value.



(a) GLM (b) BML (c) SIMBA-Gibbs (d) SIMBA-VI

Figure 2.3: Posterior predictive checking curves for Bayesian models for real fMRI data (Ex. 2, NARPS). The yellow lines represent 150 posterior draws of predictive data distribution, and the black line shows the observed data density for all voxels. Having a larger overlap and centering of posterior draws over the observed data fit (black line) suggests better model fit. Here, the SIMBA models have much better posterior check results than the other two models.

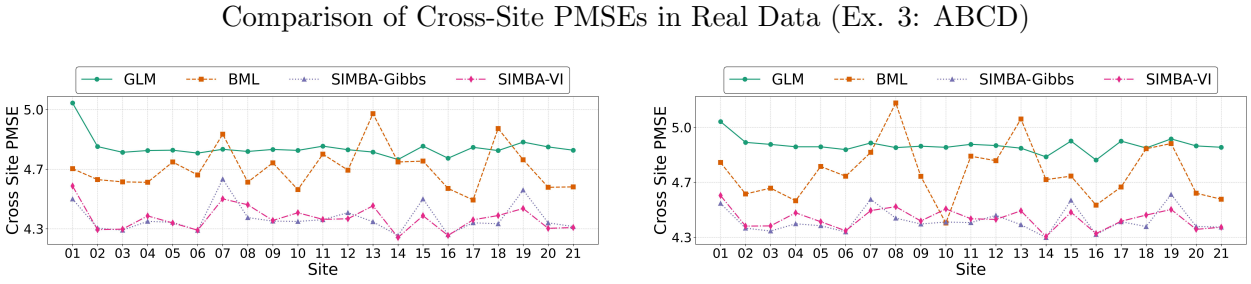
using n-back cortical surface fMRI data. Specifically, we analyze the 2-back versus 0-back contrast maps in z-statistics scale, derived from the fMRI data collected at 21 study sites in the baseline study cohort.

The cortical surface data are stored on two-dimensional meshes, which contain $V = 29{,}696$ and $V = 29{,}716$ spatial locations for the left and right hemispheres, respectively. In the present analysis, we represent these vertices using their three-dimensional geometric coordinates to capture spatial positions in the brain, but we do not explicitly incorporate surface topology or mesh adjacency into the modeling framework (see the Discussion for future extensions with surface datasets). The processed dataset includes 2,497 individuals with complete imaging and covariate information for analysis. Our group-level model specification[1] includes one spatially varying intercept along with $q = 4$ covariates: the general cognitive ability component score [Sripada et al., 2020], which is our primary predictor, and the psychiatric diagnostic score, age, and sex. For BML and SIMBA-Gibbs models, we ran 4,000 burn-in iterations with 1,000 posterior samples in three MCMC chains. The SIMBA model takes approximately 3.5 minutes per 1,000 iterations for Gibbs sampling and 1.4 minutes for the VI algorithm to converge.

In multi-site neuroimaging studies, scanner differences, acquisition protocols, and site-specific demographics can introduce variability that challenges model generalization. To evaluate whether the model produces stable and accurate estimates that can be generalized across study sites, we adopt a cross-site prediction framework. For each site, we fit the model

[1]Note that our primary interest in the present work is testing the model approach with a realistic setup of fMRI data; to address specific questions of detailed brain function and cognitive ability in practice, one would need to perform a more rigorous procedure for model testing, such as through a causal inference framework.

using only its data, then predict imaging outcomes for participants from all other sites. We evaluate the performance by the out-of-site prediction MSEs as an external validation metric. Figure 2.4 presents the cross-site prediction MSEs for both hemispheres among the baseline cohort. Across both hemispheres, the GLM typically shows the highest prediction MSE and the largest variability across sites, indicating limited generalizability. BML mostly performs better than GLM, but it is generally outperformed by our proposed SIMBA methods. Both SIMBA models with Gibbs sampler and VI implementations achieve consistently lower or comparable MSEs in nearly all sites. These results demonstrate that SIMBA better captures the spatial structure to achieve more accurate and stable estimation results and cross-site predictions, highlighting its robustness to site heterogeneity.



(a) Left hemisphere

(b) Right hemisphere

Figure 2.4: Cross-site prediction mean squared errors (MSEs) for single-site analyses on the ABCD cortical surface fMRI data, shown separately for the left hemisphere (left) and right hemisphere (right).

In Figure 2.5, we visualize the estimated associations between brain activity and cognitive function across both hemispheres using Nilearn [Nilearn contributors, 2025]. Results are again thresholded at $|E_s| > 0.95$. In this case, the extent of positive and negative regions are broadly similar across models, but detailed features differ. Compared with other models, SIMBA produces more spatially contiguous activation patterns with smooth boundaries and fewer very small islands, which makes them more interpretable. Notably, positive associations with the cognitive scores are observed in the dorsal attention network, fronto-parietal network, and salience network, which are the brain regions well known for supporting attentional control, cognitive flexibility, and integration of task-relevant information that are also align with established findings [Gordon et al., 2016, Menon, 2011]. Negative associations are identified in the cingulo-opercular, default mode, and parieto-occipital networks, consistent with suppression of task-irrelevant processes during working memory tasks.

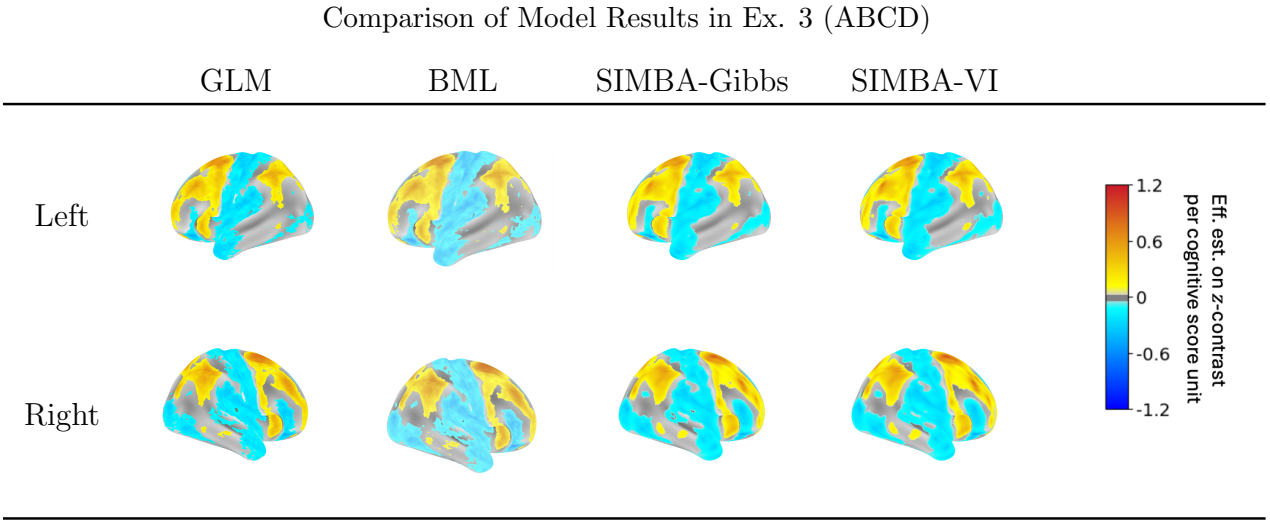


Figure 2.5: Estimated associations and selected activation regions between fMRI data and cognitive scores in the ABCD study for the left and right hemispheres. The values represent the expected change in the 2-back versus 0-back contrast (expressed as z-statistics) for a one-unit increase in cognitive score, while holding all other demographic information.

2.6 Discussion and Conclusion

In this work, we propose SIMBA, a fully Bayesian spatial hierarchical model for voxel-wise whole brain analysis of task-based fMRI data. Our framework models population-level covariate effects and individual-specific deviations using Gaussian process priors, capturing spatially smooth and structured activation patterns across the brain. To mitigate the computational challenges of GP-based modeling for high-dimensional neuroimaging data, we utilize low-rank kernel approximation and a reparameterized representation that enables scalable posterior inference. We develop both a Gibbs sampling algorithm and an alternative based on variational inference, demonstrating accurate posterior estimation and significant gains in computational efficiency. Comparing these two implementations of SIMBA, the latter shows significantly reduced computational runtime, with similar and in several cases slightly improved model performance.

Through simulation studies, we show that the proposed SIMBA model outperforms classical voxel-wise GLM (i.e., a MUA) with standard FDR adjustment and BML models with information pooling in both estimation accuracy and signal detection power. This was quantitatively assessed using key relevant metrics, including both true positive and false discovery rates. Importantly, SIMBA typically had the leading counts for both simultaneously. SIMBA has the advantages of pooling spatial information at multiple scales, and unlike standard MUA-based approaches incorporates effect estimates (rather than being limited to

solely adjusting statistics). Our model appears to more accurately reconstruct true spatial effects while also being more robust to the presence of noise.

Applications to real task-based fMRI datasets from the NARPS and ABCD studies further highlight the efficacy of SIMBA. Across both whole brain and cortical data, SIMBA identifies interpretable activation patterns that align with established findings, while also achieving improved model fitting that more align with true data distribution examined by PPC curves and predictive performance in out-of-sample evaluation metrics. These results demonstrate that SIMBA offers an accurate and computationally efficient approach for large-scale spatial modeling in neuroimaging, providing a powerful tool for detecting, quantifying, and interpreting brain activities across the whole brain volume.

One limitation of our proposed method is that the estimated effects are highly dependent on the choice of kernel functions. Although the Matérn kernel is widely used for modeling activation functions, a parametric kernel function could still limit its flexibility. Exploring more flexible or nonparametric kernel representations could potentially improve the ability to capture complex, spatially varying functions in brain activity. Additionally, our use of low-rank kernel approximation is critical for computational scalability but inevitably introduces approximation error. While cross-validation helps in selecting an appropriate number of basis functions, an over-truncation may lead to underfitting, whereas including more bases increases computation demands or overfitting. In addition, for computational purposes, our model reparametrization requires the assumption that the main spatially varying effects share the same kernel structure with the spatial noise. This may limit the ability to capture different spatial dependence structures that could exist across these spatial components. Extending the model to incorporate more flexible or adaptive kernels would enhance its expressiveness but would also lead to additional computational costs.

Another limitation arises in our current treatment of surface-based cortical gray matter fMRI data. Although the method is applied to data stored as meshes, we currently rely on geometric distances between vertices rather than explicitly modeling surface topology. This choice does not fully explore the rich structural information of surface-based data. Future work will aim to extend the framework to explicitly incorporate cortical surface topology into spatial dependencies, enabling more accurate modeling of brain activity on cortical manifolds.

There are several potential extensions of this work in both methodology and application that are worth exploring. Integrating SIMBA into the AFNI software suite would greatly increase its accessibility and practical utility for the neuroimaging community for a scalable Bayesian image regression pipeline. Work is currently underway for developing a command line tool for such distribution. Methodologically, one direction is to extend the SIMBA

framework to incorporate within-individual or repeated-measures variables, enabling richer modeling of longitudinal and task-based designs. Another is to account for measurement error in the response variable, for example by incorporating voxel-wise standard errors from subject-level GLMs directly into the model. Finally, SIMBA could be adapted for voxel-level test–retest reliability estimation, providing a principled Bayesian approach to quantifying reproducibility in fMRI studies.

CHAPTER 3

Deep Kernel Learning Based Gaussian Processes for Bayesian Image Regression Analysis

3.1 Introduction

Statistical inference in neuroimaging data plays a critical role in medical research and clinical applications. Common imaging modalities, such as sMRI, fMRI, and EEG, often involve complex spatial and temporal relationships. Nonparametric regression models for functional data analysis [Ramsay and Silverman, 2005] provide a flexible framework for these challenges. Regression models in neuroimaging can be categorized as image-on-scalar, scalar-on-image, or image-on-image regression, each addressing different research objectives by assigning images and clinical covariates as predictors or responses.

Image-on-scalar regression, such as analyzing fMRI activation patterns, faces challenges from both spatial and inter-subject heterogeneity. Brain regions often exhibit varying relationships with outcomes, requiring flexible modeling to capture localized effects. Additionally, inter-subject variability in anatomy and functional organization complicates the pooling of data across participants. High dimensionality and complex noise structures, including physiological artifacts, further complicate analysis, demanding methods that balance interpretability, robustness, and computational feasibility.

Scalar-on-image regression, as seen in EEG-based brain-computer interface studies, involves modeling dynamic neuroimaging features to predict scalar outcomes. EEG's non-stationary, noisy signals necessitate extensive preprocessing and careful handling of inter-channel correlations and temporal dependencies. These challenges highlight the need for scalable, flexible, and interpretable regression approaches to address the complexities of neuroimaging data.

In this article, we introduce a data-driven deep kernel learning process (DKLP) that en-

riches the Gaussian Process (GP) priors in Bayesian nonparametric inferences across diverse image regression models. Leveraging deep neural networks (DNNs) to flexibly learn kernel functions from data, DKLP circumvents the need for traditional kernel pre-specification and tuning. Our kernel approximation technique is inspired by Ma et al. [2023], who propose a multi-stage deep kernel estimation method for image-on-image regression models. However, we aim to create a unified prior model rather than estimating parameters by a multi-stage nature. Moreover, our method is generally adaptable to any regression model that integrates GP priors. In addition to kernel learning, we aim for more accurate posterior inferences on the unknown function of interest, facilitated by the kernel estimated from DNNs. This approach circumvents the need for direct interpretation of the DNN while allowing for posterior inference on latent functions.

Our work is motivated by addressing the limitations of existing GP-based nonparametric methods in capturing the heterogeneity of neuroimaging data. GP is a widely studied Bayesian nonparametric approach for modeling unknown regression functions. In imaging studies, GP priors are commonly used to infer posterior distributions on real-valued unknown functions, such as imaging signal intensities or spatially varying coefficients. Most studies assume zero-mean GP priors, whose performance is heavily influenced by the choice of covariance kernel function. Traditional GP models rely on fixed parametric kernels, such as the radial basis function (RBF) or Matérn kernels, which have limited flexibility for capturing the complex and heterogeneous nature of imaging data. Furthermore, generalizing a specific kernel across datasets with diverse spatial patterns often requires extensive tuning in hyperparameters and involves high computational costs.

DNNs, on the other hand, have proven to be powerful machine learning tools for modeling complex and non-linear functions. Many studies have demonstrated the superiority of DNNs in medical imaging analysis tasks, including pattern recognition, feature extraction, and prediction [Shen et al., 2017, Lee et al., 2017, Zhou et al., 2021]. Recent research has also emphasized bridging the gap between theoretical foundations and practical applications of DNNs as a nonparametric approach. The universal approximation theorem [Scarselli and Tsoi, 1998, Hornik et al., 1989] guarantees that DNNs with sufficient network depth and complexity can approximate any continuous function with arbitrary precision. This theoretical property supports the successful applications of DNNs in learning complex spatial or temporal relationships in neuroimaging data. Furthermore, several Bayesian nonparametric neural network models have been established and their theoretical properties, including posterior consistency and concentrate rate, have been studied by Lee [2000] for single-layer neural networks and Polson and Ročková [2018] for sparse DNNs with spike-and-slab priors.

In recent years, a variety of advanced GP-based Bayesian nonparametric models have been

proposed to solve image regression tasks [Kang et al., 2018, Zeng et al., 2022, Guo et al., 2022, Xu and Kang, 2023]. However, these methods rely on parametric kernel functions, which can limit their flexibility in capturing the complexity of image data and reduce their applicability to diverse datasets. In the frequentist framework, Zhang et al. [2023a] introduces a deep learning-based image-on-scalar regression model to estimate spatially varying functions by DNNs. While this method demonstrates the strengths of DNNs in learning complicated image regression functions, it falls short in model inference and could be subject to overfitting. In particular, all these methods are restricted to either image-on-scalar or scalar-on-image frameworks, requiring substantial modifications for applications across different types of image regression models.

Different kernel learning techniques have been proposed to develop more generally applicable and flexible kernel functions. One major approach considers multiple kernel learning, such as additive and multiplicative kernel composition [Duvenaud et al., 2013, Gönen and Alpaydın, 2011, Williams and Rasmussen, 2006]. Wilson et al. [2016] and Damianou and Lawrence [2013] integrate kernel composition with DNNs under a probabilistic framework. However, these methods still require choosing parametric kernel functions. Another direction focuses on nonparametric kernel learning techniques, including extensions of spectral mixture kernels [Jang et al., 2017, Oliva et al., 2016, Benton et al., 2019], and low-rank decomposition methods to derive a positive semi-definite covariance kernel matrix [Zhu et al., 2004, Kulis et al., 2006, Zhuang et al., 2011]. Although these methods mark improvements over traditional parametric approaches, they often require carefully specified prior distributions and involve considerable tuning and manual input. Moreover, many of these methods face high computational costs and have limited applicability to large-scale imaging data or diverse image regression models.

By contrast, our proposed DKLP framework avoids explicit kernel pre-specification and excessive tuning while retaining the interpretability and rigorous inference benefits of a Bayesian GP model. Our unified approach seamlessly applies to various image regression models (image-on-scalar, scalar-on-image, and image-on-image), and the deep kernel structure learned by the DNN serves as a flexible, data-driven covariance function. Additionally, we develop a scalable posterior update algorithm, which adopts a hybrid Markov chain Monte Carlo (MCMC) based algorithm, using stochastic gradient Langevin dynamics (SGLD) and Gibbs sampling for computational efficiency. We perform theoretical analysis of DKLP by establishing the posterior concentration rates for regression and kernel function estimation. We evaluate the performance of DKLP on image regression models through simulation studies and applications on neuroimaging data.

The remainder of the paper is organized as follows. In Section 3.2, we introduce our

kernel learning process and its application to three regression models. Section 3.3 establishes the theoretical properties of DKLP, focusing on posterior concentration rates. Section 3.4 introduces an efficient and scalable posterior computation algorithm. Sections 3.5 and 3.6 show the empirical performance of our method in various regression frameworks through simulations and brain imaging applications. Finally, we conclude our paper in Section 3.7.

3.2 Methods

3.2.1 Kernel learning via deep neural network

Suppose that we collect d-dimensional imaging data at V spatial locations from a compact space $\mathcal{S} \subset \mathbb{R}^d$. Let $\mathbf{s} = \{s_v\}_{v=1}^V$ be the collection of spatial location, where $s_v \in \mathcal{S}, v = 1, \ldots, V$. The true image value at each location s_v is generated from an unknown function $w : \mathbb{R}^d \to \mathbb{R}$. We introduce a prior probability on $w(s_v)$ by a GP prior with mean zero and unknown covariance kernel $K(\cdot, \cdot)$, denoted as $w(s_v) \sim \mathcal{GP}(0, K)$. The covariance kernel $K : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a continuous positive semi-definite function. By Mercer's theorem, the kernel function can be expressed as an infinite series: $K(s_v, s_{v'}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s_v)\phi_j(s_{v'})$, where $\{\phi_j\}_{j=1}^{\infty}$ is a set of orthonormal eigenfunctions associated with non-negative eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$. Here, $\{\phi_j\}_{j=1}^{\infty}$ represent the kernel basis functions in $L^2(\mathcal{S})$ and $\{\lambda_j\}_{j=1}^{\infty}$ are absolutely summable and in decreasing order, i.e., $\sum_{j=1}^{\infty} \lambda_j < \infty$ and $\lambda_1 > \cdots > \lambda_\infty \geq 0$. We can further apply the Karhunen-Loève (KL) expansion and decompose the function as $w(s_v) = \sum_{j=1}^{\infty} \phi_j(s_v)\theta_j$, with each basis coefficient θ_j following a normal distribution with mean zero and variance λ_j, denoted as $\theta_j \sim \mathcal{N}(0, \lambda_j)$. We denote a truncated kernel as $K_J(s_v, s_{v'}) = \sum_{j=1}^{J} \lambda_j \phi_j(s_v)\phi_j(s_{v'})$ along with the truncated function as $w_J(s_v) = \sum_{j=1}^{J} \phi_j(s_v)\theta_j$ by some truncation number $J \in \mathbb{N}^+ \cup \infty$.

We propose to approximate the kernel function by separately estimating eigenvalues and eigenfunctions from the data. The kernel series expansion indicates that the eigenvalues can be incorporated in Bayesian modeling as the variances of the basis coefficients. We assign the square root of each eigenvalue with a half-Cauchy prior, denoted as $\sqrt{\lambda_j} \sim C^+(A)$, for some $A > 0$. For basis functions, we borrow the approximation power from DNNs to ensure high flexibility in estimated kernels. We consider deep feed-forward DNNs with $L \in \mathbb{N}^+$ hidden layers and a vector $\mathbf{h} = (h_0, \ldots, h_{L+1}) \in \mathbb{N}^{L+2}$ representing the number of neurons in the input, hidden, and output layers. For simplicity, we assume that all hidden layers have an equal number of neurons $h = h_1 = \cdots = h_L$. Then, a DNN $f^{\text{DNN}} : \mathbb{R}^{h_0} \to \mathbb{R}^{h_{L+1}}$ is any function of the form: $f_{\mathbf{B}}^{\text{DNN}}(\mathbf{x}) = \mathbf{W}_{L+1}\sigma_{\mathbf{b}_L}(\mathbf{W}_l\sigma_{\mathbf{b}_{L-1}} \ldots \sigma_{\mathbf{b}_1}(\mathbf{W}_1\mathbf{x}))$, characterizing by the set of parameters $\mathbf{B} = \{\mathbf{W}_1, \mathbf{b}_1, \ldots, \mathbf{b}_L, \mathbf{W}_{L+1}\}$, and the element-wise rectified linear unit

(ReLU) function $\sigma_{\mathbf{b}} : \mathbb{R}^h \to \mathbb{R}^h$ as the between-layer activation function. We denote the total number of DNN parameters $|\mathbf{B}| = h_{\mathbf{B}}$. Here, $\mathbf{W}_l \in \mathbb{R}^{h_l \times h_{l-1}}$ are weight matrices connecting $(l-1)^{th}$ to l^{th} layer and $\mathbf{b}_L \in \mathbb{R}^{h_l}$ are bias vectors in each layer. We assign independent normal priors with mean zero and unknown variance σ_f^2, i.e., $\mathbf{B} \sim \mathcal{N}(0, I_{h_{\mathbf{B}}}\sigma_f^2)$. To represent J kernel basis functions simultaneously, we configure the DNN with $h_0 = d$ and $h_{L+1} = J$. We define the class of J-dimensional uniformly bounded DNN functions with bounded parameters as $\mathcal{F}^{\text{DNN}}(L, \mathbf{h}) := \left\{ f : f = f_{\mathbf{B}}^{\text{DNN}} \text{ with } (L, (d, h, \ldots, h, J)^\top) \text{ architecture, } |\mathbf{B}|_\infty \leq \infty \right\}$. Since DNN outputs are not inherently orthonormal, we apply a differentiable normalizing orthogonalization transformation operator $\mathcal{T} : \mathbb{R}^J \to \mathbb{R}^J$ that transforms a set of linearly independent vectors into a set of orthonormal vectors spanning the same subspace. In practice, the operator $\mathcal{T}$ can be realized through the Gram-Schmidt process or singular value decomposition.

Definition 1. *Let $J \in \mathbb{N}^+ \cup \{\infty\}$. Suppose there exists a sequence of nonnegative, non-increasing values $\{\lambda_j\}_{j=1}^J$ with prior probabilities $\sqrt{\lambda_j} \overset{i.i.d.}{\sim} C^+(A)$ for some scale parameter $A > 0$. Let $f_{\mathbf{B}}^{\text{DNN}} \in \mathcal{F}^{\text{DNN}}(L, h, J)$ be a J-dimensional output DNN function with parameter $\mathbf{B} \sim \mathcal{N}(0, I_{h_{\mathbf{B}}}\sigma_f^2)$. Define $\{\phi_j\}_{j=1}^J = \mathcal{T}(f_{\mathbf{B}}^{\text{DNN}})$, where $\mathcal{T} : \mathbb{R}^J \to \mathbb{R}^J$ is an orthonormal transformation operator. Then the kernel function $K(s_v, s_{v'}) = \sum_{j=1}^J \lambda_j \, \phi_j(s_v) \, \phi_j(s_{v'})$ is said to follow a deep kernel learning prior (DKLP), denoted $K \sim \mathcal{DKLP}(f_{\mathbf{B}}^{\text{DNN}}, \sigma_f^2, A)$.*

3.2.2 Image regression models

We proceed by integrating the DKLP prior into image regression models through three model specifications. We emphasize the adaptability of our kernel learning process across various nonparametric regression models with GP priors under different study contexts. Suppose that we collect n image data at spatial location $\{s_v\}_{v=1}^V$ for each individual $i = 1, \ldots, n$. Let $y_i(s_v) \in \mathbb{R}$ denote the observed noisy image intensity at location s_v for individual i.

Nonparametric regression (NPR) model is formulated as:

$$y_i(s_v) = w_i(s_v) + \epsilon_i(s_v), \quad \epsilon_i(s_v) \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{3.1}$$

where $w_i(s_v)$ stands for an individual-specific spatially varying image function. The spatial Gaussian noise $\epsilon_i(s_v)$ is independent across locations and individuals with mean zero and unknown variance σ_ϵ^2.

Image-on-scalar regression aims to study the associations between image outcomes and covariates. Suppose we further collect q covariates: $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top \in \mathbb{R}^q$. We consider

the following model specification:

$$y_i(s_v) = \sum_{r=1}^{q} x_{ir}\beta_r(s_v) + \eta_i(s_v) + \epsilon_i(s_v), \quad \epsilon_i(s_v) \sim \mathcal{N}(0, \sigma_y^2). \tag{3.2}$$

Together with previously defined parameters, $\beta_r(s_v) \in \mathbb{R}$ represents the main spatially varying effect of the covariate x_{ir} at location s_v. In addition, we include $\eta_i(s_v)$ to account for individual-specific deviations. We assume the covariates matrix $\mathbf{x}_i$ is well-conditioned and impose the constraint $\sum_{i=1}^{n} \eta_i(s_v) = 0$ to ensure identifiability.

Scalar-on-image regression takes images as V-dimensional spatial predictors to model and predict a scalar outcome $u_i \in \mathbb{R}$. Similarly to image-on-scalar regression, we collect q confounding variables denoted as $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})$. The scalar-on-image regression model is defined by:

$$u_i = \frac{1}{V}\sum_{v=1}^{V} y_i(s_v)\alpha(s_v) + \sum_{r=1}^{q} x_{ir}\xi_r + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_u^2), \tag{3.3}$$

where $\alpha(s_v)$ is the spatially varying coefficient for each image location. Let ξ_r represent the effect of the r-th confounder. We normalize the total image effect by V, preventing an unbounded total effect as imaging resolution increases. In SIR, $\alpha(s_v)$ is typically treated as a single spatial realization, making it infeasible to learn its covariance kernel structure. Therefore, we jointly model the latent image functions $w_i(s_v)$ through the NPR framework in (3.1). We assume that $\alpha(s_v)$ and the underlying image function $w_i(s_v)$ share the same spatial dependencies. Under this assumption, the kernel parameters are jointly learned from the likelihoods of both models, allowing the spatial information inferred from the observed images to inform the estimation of $\alpha(s_v)$.

3.2.3 Prior specification and model representation

For latent individual-level spatially varying regression functions from models above, we assign independent GP priors along with a DKLP prior on the unknown kernel function as $w_i(s_v), \eta_i(s_v) \sim \mathcal{GP}(0, K)$ and $K \sim \mathcal{DKLP}(f_{\mathbf{B}}^{\text{DNN}}, \sigma_f^2, A)$ as in Definition 1. The basis functions $\{\phi_j\}_{j=1}^{J}$, orthogonalized from the DNN outputs, inherit prior probabilities from DNN parameters.

In both ISR and SIR regression, we often assume the spatially varying components are sparse. To impose sparsity in regression framework, we define the covariate effects in ISR and image predictors in SIR by $\beta_r(s_v) = \delta_r(s_v)\tilde{\beta}_r(s_v)$ and $\alpha(s_v) = \delta_\alpha(s_v)\tilde{\alpha}(s_v)$, where $\delta(s_v)$ represents a binary spatial selection indicator that takes values in $\{0, 1\}$ at each location. For

covariate effects, we assign the same GP priors $\tilde{\beta}_r(s_v), \tilde{\alpha}(s_v) \sim \mathcal{GP}(0, K)$ where the unknown kernel function follow a DKLP prior. For the selection variable, we employ independent Bernoulli priors for each selection parameter and incorporate the spatial information learned from the kernel functions, denoted as $\delta(s_v) \sim \text{Ber}\{\rho(s_v)\}$, where $\text{logit}\{\rho(s_v)\} \sim \mathcal{GP}(0, K)$. This approach can be interpreted as adding one additional classification layer to the original DNN output.

We represent nonparametric regression as $y_i(s_v) = \sum_{j=1}^{J} \phi_j(s_v)\theta_{ij} + \epsilon_i(s_v)$, image-on-scalar regression as $y_i(s_v) = \sum_{r=1}^{q} x_{ir} \sum_{j=1}^{J} \phi_j(s_v)\theta_{\beta_{rj}}\delta_j(s_v) + \sum_{j=1}^{J} \phi_j(s_v)\theta_{\eta_{ij}} + \epsilon_i(s_v)$, and scalar-on-image regression as $u_i = V^{-1} \sum_{v=1}^{V} y_i(s_v) \left(\sum_{j=1}^{J} \phi_j(s_v)\theta_{\alpha_j}\right) \delta_\alpha(s_v) + \sum_{r=1}^{q} \xi_r x_{ir} + \epsilon_i$. Similarly, we write the logit function of the selection probability as $\text{logit}\{\rho(s_v)\} = \sum_{j=1}^{J} \phi_j(s_v)\theta_{\rho_j}$. Each basis coefficient follows an independent normal distribution $\mathcal{N}(0, \lambda_j)$, where $\sqrt{\lambda_j} \sim C^+(A)$ with some hyperparameter $A > 0$. We impose the prior for each confounder effect in the scalar-on-image model as $\xi_r \sim \mathcal{N}(0, \sigma_\xi^2)$. For all other variance parameters, the inverse gamma (IG) priors are imposed: $\sigma_\epsilon^2 \sim \text{IG}(a_\epsilon, b_\epsilon)$, $\sigma_f^2 \sim \text{IG}(a_f, b_f)$, and $\sigma_\alpha^2 \sim \text{IG}(a_\alpha, b_\alpha)$. We set $A = 100$ and $a_\epsilon, b_\epsilon, a_f, b_f, a_\alpha, b_\alpha = 0.001$ to induce weakly informative priors.

3.3 Theory

In this section, we establish the posterior convergence rates for regression and kernel function estimation using DKLP prior within regression models. Denote $a_n \lesssim b_n$ for inequality up to a constant and $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. For any function $f : \mathcal{S} \to \mathbb{R}$ on the spatial grid $\{s_v\}_{v=1}^{V} \subset \mathcal{S}$, we define the empirical L^2-norm as $\|f\|_V^2 = (V)^{-1} \sum_{v=1}^{V} f(s_v)^2$. Denote $\mathcal{H}_d^\beta$ as the β-Hölder smooth function in $L_2(\mathcal{S})$. For a kernel function K, we define $\|K\|_{VV}^2 = (V)^{-2} \sum_{v=1}^{V} \sum_{v'=1}^{V} K(s_v, s_{v'})^2$. Let $\|f\|_\infty$ be the supremum norm of the function f and $|S|$ denote the cardinality of a set S.

We establish posterior contraction rates for the three image regression models in (3.1)–(3.3). Throughout this section, we assume that the spatial locations $\mathcal{S} \subset [0, 1]^d$ are fixed grid points from a compact set and shared across all individuals. Without loss of generality, we assume that the residual noise variance $\sigma_\epsilon^2, \sigma_y^2, \sigma_u^2$ and, for the SIR model, the scalar effect $\{\xi_r\}_{r=1}^{q}$ are known and fixed. As our primary interest is on the asymptotic behavior of the posterior over the infinite-dimensional spatial functions, posterior contraction rates are unaffected by these scalar parameters with standard priors.

We begin by defining the functional spaces that characterize the true spatially varying functions and the admissible kernels under the DKLP prior.

Assumption 1 (True function space)**.** *We assume that the true underlying function belongs*

to the class of functions Θ*, which contains functions* $f \in L^2(\mathcal{S})$ *that can be represented as:* $f(s_v) = \sum_{j=1}^{\infty} \theta_j \lambda_j^{1/2} \phi_j(s_v)$*, where* $\{\phi_j\}_{j=1}^{\infty}$ *are orthonormal basis functions and* $\{\lambda_j\}_{j=1}^{\infty}$ *are non-negative eigenvalues satisfying the following conditions: (a)(Smoothness)* $\phi_j \in \mathcal{H}_d^{\beta}$ *for some* $\beta > 0$*, and* $|\phi_j|_{\infty} \leq C_{11}$ *for constant* $C_{11} > 0$*; (b)(Eigenvalue decay)* $\lambda_j \asymp j^{-2\alpha/d-1}$ *for some* $\alpha > 0$*; (c)(Coefficient control)* $|\theta|_{\infty} \leq C_{12}$ *and* $\sum_{j>J} \theta_j^2 \leq J^{-C_{13}}$ *for some constant* $C_{12}, C_{13} > 0$*.*

In addition, we define a space of valid kernel functions that can be constructed from the basis functions and eigenvalues above as

Definition 2 (Kernel space)**.** *Let* Θ_K *denote the class of positive semi-definite kernels* $K : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ *such that* $K(s_v, s_v') = \sum_{j=1}^{\infty} \phi_j(s_v)\phi_j(s_v')\lambda_j.$*, where* $\{\phi_j\}_{j=1}^{\infty}$ *and* $\{\lambda_j\}_{j=1}^{\infty}$ *satisfy the conditions in Assumption 1.*

Assumption 1 restricts the smoothness of the true spatially varying functions by the eigenvalue decay rate and the Hölder smoothness of the eigenfunctions. In this work, we focus on kernels with polynomial eigenvalue decay, which provides greater flexibility for modeling complex spatial correlations. These kernels distribute information more evenly across the spectrum compared to those with exponentially decaying eigenvalues, making them better suited for high-dimensional and heterogeneous data. For example, Matérn kernels with length-scale $\nu > 1/2$ are kernels with eigenvalues that decay in polynomial order, with a rate of $j^{-2\nu/d-1}$ [Santin and Schaback, 2016]. It is possible to consider other types of eigenvalue decaying conditions similarly as in Sections 4.4 and 4.5 in van der Vaart and van Zanten [2008].

Proposition 1. *Let* $f(s_v) = \sum_{j=1}^{\infty} \theta_j \lambda_j^{1/2} \phi_j(s_v)$ *and* $f'(s_v) = \sum_{j=1}^{\infty} \theta_j' \lambda_j^{1/2} \phi_j'(s_v)$ *be two series representations in the function space* Θ *satisfying Assumption 1 such that:* $f(s_v) = f'(s_v)$ *for all* $s \in \mathcal{S}$*. Then, there exist sign constants* $\nu_j \in \{\pm 1\}$ *such that* $\phi_j' = \nu_j \phi_j, \theta_j' = \nu_j \theta_j$*. That is, the decomposition is identifiable up to element-wise sign changes.*

To approximate the basis functions $\{\phi_j\}$, we consider the DNN function $f_{\mathbf{B}}^{\text{DNN}} \in \mathcal{F}_{\mathbf{B}}^{\text{DNN}}$ with the $(L_V, \mathbf{h}_V)$ architecture DNN from Definition 1. Following Theorem 1 of Kong and Kim [2024], we specify L_V, h_V by

$$L_V := \lceil C_{21} \log V \rceil, \quad h_V := \lceil C_{22} V^{\frac{d}{2(2\beta+d)}} \rceil \tag{3.4}$$

for constant $C_{21}, C_{22} > 0$.

Let the observed data be denoted by $\mathcal{D}^{(V,n)}$, which includes imaging observations and, when applicable, associated clinical covariates or outcomes, collected according to the rel-

evant model specifications. We begin with the NPR model (3.1) to establish its posterior contraction rate under the DKLP prior.

Theorem 1. *Assume data are generated from Model* (3.1) *and the true spatially varying functions* $\{w_{i,0}\}_{i=1}^n \subset \Theta$ *that satisfies Assumption 1. Consider zero-mean GP priors where* $w_1, \ldots, w_n \overset{i.i.d}{\sim} \mathcal{GP}(0, K)$ *with a kernel prior* $K \sim \mathcal{DKLP}(f_{\mathbf{B}}^{\mathrm{DNN}}, \sigma_f^2, A)$ *as in Definition 1 with a* $(L_V, \mathbf{h}_V)$ *architecture in* (3.4) *and* $J \gtrsim \epsilon_{Vn}^{-d/(\alpha\wedge\beta)}$. *Then, the posterior distribution concentrates around truth at the rate* $\epsilon_{Vn} = \max\left(V^{\frac{-2\beta^2}{(2\beta+d)^2}}\log^\gamma V, n^{-\frac{\beta}{2\beta+d}}\right)$ *for* $\gamma > 2$, *in the sense that*

$$\Pi\left\{(w_i)_{1\le i\le n} : n^{-1}\sum_{i=1}^n \|w_i - w_{i,0}\|_V > M\epsilon_{Vn} \mid \mathcal{D}^{(V,n)}\right\} \to 0$$

for any $M > 0$ *as* $V \to \infty, n \to \infty$ *under* P_0, *which is the probability measure of the observed data.*

Next, we establish posterior contraction rates for the ISR and SIR models, given in Model (3.2) and (3.3), respectively. In neuroimaging applications, spatially varying coefficients, such as activation maps or voxel-level associations, are often sparse and piecewise smooth. Therefore, we assume true spatial varying functions β_0 in the ISR model (3.2) and α_0 in the SIR model (3.3) belong to a sparse function class Θ_{SP} defined below, which is adopted from the definitions in [Xu and Kang, 2023, Kang et al., 2018]. To capture such sparsity through a location-wise selection indicator, we introduce an auxiliary selection function $\zeta(s_v)$ that determines the sparsity via a probabilistic thresholding function.

Assumption 2 (Selection functional space)**.** *Let* $\Theta_{\mathrm{SP}} \subset L^2(\mathcal{S})$ *denote the collection of functions that* $\Theta_{\mathrm{SP}} = \{f : f(s_v) = \tilde{f}(s_v)\,\mathbf{1}\{\zeta(s_v) \ge 0\},\ (\tilde{f}, \zeta) \in \Theta\}$ *for an active set* $R_1 = \{s \in \mathcal{S} : \zeta(s_v) \ge 0\}$ *with nonempty interior and* $R_0 := \mathcal{S} \setminus R_1$. *Then, we assume the following holds:*

(a) There exists some $t > 0$ *and* $C_{31} > 0$ *such that for* $B_t = \{s \in \mathcal{S} : |\zeta(s_v)| \le t\}$, *we have* $|B_t| \lesssim tV$, *for some* $t > 0$;

(b) Let $\rho(s_v) = \mathrm{logit}^{-1}\{\zeta(s_v)\}$. *For some sequence* $\varepsilon_V \to 0$ *as* $V \to \infty$, *we have* $\sup_{s\in R_1\setminus B_t}\big(1 - \rho(s_v)\big) \le \varepsilon_V$, *and* $\sup_{s\in R_0\setminus B_t}\rho(s_v) \le \varepsilon_V$;

(c) There exists constant $C_{32} > 0$ *such that* $\sum_{s\in R_0}\tilde{f}^2(s_v) \le C_{32}|B_t|$.

Assumption 2 provides a flexible and probabilistically smooth representation of spatial sparsity by decomposing a sparse function into a smooth coefficient function $\tilde{f}(s_v)$ and a smooth selection function $\zeta(s_v)$, which defines the active region. The assumptions ensure

that the transition between active and inactive regions becomes sharp as V grows, with minimal ambiguity and negligible signal outside the support. This formulation enables efficient modeling of sparse, piecewise smooth signals and facilitates theoretical analysis under our prior specifications. Similarly to Theorem 1, we present the posterior concentration rates for the spatially varying coefficients in the ISR and SIR models, respectively.

Theorem 2. *Assume data are generated from Model* (3.2) *and the true spatially varying functions* $\{\eta_{i,0}\}_{i=1}^n \subset \Theta$ *and* $\beta_0 \subset \Theta_{\text{SP}}$ *that satisfies Assumption 1 and 2. Consider the same GP priors on each regression function. Then, the posterior distributions concentrate around the truth at the rate* $\epsilon_{Vn} = \max\left(V^{\frac{-2\beta^2}{(2\beta+d)^2}}\log^\gamma V, n^{-\frac{\beta}{2\beta+d}}\right)$ *for* $\gamma > 2$*, in the sense that*

$$\Pi\left\{(\eta_i, \beta_r)_{i=1,r=1}^{n,q} : n^{-1}\sum_{i=1}^n \|\eta_i - \eta_{i,0}\|_V + \sum_{r=1}^q \|\beta_{0r} - \beta_r\|_V > M\epsilon_{Vn} \mid \mathcal{D}^{(V,n)}\right\} \to 0$$

for any $M > 0$ *as* $V \to \infty, n \to \infty$ *under* P_0*, which is the probability measure of the observed data.*

Theorem 3. *Assume the scalar outcome* u_i *is generated from Model* (3.3) *given image data generated under the conditions in Theorem 1. Suppose the true spatially varying functions* $\alpha_0 \subset \Theta_{\text{SP}}$ *that satisfies Assumption 2. Consider the same GP priors on each regression function. Then, the posterior distributions concentrate around the truth at the rate* $\epsilon_{Vn} = \max\left(V^{\frac{-2\beta^2}{(2\beta+d)^2}}\log^\gamma V, n^{-\frac{\beta}{2\beta+d}}\right)$ *for* $\gamma > 2$*, in the sense that*

$$\Pi\left\{(w_i, \alpha)_{1\le i\le n} : n^{-1}\sum_{i=1}^n \|w_i - w_{i,0}\|_V + \|\alpha_0 - \alpha\|_V > M\epsilon_{Vn} \mid \mathcal{D}^{(V,n)}\right\} \to 0$$

for any $M > 0$ *as* $V \to \infty, n \to \infty$ *under* P_0*, which is the probability measure of the observed data.*

Similarly to the general posterior concentration results in Ghosal and van der Vaart [2007], the contraction rate in the above theorems is derived by verifying the three conditions, including prior mass, entropy, and the size of the constructed sieve. The proof is given in the supplementary materials. This result provides a theoretical justification for the excellent numerical performance of the regression function estimation, especially when the number of observed spatial locations and the number of collected sample functions are large. The posterior convergence rate is decided by the maximum of two terms, where the first term, $V^{\frac{-2\beta^2}{(2\beta+d)^2}}\log^\gamma V$, corresponds to the approximation of J eigenfunctions by DNNs and a second term $n^{-\frac{\beta}{2\beta+d}}$ is due to the eigenvalue estimation based on n independent samples under

the β-regularity norm. The sufficient number of J eigenfunctions and eigenvalues needed to achieve a desired concentration rate is determined by $\min(\alpha, \beta)$. The regularity order α controls the rate of eigenvalues decay while β represents the smoothness of the Hölder continuous basis function. A larger value of α and β indicates a smoother kernel function and a smaller number of truncations needed for the same number of spatial locations and sample functions collected to achieve the same accuracy.

Corollary 1. *Suppose the assumptions and conditions in Theorem 1 - 3 are satisfied. We have* $\Pi\left(K \in \Theta_K : \|K - K_0\|_{VV} > M\epsilon^*_{Vn} | \mathcal{D}^{(V,n)}\right) \to 0$ *under* P_0 *for* $\gamma > 2$ *and any* $M > 0$ *as* $V \to \infty, n \to \infty$, *where* $\epsilon^*_{Vn} = \sqrt{J}\left(V^{\frac{-\beta}{2\beta+d}} \log^{\gamma} V + n^{-1/2}\right)$.

Corollary 1 can be viewed as a natural consequence of the accurate estimation of the eigenvalues and eigenfunctions. By adopting the DNN method, the covariance kernel can be approximated sufficiently well as V and n both go to infinity. The rate ϵ^*_{Vn} is decided by the eigenfunction estimation error, the usual parametric rate for eigenvalue estimation, and a multiplicative factor J due to the number of terms in the truncated series.

To better understand the rates, consider a special case where $V = n$ and $\alpha = \beta$. Then $n^{\frac{2\beta d}{(2\beta+d)^2}} \lesssim J$ and $\epsilon_{Vn} = n^{-\frac{2\beta^2}{(2\beta+d)^2}}$ for the theorems above. The rate ϵ_{Vn} is slightly worse than the usual nonparametric rate of $n^{-\frac{\beta}{2\beta+d}}$. This is due to the additional step of estimating covariance kernel eigen-terms. For Corollary 1, the kernel estimation error is characterized by $\epsilon^*_{Vn} = n^{-\frac{2\beta^2}{(2\beta+d)^2}}$ by choosing $J \asymp n^{\frac{2\beta d}{(2\beta+d)^2}}$. Another scenario is $V \gg n$, which means that the number of spatial locations is much larger than the number of study subjects. Assume $V = n^{\tau}$ for $\tau > (2\beta + d)/2\beta$ and $\alpha = \beta$. Then $\epsilon_{Vn} = n^{-\frac{\beta}{2\beta+d}}$. We can choose $J = n^{\frac{d}{2\beta+d}}$, then $\epsilon^*_{Vn} \asymp Jn^{-1/2} = n^{-\frac{2\beta-d}{2(2\beta+d)}}$ if we further assume $d < 2\beta$. As expected, both ϵ_{Vn} and ϵ^*_{Vn} only depend on n.

3.4 Posterior Computation

Fully Bayesian inference in imaging studies can be computationally challenging due to large sample sizes, high-resolution images, and high-dimensional parameter spaces. To address these issues, we propose a scalable and efficient hybrid MCMC algorithm that combines SGLD sampling for parameters with non-conjugate priors and Gibbs sampling for parameters with conjugate or conditionally conjugate priors. This hybrid approach combines the advantages of the scalability of gradient-based methods and the efficiency of Gibbs sampling to handle the complexities of our proposed model.

Gradient-based MCMC sampling algorithms such as Metropolis-Hastings or Hamiltonian Monte Carlo are feasible for posterior updates of non-conjugate parameters. We adopt

the stochastic gradient Langevin dynamics algorithm [Welling and Teh, 2011], which is scalable to large datasets. Denote $\mathcal{D}$ as the collection of input data and $\mathcal{D}_{\mathcal{I}}$ be a subset of data with index set $\mathcal{I} \subset \{1, \ldots, n\}$, where n is the total sample size and $|\mathcal{I}|$ denotes the mini-batch size. At the tth iteration, SGLD updates DNN parameters as:

$$\mathbf{B}^{(t+1)} = \mathbf{B}^{(t)} + \frac{\epsilon_t}{2}\left\{\nabla \log \Pi(\mathbf{B}^{(t)}) + \frac{n}{|\mathcal{I}|}\sum_{i=1}^{|\mathcal{I}|}\nabla \log \Pi(\mathcal{D}_{\mathcal{I}}|\mathbf{B}^{(t)}, \cdot)\right\} + \zeta_t, \quad \zeta_t \sim \mathcal{N}(0, \epsilon_t),$$

where ϵ_t is a sequence of step sizes decaying to zero such that $\sum_{t=1}^{\infty}\epsilon_t = \infty$ and $\sum_{t=1}^{\infty}\epsilon_t^2 < \infty$.
Pólya-Gamma data augmentation [Polson et al., 2013] is used to implement the Gibbs sampler for updating the basis coefficients $\{\theta_{\rho_j}\}_{j=1}^{J}$ for the selection indicators. Since the selection indicator $\delta(s_v)$ has a binomial likelihood, we can view $\{\theta_{\rho_j}\}_{j=1}^{J}$ as parameters in a logistic regression model with the selection indicators as outcomes and a design matrix $\{\phi_j\}_{j=1}^{J}$, given all other parameters. By introducing auxiliary variables $\omega(s_v) \sim PG(0, 1)$ following the Pólya-Gamma distribution for each $\delta(s_v)$, we achieve conditional conjugacy priors, allowing for efficient Gibbs sampling.
A hybrid MCMC algorithm is proposed by integrating SGLD updates and Gibbs sampling. This algorithm requires pre-specifications including DNN architecture, prior hyperparameters, and a decaying learning rate. Parameters are initialized from prior distributions. We partition data into $M \in \mathbb{N}^+$ equal-sized mini-batches to accommodate individual-specific parameters and iteratively update parameters through each batch. Let $\mathcal{I}_m$ be the mth index set for batch data such that $\bigcup_{m=1}^{M}\mathcal{I}_m = \{1, \ldots, n\}$ and $\mathcal{I}_m \bigcap \mathcal{I}_{q'} = \emptyset$ for all $m \neq m'$, $1 \leq m, m' \leq M$. This partitioning ensures efficient updates across all parameters. In each MCMC iteration and each data batch, we update the DNN parameters using SGLD and all other parameters by Gibbs sampling to achieve an efficient and scalable sampling algorithm, as shown in Algorithm 1.

3.5 Simulations

We evaluate the numerical performance of our proposed model under three image regression frameworks outlined in Section 3.2.2. For the nonparametric regression setting, we present empirical comparisons validating our theoretical results on function and kernel estimation accuracy. In the image-on-scalar and scalar-on-image regression frameworks, we compare the DKLP model to other competing methods using two-dimensional image data, focusing on the estimation and selection accuracy of key spatially varying effects. For each simulation scenario, we generate 50 data replications and report the mean and standard deviation of the

Algorithm 1: Hybrid MCMC algorithm for DKLP in regression models

Data: Input data collection $\mathcal{D}$ with size n in M mini-batches
Pre-specification: DNN function $f_{\mathbf{B}}^{\text{DNN}}$ with output dimension J, learning rate ϵ, number of iterations T, hyperparameters in priors, orthonormal operator $\mathcal{T}$;
Initialization: $\mathbf{B}^{(0)}, \boldsymbol{\theta}_J^{(0)} = \{\phi_j^{(0)}, \lambda_j^{(0)}, \theta_j^{(0)}\}_{j=1}^J$, and other parameters $\boldsymbol{\gamma}^{(0)}$ in regression;
Result: Posterior samples for all parameters.
for $t = 1$ *to* T **do**
  for $m = 1$ *to* M **do**
    $\nu_t \sim \mathcal{N}(0, \epsilon_t)$;
    $\Delta\mathbf{B}^{(t)} = \frac{\epsilon_{t-1}}{2}\left\{\nabla \text{log}p(\mathbf{B}^{(t-1)}) + \frac{n}{|\mathcal{I}_m|}\sum_{i=1}^{|\mathcal{I}_m|}\nabla \text{log}p(\mathcal{D}_{\mathcal{I}_m}|\mathbf{B}^{(t-1)}, \boldsymbol{\theta}_J^{(t-1)}, \boldsymbol{\gamma}^{(t-1)})\right\} + \nu_t$;
    $\{\phi_j^{(t)}\}_{j=1}^J = \mathcal{T}(f_{\mathbf{B}^{(t)}}^{\text{DNN}})$;
    Sample $\{\lambda_j^{(t)}, \theta_j^{(t)}\}_{j=1}^J, \boldsymbol{\gamma}^{(t)}$ from posterior distributions;
  end
end

evaluation metrics. Detailed specifications of the DNN architectures for each scenario, along with additional simulation comparison results, are included in the supplementary materials.

3.5.1 Nonparametric regression

We evaluate the estimation accuracy of spatially varying functions and kernel functions for DKLP with respect to spatial resolution and sample sizes under the nonparametric regression framework. Data are generated based on model (3.1) with image resolutions 50×50 and sample sizes $n = 50$. The true spatially varying functions w_i are sampled from GPs with RBF and Matérn kernels as the true kernel functions. The kernel hyperparameters are set with a length scale of 0.2 and kernel variance of 1. To mimic the high noise feature in neuroimaging data, we employ additive Gaussian noise with $\sigma_\epsilon = 4$. To evaluate model performance, we adopt the function mean squared error (FMSE) $n^{-1}\sum_{i=1}^n \|w_i - w_{i0}\|_V$, the kernel mean squared error (KMSE) $\|K - K_0\|_{VV}$, where the norms are defined in Section 3.3, and the prediction interval coverage (CVG), which measures the percentage of intervals containing the true value. DKLP is compared with other nonparametric models for large spatial data, including standard Gaussian process regression under RBF (GPR-R) and Matérn (GPR-M) kernels, where we fix the kernel hyperparameters at the true values, nearest neighbor Gaussian process (NNGP) Datta et al. [2016], and stochastic partial differential equation approach (SPDE) [Lindgren et al., 2011] implemented by INLA [Rue et al., 2009]. All the methods are trained on the subsets of spatial observations $V \in \{30^2, 40^2, 50^2\}$, with metrics evaluated on the full spatial data. Table 3.1 demonstrates the performance of each method, with the two best performance metrics highlighted in bold. Across all scenarios, both SPDE

and DKLP consistently achieve the lowest FMSE and KMSE values, achieving improved accuracy for both function and kernel estimation as the number of observed spatial locations increase. Compared the SPDE, DKLP attains higher coverage, with close or higher than the nomial rate of 0.95, suggesting that it provides better uncertainty quantification while preserving high estimation accuracy. In contrast, other nonparametric regression models show larger estimation errors under the low signal-to-noise ratios. The robust performance of DKLP across spatial resolution and true kernel settings highlights the effectiveness of DKLP in learning complex spatial dependencies compared to baseline nonparametric models.

Table 3.1: Evaluation results on nonparametric regression, comparing the estimation accuracy of image functions and kernel functions. Mean values with standard deviations (in parentheses) are reported for each metric and scaled by factors of 0.01. The two best-performing metrics are highlighted in bold.

		RBF			Matérn		
V	Method	FMSE	CVG	KMSE	FMSE	CVG	KMSE
30^2	GPR-R	96(02)	**100**(00)	07(01)	106(02)	**100**(00)	06(00)
	GPR-M	135(02)	**100**(00)	10(01)	140(03)	**100**(00)	09(00)
	SPDE	**61**(01)	76(01)	**02**(00)	**69**(01)	71(01)	**01**(00)
	NNGP	102(03)	**100**(00)	03(00)	103(03)	**100**(00)	02(00)
	DKLP	**62**(01)	**93**(00)	**02**(00)	**68**(01)	**91**(00)	**02**(00)
40^2	GPR-R	64(01)	**100**(00)	05(00)	74(01)	**100**(00)	04(00)
	GPR-M	102(01)	**100**(00)	08(00)	106(01)	**100**(00)	07(00)
	SPDE	**45**(01)	81(01)	**01**(00)	**56**(01)	74(01)	**01**(00)
	NNGP	107(03)	**100**(00)	03(00)	107(03)	**100**(00)	03(00)
	DKLP	**45**(01)	**97**(00)	**02**(00)	**53**(01)	**96**(00)	**01**(00)
50^2	GPR-R	46(01)	**100**(00)	04(00)	55(01)	**100**(00)	03(00)
	GPR-M	81(01)	**100**(00)	06(00)	84(01)	**100**(00)	06(00)
	SPDE	**34**(01)	87(01)	**01**(00)	**46**(01)	78(01)	**01**(00)
	NNGP	113(03)	100(00)	04(00)	113(02)	**100**(00)	03(00)
	DKLP	**34**(01)	**99**(00)	**01**(00)	**42**(01)	**98**(00)	**01**(00)

3.5.2 Image-on-scalar regression

We simulate data from (3.2) with image resolutions of 32×32 and 64×64. The model includes $q = 5$ independent covariates sampled from $\mathcal{N}(0, I_q)$. We test different signal-to-noise ratio (SNR) settings with sample sizes $n \in \{20, 50, 200\}$ and noise levels $\sigma_y \in \{2, 4, 10\}$, mimicing typical imaging study conditions. The true spatial effects $\{\beta_r(s_v)\}$ include five distinct shapes, four with radial intensity changes in $[0.5, 2]$ and one with a complex constant pattern, with zero effects outside these regions. The selection indicator $\delta_r(s_v) = I\{\beta_r(s_v) \neq 0\}$ tests the method's ability to recover diverse spatial patterns and distinguish smooth and

sharp boundaries. Individual effects $\eta_i(s_v)$ follow a random radial pattern with intensity in $[-1, 1]$. DKLP is compared with mass univariate analysis (MUA) and statistical parametric mapping (SPM) [Friston, 2003], where SPM applies MUA after pre-smoothing and uses random field theory for selection. Additionally, we compare with the spatially varying coefficient model with jump discontinuities (SVCM) [Zhu et al., 2014], bivariate splines over triangulation-based spatially varying coefficient model (BST) [Li et al., 2021], and image response regression via neural networks (IRRNN) [Zhang et al., 2023a]. For MUA and SVCM, we control the false discovery rate (FDR) at 0.05 and employ corresponding selection methods for other methods and set the selection threshold at 0.05. For our methods, we control the FDR at 5% using Bayesian FDR control [Morris et al., 2008].

Table 3.2 presents the quantitative evaluation results. Across all simulation settings, DKLP consistently attains the lowest mean squared error (MSE) and maintains a high true positive rate (TPR) with a low false discovery rate (FDR), demonstrating both accurate estimation and reliable detection of true activation regions. In contrast, classical voxel-wise models such as MUA show increased MSEs and reduced TPR under noisy or high-dimensional conditions, reflecting the limited power in signal selection. While SPM and BST achieve high TPRs, they tend to over-select regions, resulting in elevated FDRs. From the visual comparison in Figure 3.1a, DKLP provides the most accurate reconstruction of the true spatial effects, accurately preserving both geometric boundaries and intensity contrasts of the activation patterns. By contrast, MUA produces scattered, spatially discontinuous detections, while SPM and BST yield smoother but still blurred estimates with loss of structural sharpness. The SVCM model identifies relatively sharp boundaries but introduces noisy dots around true shapes, while the IRRNN detects fewer activation locations overall, likely due to its sensitivity to the limited number of spatial observations available. Overall, DKLP exhibits robustness and adaptability, achieving spatially accurate estimation with high sensitivity and specificity, thereby validating its advantage in modeling complex spatial correlations across neuroimaging data of varying scales and noise levels.

3.5.3 Scalar-on-image regression

We generate data according to model (3.3) with image resolution $V = 32 \times 32$. The true spatially varying coefficient $\beta(s_v)$ contains two regions with different shapes and varying image intensities, as shown in Figure 3.1b. The image predictors are sampled as $y_i(s_v) \sim \mathcal{N}(w_i(s_v), \sigma_\epsilon^2)$, where each $w_i(s_v)$ contains variants of the geometric shapes in $\beta(s_v)$ with differing intensities, along with two additional regions, ensuring that the images share a similar kernel with $\beta_s(v)$ while maintaining sufficient variability. We study continuous

Table 3.2: Evaluation results for image-on-scalar regression simulations. Mean values with standard deviations (in parentheses) are reported for each metric in unit 0.01. The best performing metric is highlighted in bold.

V	n	σ_ϵ	MUA	SPM	BST	SVCM	IRRNN	DKLP
Mean squared error (Unit: 0.01)								
32^2	20	2	26(05)	05(01)	05(01)	10(03)	16(06)	**03**(01)
	20	4	57(01)	**10**(02)	12(03)	42(05)	49(03)	12(04)
	50	2	04(01)	02(00)	03(00)	02(00)	**01**(01)	02(00)
	50	4	36(03)	05(01)	05(01)	11(02)	14(04)	**04**(01)
	200	10	49(01)	**06**(00)	**06**(00)	13(01)	21(04)	**06**(01)
64^2	20	2	26(04)	**03**(00)	**03**(00)	09(02)	14(06)	**03**(00)
	20	4	57(01)	**05**(01)	**05**(01)	40(04)	51(04)	**05**(00)
	50	2	04(01)	**01**(00)	03(00)	02(00)	**01**(01)	**01**(00)
	50	4	36(04)	**03**(00)	04(00)	10(01)	10(03)	**03**(00)
	200	10	49(01)	**04**(00)	**04**(00)	13(01)	16(02)	**04**(00)
True positive rate (Unit: 0.01)								
32^2	20	2	60(08)	**98**(01)	97(03)	89(04)	66(11)	**98**(02)
	20	2	03(02)	**90**(04)	86(09)	42(08)	08(05)	84(07)
	50	2	96(01)	99(00)	**100**(00)	99(00)	99(01)	**100**(00)
	50	4	41(06)	**97**(01)	**97**(02)	88(02)	72(07)	96(01)
	200	10	18(03)	96(01)	**97**(01)	84(02)	58(07)	93(01)
64^2	20	2	60(07)	**99**(01)	**99**(00)	90(02)	69(10)	**99**(01)
	20	4	03(02)	95(05)	96(01)	45(07)	06(06)	**96**(01)
	50	2	96(01)	**100**(00)	**100**(00)	99(00)	99(00)	**100**(00)
	50	4	40(06)	99(01)	**100**(00)	89(02)	78(06)	98(00)
	200	10	19(02)	99(00)	**100**(00)	85(01)	64(03)	96(00)
False discovery rate (Unit: 0.01)								
32^2	20	2	09(01)	12(03)	11(04)	09(01)	**01**(01)	06(01)
	20	4	13(06)	14(02)	11(06)	10(01)	**01**(02)	10(02)
	50	2	05(01)	03(01)	12(01)	06(01)	**01**(02)	06(00)
	50	4	06(01)	09(02)	08(02)	08(01)	**01**(01)	05(01)
	200	10	04(01)	09(02)	12(02)	06(01)	**01**(01)	06(01)
64^2	20	2	09(00)	08(02)	10(01)	08(00)	**01**(01)	06(01)
	20	4	13(04)	08(01)	08(01)	09(01)	**00**(01)	08(01)
	50	2	05(00)	02(00)	17(01)	06(00)	**01**(01)	**01**(00)
	50	4	06(00)	06(01)	14(01)	07(01)	**01**(01)	06(01)
	200	10	04(01)	09(01)	17(01)	05(00)	**01**(01)	07(01)

outcomes generated as $u_i \sim \mathcal{N}(\sum_{v=1}^{V} y_i(s_v)\beta(s_v), \sigma_u^2)$. We evaluate sample size $n = 300$ under image noise standard deviations $\sigma_\epsilon \in \{3, 5\}$ and measurement error standard deviations $\sigma_u \in \{2, 4\}$. We split the data with 80% as training data and 20% for testing data to evaluate prediction performance by prediction MSEs. DKLP is compared with the least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996], sparse fused LASSO [Tibshirani, 2011, Wang et al., 2016], graph OSCAR (GOSCAR)[Yang et al., 2012], soft-thresholded Gaussian processes (STGP) [Kang et al., 2018], and the Tree-based Low-rank Horseshoe model (T-LoHo) [Lee et al., 2021]. STGP shares the same model specification as our method but adopts a prespecified Matérn kernel function in the GP prior. For both STGP and DKLP methods, we apply the Bayesian FDR control method [Morris et al., 2008] to restrict the expected FDR to 5%.

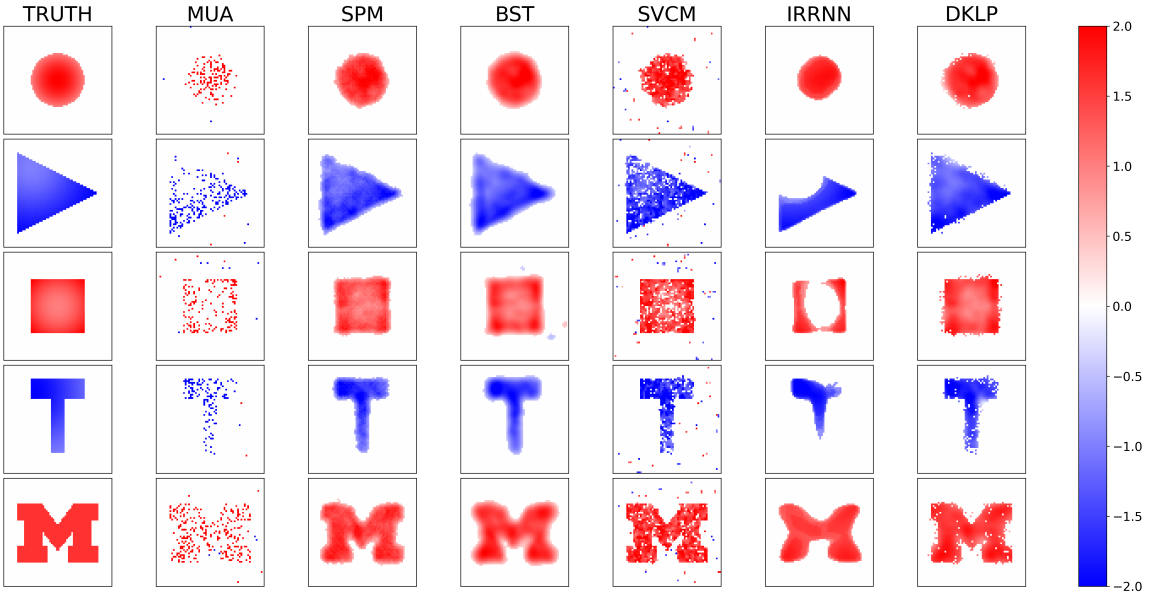
We report all empirical performance in Table 3.3. LASSO and GOSCAR fail to recover any spatial pattern or achieve reasonable estimation accuracy in all simulation settings. FL, STGP, and T-LoHo gain high estimation accuracy, while DKLP outperforms these methods with the smallest MSEs. We observe that STGP and FL are often subject to high FDR, resulting in overly selecting the regions. T-LoHo achieves reasonable selection accuracy and robustness in FDR, but it is conservative in region selection and cannot capture continuously varying image intensities. In terms of predictive performance, FL and DKLP both show consistently low predictive MSEs on the testing data, while DKLP outperforms with smaller standard deviations across scenarios. Figure 3.1b shows that DKLP can correctly identify signal regions in spatial varying function with accurate spatial varying effects while controlling the FDR below 10%.

3.6 Brain Imaging Applications

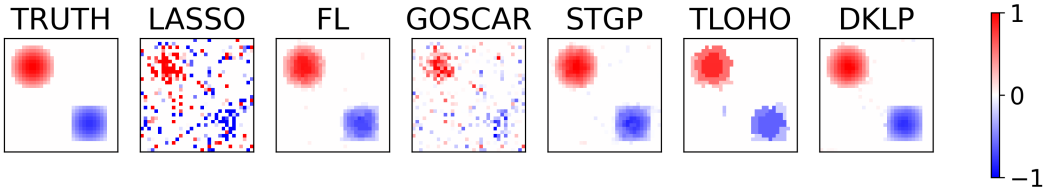
In this section, we illustrate the efficacy of our proposed DKLP method through two medical brain imaging studies, utilizing image-on-scalar regression and scalar-on-image regression models, respectively. The image-on-scalar regression study focuses on assessing estimation accuracy and selection consistency, while the scalar-on-image regression study evaluates prediction accuracy.

3.6.1 Cortical surface task fMRI

We apply DKLP to analyze fMRI data in the Adolescent Brain Cognitive Development (ABCD) study. We adopt the image-on-scalar regression model to investigate the association between cognitive functions and brain development among children [Casey et al., 2018]. Our



(a) Image-on-scalar regression under scenario $N = 200, \sigma_\epsilon = 10, V = 64^2$



(b) Scalar-on-image regression under scenario $\sigma_u = 2, \sigma_\epsilon = 5$

Figure 3.1: Comparison of true spatially varying functions with estimated effects from DKLP and other competing methods in simulation studies across two image regression frameworks.

Table 3.3: Evaluation results from scalar-on-image regression reported in mean and standard deviation (in parentheses) in scaled units. The two best-performing metrics are highlighted in bold.

σ_u	σ_ϵ	LASSO	FL	GOSCAR	STGP	T-LoHo	DKLP
Mean squared error (Unit: 0.001)							
2	3	290(03)	**02**(00)	56(01)	**02**(20)	06(00)	**01**(00)
2	5	827(07)	**01**(00)	56(01)	02(01)	05(00)	**00**(00)
4	3	275(03)	**03**(00)	58(01)	**06**(01)	08(00)	**03**(00)
4	5	811(07)	**02**(00)	57(01)	**02**(00)	06(00)	**01**(00)
True positive rate (Unit: 0.01)							
2	3	39(02)	**99**(01)	52(06)	**99**(02)	88(04)	**97**(02)
2	5	40(02)	**100**(00)	52(05)	**99**(04)	90(03)	**100**(01)
4	3	37(02)	**98**(01)	52(05)	**97**(11)	87(04)	90(03)
4	5	39(02)	**99**(01)	52(05)	**99**(01)	89(04)	**96**(02)
False discovery rate (Unit: 0.01)							
2	3	62(02)	09(03)	59(04)	29(05)	**02**(02)	**06**(04)
2	5	62(02)	05(02)	59(04)	25(08)	**02**(02)	**07**(03)
4	3	61(03)	17(05)	59(04)	35(08)	**08**(08)	**10**(06)
4	5	62(02)	11(03)	59(04)	31(05)	**05**(06)	**07**(04)
Prediction mean squared error (Unit: 0.01)							
2	3	198(176)	**51**(49)	60(11)	**51**(49)	**50**(43)	**51**(48)
2	5	1210(1240)	**51**(48)	106(43)	**52**(48)	55(40)	**52**(48)
4	3	210(219)	**52**(47)	61(10)	54(46)	**51**(42)	**52**(43)
4	5	1183(1209)	**53**(46)	107(43)	**54**(46)	57(38)	**54**(44)

specific focus is on n-back cortical surface fMRI data from 21 study sites, as featured in Release 4.0. The cortical surface data are collected on the two-dimensional manifold surfaces, resulting in three-dimensional imaging grids on $V = 29{,}696$ and $V = 29{,}716$ spatial locations for the left and right hemispheres, respectively.

Our analysis considers $q = 4$ covariates, including the general cognitive ability component score [Sripada et al., 2020], which is our primary predictor, along with the psychiatric diagnostic score, age, and sex. We add a spatial varying intercept term $\beta_0(s_v)$, without imposing a selection indicator. The imaging data consist of complete data from 2,497 individuals at baseline and 1,691 individuals at a two-year follow-up.

Our DKLP model is compared to MUA, SPM, IRRNN, and SIMBA, which is proposed in Chapter II, in terms of estimation accuracy and selection reproducibility. We control the FDR at 5% for all methods. Since the true spatially varying functions are unknown, we validate the model fitting results through cross-site analyses. For estimation accuracy, each method is fitted to data from one study site and used to predict the image data from all other sites. Prediction MSEs are reported for each single-site analysis, as shown in Figure

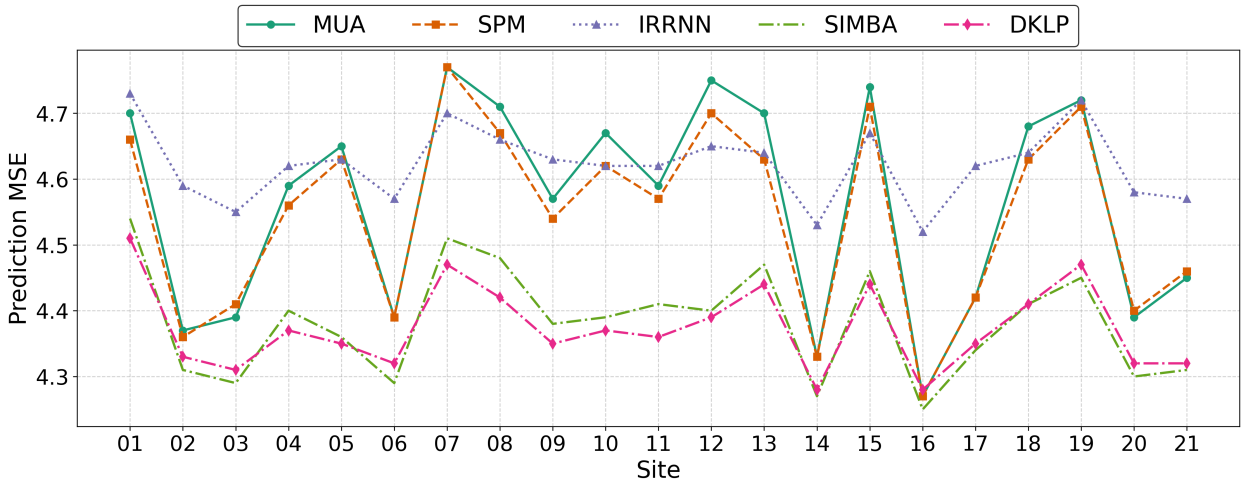


Figure 3.2: Cross-site prediction MSEs for cortical surface fMRI data collected at baseline on the left hemisphere in the ABCD study.

3.2. Our method consistently achieves smaller prediction MSEs in all single-site analyses. Here, we present evaluation results from the baseline time point on the left hemisphere for demonstration purposes. Similar results are observed for the right hemisphere and at the two-year follow-up time point, which are included in the supplementary materials.

We apply the same single-site analyses to examine the reproducibility of the identified brain regions. Following the criteria established by Zhang et al. [2023a], a voxel is considered reproducible if it is selected in at least five single-site results. Figure 3.3 illustrates the selected regions for the left and right hemispheres at two time points. Dark regions indicate voxels selected using all-site data, while light regions represent reproducible regions identified from single-site analyses. The results show that all four methods produce similar region selection patterns in the all-site analyses. However, MUA, SPM, and IRRNN demonstrate a higher proportion of regions with low reproducibility, likely due to their limited approximation capabilities, particularly under the small sample size settings of single-site studies. Among these methods, IRRNN performs better than MUA and SPM with more reproducible locations. DKLP, however, yields the most robust selection results and the minimal presence of unreproducible regions. Moreover, the reproducible brain regions identified by DKLP are consistently selected by the other three methods in the all-site analyses, further corroborating the accuracy and reliability of DKLP's selection results.

3.6.2 EEG brain-computer interfaces

Electroencephalogram (EEG)-based brain-computer interfaces (BCIs) enable communication between the brain and external devices by decoding neural signals. This technology benefits

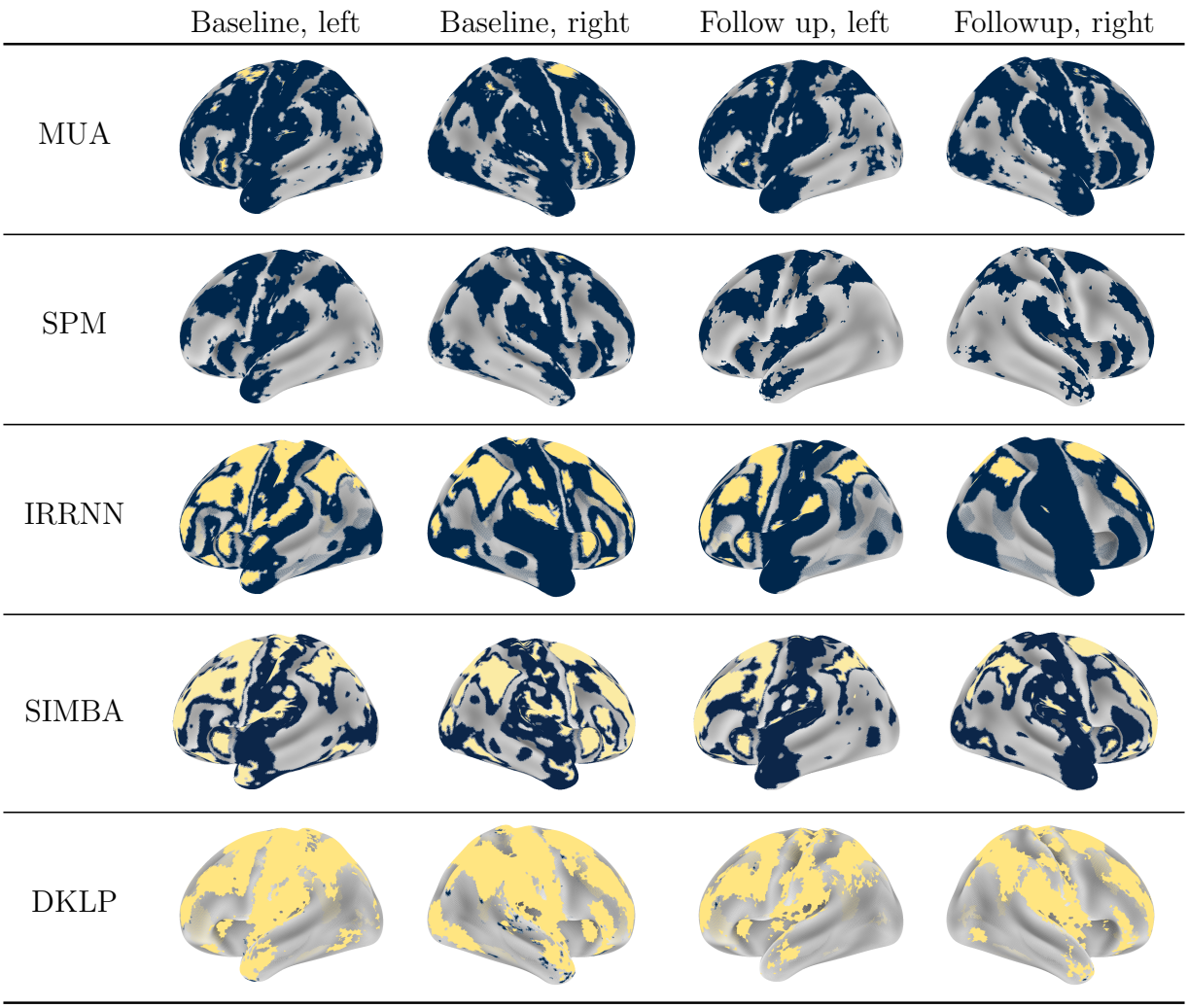


Figure 3.3: Selected brain regions using the ABCD cortical surface fMRI data in left and right hemispheres at baseline and follow-up time points. Blue regions are voxels selected using all-site data and yellow regions are reproducible voxels selected from 21 single-site analyses.

patients with motor disability and supports neuroscientific research. In this study, we focus on the P300 event-related potential (ERP)-BCI system, which detects brain responses to target and non-target stimuli. Detailed information can be found in Farwell and Donchin [1988].

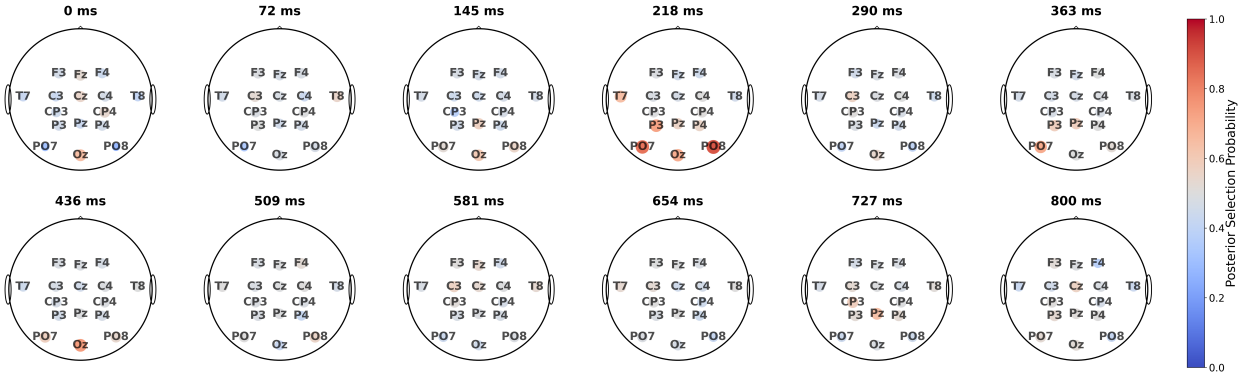
In the P300-BCI system, a participant-specific classifier is developed to decode intended characters using multi-channel EEG data. The BCI device displays a 6×6 grid, where rows or columns flash as stimuli, and participants mentally count when their target character appears. Each complete set of 12 flashes forms one sequence, and multiple sequences make up a super-sequence for a target character.

We apply DKLP under a scalar-on-image regression framework with binary outcomes for each participant to predict the intended characters in testing sessions based on data collected during training sessions. In the training data, participants use the BCI2000 software platform to mentally type 19 characters, resulting in 19 super-sequences. Each super-sequence comprises 15 sequences, each including 12 flashing stimuli (one column, one row, and ten non-target stimuli). EEG data are recorded from 16 channels, each corresponding to different brain regions, with each super-sequence lasting approximately 29,000ms at a sampling rate of 256Hz. After preprocessing with notch and band-pass filters, each stimulus is followed by EEG signals from $T = 205$ time points in each channel. In total, each participant receives 3,420 stimuli in the training session, with each stimulus corresponding to a binary outcome: $u_i = 1$ for target stimuli and $u_i = 0$ for non-target stimuli. The space-time varying image predictor, $y_{ie}(s_t)$, represents the collected EEG signal for the ith stimulus on the eth channel at time point $s_t \in [0, 800]$ms. Our model is defined as $\text{logit}\{P(u_i = 1)\} = \sum_{e=1}^{E} \sum_{t=1}^{T} y_{ie}(s_t)\beta_e(s_t)\delta_e(s_t)$, where $\beta_e(s_t) \sim \mathcal{GP}(0, K)$ for all $e = 1, \ldots, 16$ and $K \sim \mathcal{DKLP}(f_{\mathbf{B}}^{\text{DNN}}, \sigma_f^2, A)$. With binary outcomes, we employ the same data augmentation strategy as in Section 3.4. We compare DKLP with six popular machine learning algorithms, including logistic regression (LR) [Viana et al., 2014], support vector machine (SVM) [Kaper et al., 2004], random forest (RF) [Okumuş and Aydemır, 2017], XGBoost (XGB) [Leoni et al., 2021], stepwise linear discriminant analysis (swLDA) [Donchin et al., 2000] and the EEGNet implemented by convolutional neural network [Lawhern et al., 2018], all of which have demonstrated promising performance in EEG predictions. We fit each model to both Amyotrophic Lateral Sclerosis (ALS) and healthy control participants and evaluate the classifier performance based on character-level prediction accuracy. Table 3.4 summarizes the prediction results in percentage for the two participant groups, averaged across available testing sessions. Participants are included if any classifier achieves an accuracy above 70%. Although both swLDA and EEGNet achieve high accuracy for several participants, DKLP consistently demonstrates higher or comparable testing accuracy across

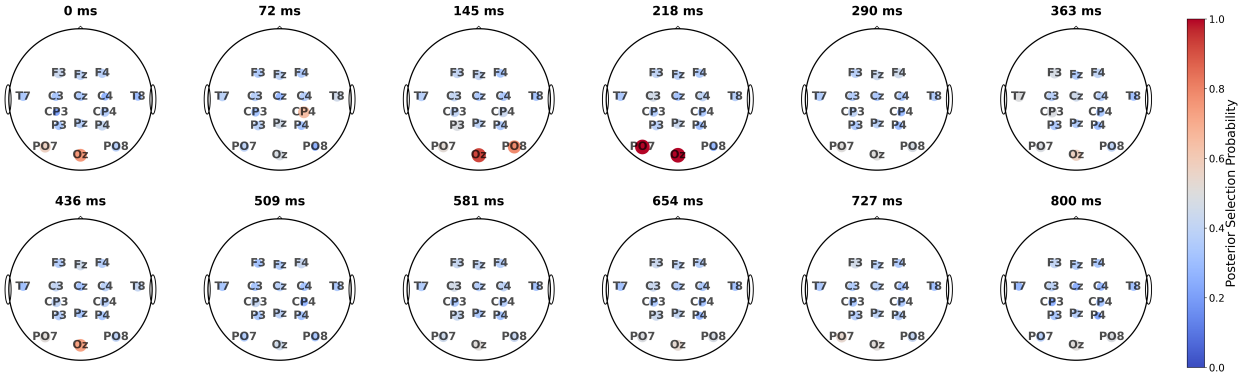
all participants. Particularly for ALS participants, we increase the character-level prediction accuracy from 75% to 80% in average. To further investigate the significance of different brain channels in the character prediction task, we analyze the posterior selection probabilities of neural signals along the 205 time points for both ALS and healthy participants. Figure 3.4 presents the posterior probabilities at 12 equally spaced time points across 16 brain channel locations, with node size and color indicating the magnitude of the probabilities. Notably, channels PO7, PO8, and Oz show higher contributions to character-level prediction. This observation aligns well with existing medical knowledge, as these channels correspond to optical regions associated with visual stimuli processing. These results are further supported by other statistical methods [Zhao et al., 2023, Chenane et al., 2019], validating that these three channels consistently hold higher weights in neural activity patterns, enhancing the modeling performance of the BCI studies.

Table 3.4: Individual-level BCI character prediction accuracy percentage averaged across testing sessions for ALS and healthy control participants.

Group	ID	LR	SVM	RF	XGB	swLDA	EEGNet	DKLP
Control	106	81	49	15	25	**87**	85	**87**
	108	83	80	40	69	92	**94**	92
	111	66	58	30	39	79	79	**83**
	114	88	90	63	82	93	91	**97**
	115	83	90	47	60	**95**	89	94
	117	92	90	72	82	**99**	93	98
	119	46	53	35	32	61	**85**	66
	120	81	76	58	68	87	86	**88**
	121	83	73	45	66	**92**	88	90
	122	69	48	19	31	84	78	**85**
	123	92	84	33	66	**99**	94	98
	171	83	63	44	59	89	88	**90**
	172	61	61	32	48	77	79	**82**
	177	81	83	46	75	**93**	90	92
	183	83	74	33	58	94	89	**95**
ALS	143	79	76	49	59	84	**88**	**88**
	145	61	42	09	25	69	57	**74**
	146	85	77	45	69	**93**	87	90
	147	81	78	31	56	**93**	**93**	**93**
	151	91	90	53	77	94	93	**95**
	152	47	51	40	47	59	**75**	66
	156	71	64	23	35	81	**90**	83
	158	56	39	29	37	72	70	**74**
	159	59	26	13	13	70	67	**77**



(a) ID 147: ALS participant



(b) ID 171: Healthy participant

Figure 3.4: Posterior selection probabilities of EEG brain signals across sixteen brain channels at twelve equally spaced time points. Node sizes represent the magnitude of the selection probabilities on each channel.

3.7 Conclusion

In this paper, we introduce DKLP, a deep learning-based Bayesian nonparametric model for Gaussian process kernel learning within the image regression framework. Through nonparametric, image-on-scalar, and scalar-on-image regression models, we demonstrate the adaptability of DKLP to various Bayesian nonparametric models with GP priors. By integrating the approximation power of DNNs, DKLP significantly enhances the kernel learning flexibility, allowing it to capture complex spatial and temporal relationships in imaging data. We study the theoretical properties of the DKLP prior and establish the posterior contraction rate of function and kernel estimation in regression models, with an efficient and scalable posterior computation algorithm for large-scale datasets. Empirical evaluations in both simulation and real data analysis show that DKLP consistently improves the model performance in complex image datasets over other established image regression methods. However, our method is limited to images with aligned coordinates, restricting its applicability to datasets with misaligned or heterogeneous spatial structures. Future work could extend DKLP to handle unaligned imaging data and time-varying kernels for longitudinal studies, broadening its utility in scientific and clinical applications.

CHAPTER 4

Activation-guided Diffusion Models for fMRI Synthesis: Integrating Image Regression and Generation for Statistical Inference

4.1 Introduction

In the field of neuroimaging studies, task-based functional magnetic resonance imaging (fMRI) contrast maps provide individual-specific voxelwise summaries of brain activation differences between experimental conditions. These contrast maps are widely utilized in downstream statistical inference, including group-level analyses to identify brain regions associated with clinical factors, as well as in predictive modeling for diagnosing or prognosing neurological diseases. However, obtaining large-scale, high-quality fMRI data is costly and resource-intensive, requiring not only image acquisition but also carefully designed task paradigms and extensive preprocessing. Consequently, many neuroimaging studies are constrained by relatively small sample sizes, which limit statistical power and reproducibility in subsequent data analyses.

Recent advances in generative models provide promising directions for learning complex data distributions in image synthesis. While early approaches such as generative adversarial networks (GANs) [Goodfellow et al., 2014] and variational autoencoders (VAEs) [Kingma et al., 2013] demonstrated feasibility in learning high-dimensional data distributions, denoising diffusion probabilistic models (DDPMs) [Ho et al., 2020] have emerged as state-of-the-art methods, offering stable training and high-quality image generation [Dhariwal and Nichol, 2021]. Extensions such as latent diffusion models (LDMs) [Rombach et al., 2022] and diffusion transformers (DiTs) [Peebles and Xie, 2023] further improve their scalability to high-resolution data. Motivated by these successes, diffusion models have gained increas-

ing attention in the neuroimaging community and have recently been applied to various data types, including structural MRI generation [Peng et al., 2023, Pinaya et al., 2022], functional connectivity synthesis in resting-state fMRI [Ajith and Calhoun, 2024], and physiologically realistic fMRI time series [Hu et al.]. Despite these developments, there remains limited work on applying diffusion models to task-based fMRI, particularly for synthesizing individualized contrast maps.

Task-based fMRI generation presents unique challenges compared to other MRI data. Unlike structural MRI, fMRI data are characterized by high levels of noise and large individual variability from both measurement error and inherent heterogeneity in functional activation patterns. Task-related activation signals are typically sparse and weak, while the background contains complex and structured noise. Moreover, available subject-level information, such as age and gender, or clinical characteristics like cognitive scores, provides only coarse information and fails to account for fine-grained individual differences. Therefore, standard conditional generation frameworks based only on low-dimensional labels may yield anatomically plausible but functionally misaligned with target task-related activity patterns, highlighting the need for methods that explicitly model individual brain activation profiles.

A key limitation in existing fMRI generation approaches lies in the lack of informative conditioning signals that contain spatial and functional prior information. Studies in computer vision and medical imaging have shown that spatial conditioning mechanisms are essential for controllable and functionally relevant generation. One influential development is the ControlNet architecture [Zhang et al., 2023b], which extends diffusion models by introducing spatially localized conditioning inputs while preserving the pretrained generative backbone. ControlNet achieves this by freezing the original model parameters and attaching trainable encoder copies linked through zero-initialized convolution layers, enabling stable fine-tuning and reliable incorporation of structural guidance. This design aligns well with fMRI data, where the underlying subject-specific activation patterns can provide critical spatial priors for noisy observed images. Recent work has demonstrated the effectiveness of ControlNet in neuroimaging applications, such as MRI-to-PET synthesis [Sargood et al., 2025] and anatomically constrained counterfactual MRI generation [Li et al., 2025], illustrating its potential to embed domain-specific priors in generative models.

In this work, we propose a brain activation controlled network (BACNet) framework for individualized task fMRI contrast map generation. Our base model is a standard diffusion model with classifier-free guidance, incorporating demographic covariates as conditioning factors. To capture latent participant-specific activation profiles, we employ Bayesian nonparametric regression (BNPR) with a deep kernel learning process prior (DKLP) introduced in Chapter 3 to estimate individualized activation maps, which are then incorporated as

spatial conditioning through the ControlNet mechanism. This hybrid framework integrates statistical modeling with generative learning, enabling the synthesis of contrast maps that preserve individual-level activation signals while maintaining biological plausibility. We evaluate our approach on both synthetic and real data. Using noisy MNIST images, we illustrate that BNPR effectively recovers sparse activation patterns under noise, providing reliable spatial priors for generation. We then apply the framework to cortical surface fMRI contrast maps from the Adolescent Brain Cognitive Development (ABCD) study [Casey et al., 2018], showing that our method produces biologically plausible individualized maps that preserve both brain activation patterns and task-related associations with clinical covariates. Finally, we demonstrate that the generative model can be used for downstream statistical analysis, including the construction of empirical null distributions for hypothesis testing in small-sample studies, thereby improving robustness and interpretability in statistical findings.

The remainder of the paper is organized as follows. Section 4.2 introduces the proposed generative framework, detailing each of the modeling steps in three stages. Section 4.3 demonstrates results from synthetic MNIST data and real cortical surface fMRI data from the ABCD study, and illustrates how the proposed generative model can be applied to statistical inference. Finally, Section 4.4 concludes with a discussion and outlines directions for future research.

4.2 Methods

In this section, we introduce our activation-based diffusion model for generating individualized fMRI contrast maps. We first define the notation for the imaging data and covariate labels. We then present the three main components of our approach, including estimation of subject-specific activation patterns from observed contrast maps using a Bayesian regression model, conditional diffusion modeling with classifier-free guidance, and incorporation of spatial conditioning through the ControlNet scheme. An overview of the BACNet framework is provided in Figure 4.1.

Data Description Let $y_i(s_v)$ denote the observed fMRI contrast map value at spatial location $s_v \in \mathcal{S}$ for subject $i = 1, \ldots, n$, where $\mathcal{S} = \{s_1, \ldots, s_V\}$ is the set of voxel locations in the three-dimensional brain volume. The observed contrast maps $\mathbf{y}_i = (y_i(s_1), \ldots, y_i(s_V))$ reflect noisy measurements of underlying task-related activation patterns. In addition, each subject i is associated with a vector of q covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})$, such as age, gender, or clinical assessments, which are used as conditioning variables in the generative model.

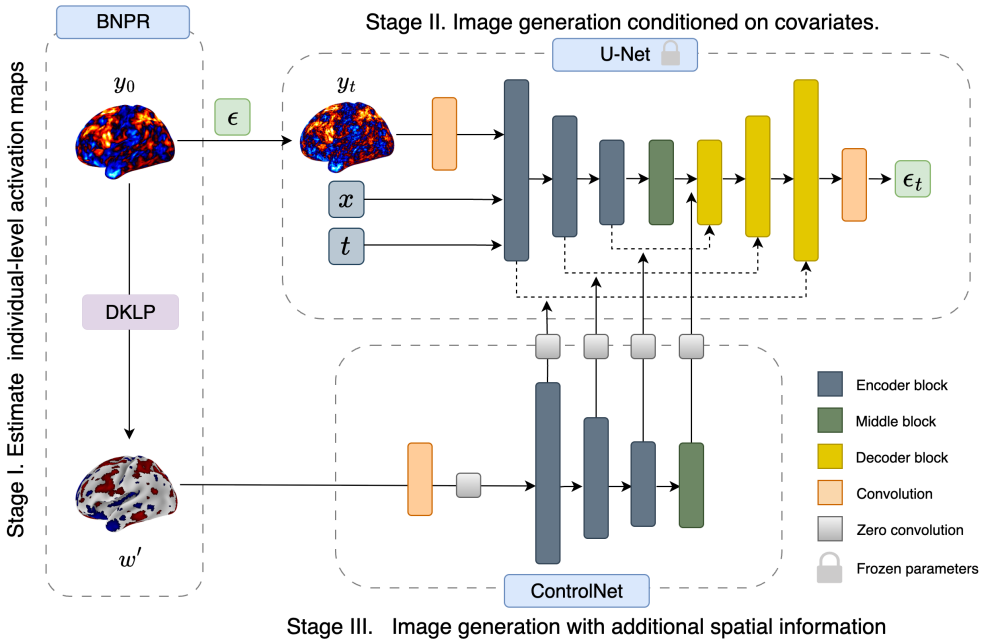


Figure 4.1: Overview of the proposed BACNet framework. Stage I applies BNPR with a DKLP prior to estimate subject-specific activation maps from observed noisy fMRI contrast maps. Stage II implements a conditional diffusion model with a U-Net backbone, learning the predicted noises at each denoising step. Stage III introduces ControlNet to incorporate the estimated activation maps as spatial priors, which are injected into the U-Net through zero-initialized convolutional layers.

Stage 1: Activation map estimation via Bayesian regression model In the first stage, we use a Bayesian nonparametric regression (BNPR) model to denoise observed task-based fMRI contrast maps. The goal of this step is to separate subject-specific, task-related activation signals from structured background noise, thereby producing clean and interpretable spatial representations that can guide the generative model. We model each observed contrast map as $y_i(s_v) = w_i(s_v) + \epsilon_i(s_v)$, where $w_i(s_v)$ denotes the latent activation function of interest and $\epsilon_i(s_v)$ represents the complex residual noise process. To flexibly capture the complex spatially varying features of $w_i(s_v)$, we place a deep kernel learning process (DKLP) prior, which integrates Gaussian process (GP) priors with kernels parameterized by deep neural networks. This construction enables data-adaptive modeling of spatial dependencies in neuroimaging data while maintaining the nonparametric interpretability of GP priors. Details of the DKLP prior and posterior computation are provided in Chapter 3. To emphasize biologically meaningful activation patterns, we apply a thresholding function $w_i'(s_v) = w_i(s_v) \cdot I(|w_i(s_v)| > \gamma)$ to derive a sparse activation map $\mathbf{w}_i' = (w_i'(s_1), \ldots, w_i'(s_V))$, where γ is a fixed scalar threshold. This regression-based activation learning yields individualized denoised maps that highlight task-relevant brain regions, which improve interpretability at the subject level and, more importantly, serve as spatial conditioning inputs for the diffusion model in Stage 3.

Stage 2: Conditional DDPM with Classifier Free Guidance In the second stage, we employ a denoising diffusion probabilistic model (DDPM) to generate the noisy contrast maps y_i conditioned on covariates $\mathbf{x}_i$. DDPMs synthesize data by learning to reverse a forward noising process. For simplicity, we drop the subject index i below. In the forward process, a clean image $\mathbf{y}_0$ is gradually corrupted over $t = 1, \ldots, T$ steps according to

$$q(\mathbf{y}_t \mid \mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t; \sqrt{1-\eta_t}, y_{t-1}, \eta_t \mathbf{I}),$$

with a predefined noise schedule $\{\eta_t\}_{t=1}^T$. The reverse process is parameterized by a neural network $\epsilon_\theta(\mathbf{y}_t, t)$ trained to predict the Gaussian noise at each step. The training steps minimize the objective $\mathcal{L} = \mathbb{E}_{\mathbf{y}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{y}_t, t)\|^2\right]$, where $\mathbf{y}_t = \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ and $\bar{\alpha}_t = \prod_{s=1}^t (1-\eta_s)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

To incorporate conditional generation, we adopt classifier-free guidance (CFG) [Ho and Salimans, 2022]. During training, covariate labels $\mathbf{x}$ are randomly replaced with null labels $\emptyset$ at a fixed probability, allowing the model to learn both conditional $\epsilon_\theta(\mathbf{y}_t, t, \mathbf{x})$ and unconditional $\epsilon_\theta(\mathbf{y}_t, t, \emptyset)$ distributions. At inference, these predictions are combined as $\hat{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{x}) = (1+a) \cdot \epsilon_\theta(\mathbf{y}_t, t, \mathbf{x}) - a \cdot \epsilon_\theta(\mathbf{y}_t, t, \emptyset)$, where $a \geq 0$ is the guidance scale that

controls the strength of conditioning.

In this work, the denoising network ϵ_θ is implemented as a U-Net–based diffusion model [Ronneberger et al., 2015] with residual and attention mechanisms adapted from Cardoso et al. [2022]. Conceptually, the U-Net consists of an encoder path that extracts hierarchical spatial features and a decoder path that reconstructs the image from these features through symmetric upsampling. Timestep information is encoded through sinusoidal features and passed through a multilayer perceptron. Clinical or demographic covariates are projected into label embeddings, which are then concatenated with the time embedding and forwarded to the residual blocks. The encoder path sequentially downsamples features using ResNet blocks and attention modules, while the decoder mirrors this structure with skip connections, additional residual blocks, and progressive upsampling. Further details of the U-Net design can be found in Ronneberger et al. [2015].

An additional advantage of CFG is its ability to handle partially labeled data, a common issue in neuroimaging studies due to incomplete covariate collection or participant dropout. Missing labels can be treated as null labels during training, allowing the model to simultaneously learn conditional and marginal distributions. This improves sample efficiency and ensures that both labeled and unlabeled data contribute to training.

Stage 3: Brain Activation Guided DDPM with ControlNet While the conditional DDPM in Stage 2 learns the distribution of contrast maps given subject covariates, label conditioning alone is insufficient to capture fine-grained, subject-specific activation patterns. This is because covariates provide only low-dimensional information, while the generated images have high-dimensional complex spatial patterns and are subject to large inter-individual variability. As a result, label-conditioned DDPMs often produce contrast maps that are visually plausible under the data distribution but misaligned with the target subject-specific activation signals.

To impose additional guidance besides demographic or clinical labels, we adapt the ControlNet scheme [Zhang et al., 2023b], which introduces a conditioning pathway for injecting spatial conditioning signals into the U-Net architecture. The key idea is to mirror each encoder block of the pretrained U-Net with a trainable copy, allowing spatial priors, i.e., the BNPR-estimated activation maps, to be incorporated without disrupting the pretrained network. Following Zhang et al. [2023b], let the denoising network be composed of L neural blocks, such as a ResNet unit or an attention mechanism. Denote the ℓ-th block of the pretrained U-Net as $\mathcal{F}^\ell(\cdot;\Theta_u^\ell)$ with parameter Θ_u^ℓ for $\ell = 1,\ldots,L$ that mapping an input feature map $\mathbf{v}^\ell$ from the previous layer to an output $\mathbf{u}^\ell = \mathcal{F}^\ell(\mathbf{v}^\ell;\Theta_u^\ell)$. Additionally, we define a zero-initialized convolution layer $\mathcal{Z}(\cdot;\Theta_z^\ell)$ parameterized by $\Theta_z^\ell = (\mathbf{W}^\ell, \mathbf{b}^\ell)$. For input feature map

$\mathbf{v}^\ell$, it performs convolution as $\mathcal{Z}(\mathbf{v};\Theta_z^\ell) = \mathbf{W}^\ell \star \mathbf{v}^\ell + \mathbf{b}^\ell$, where $\mathbf{W}^\ell \in \mathbb{R}^{C_{\text{out}}^\ell \times C_{\text{in}}^\ell \times 1 \times 1}$, $\mathbf{b} \in \mathbb{R}^{C_{\text{out}}^\ell}$, $C_{\text{out}}^\ell, C_{\text{in}}^\ell$ represent the numbers of the ℓ-th input and output feature channels, respectively, and $\star$ denote the convolution operator applied over the input feature maps. All the parameters in Θ_z^ℓ for every $\ell = 1, \ldots, L$ are initialized with zeros and gradually learns nonzero weights and biases, enabling the conditioning pathway to progressively incorporate spatial information into the U-Net backbone.

ControlNet augments each encoder block by introducing a trainable copy of $\mathcal{F}^\ell(\cdot;\Theta_c^\ell)$, which contains the same neural block as in the pretrained U-Net but parameterized by a different set of parameters Θ_c^ℓ. The corresponding neural blocks are then linked through the zero-initialized 1×1 convolutional layers $\mathcal{Z}(\cdot;\Theta_z^\ell)$. At the beginning of training, the ControlNet pathway does not influence the pretrained network due to zero-initialized neural network parameters, and the behavior of the base diffusion model is fully preserved. Given input feature $\mathbf{v}^\ell$ and a spatial conditioning signal $\mathbf{w}'$, the activation map estimated from Stage 1, the output of the ℓ-th ControlNet-augmented block is defined as

$$\mathbf{u}_C^\ell = \mathcal{F}^\ell(\mathbf{v}^\ell;\Theta_u^\ell) + \mathcal{Z}\Big\{\mathcal{F}^\ell(\mathbf{v}^\ell + \mathcal{Z}(\mathbf{w}';\Theta_{z1}^\ell);\Theta_c^\ell);\Theta_{z2}^\ell\Big\},$$

where Θ_{z1}^ℓ and Θ_{z2}^ℓ are learnable parameters of the zero-initialized convolution layers for the l-th block. The overall predicted noise from the ControlNet-augmented model is denoted by $\epsilon_{\theta,\phi}(\mathbf{y}_t, t, \mathbf{x}; \mathbf{w}')$, where θ are frozen U-Net parameters and $\phi = \{\Theta_c^\ell, \Theta_{z1}^\ell, \Theta_{z2}^\ell\}_{\ell=1}^{L_e}$ are the trainable ControlNet parameters where L_e denotes the number of all the neural blocks in the encoder and the middle blocks. The training loss following the loss function in diffusion model becomes $\mathcal{L}_{CN} = \mathbb{E}_{\mathbf{y}_t,\epsilon,t}\left[\|\epsilon - \epsilon_{\theta,\phi}(\mathbf{y}_t, t, \mathbf{x}; \mathbf{w}')\|^2\right]$. This mechanism allows the diffusion process to align synthesized images with given activation maps while leveraging the pretrained backbone. In practice, ControlNet ensures that the generated contrast maps retain global covariate effects from Stage 2 while achieving finer spatial alignment to individualized task-related activations provided by Stage 1.

4.3 Experiments

In this section, we conduct experiments on the synthetic dataset and real fMRI data to demonstrate the efficiency of our proposed generative framework. To quantitatively compare our model to other baseline models, we evaluate the generative performance between the generated image $\widehat{\mathbf{y}}_i = \{\widehat{y}_i(s_v)\}_{v=1}^V$ and the target test image $\mathbf{y}_i = \{y_i(s_v)\}_{v=1}^V$ at the spatial locations $\mathcal{S} = \{s_v\}_{v=1}^V$ using image fidelity and similarity metrics, including the Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio

(PSNR), and Mean Squared Error (MSE). First, the MSE measures the voxelwise discrepancy between the generated and true images on the spatial locations, which is defined as $\text{MSE}(\widehat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{V}\sum_{v=1}^{V}\left(\widehat{y}_i(s_v) - y_i(s_v)\right)^2$. Additionally, we define a MSE_A on the activation region by $\text{MSE}_A(\widehat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{|R|}\sum_{s_v \in R}\left(\widehat{y}_i(s_v) - y_i(s_v)\right)^2$, where $R = \{s_v \in \mathcal{S} : w_i'(s_v) \neq 0\}$ defines the image activation region and $|R|$ is the cardinality of the activation set. This metric provides a direct measure of pixel-level reconstruction accuracy in the activation regions. Building on this, the PSNR expresses the logarithmic ratio between the maximum signal intensity and reconstruction error, offering an interpretable scale-invariant assessment of image quality, which is defined by $\text{PSNR}(\widehat{\mathbf{y}}_i, \mathbf{y}_i) = 10\log_{10}\left(\frac{\left(\max_v y_i(s_v)\right)^2}{\text{MSE}(\widehat{\mathbf{y}}_i, \mathbf{y}_i)}\right)$. A higher PSNR value corresponds to more accurate reconstructions. The SSIM [Wang et al., 2004] further evaluates structural similarity by comparing local intensity, contrast, and texture between $\widehat{\mathbf{y}}_i$ and $\mathbf{y}_i$. Let $\mu_{\hat{y}}, \mu_y$ denote the local means, $\sigma_{\hat{y}}^2, \sigma_y^2$ the local variances, and $\sigma_{\hat{y}y}$ the covariance between the generated and target images. The SSIM is then defined as $\text{SSIM}(\widehat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{(2\mu_{\hat{y}}\mu_y + C_1)(2\sigma_{\hat{y}y} + C_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2)}$, where C_1 and C_2 are small stability constants. SSIM values range from -1 to 1, with higher scores indicating better structural consistency. Finally, the FID score [Heusel et al., 2017] measures the distributional similarity between generated and real images in a learned feature space. Let $\mathcal{N}(\mathbf{m}_r, \mathbf{C}_r)$ and $\mathcal{N}(\mathbf{m}_g, \mathbf{C}_g)$ represent Gaussian distributions to the embeddings of real and generated images extracted from a pretrained encoder. The FID is computed as $\text{FID}^2 = \|\mathbf{m}_r - \mathbf{m}_g\|_2^2 + \text{Tr}\Big(\mathbf{C}_r + \mathbf{C}_g - 2(\mathbf{C}_r^{1/2}\mathbf{C}_g\mathbf{C}_r^{1/2})^{1/2}\Big)$. Lower FID values indicate that the generated images more closely match the statistical distribution of real neuroimaging data. Together, these metrics comprehensively evaluate fidelity, perceptual quality, and statistical realism of the generated images.

4.3.1 Synthetic MNIST Data with Gaussian noise

In the real fMRI setting, true individual-level activation maps are unobserved, making it difficult to validate the Bayesian nonparametric regression (BNPR) model in recovering sparse activation patterns from noisy data. Therefore, we first assess the BNPR performance in whether the estimation provides effective subject-level guidance for image generation through a simulation study using the MNIST dataset under different noise structures. MNIST consists of 28×28 grayscale images of handwritten digits from zero to nine. To mimic the key characteristic of fMRI contrast maps, we treat the handwritten digit shapes as ground-truth activation patterns $\mathbf{w}_i'$. We first corrupt each image by adding Gaussian noise with mean zero and standard deviation $\sigma = 0.5$ to mimic the high noise level in fMRI data. The resulting noisy images $\mathbf{y}_i$ represent observed contrast maps, while the original digits $\mathbf{w}_i'$ serve as the true underlying activation signals. However, in real fMRI data, background noises are

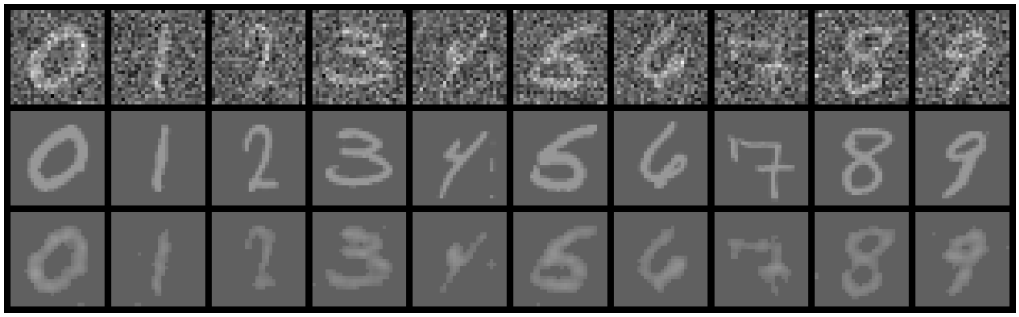
not purely independent Gaussian noise but instead contain complex spatial correlations. To better mimic this structure, we design an alternative scenario where spatial noise is generated by adding a scaled and sign-flipped combination of other digit images to each target image. Specifically, for each digit, we sample a random subset of other digits, flip their pixel intensities in sign, rescale them to 20% of their original intensity, and add them to the target image. This design introduces spatially structured interference patterns that more closely resemble the correlated background processes present in real fMRI data, thereby providing a more realistic evaluation of BNPR's ability to recover underlying activation maps under complex noise conditions.

From the complete MNIST dataset, we sample 500 training and 50 testing images from each digit class (0–9), yielding 5,000 training images in total. For each test sample, we generate 100 replicates, resulting in 50,000 synthetic images conditioned jointly on the categorical digit labels and the BNPR-estimated activation profiles. This design mimics the small-sample regime frequently encountered in fMRI studies and enables evaluation of both denoising performance and generative quality. We compare our proposed BACNet against several baselines, including a conditional DDPM using only digit labels (C-DDPM), a conditional GAN (cGAN)[Goodfellow et al., 2014], and the pix2pix image-to-image translation GAN model [Isola et al., 2017]. For implementation, we use the official publicly available code for the cGAN baselines, and adapt the diffusion-based models with U-Net network backbones from Cardoso et al. [2022].

Figure 4.2 illustrates representative training data, including noisy observations, ground-truth digits, and BNPR-estimated activation maps under two distinct noise structures. As shown in the bottom row of each scenario, BNPR effectively suppresses both Gaussian and spatially structured noise, recovering sparse and well-defined digit patterns that serve as interpretable spatial priors for subsequent generative modeling. Figure 4.3 further compares the generated test images from BACNet with those from baseline generative models under Gaussian noise settings. While baseline methods such as C-DDPM and cGAN reproduce images that align with class labels, their generated spatial features deviate from the true digit shapes. The pix2pix model, although capable of conditioning on spatial input, produces overly smoothed and blurred outputs that fail to capture realistic background variability. In contrast, BACNet generates images that most closely resemble the ground truth, accurately reproducing both signal structures and noise textures. These results demonstrate the advantage of activation-driven conditioning in preserving subject-specific spatial details and highlight the effectiveness of BNPR-guided priors for interpretable and biologically meaningful generative modeling.

Table 4.1 presents quantitative comparisons of generative performance evaluated using the

metrics including FID, SSIM, PSNR, and MSE. Across all metrics, the incorporation of spatial conditioning markedly improves generation fidelity to the target images. The proposed BACNet achieves the lowest FID and competitive SSIM and PSNR values, indicating that activation-driven conditioning enhances both global realism and local structural accuracy. Compared to other baselines, pix2pix yields relatively low MSE and high SSIM because it produces visually clean digit structures; however, its outputs are overly smoothed and fail to reconstruct realistic background noise. This limitation arises partly from its official implementation is for RGB images at 256×256 resolution, where downscaling and interpolation to the MNIST setting inevitably degrade spatial granularity. In contrast, BACNet successfully reproduces both the fine spatial details and the inherent noise characteristics of the data, generating individualized and structurally consistent synthetic images. Detailed model specifications, including BNPR conditioning design, diffusion architecture, and training hyperparameters, are provided in the Supplementary Materials.

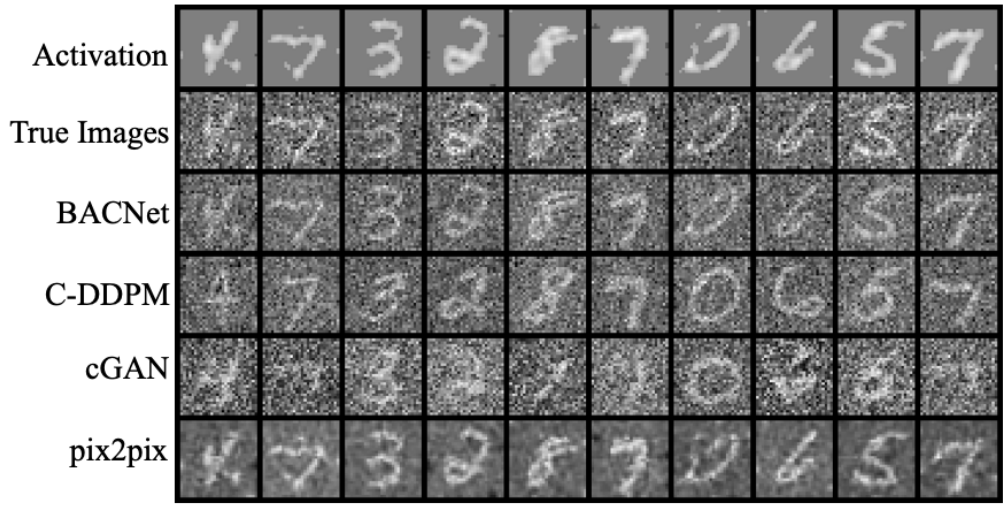


(a) Gaussian noise ($\sigma = 0.5$)

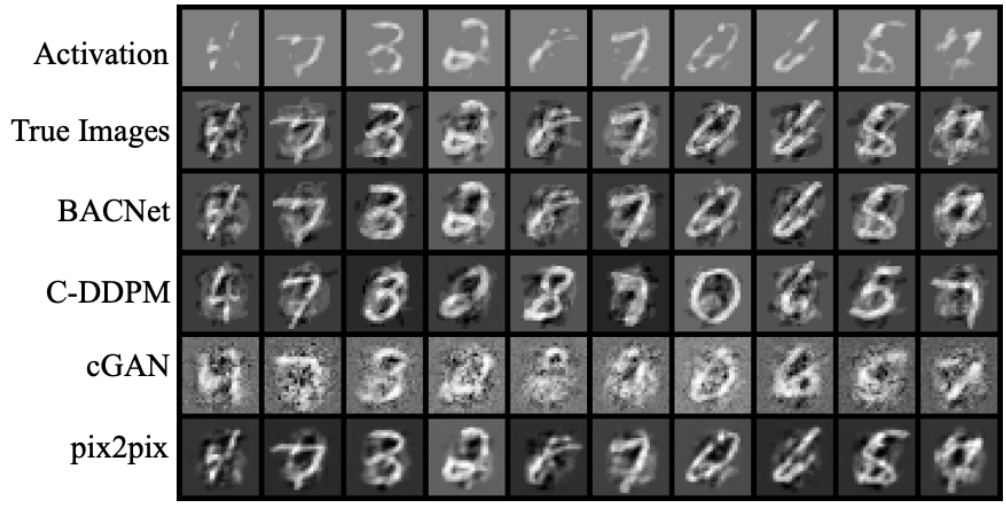


(b) Spatial Noise

Figure 4.2: MNIST synthetic images added by Gaussian noise with $\sigma = 0.5$ and spatial digit noises. Each figure shows the observed noisy images (top), original clean images with sparse continuous patterns (middle), and estimated activation region from Bayesian nonparametric regression (bottom).



(a) MNIST data with Gaussian noise ($\sigma = 0.5$)



(b) MNIST data with spatial noise

Figure 4.3: Comparison of reconstructed images from different generative models on the MNIST dataset with independent Gaussian noise and spatially digit noise. The top row shows the estimated signal maps, ground-truth images, followed by outputs from the BACNet and other comparative baseline generative models.

Table 4.1: Metric evaluation of generative performance on the MNIST simulation study, comparing the proposed model framework with baseline models. MSEs are evaluated on estimated sparse activation regions. Lower FID, MSE, and MSE_A and higher SSIM and PSNR indicate better performance.

Model	FID ↓	SSIM ↑	PSNR ↑	MSE ↓	MSE_A ↓
Gaussian noise					
cGAN	23.78	0.091	27.90	0.196	0.243
pix2pix	218.0	**0.466**	27.90	**0.077**	0.046
C-DDPM	8.280	0.099	27.93	0.162	0.226
BACNet	**7.043**	0.418	**28.04**	0.110	**0.045**
Spatial noise					
cGAN	272.1	0.246	27.91	0.037	0.227
pix2pix	58.52	**0.797**	28.65	0.178	**0.007**
C-DDPM	36.86	0.316	28.39	0.133	0.344
BACNet	**35.42**	0.771	**29.17**	**0.028**	**0.007**

4.3.2 Cortical surface fMRI generation

We next evaluate our framework on real task-based cortical surface fMRI data from the Adolescent Brain Cognitive Development (ABCD) study [Casey et al., 2018], which is a large-scale longitudinal cohort examining brain development and cognition among children. We focus on the n-back working memory task contrast maps from Release 4.0, which is widely used for studying brain–behavior associations with cognitive abilities. In this study, we combine the full data from two time points and 21 study sites, which in total gives 4,188 cortical surface fMRI contrast maps. Since cortical surface data are defined on a two-dimensional mesh, we mapped vertex-wise signals onto a 128×128 regular two-dimensional grid using surface-mesh coordinates provided by Human Connectome Project utilities [Glasser et al., 2016].

To provide conditioning information, we discretize four subject-level covariates into categorical labels: age is grouped into three categories (< 10, $[10, 11)$, and ≥ 11); sex is encoded as male or female; general cognitive ability scores are categorized into three levels (< -0.5, $[-0.5, 0.5)$, and ≥ 0.5); and psychiatric diagnostic scores are dichotomized into negative and positive values. These labels, along with the estimated activation maps, are utilized to condition the diffusion models. For evaluation, we randomly selected two samples from each unique covariate combination as test cases, and generated 100 replicates for each target image, yielding 9,600 synthetic images. Similarly as the MNIST experiment, we compare our generated images to other baseline generative models and evaluate image-based metrics

including FID, SSIM, PSNR, and MSE. Since the official implementation of cGAN only supports single categorical labels, we use GAN instead of cGAN as the baseline method in learning the fMRI data distribution.

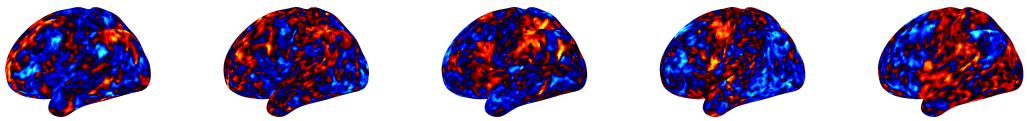
Figure 4.4 provides qualitative comparisons between the observed fMRI contrast maps, BNPR-estimated activations, and generated images obtained from various competing models. The C-DDPM results capture broad distributional features of the data but frequently generate spatial patterns that deviate from true activation regions, likely due to high inter-subject variability and the lack of spatially informative conditioning. The GAN and pix2pix models produce visually coherent but overly simplified reconstructions. While pix2pix retains a degree of spatial guidance that helps preserve the general activation patterns, its outputs tend to be smoothed, exhibiting blurred activation boundaries and diminished spatial heterogeneity. In contrast, the activation-guided DDPM generation framework consistently preserves localized activation patterns given by the BNPR-estimation patterns as spatial guidance, resulting in synthetic contrast maps that closely resemble the true target maps in both spatial extent and intensity distribution.

Table 4.2 summarizes the quantitative evaluation metrics, which support our visual findings. The MSE values are calculated averaged across all spatial locations. Among all competing methods, BACNet achieves the lowest FID and MSEs, as well as the highest SSIM and PSNR, indicating superior generative fidelity and structural accuracy. These results demonstrate that integrating subject-specific activation priors substantially enhances both the realism of synthesized fMRI contrast maps and their preservation of biologically meaningful activation patterns. Compared to C-DDPM, GAN, and pix2pix, BACNet produces more faithful reconstructions that better align with the true cognitive association effects, highlighting the importance of activation-guided conditioning for individualized, functionally interpretable fMRI generation. Collectively, these findings underscore the potential of the proposed framework as a generative tool to augment statistical inference and enable biologically grounded simulation in neuroimaging studies. We include additional generation results under different initial values in the Supplementary Materials, showing that BACNet is capable for preserving both spatial condition and fidelity in the generation process.

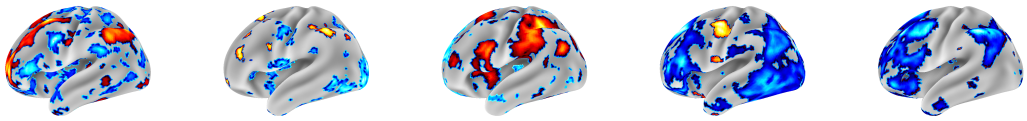
4.3.3 Hypothesis testing via synthetic null distributions

A major motivation for developing our generative framework is downstream statistical analysis, such as improving statistical inference in neuroimaging studies with limited sample size. As shown in Chapter 2, while standard hypothesis testing in massive univariate analysis can recover biologically meaningful associations in large datasets, it has substantially

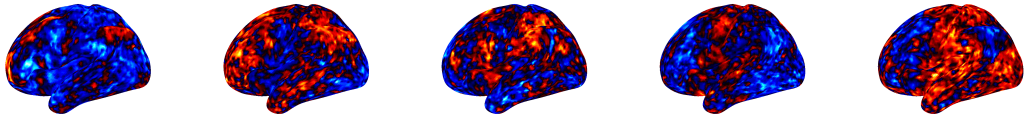
(a) True images



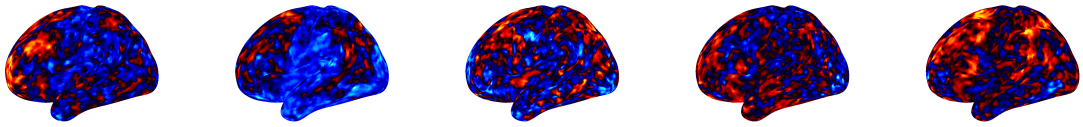
(b) Learned activation



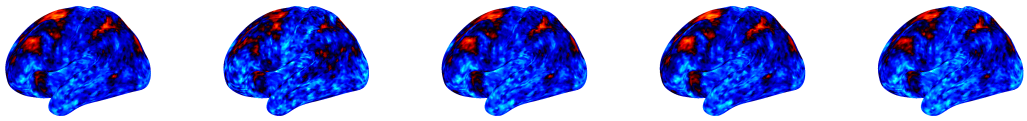
(c) BACNet



(d) C-DDPM



(e) GAN



(f) pix2pix

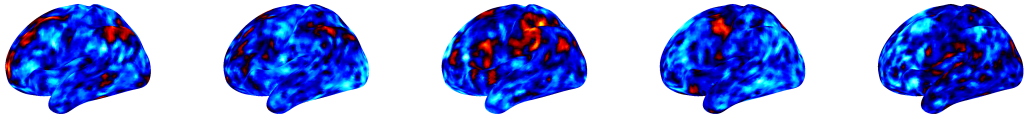


Figure 4.4: Comparison of generated fMRI contrast maps with the true target contrast maps across different generative models.

Table 4.2: Metric evaluation of generative performance on real fMRI data in the ABCD study, comparing the proposed model framework with baseline models. Lower FID, MSE, and MSE_A and higher SSIM, PSNR, and PCC indicate better performance.

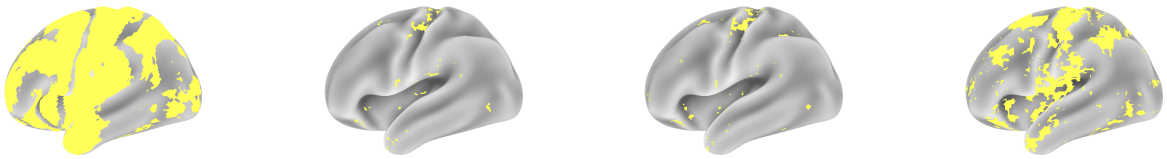
Model	FID ↓	SSIM ↑	PSNR ↑	MSE ↓	MSE_A ↓	PCC ↑
GAN	112.1	0.10	28.0	0.11	0.17	0.02
pix2pix	54.1	0.30	28.5	0.03	0.04	0.77
C-DDPM	8.85	0.41	29.9	0.02	0.04	0.95
BACNet	**7.20**	**0.51**	**30.1**	**0.01**	**0.02**	**0.96**

worse performance when applied to smaller cohorts. In such contexts, the combination of high noise levels, limited sample sizes, and the need for multiple-testing correction results in a significant loss of statistical power. Consequently, as illustrated in Chapter 3, spatially varying effects identified by classic GLM under MUA framework in cross-site analyses appear scattered, unstable, with poor reproducibility, limiting interpretability and undermining scientific conclusions.

To address this problem, we propose a generative-based hypothesis testing procedure that leverages synthetic null distributions to increase power while maintaining false discovery control. The key idea is to apply our diffusion framework to learn the distribution of contrast maps under the null hypothesis of no association. Specifically, we first estimate the group-level association effect between brain activation and cognitive ability using GLM on the full ABCD dataset, yielding an effect map $\beta_g(s_v)$ that serves as the reference true effect. To generate data under the null, we subtract this estimated effect from the observed images, thereby constructing a dataset where $\beta_g(s_v) = 0$ by design. We then train BACNet on this null dataset to learn the distribution of contrast maps under the null hypothesis.

To evaluate the method under small-sample conditions, we focus on a single study site with only 56 individuals. For each observed image, we generate $B = 1,000$ synthetic replicates under the null distribution by the proposed BACNet, producing an empirical reference distribution of test statistics. For each replicate dataset, we compute voxelwise T-statistics $T^{(b)}_{\text{null}}(s_v) = \frac{\widehat{\beta}^{(b)}_g(s_v)}{\text{SE}\left(\widehat{\beta}^{(b)}_g(s_v)\right)}$, where $\widehat{\beta}^{(b)}_g(s_v)$ is the estimated effect at voxel s_v from the b-th replication and $\text{SE}\left(\widehat{\beta}^{(b)}_g(s_v)\right)$ is its estimated standard error. These statistics collectively can used as null distributions to compare the observed voxelwise $T_{\text{obs}}(s_v) = \frac{\widehat{\beta}_g(s_v)}{\text{SE}\left(\widehat{\beta}_g(s_v)\right)}$ from the real data. Voxelwise p-values are then obtained by the empirical probability as $p(s_v) = \frac{1}{B}\sum_b \mathbb{1}\{|T^{(b)}_{\text{null}}(s_v)| > |T_{\text{obs}}(s_v)|\}$, and significance is determined using the Benjamini–Hochberg procedure at a FDR level of 0.01.

Figure 4.5 compares the results of the proposed synthetic hypothesis testing framework



(a) All-site GLM (b) Single-site GLM (c) Single-site BS-null (d) Single-site Gen-null

Figure 4.5: Identified activation regions associated with cognitive score under different analysis frameworks, including (a) reference results from the all-site full ABCD data; (b) single-site analysis; (c) single-site analysis using a bootstrap-based null distribution (d) single-site analysis using the proposed generative null approach, where null distributions are generated from the BACNet model.

under generated null distribution (Gen-null) with those obtained using a conventional GLM applied directly to a single-site dataset, and a bootstrap-based null (BS-null) resampling approach. Specifically, we repeatedly resample from the pool of constructed null images to generate 1,000 bootstrap datasets. For each resampled dataset, voxelwise test statistics are computed similarly to our proposed approach. Table 4.3 further summarizes the true positive rate (TPR) and false discovery rate (FDR) of the single-site inference results, using the all-site GLM analysis as the reference truth. While the standard GLM detects only 3% of the scattered voxel regions from the results in all-site analysis with limited spatial coherence, the proposed method identifies around 44% of those regions and are spatially contiguous while biologically plausible. Notably, these clusters substantially overlap with those detected by the conventional GLM applied to the full ABCD dataset, indicating that the synthetic null approach successfully recovers reliable activation patterns even under limited sample conditions. The additional comparison with BS-null resampling approach further confirms that the observed improvements are not attributable to random variation. Overall, these findings demonstrate that the proposed synthetic null framework provides a principled and practical approach to enhance voxelwise inference in small-sample neuroimaging studies, yielding greater statistical power, reproducibility, and spatial interpretability than standard hypothesis testing procedures.

Table 4.3: Metric comparison of brain activation detection across inference methods, using all-site GLM results as the reference truth.

Metric	GLM	BS-null	Gen-null
TPR	0.031	0.067	**0.443**
FDR	**0.014**	0.025	0.037

4.4 Discussion

In this work, we developed an activation-guided generative framework for task-based fMRI contrast maps that integrates Bayesian nonparametric regression with diffusion probabilistic models assisted by ControlNet for additional conditions. The BNPR component enables subject-level denoising and estimation of individualized activation patterns, which serve as spatial priors for generative modeling. By incorporating the spatial guidance through the ControlNet mechanism, the proposed BACNet framework produces synthetic contrast maps that preserve both individual-specific features and task-relevant functional signals, surpassing conventional label-conditioned generative approaches and providing individual-level customized generation. Through experiments on synthetic MNIST data, we demonstrated that BNPR effectively recovers sparse activation patterns under high noise and improves generation fidelity when used for spatial conditioning. Applying the method to cortical surface fMRI data from the ABCD study, we further showed that the proposed framework not only improves image fidelity and similarity to target generation but also preserves meaningful brain–behavior associations, such as the relationship between brain activity and cognitive ability. Finally, we demonstrated how the generator can be used to construct synthetic null distributions for hypothesis testing, thereby enhancing statistical power and reproducibility in small-sample studies.

Many potential directions can be built upon the proposed generation framework in future work. Methodologically, integrating more advanced conditioning mechanisms, such as cross-attention with multimodal covariates like clinical measurements or neuroimaging data such as PET scans or electroencephalography data, may further improve the alignment between generated images and subject-level characteristics. From the generative modeling perspective, incorporating recent advances such as stable diffusion or DDIM-based sampling could improve both the stability and efficiency of training while enabling high-resolution volumetric generation. From a statistical perspective, combining the generator with hierarchical regression frameworks could allow simultaneous modeling of individual- and group-level effects, bridging generative modeling with population-level inference. Forapplications, extending the generator to simulate counterfactual scenarios, such as artificial activation maps designed under certain hypothesis, augment underrepresented populations, or evaluate reproducibility across statistical methods or study populations could provide powerful tools for both clinical and methodological neuroimaging research. In brief, by integrating regression-based activation modeling with generative diffusion models, our framework demonstrates how statistical inference and generative modeling can inform and reinforce each other, leading the way for more interpretable and reproducible analyses in neuroimaging studies.

CHAPTER 5

Conclusion

This dissertation is motivated by the fundamental challenges of neuroimaging data analysis, including extreme high dimensionality, low signal-to-noise ratios, complex spatial correlations, and substantial inter-subject heterogeneity. Within this context, both Bayesian image regression and deep learning models have shown promising advancements but remain limited in complementary ways. Bayesian nonparametric regression models provide a principled modeling framework and interpretable inference but often suffer from computational costs for whole-brain analyses and limited flexibility in learning complex functions. Deep learning methods, on the other hand, can well capture the nonlinear relationships in high-dimensional imaging data, but often lack formal statistical interpretability. This work contributes to bridging this gap by developing unified frameworks that integrate Bayesian statistical inference with deep neural networks to advance both the theoretical understanding and practical applicability of statistical models in large-scale neuroimaging research.

In Chapter II, we proposed the Scalable Image Modeling using a Bayesian Approach (SIMBA), which is a spatial Bayesian hierarchical model with Gaussian process (GP) priors for whole-brain fMRI analysis. SIMBA overcomes the computational bottlenecks of traditional GP-based Bayesian models through low-rank kernel approximations, model reparameterization, and efficient Gibbs and variational inference algorithms. The model provides full posterior inference and spatial uncertainty quantification at voxel-level resolution within minutes, scaling Bayesian inference to previously infeasible whole-brain analyses. Empirical applications to the NARPS and ABCD datasets demonstrated improved spatial smoothing, interpretability, and robustness compared with classical massive univariate and hierarchical models. The implementation within neuroimaging software further ensures direct applicability in large-scale neuroimaging pipelines.

In Chapter III, we introduced the Deep Kernel Learning Process (DKLP), which integrates deep neural networks and Bayesian image regression models with GP priors. DKLP adaptively learns GP kernel representations from imaging data, replacing fixed parametric covariance functions with flexible, data-driven kernel bases that are parameterized by neural

networks. We further establish theoretical results in posterior contraction rates for both the learned kernels and spatially varying functions. Across nonparametric, image-on-scalar, and scalar-on-image regression frameworks, DKLP substantially improves estimation and spatial region selection in comparison to existing GP and other nonparametric models. Applications to ABCD cortical-surface fMRI and EEG BCI data further validate its robustness across imaging modalities and tasks, highlighting the framework's flexibility in neuroimaging studies.

In Chapter IV, we developed an activation-guided diffusion framework that bridges Bayesian regression and deep generative modeling for fMRI synthesis. We follow the denoising diffusion probabilistic models (DDPM) generation framework, augmented by ControlNet. Specifically, this method utilized the individual-specific activation maps, which are estimated via Bayesian nonparametric regression, as spatial conditions to inform diffusion models. This design introduces spatially guided generation, enabling the synthesis of individualized and anatomically consistent fMRI contrast maps. Experimental results on MNIST-based simulations and ABCD cortical-surface data show significant gains in image fidelity and preserve the biologically meaningful activation associations. Furthermore, we proposed a new synthetic null-distribution framework for hypothesis testing, which benefits from the generative model to enhance statistical power and reproducibility in small-sample neuroimaging studies.

Future research can extend this work in several important directions. First, for both volumetric and cortical data analyses, our current models rely on Euclidean distance to define spatial relationships, which may inadequately capture the intrinsic topological geometry of the cortical manifolds. Incorporating topological or geodesic distances on cortical surfaces would more accurately model spatial correlations and fully use the anatomical information of the brain structure. Second, while the proposed DKLP framework assumes spatial images are well aligned across subjects, future work could address spatial misalignment and coordinate variability. Moreover, extending the diffusion-based generative framework to volumetric fMRI data using advanced framework such as latent diffusion models and denoising diffusion implicit models could enable higher-resolution synthesis and improved representation of 3D brain structures. Finally, an important direction is to use this generative modeling framework as a benchmarking tool for downstream analyses, such as simulating realistic imaging data or construct hypothesis tests to systematically assess the robustness, calibration, and reproducibility of different statistical models under controlled conditions. Together, these directions would further extend the theoretical understanding and practical utility of Bayesian regression and deep learning modeling integration in neuroimaging studies.

APPENDIX A

Supplementary Materials for Chapter II

This supplementary document includes details of the posterior distributions of the full set of parameters: $\boldsymbol{\Theta} = \{(\alpha_r)_{r=0}^q, (\boldsymbol{\theta}_\beta)_{r=0}^q, (\boldsymbol{\theta}_{\eta_i})_{i=1}^N, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\eta^2, \sigma_\epsilon^2, a_\alpha, a_\beta, a_\eta, a_\epsilon\}$, that are used to draw samples from Gibbs sampling algorithm in Section 2.3.2, and the variational distributions in Section 2.3.3. The full code are implemented in Pytorch and included in the GitHub repository https://github.com/y1zhong/apaper_simba.

Gibbs sampler

Given the reparameterized model in (2.4) and the specifications of the conjugate priors, we present the posterior distributions for each parameter below. Denote $\mathbf{1}_V \Phi = \widetilde{\Phi}$.
The full conditional distribution of each α_r, for $r = 0, \ldots, q$ is

$$\begin{aligned}
\alpha_r \mid \text{rest} &\sim N\left(\mu_{\alpha_r}, \nu_{\alpha_r}\right); \\
\nu_{\alpha_r} &= \left(\frac{\sum_{i=1}^N x_{ir}^2 \|\widetilde{\Phi}\|_F^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_\alpha^2} \right)^{-1} \\
\mu_{\alpha_r} &= \nu_{\alpha_r} \cdot \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^N x_{ir} (\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top \boldsymbol{\theta}_\beta - \boldsymbol{\theta}_{\eta_i} \Phi_\eta) \widetilde{\Phi}^\top.
\end{aligned}$$

The full conditional distribution of each $\boldsymbol{\theta}_{\beta_r}$ for $r = 0, \ldots, q$ is

$$\begin{aligned}
\boldsymbol{\theta}_{\beta_r} \mid \text{rest} &\sim N\left(\mu_{\beta_r}, \Sigma_{\beta_r}\right); \\
\Sigma_{\beta_r} &= \left(\frac{\sum_{i=1}^N x_{ir}^2}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2} \right)^{-1} I_J \\
\mu_{\beta_r} &= \Sigma_{\beta_r} \cdot \frac{1}{\sigma_\epsilon^2} \sum_{i=1}^N x_{ir} \left(\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top \boldsymbol{\alpha} \widetilde{\Phi} - \boldsymbol{\theta}_{\eta_i} \Phi_\eta\right)
\end{aligned}$$

The full conditional distribution of each $\boldsymbol{\theta}_{\eta_i}$ for $i = 1, \ldots, N$ is

$$\begin{aligned}
\boldsymbol{\theta}_{\eta_i} \mid \text{rest} &\sim N\left(\mu_{\eta_i}, \Sigma_{\eta_i}\right); \\
\Sigma_{\eta_i} &= \left(\frac{\Phi_\eta \Phi_\eta^\top}{\sigma_\epsilon^2} + \frac{1}{\sigma_\eta^2}\right)^{-1} \\
\mu_{\eta_i} &= \Sigma_{\eta_i}\left(\frac{1}{\sigma_\epsilon^2}\left(\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top \boldsymbol{\alpha}\widetilde{\Phi} - \mathbf{x}_i^\top \boldsymbol{\theta}_\beta\right)\right)\Phi_\eta^\top
\end{aligned}$$

The full conditional distribution of each variance parameter is

$$\begin{aligned}
\sigma_\epsilon^2 \mid \text{rest} &\sim \text{IG}\left(\frac{1+NJ}{2},\ \frac{1}{2}\|\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top \boldsymbol{\alpha}\widetilde{\Phi} - \mathbf{x}_i^\top \boldsymbol{\theta}_\beta - \boldsymbol{\theta}_{\eta_i}\Phi_\eta\|_2^2 + \frac{1}{a_\epsilon}\right) \\
a_\epsilon \mid \text{rest} &\sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{\sigma_\epsilon^2}\right) \\
\sigma_\eta^2 \mid \text{rest} &\sim \text{IG}\left(\frac{1+NJ_\eta}{2},\ \frac{\sum_{i=1}^N \|\theta_{\eta_i}\|_2^2}{2} + \frac{1}{a_\eta}\right) \\
a_\eta \mid \text{rest} &\sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{\sigma_\eta^2}\right) \\
\sigma_\beta^2 \mid \text{rest} &\sim \text{IG}\left(\frac{1+(q+1)L}{2},\ \frac{\sum_{r=0}^q \|\theta_{\beta_r}\|_2^2}{2} + \frac{1}{a_\beta}\right), \\
a_\beta \mid \text{rest} &\sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{\sigma_\beta^2}\right) \\
\sigma_\alpha^2 \mid \text{rest} &\sim \text{IG}\left(\frac{1+q+1}{2},\ \frac{\sum_{r=0}^q \alpha_r^2}{2} + \frac{1}{a_\alpha}\right), \\
a_\alpha \mid \text{rest} &\sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{\sigma_\alpha^2}\right)
\end{aligned}$$

Mathematical details in VI algorithm

Under the framework of mean field variational inference, we approximate the posterior distributions using

$$\begin{aligned}
p(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y}) \approx q(\boldsymbol{\Theta}) &= \prod_{m=1}^M q(z_m) \\
&= \left(\prod_{r=0}^q q(\alpha_r) q(\boldsymbol{\theta}_{\beta_r})\right)\left(\prod_{i=1}^N q(\boldsymbol{\theta}_{\eta_i})\right) q(\sigma_\beta^2) q(\sigma_\eta^2) q(\sigma_\epsilon^2) q(a_\beta) q(a_\eta) q(a_\epsilon)
\end{aligned}$$

For any m-th latent variable, the optimal $q(z_m)$ is given by:

$$q(z_m) \propto \exp\left\{E_{-m}[\log p(z_m|z_{-m}, \mathbf{X}, \mathbf{Y})]\right\}.$$

With derivation, we state the distributions under $q(\cdot)$ for each parameter as below:

$$\begin{aligned}
&\alpha_r \mid \text{rest} \sim N\left(\mu_{\alpha_r}, \nu_{\alpha_r}\right); \\
&\qquad \nu_{\alpha_r} = \left(\frac{\sum_{i=1}^{N} x_{ir}^2 \|\widetilde{\Phi}\|_F^2}{E_q(\sigma_\epsilon^2)} + \frac{1}{E_q(\sigma_\alpha^2)}\right)^{-1} \\
&\qquad \mu_{\alpha_r} = \nu_{\alpha_r} \cdot \frac{1}{E_q(\sigma_\epsilon^2)} \sum_{i=1}^{N} x_{ir}(\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top E_q(\boldsymbol{\theta}_\beta) - E_q(\boldsymbol{\theta}_{\eta_i})\Phi_\eta)\widetilde{\Phi}^\top
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{\theta}_{\beta_r} \mid \text{rest} \sim N\left(\mu_{\beta_r}, \Sigma_{\beta_r}\right); \\
&\qquad \Sigma_{\beta_r} = \left(\frac{\sum_{i=1}^{N} x_{ir}^2}{E_q(\sigma_\epsilon^2)} + \frac{1}{E_q(\sigma_\beta^2)}\right)^{-1} I_J \\
&\qquad \mu_{\beta_r} = \Sigma_{\beta_r}\left(\frac{1}{E_q(\sigma_\epsilon^2)} \sum_{i=1}^{N} x_{ir}\left(\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top E_q(\boldsymbol{\alpha})\widetilde{\Phi} - E_q(\boldsymbol{\theta}_{\eta_i})\Phi_\eta\right)\right)
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{\theta}_{\eta_i} \mid \text{rest} \sim N\left(\mu_{\eta_i}, \Sigma_{\eta_i}\right); \\
&\qquad \Sigma_{\eta_i} = \left(\frac{\Phi_\eta \Phi_\eta^\top}{E_q(\sigma_\epsilon^2)} + \frac{1}{E_q(\sigma_\eta^2)}\right)^{-1} \\
&\qquad \mu_{\eta_i} = \Sigma_{\eta_i}\left(\frac{1}{E_q(\sigma_\epsilon^2)}\left(\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top E_q(\boldsymbol{\alpha})\widetilde{\Phi} - \mathbf{x}_i^\top E_q(\boldsymbol{\theta}_\beta)\right)\right)\Phi_\eta^\top
\end{aligned}$$

$$\sigma_\epsilon^2 \mid \text{rest} \sim \text{IG}\left(\frac{1+NJ}{2},\ \frac{1}{2}E_q\left(\|\widetilde{\mathbf{y}}_i - \mathbf{x}_i^\top \boldsymbol{\alpha}\widetilde{\Phi} - \mathbf{x}_i^\top \boldsymbol{\theta}_\beta - \boldsymbol{\theta}_{\eta_i}\Phi_\eta\|_2^2\right) + \frac{1}{E_q(a_\epsilon)}\right)$$

$$a_\epsilon \mid \text{rest} \sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{E_q(\sigma_\epsilon^2)}\right)$$

$$\sigma_\eta^2 \mid \text{rest} \sim \text{IG}\left(\frac{1+NJ_\eta}{2},\ \frac{\sum_{i=1}^N E_q(\|\theta_{\eta_i}\|_2^2)}{2} + \frac{1}{E_q(a_\eta)}\right)$$

$$a_\eta \mid \text{rest} \sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{E_q(\sigma_\eta^2)}\right)$$

$$\sigma_\beta^2 \mid \text{rest} \sim \text{IG}\left(\frac{1+(q+1)J}{2},\ \frac{\sum_{r=0}^q E_q(\|\theta_{\beta_r}\|_2^2)}{2} + \frac{1}{E_q(a_\beta)}\right)$$

$$a_\beta \mid \text{rest} \sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{E_q(\sigma_\beta^2)}\right)$$

$$\sigma_\alpha^2 \mid \text{rest} \sim \text{IG}\left(\frac{1+q+1}{2},\ \frac{\sum_{r=0}^q E_q(\alpha_r^2)}{2} + \frac{1}{E_q(a_\alpha)}\right)$$

$$a_\alpha \mid \text{rest} \sim \text{IG}\left(1,\ \frac{1}{A^2} + \frac{1}{E_q(\sigma_\alpha^2)}\right)$$

APPENDIX B

Supplementary Materials for Chapter III

B.1 Proof for Theoretical Results

In this section, we present the proof of Theorem 1, following the general theory of posterior convergence rates for Gaussian process (GP) priors developed by van der Vaart and van Zanten [2008] in regression models. We denote ϑ, ϑ_0 as the collection of all parameters and true parameters, respectively, which are defined in more detail in each proof in the later sections. Additionally, we define the Kullback-Leibler (KL) divergence and its variance at each observation as

$$\begin{aligned}
\Lambda_i(\pi, y_i; \vartheta_0, \vartheta) &= \log\{\pi(y_i; \vartheta_0)/\pi(y_i; \vartheta)\} \\
K_i(\pi, y_i; \vartheta_0, \vartheta) &= \mathbb{E}_{\vartheta_0}\{(\Lambda_i(\pi, y_i; \vartheta_0, \vartheta)\} \\
KV_i(\pi, y_i; \vartheta_0, \vartheta) &= \mathrm{Var}_{\vartheta_0}\{\Lambda_i(\pi, y_i; \vartheta_0, \vartheta)\}
\end{aligned}$$

B.1.1 Additional lemmas

First, we show that the orthonormal transformation from the Gram-Schmidt (GS) process preserves the estimation error rates by the following proposition.

Lemma 1. *Let $(f_1, \ldots, f_J)$ and $(f'_1, \ldots, f'_J)$ be two collections of functions in $L^2(\mathcal{S})$ and let $(\phi_1, \ldots, \phi_J) = \mathcal{GS}(f_1, \ldots, f_J)$, $(\phi'_1, \ldots, \phi'_J) = \mathcal{GS}(f'_1, \ldots, f'_J)$ be the orthonormal functions obtained from the Gram-Schmidt process. Let R_{j-1} denote the orthogonal projector of f_j onto $\mathrm{span}\{\phi_1, \ldots, \phi_{j-1}\}$ for each $j = 1, \ldots, J$ and define R'_{j-1} and r'_j analogously. Assume there exist constants $c_{11}, c_{12} > 0$ and such that for all $j = 1, \ldots, J, \|f_j\|_2 \leq c_{11}, \|r_j\|_2 \geq c_{12}$, where $r_j = (I - R_{j-1})f_j$. Then, if $\max_{1\leq j\leq J} \|f_j - f'_j\|_2 \lesssim \epsilon$ for some $\epsilon > 0$, we have $\max_{1\leq j\leq J} \|\phi_j - \phi'_j\|_2 \lesssim \epsilon$.*

Proof. Define $R_j = \sum_{k=1}^{j} \phi_k \phi_k^\top$ and $R'_j = \sum_{k=1}^{j} \phi'_k \phi_k'^\top$. We prove this by induction with two statements: (A_j) $\|\phi'_k - \phi_k\|_2 \lesssim \varepsilon$ for all $k \leq j$; (B_j) $\|R'_j - R_j\| \lesssim \varepsilon$.

Base case ($j = 1$): We have $r_1 = f_1$, $r'_1 = f'_1$, and hence $\phi_1 = f_1/\|f_1\|_2$, $\phi'_1 = f'_1/\|f'_1\|_2$. and we can obtain $\|\phi'_1 - \phi_1\| \leq \frac{2\|f_1 - f'_1\|}{\|f_1\| \wedge \|f'_1\|} \lesssim \varepsilon$. Additionally, because $\|R'_1 - R_1\| = \|\phi'_1 {\phi'_1}^\top - \phi_1 \phi_1^\top\| \leq 2\|\phi'_1 - \phi_1\|$, ($\mathrm{B}_1$) holds as well.

Inductive step : Assume (A_j) and (B_j) hold for all $k \leq j$. We define

$$r_{j+1} = (I - R_j) f_{j+1}, \quad r'_{j+1} = (I - R'_j) f'_{j+1}.$$

Then

$$r'_{j+1} - r_{j+1} = (I - R'_j)(f'_{j+1} - f_{j+1}) + (R_j - R'_j) f_{j+1}.$$

$$\|r'_{j+1} - r_{j+1}\|_2 \leq \|f'_{j+1} - f_{j+1}\|_2 + \|R'_j - R_j\| \, \|f_{j+1}\|_2 \lesssim \varepsilon + \|R'_j - R_j\|.$$

By normalization, we have $\phi_{j+1} = \frac{r_{j+1}}{\|r_{j+1}\|_2}$, $\phi'_{j+1} = \frac{r'_{j+1}}{\|r'_{j+1}\|_2}$. Using again the bound for normalized differences,

$$\|\phi'_{j+1} - \phi_{j+1}\|_2 \lesssim \|r'_{j+1} - r_{j+1}\|_2 \lesssim \varepsilon + c_{11} \|R'_j - R_j\| \lesssim \varepsilon,$$

Finally, we have

$$R_{j+1} = R_j + \phi_{j+1} \phi_{j+1}^\top, \quad R'_{j+1} = R'_j + \phi'_{j+1} \phi'^\top_{j+1},$$

Then,

$$\|R'_{j+1} - R_{j+1}\| \leq \|R'_j - R_j\| + \|\phi'_{j+1} \phi'_{j+1} - \phi_{j+1} \phi_{j+1}^\top\| \leq \|R'_j - R_j\| + 2\|\phi'_{j+1} - \phi_{j+1}\|_2.$$

Applying the inductive bounds (A_{j+1}) and (B_j), we get $\|R'_{j+1} - R_{j+1}\| \lesssim \varepsilon$, which completes the proof. □

For any orthonormal system $(\phi_{01}, \ldots, \phi_{0J}) \subset \mathcal{H}_d^\beta(\mathcal{S}) \subset L^2(\mathcal{S})$, there exists a sequence $(f_{01}, \ldots, f_{0J}) \subset \mathcal{H}_d^\beta(\mathcal{S})$ such that $\mathcal{GS}(f_{01}, \ldots, f_{0J}) = (\phi_{01}, \ldots, \phi_{0J})$. This follows directly from the construction of the Gram-Schmidt process, which involves only finite linear combinations and scalar normalizations that preserve Hölder regularity. In subsequent proofs, we approximate the generating functions $(f_{01}, \ldots, f_{0J})$ by deep neural network (DNN) functions and then apply Lemma 1 to induce the resulting approximation rates to the true orthonormal basis functions.

Lemma 2. *Let $f_0 = \sum_{j=1}^\infty \theta_{0j} \lambda_{0j}^{1/2} \phi_{0j}(s_v) \in \Theta$ that satisfies Assumption 1. Then, for any well-defined function $f = \sum_{j=1}^J \phi_j \lambda_j^{1/2} \theta_j$ in $L^2(\mathcal{S})$ with a truncation integer $J \in \mathbb{N}^+$, we have*

$$\|f_0 - f\|_V \lesssim \sum_{j \leq J} \left(\|\phi_{0j} - \phi_j\|_V + |\lambda_{0j}^{1/2} - \lambda_j^{1/2}| + |\theta_{0j} - \theta_j| \right) \; + \; J^{-\alpha/d}.$$

Proof. By definition of f_0, we decompose the functional difference by the different in each basis terms and the tail as

$$\|f_0 - f\|_V \leq \underbrace{\left\|\sum_{j=1}^{J}\left(\theta_{0j}\lambda_{0j}^{1/2}\phi_{0j} - \theta_j\lambda_j^{1/2}\phi_j\right)\right\|_V}_{=:T_1} + \underbrace{\left\|\sum_{j>J}\theta_{0j}\lambda_{0j}^{1/2}\phi_{0j}\right\|_V}_{=:T_2}.$$

For each $j \leq J$, expand

$$\theta_{0j}\lambda_{0j}^{1/2}\phi_{0j} - \theta_j\lambda_j^{1/2}\phi_j = (\theta_{0j} - \theta_j)\lambda_{0j}^{1/2}\phi_{0j} + \theta_j(\lambda_{0j}^{1/2} - \lambda_j^{1/2})\phi_{0j} + \theta_j\lambda_j^{1/2}(\phi_{0j} - \phi_j).$$

Therefore,

$$\begin{aligned}\|T_1\|_V &\leq \left\|\sum_{j=1}^{J}(\theta_{0j} - \theta_j)\lambda_{0j}^{1/2}\phi_{0j}\right\|_V + \left\|\sum_{j=1}^{J}\theta_j(\lambda_{0j}^{1/2} - \lambda_j^{1/2})\phi_{0j}\right\|_V + \left\|\sum_{j=1}^{J}\theta_j\lambda_j^{1/2}(\phi_{0j} - \phi_j)\right\|_V \\ &\lesssim \sqrt{J}\max_{1\leq j\leq J}\left(\|\phi_j - \phi_{0j}\|_V + |\lambda_j^{1/2} - \lambda_{0j}^{1/2}| + |\theta_j - \theta_{0j}|\right)\end{aligned}$$

From Assumption 1, the eigenvalues have a decay rate $\lambda_{0j} \sim j^{-2\alpha/d-1}$ and bounded basis functions and coefficients, the tail is bounded by:

$$T_2 = \left\|\sum_{j>J}\theta_{0j}\lambda_{0j}^{1/2}\phi_{0j}\right\|_V \lesssim \left(\sum_{j>J} j^{-2\alpha/d-1}\right)^{1/2} \lesssim J^{-\alpha/d}.$$

□

Lemma 2 decomposes the approximation errors into four components, including basis error $\|\phi_{0j} - \phi_j\|_V$, eigenvalues error $|\lambda_{0j}^{1/2} - \lambda_j^{1/2}|$, basis coefficient error $|\theta_{0j} - \theta_j|$, and the truncation tail $\left(\sum_{j>J}\lambda_{0j}\theta_{0j}^2\right)^{1/2} \lesssim J^{-\alpha/d}$. Setting $J^{-\alpha/d} \lesssim \epsilon$ leads to the requirements of $J \gtrsim \epsilon^{-d/\alpha}$ such that the truncation tail is bounded.

Lemma 3. *Fix $c_2 \in (0, 1/2)$ and let $\rho_0 = \text{logit}^{-1}(\zeta_0)$ and $\rho = \text{logit}^{-1}(\zeta)$ with $\rho_0, \rho \in [c_2, 1 - c_2]$. For $\Lambda(Y; \rho_0, \rho) = Y\log\frac{\rho_0}{\rho} + (1 - Y)\log\frac{1-\rho_0}{1-\rho}$, where $Y \sim \text{Bern}(\rho_0)$, we have,*

$$\mathbb{E}_p[\Lambda(Y; \zeta_0, \zeta)] \lesssim (\zeta - \zeta_0)^2,$$

$$\text{Var}_p\big(\Lambda(Y; \zeta_0, \zeta)\big) \lesssim (\zeta - \zeta_0)^2.$$

Proof. Define a function $g(\cdot)$ as

$$\begin{aligned} g(\zeta) &= \log(1+e^{\zeta}) = -\log(1-\rho) \\ g'(\zeta) &= \frac{e^{\zeta}}{1+e^{\zeta}} = \rho \\ g''(\zeta) &= \frac{e^{\zeta}}{1+e^{\zeta}}\frac{1}{1+e^{\zeta}} = \rho(1-\rho) \end{aligned}$$

We have

$$\begin{aligned} \mathbb{E}_{\rho_0}[\Lambda(Y;\rho_0,\rho)] &= \rho_0 \log\Big(\frac{\rho_0}{\rho}\Big) + (1-\rho_0)\log\Big(\frac{1-\rho_0}{1-\rho}\Big) \\ &= \log(1-\rho_0) - \log(1-\rho) - \rho_0 \log\Big(\frac{\rho}{1-\rho}\Big) + \rho_0 \log\Big(\frac{\rho_0}{1-\rho_0}\Big) \\ &= g(\zeta) - g(\zeta_0) - g'(\zeta_0)(\zeta-\zeta_0) \end{aligned}$$

By Taylor's expansion, we have

$$g(\zeta) = g(\zeta_0) + g'(\zeta_0)(\zeta-\zeta_0) + \frac{1}{2}g''(\xi)(\zeta-\zeta_0)^2$$

for some ξ between ζ and ζ_0. Since $g''(\xi) = \xi(1-\xi)$ and $\xi \in (0,1)$. Then $g''(\xi) \le \frac{1}{4}$. Therefore

$$\mathbb{E}_{\rho_0}[\Lambda(Y;\rho_0,\rho)] = \frac{1}{2}g''(\xi)(\zeta-\zeta_0)^2 \lesssim (\zeta-\zeta_0)^2$$

Since we have $\log\frac{\rho_0}{1-\rho_0} = \zeta_0$ and $\log\frac{\rho}{1-\rho} = \zeta$,

$$\log\frac{\rho_0}{\rho} - \log\frac{1-\rho_0}{1-\rho} = \log\frac{\rho_0(1-\rho)}{(1-\rho_0)\rho} = \zeta_0 - \zeta = -(\zeta-\zeta_0).$$

Hence

$$\Lambda(Y;\zeta_0,\zeta) = (Y-\rho_0)\,(\zeta_0-\zeta) + \mathbb{E}_{\rho_0}\big[\Lambda(Y;\zeta_0,\zeta)\big].$$

Since $\mathrm{Var}_{\rho_0}(Y) = \rho_0(1-\rho_0) \le 1/4$, we obtain

$$\mathrm{Var}_{\rho_0}(\Lambda(Y;\zeta_0,\zeta)) = \mathrm{Var}_{\rho_0}(Y)\,(\zeta-\zeta_0)^2 \;\le\; \frac{1}{4}\,(\zeta-\zeta_0)^2.$$

□

To approximate $f \in \Theta$ in the parameter space, we define the sieve as $\mathcal{F}_{(V,\lambda,\theta)} = \mathcal{F}_V^{\mathrm{DNN}} \times$

$\mathcal{F}_\lambda \times \mathcal{F}_\theta$, where

$$\mathcal{F}_V^{\text{DNN}} = \left\{ f_{\mathbf{B}}^{\text{DNN}} : f_{\mathbf{B}}^{\text{DNN}} \in \mathcal{F}^{\text{DNN}}(L_V, \mathbf{h}_V) \right\}$$
$$\mathcal{F}_\lambda = \left\{ (\lambda_1, \ldots, \lambda_J) : \ c_{31}\, \lambda_0 \ \leq \ \lambda_j \ \leq \ c_{32} \lambda_0, \text{ for all } j \leq J \right\}$$
$$\mathcal{F}_\theta = \left\{ (\theta_1, \ldots, \theta_J) : \ \sum_{j=1}^{J} \lambda_j \theta_j^2 \leq c_{33}, \quad \|\theta\|_\infty \leq c_{34} \right\}$$

Then, the induced function sieve is

$$\mathcal{F}_V = \left\{ f : f = \sum_{j=1}^{J} \phi_j \lambda_j^{1/2} \theta_j, \quad (\phi_1, \ldots, \phi_J) = \mathcal{GS}(f_{\mathbf{B}}^{\text{DNN}}), \quad (f_{\mathbf{B}}^{\text{DNN}}, \lambda, \theta) \in \mathcal{F}_{(V,\lambda,\theta)} \right\}$$

B.1.2 Proof of Theorem 1

We prove the theorem by verifying the three conditions from Ghosal and van der Vaart [2007], and following the strategy outlined in Kong and Kim [2024] to show the contraction rate for DNN functions, while specifically adapting it to our model framework. Let $\pi(y; \vartheta)$ denote a generic density function for y given a set of parameters ϑ.

Define the sieve $\mathcal{F}_{\text{NP}} = \Pi_{i=1}^n \mathcal{F}_{w,i}$, where $\mathcal{F}_{w,i} = \{w_i : w_i \in \mathcal{F}_V\}$. For $\epsilon > 0$, we denote the ϵ-neighborhood around the true value as:

$$A_\epsilon = \left\{ (w_i)_{1 \leq i \leq n} \in \mathcal{F}_{\text{NP}} : n^{-1} \sum_{i=1}^{n} \|w_i - w_{i0}\|_V \leq \epsilon \right\}$$

Let $\vartheta_0 = \{w_{i0}\}_{i=1}^n$ under true basis parameters $\{\phi_{0j}, \lambda_{0j}, \theta_{0ij}\}_{i,j=1}^{n,J}$ and $\vartheta = \{w_i\}_{i=1}^n$ under parameters in the sieve. Denote the density

$$\pi_{\text{NP}}(y_i; \vartheta) = \sum_{v=1}^{V} \left\{ -\frac{1}{2} \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} (y_i(s_v) - w_i(s_v)) \right\}^2.$$

For a single (i, v), we have $\Lambda_{iv}(\pi_{\text{NP}}, y_i(s_v); \vartheta_0, \vartheta) = \log \frac{\pi_{\vartheta_0}(y_i(s_v))}{\pi_\vartheta(y_i(s_v))}$. For Gaussian means with known variance,

$$\mathbb{E}_{\vartheta_0}\big[\Lambda_{iv}\big] \ = \ \frac{\{w_{i0}(s_v) - w_i(s_v)\}^2}{2\sigma_\epsilon^2}, \qquad \text{Var}_{\vartheta_0}\big(\Lambda_{iv}\big) \ = \ \frac{\{w_{i0}(s_v) - w_i(s_v)\}^2}{\sigma_\epsilon^2}. \tag{B.1}$$

Lemma 4. *For $\epsilon > 0$, we have*

$$\sup_{\epsilon > \epsilon_{Vn}} \log \mathcal{N}\left(\frac{\epsilon}{36}, A_\epsilon, \|\cdot\|_\infty\right) \lesssim (V+n)\epsilon_{Vn}^2.$$

Proof. For true basis function $(\phi_{0j})_{1\le j\le J}$ that satisfies Assumption 1, we denote the generating function $(\phi_{01},\dots,\phi_{0J}) = \mathcal{GS}(f_{01},\dots,f_{0J})$ and $(\phi_1,\dots,\phi_J) = \mathcal{GS}(f_{\mathbf{B},1}^{\mathrm{DNN}},\dots,f_{\mathbf{B},J}^{\mathrm{DNN}})$, where $f_{0j} \in \mathcal{H}_d^\beta(\mathcal{S})$ for each $j = 1,\dots,J$. From Lemma 1, we have $\{(\phi_j)_{j\le J} : \|\phi_{j0} - \phi_j\|_V \lesssim \epsilon\} \subset \left\{f_{\mathbf{B}}^{\mathrm{DNN}} : \|f_{\mathbf{B},j}^{\mathrm{DNN}} - f_{0j}\|_V \lesssim \epsilon\right\}$. If $\max_{j\le J}\|f_{\mathbf{B},j}^{\mathrm{DNN}} - f_{0j}\|_V \le \frac{\epsilon}{3\sqrt{J}}$, $\max_{j\le J}|\lambda_j^{1/2} - \lambda_{0j}^{1/2}| \le \frac{\epsilon}{3\sqrt{J}}$, and $\max_{j\le J}\left|\theta_{ij} - \theta_{0ij}\right| \le \frac{\epsilon}{3\sqrt{J}}$, with $J \gtrsim \epsilon^{-d/(\alpha\wedge\beta)}$. This implies $n^{-1}\|w_i - w_{i0}\|_V \lesssim \epsilon$ by Lemma 2. Hence, we can cover A_ϵ by the product of the covers of $\mathcal{F}_V^{\mathrm{DNN}}, \mathcal{F}_\lambda, \mathcal{F}_\theta$. Therefore, we can bound the covering number of A_ϵ via the covering numbers of each component space:

$$\begin{aligned}
\log\mathcal{N}\left(\frac{\epsilon}{36}, A_\epsilon, \|\cdot\|_\infty\right) &\le \log\mathcal{N}\left(\frac{\epsilon}{108\sqrt{J}}, \mathcal{F}_V^{\mathrm{DNN}}, \|\cdot\|_\infty\right) \\
&\quad + \log\mathcal{N}\left(\frac{\epsilon}{108\sqrt{J}}, \mathcal{F}_\lambda, \|\cdot\|_\infty\right) + \log\mathcal{N}\left(\frac{\epsilon}{108\sqrt{J}}, \mathcal{F}_\theta, \|\cdot\|_\infty\right) \\
&\lesssim JL_V^2 h_V^2 \log(L_V r_V^2)\log 3\left(\frac{216\sqrt{J}eF}{\epsilon}\log\frac{324\sqrt{J}eF}{\epsilon}\right) \\
&\quad + n\epsilon^2\log\frac{108\sqrt{J}}{\epsilon} + n\epsilon^2\log\frac{108\sqrt{J}}{\epsilon}
\end{aligned}$$

By the choice of $L_V \lesssim \log V$, $\quad h_V \lesssim V^{\frac{d}{2(2\beta+d)}}$, and $J \gtrsim \epsilon^{-d/(\alpha\wedge\beta)}$, we have

$$\sup_{\epsilon > \epsilon_{Vn}} \log\mathcal{N}\left(\frac{\epsilon}{36}, A_\epsilon, \|\cdot\|_\infty\right) \lesssim (V+n)\epsilon_{Vn}^2,$$

which completes the proof. □

Lemma 5. *(Prior Mass) Define*

$$A^*(\vartheta_0, \epsilon_{Vn}) = \{\vartheta : \mathbb{E}_{\vartheta_0}\{\Lambda_{i,v}(\pi_{\mathrm{NP}}, y_i; \vartheta_0, \vartheta)\} \le \epsilon_{Vn}^2, \mathrm{Var}_{\vartheta_0}\{\Lambda_{i,v}(\pi_{\mathrm{NP}}, y_i; \vartheta_0, \vartheta) \le \epsilon_{Vn}^2\}.$$

Then, there exists $c_4 > 0$ such that

$$\Pi(A^*(\vartheta_0, \epsilon_{Vn})) \ge e^{-c_4(V+n)\epsilon_{Vn}^2}$$

Proof. For $\epsilon > 0$, we define

$$B^*(\vartheta_0, \epsilon_{Vn}) = \{\vartheta : \max_{1\le j\le J} \|f^{\text{DNN}}_{\mathbf{B},j} - f_{0j}\|_V \le \frac{\epsilon}{6\sqrt{J}}, \\ \max_{1\le j\le J} |\lambda_j - \lambda_{0j}| \le \frac{\epsilon}{6\sqrt{J}}, \max_{1\le j\le J} \|\theta_j - \theta_{0j}\|_n \le \frac{\epsilon}{6\sqrt{J}}\}$$

Then, by Lemma 1 and Lemma 2, for every parameter set $\vartheta \in B^*(\vartheta_0, \epsilon_{Vn})$, and for Gaussian regression likelihood with known variance,

$$\begin{aligned} K_{iv}(\vartheta_0, \vartheta) &\lesssim \frac{1}{2\sigma_\epsilon^2}(w_{i0}(s_v) - w_i(s_v))^2 \\ &\lesssim \sum_{j=1}^{J} \left(\|f_{0j} - f^{\text{DNN}}_{\mathbf{B},j}\|_V^2 + |\lambda_{0j}^{1/2} - \lambda_j^{1/2}|^2 + \|\theta_{0ij} - \theta_{ij}\|_2^2 \right) + J^{-\alpha/d} \lesssim \epsilon^2 \\ KV_{iv}(\vartheta_0, \vartheta) &\lesssim \text{Var}_{\vartheta_0} \left(-\frac{(y_i(s_v) - w_{i0}(s_v))^2}{2\sigma_\epsilon^2} + -\frac{(y_i(s_v) - w_i(s_v))^2}{2\sigma_\epsilon^2} \right) \\ &\lesssim \text{Var}_{\vartheta_0} \left((w_{i0}(s_v) - w_i(s_v) Z_i \right) \lesssim \epsilon^2 \end{aligned}$$

where $Z_i = \frac{y_i(s_v) - w_{i0}(s_v)}{\sigma_\epsilon^2} \sim N(0,1)$. Therefore, we have $B^*(\vartheta_0, \epsilon_{Vn}) \subset A^*(\vartheta_0, \epsilon_{Vn})$.

It remains to lower bound the prior mass on $B^*(\vartheta_0, \epsilon_{Vn})$. We borrow Theorem 1 from Kong and Kim [2024] and standard results for fully-connected ReLU networks imply that for any J-dimensional true functions $\{\phi_{0j}\}$ in the Hölder space, there exists $\widehat{\mathbf{B}}$ with $\max_{j\le J} \|f^{\text{DNN}}_{\widehat{\mathbf{B}},j} - f_{0j}\|_V \le \epsilon/(12\sqrt{J})$ and the parameter ball

$$C_V^*(\hat{\mathbf{B}}) = \left\{ \mathbf{B} : \ \|\mathbf{B} - \widehat{\mathbf{B}}\|_\infty \le \frac{\epsilon}{12\sqrt{J}\, a(d+1)(h_V+1)^{L_V} c_5^{L_V}(L_V+1)} \right\}.$$

Since each parameter in $\mathbf{B}$ follow independent normal distribution, there exists a constant $\gamma_1 > 0$ such that

$$\Pi\left(C_V^*(\hat{\mathbf{B}})\right) \ge \gamma_1^{p_V} \left(\frac{\epsilon}{12\sqrt{J} a(d+1)(h_V+1)^{L_V} c_5^{L_V}(L_V+1)} \right)^{p_V}$$

for some $\gamma_1 > 0$. The product priors on $(\lambda_j)_{j\le J}$ and (θ_{ij}) give polynomial small-ball probabilities in $\epsilon/\sqrt{J}$. Combining these factors and the growth regimes for $L_V, \mathbf{h}_V, J$ as in Lemma 4, we obtain the joint lower bound for some $c_4 > 0$

$$\Pi\big(A^*(\vartheta_0, \epsilon_{Vn})\big) \ \ge \ \Pi(B^*(\vartheta_0, \epsilon_{Vn})) \ge \exp\left\{-c_4\,(V+n)\,\epsilon_{Vn}^2\right\} \tag{B.2}$$

□

Lemma 6. *We have*

$$\frac{\Pi(\mathcal{F}_V^C)}{\Pi(A^*(\vartheta_0, \epsilon_{Vn}))} = o(e^{-2(V+n)\epsilon_{Vn}^2})$$

Proof. We can obtain this result from proofs in Lemma 5. □

B.1.3 Proof of Theorem 2

In this section, we study the posterior concentration rate for the ISR model. For simplicity, we assume $q = 1$ for both ISR and SIR models. The results can be easily extended to $q > 1$, as q is typically finite and small in practice. Denote $\varphi(\mu, \sigma^2)$ as the density for a normal distribution with mean μ and variance σ^2. Since the ISR framework models the sparsity of $\beta(s_v) = \tilde{\beta}(s_v)\delta(s_v)$ through unobserved selection indicator, we work on the density function that integrates $\delta(s_v)$ and get the its probability function $\rho(s_v)$.

$$\pi_{(i,v)}^{\text{IS}}(y_i) = \rho_v \varphi(y_i(s_v); x_i\tilde{\beta}_v + \eta_{iv}, \sigma_\epsilon^2) + (1 - \rho_v)\varphi(y_i(s_v); \eta_{iv}, \sigma_\epsilon^2) \tag{B.3}$$

Let $\vartheta = \{\tilde{\beta}, \rho, \{\eta_i\}_{i=1}^n\}$ and the set of true parameters are denoted by $\vartheta_0 = \{\tilde{\beta}_0, \rho_0, \{\eta_{i0}\}_{i=1}^n\}$. In addition, let μ_i denote the mean function under θ, and μ_{i0} denote the mean function under the true parameters.

We define the sieve as $\mathcal{F}_{\text{IS}} = \mathcal{F}_{\tilde{\beta}} \times \mathcal{F}_{\zeta_r} \times \left(\Pi_{i=1}^N \mathcal{F}_{\eta,i}\right)$

$$\mathcal{F}_{\tilde{\beta}} = \{\tilde{\beta} : \tilde{\beta} \in \mathcal{F}_V\}, \quad \mathcal{F}_\zeta = \{\zeta : \zeta \in \mathcal{F}_V\}, \quad \mathcal{F}_{\eta_i} = \{\eta_i : \eta_i \in \mathcal{F}_V\}$$

Similarly as Theorem 1, we prove the concentration rate by verifying the three conditions. For $\epsilon > 0$, we denote the ϵ-neighborhood around the true value as:

$$A_\epsilon^{\text{IS}} = \{(\tilde{\beta}, \zeta, \eta_i)_{1\le i\le n} \in \mathcal{F}_{\text{IS}} : n^{-1}\|\eta_i - \eta_{i,0}\|_V + \|\tilde{\beta} - \tilde{\beta}_0\|_V + \|\zeta - \zeta_0\|_V \le \epsilon\}$$

Since $\tilde{\beta}$, ζ, and η_i share the same sieve, we can treat these $\tilde{\beta}$ and ζ two additional functions such that we can directly apply Lemma 5 with $n+2$ functions. Therefore, the entropy bound remains the same as Lemma 4.

Lemma 7. *(Prior mass)*

$$A_{Vn}^{*\text{IS}}(\vartheta_0, \epsilon_{Vn}) = \Big\{\vartheta : (Vn)^{-1}\sum_{i=1}^N\sum_{v=1}^V \mathbb{E}_{\vartheta_0}\{\Lambda_{i,v}(\vartheta_0, \vartheta) \le \epsilon_{Vn}^2, \\ (Vn)^{-1}\sum_{i=1}^N\sum_{v=1}^V \text{Var}_{\vartheta_0}\{\Lambda_{i,v}(\vartheta_0, \vartheta) \le \epsilon_{Vn}^2\Big\}.$$

Then, we have

$$\Pi(A_{\epsilon_{Vn}}) \geq e^{-(V+n)\epsilon_{Vn}^2}$$

Similarly, for $\epsilon > 0$, we define

Let B_ρ denote the density for Bernoulli distribution with probability ρ. By the chain rule of KL-divergence, we have

$$\begin{aligned} K_i(\vartheta_0, \vartheta) &\leq KL_i(B_\rho || B_{\rho_0}) \\ &\quad + \|\rho_0\| \|\tilde{\beta}_0(s_v) - \tilde{\beta}(s_v)\|_V^2 \|x_i\|^2 + \frac{1}{\sigma_\epsilon^2} \|\eta_{i0} - \eta_i\|_V^2 \\ &\lesssim \|\zeta - \zeta_0\|_V^2 + \|\tilde{\beta}_0(s_v) - \tilde{\beta}(s_v))\|_V^2 + \|\eta_{i0} - \eta_i\|_V^2 \end{aligned}$$

And since all parameter set $\zeta, \tilde{\beta}, \eta_i$ share the same sieve and prior probabilities, we obtain the results similar as Lemma5.

B.1.4 Proof of Theorem 3

Proof. Similarly as ISR model, we consider the integrated likelihood of u_i given $y_i(s_v)$ and $\delta_\alpha(s_v)$. Define the density of SIR model is

$$\begin{aligned} \pi_i^{\mathrm{SI}}(u_i) = &\sum_{\delta_\alpha(s_v) \in \{0,1\}^V} \left(\prod_{v=1}^V \rho_{\alpha(s_v)}^{\delta_\alpha(s_v)} (1 - \rho_{\alpha(s_v)})^{(1-\delta_\alpha(s_v))} \right) \varphi(u_i; \frac{1}{V} \sum^V y_i(s_v) \tilde{\alpha}_v \delta_\alpha(s_v) + x_i \xi, \sigma_u^2) \\ &\times \sum_{v=1}^V \left\{ -\frac{1}{2} \log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} (y_i(s_v) - w_i(s_v)) \right\}^2 \end{aligned}$$

Let $\mu_i(\delta_\alpha) = \frac{1}{V} \sum_{v=1}^V y_i(s_v) \alpha(s_v) \delta_\alpha(s_v) + x_i^\top \xi$. Let $\vartheta_0 = (\alpha_0, \zeta_0, \xi_0, \{w_{i0}\}_{i=1}^n)$ denote the true parameter, with $\rho_{0v} = \mathrm{logit}^{-1}\{\zeta_0(s_v)\}$ and $\mu_{i0}(\delta) = \frac{1}{V} \sum_{v=1}^V y_i(s_v) \alpha_0(s_v) \delta_v + x_i^\top \xi_0$. The sieve is defined as $\mathcal{F}_{Vn}^{\mathrm{SI}} = \mathcal{F}_{\tilde{\alpha}} \times \mathcal{F}_\zeta \times \left(\Pi_{i=1}^N \mathcal{F}_{w,i} \right)$

$$\mathcal{F}_{\tilde{\alpha}} = \{\tilde{\alpha} : \tilde{\alpha} \in \mathcal{F}_V\}, \quad \mathcal{F}_\zeta = \{\zeta : \zeta \in \mathcal{F}_V\}, \quad \mathcal{F}_{w_i} = \{w_i : w_i \in \mathcal{F}_V\}$$

Similarly, we only need to show the KL bound to verify prior mass condition;

$$\begin{aligned} K_i(\vartheta_0, \vartheta) \quad &\leq K_i(\pi_{NP}, y_i; \vartheta_0, \vartheta) + \sum_{v=1}^V KL(B_{\rho_{0v}} \| \, B_{\rho_v}) \\ &\quad + \mathbb{E}_{\delta \sim \prod_v \mathrm{B}(\rho_{0v})} \Big[KL_i\big(N(\mu_{i0}(\delta), \sigma_u^2) \, \| \, N(\mu_i(\delta), \sigma_u^2) \big) \Big] \end{aligned}$$

Since by Theorem 1, we already have $K_i(\pi_{NP}, y_i; \vartheta_0, \vartheta) \lesssim \epsilon_{Vn}$, we only need to bound the last two terms.

$$\begin{aligned} KL_i\big(N(\mu_{i0}(\delta), \sigma_u^2) \parallel N(\mu_i(\delta), \sigma_u^2)\big) &\lesssim \frac{1}{\sigma_u^2}(\mu_i(\delta) - \mu_i(\delta_0))^2 \\ &\lesssim \|\tilde{\alpha} - \tilde{\alpha}_0\|_V^2 \end{aligned}$$

By Lemma 3, we have $V^{-1}\sum_{v=1}^{V} KL(B_{\rho_{0v}} \parallel B_{\rho_v}) \lesssim \|\zeta_\alpha - \zeta_{\alpha 0}\|_V^2$. Therefore, we can derive the three conditions similar as the proof for Theorem 2.

□

B.1.5 Proof of Corollary 1

To prove the convergence rate of the covariance kernel function following the DKLP prior, we first introduce the notation for functions with different combinations of basis function and eigenvalues. Denote the true kernel function as $K_0(s, s') = \sum_{j=1}^{\infty} \lambda_{j0}\phi_{j0}(s)\phi_{j0}(s')$, the true kernel function truncated at J as $K_{J0}(s, s') = \sum_{j=1}^{J} \lambda_{j0}\phi_{j0}(s)\phi_{j0}(s')$, and the kernel with true eigenfunctions and estimated eigenvalues truncated at J terms as $K_{J*}(s, s') = \sum^{J} \lambda_j\phi_{j0}(s)\phi_{j0}(s')$, and the estimated kernel function following the DKLP prior as $K_J(s, s') = \sum^{J} \lambda_j\phi_j(s)\phi_j(s')$. Note that we assume the true eigenfunctions and eigenvalues are bounded and satisfy $\sum_{j=1}^{\infty} \lambda_j < \infty$. Then, for some constant M_ϕ, we have

$$\begin{aligned} \|K_J - K_0\|_2 &\leq \|K_J - K_{J*}\|_2 + \|K_{J*} - K_{J0}\|_2 + \|K_{J0} - K_0\|_2 \\ &\leq 2 \times M_\phi \left\| \sum_{j=1}^{J} (\phi_j - \phi_{j0}) \right\|_2 + M_\phi^2 \left\| \sum_{j=1}^{J} (\lambda_j - \lambda_{j0}) \right\|_2 + M_\phi^2 \sum_{j>J} \lambda_{j0} \\ &\lesssim \sqrt{J} \left(V^{\frac{-2\beta}{2\beta+d}} \log^\gamma V + n^{-1/2} \right). \end{aligned}$$

By selecting the truncation number J as above, we obtain the approximate rate of the covariance kernel function.

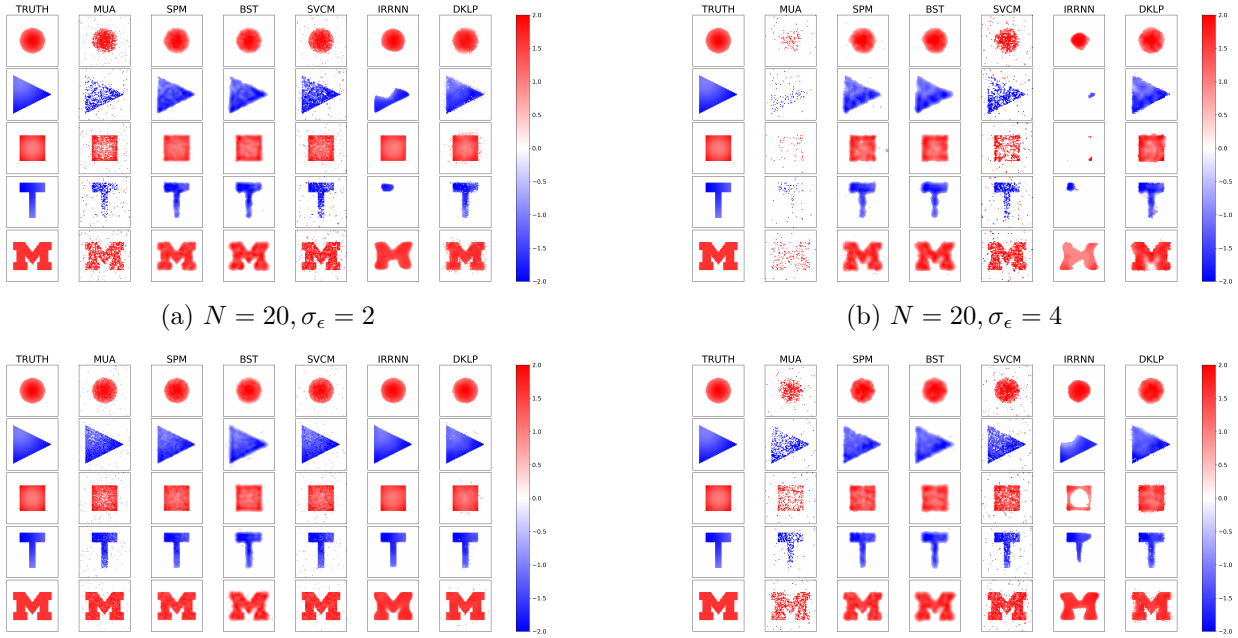
B.2 Additional Results

B.2.1 Simulations

To further evaluate the performance of DKLP, we provide additional results comparing the spatial coefficient estimation accuracy between DKLP and other competing methods across various simulation scenarios outlined in Section 2.4 of the main paper. Figure B.1 and B.2

highlight how DKLP outperforms other methods in additional image-on-scalar and scalar-on-image regression simulation settings, particularly under challenging conditions. Different from conventional methods that rely on pre-specified kernel functions, we aim to emphasize DKLP's ability to adaptively learn from data without requiring explicit kernel assumptions. This flexibility allows DKLP to robustly capture intricate spatial dependencies and varying signal structures across different analysis tasks. The results demonstrate DKLP's reliability and adaptability in identifying meaningful spatial patterns, further validating its practicality for diverse imaging regression tasks.

B.2.1.1 Image-on-scalar regression

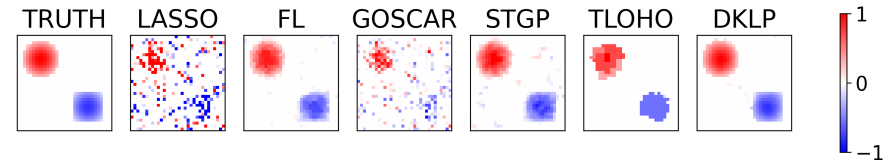


(a) $N = 20, \sigma_\epsilon = 2$ (b) $N = 20, \sigma_\epsilon = 4$

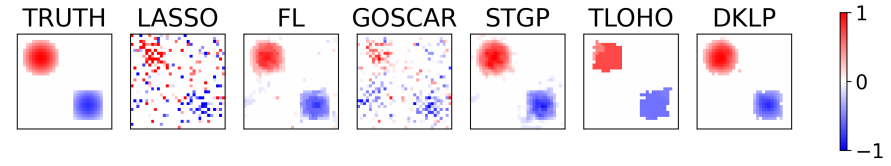
(c) $N = 50, \sigma_\epsilon = 2$ (d) $N = 50, \sigma_\epsilon = 4$

Figure B.1: Comparison of true spatially varying functions with estimated effects from DKLP and other competitive methods in image-on-scalar regression simulations with $V = 64^2$.

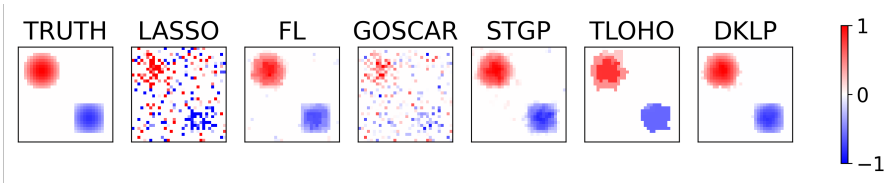
B.2.1.2 Scalar-on-image regression



(a) $\sigma_u = 2, \sigma_\epsilon = 3$



(b) $\sigma_u = 4, \sigma_\epsilon = 3$



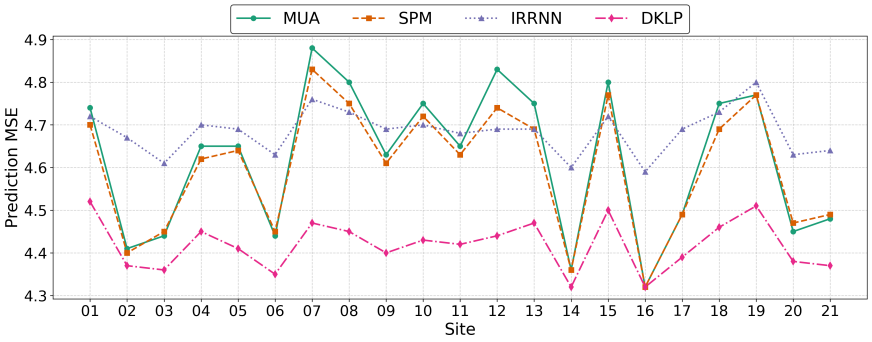
(c) $\sigma_u = 4, \sigma_\epsilon = 5$

Figure B.2: Comparison of true spatially varying functions with estimated effects from DKLP and other competitive methods in scalar-on-image regression simulations with $V = 32^2$.

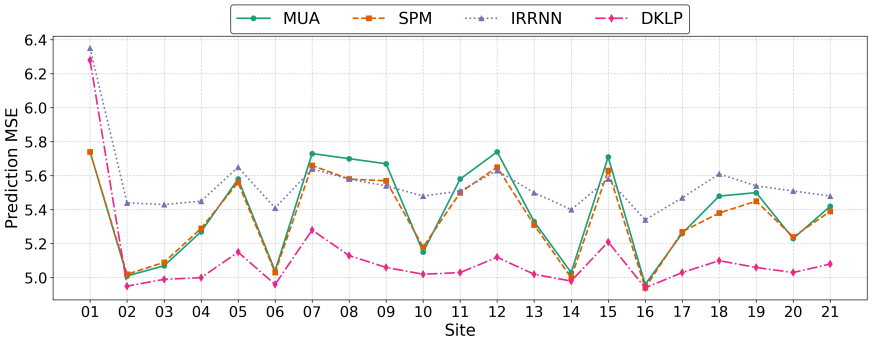
B.2.2 Real data applications

B.2.2.1 Cross-Site prediction MSEs in fMRI application

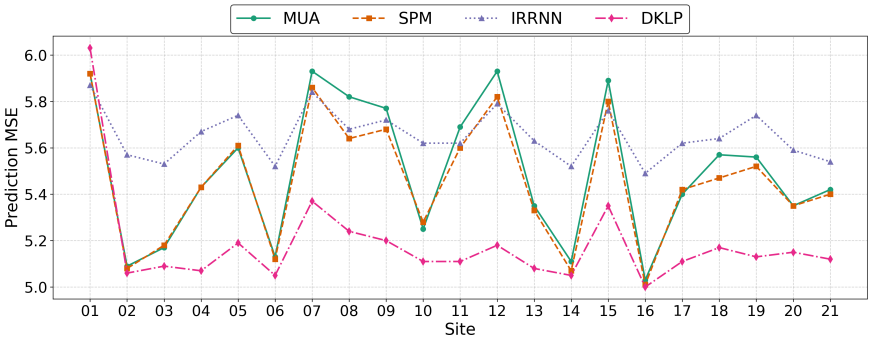
We provide additional single-site prediction MSE results using cortical surface fMRI data in the ABCD study, as shown in Figure B.3. Additional results at the follow-up time point and on the right hemisphere further confirm DKLP's robustness and adaptability, as it continues to achieve lower prediction MSEs compared to competing methods. These findings validate DKLP's ability to accurately estimate the spatially varying coefficients in studies with small sample sizes.



(a) Baseline data on the right hemisphere



(b) Followup data on the left hemisphere

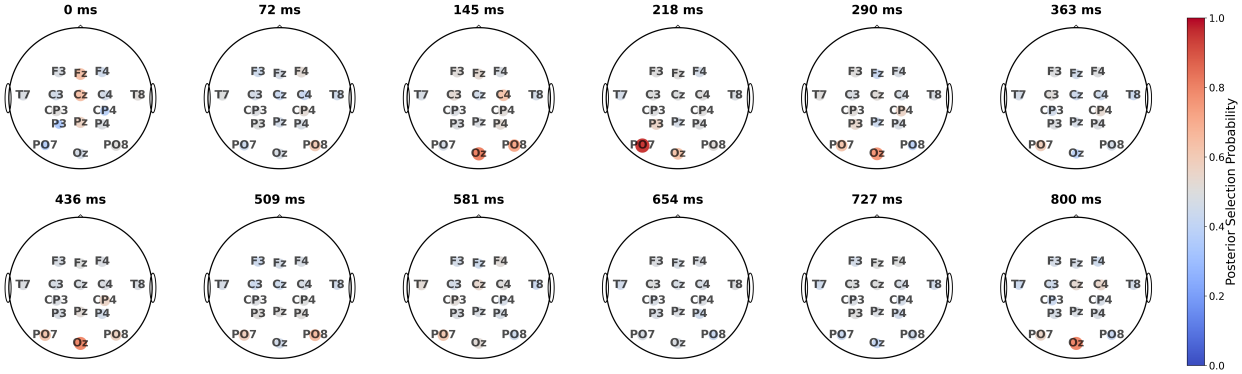


(c) Followup data on the right hemisphere

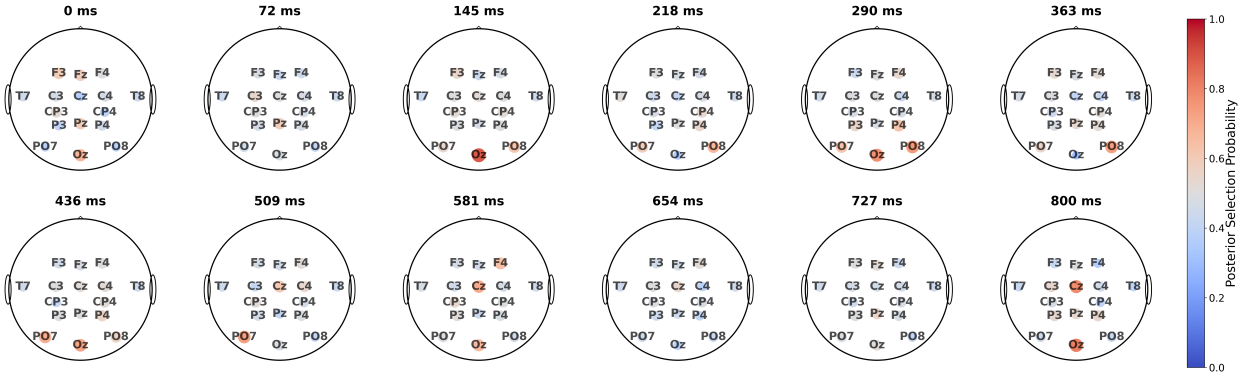
Figure B.3: Cross-site prediction MSEs on cortical surface fMRI data in the ABCD study.

B.2.2.2 Posterior EEG signal selection on additional participants

In Figure B.4, we further illustrate the posterior EEG signal selection probabilities for one ALS participant and one healthy control participant to validate the robustness of channel contributions. Specifically, channels PO7, PO8, and Oz consistently show higher significance in visual tasks for all the participants.



(a) Subject 156: ALS subject



(b) Subject 183: Healthy subject

Figure B.4: Posterior selection probabilities on additional participants at twelve equally spaced time points on sixteen EEG brain channels. Node sizes represent the magnitude of the selection probabilities on each channel.

B.3 DNN Specifications

In this section, we outline the configurations of the DNN architectures used in both the simulation studies and real data applications in Table B.1. These specifications include the number of hidden layers, the number of hidden units per layer, the activation functions, the kernel truncation numbers, and the learning rates for the Stochastic gradient Langevin dynamics (SGLD) algorithm. We observe that the truncation number and the learning rate in the SGLD sampling algorithm are tuning parameters that influence both the MCMC convergence rate and the final model fitting results. In practice, these parameters can be optimized using the deviance information criterion (DIC) [Spiegelhalter et al., 2002]. Additionally, the truncation number can be further refined based on the total variance explained by the basis functions, as discussed in Lin et al. [2024].

Section	J	h	activation	L	lr
3.5.1	4	256	ReLU	50	1e-8
3.5.2	4	256	ReLU	50	1e-4
3.5.3	4	256	ReLU	100	1e-5
3.6.1	4	512	ReLU	200	1e-5
3.6.2	4	512	ReLU	50	1e-8

Table B.1: DNN configurations in each study section, including the number of hidden layers, the number of hidden units per layer, the activation functions, the kernel truncation numbers, and the learning rates.

B.4 Sensitivity Analysis

We perform sensitivity analyses on real data applications using various DNN architectures to evaluate the robustness of the proposed method. The analysis includes testing different configurations, including the number of hidden layers, the number of units per layer, and activation functions, to assess their impact on model performance. Our findings in Figure B.5 and Table B.2 indicate that the analysis results remain robust across different DNN configurations, demonstrating the flexibility and adaptability of the proposed approach to various architectural choices.

B.4.1 Cortical surface task fMRI

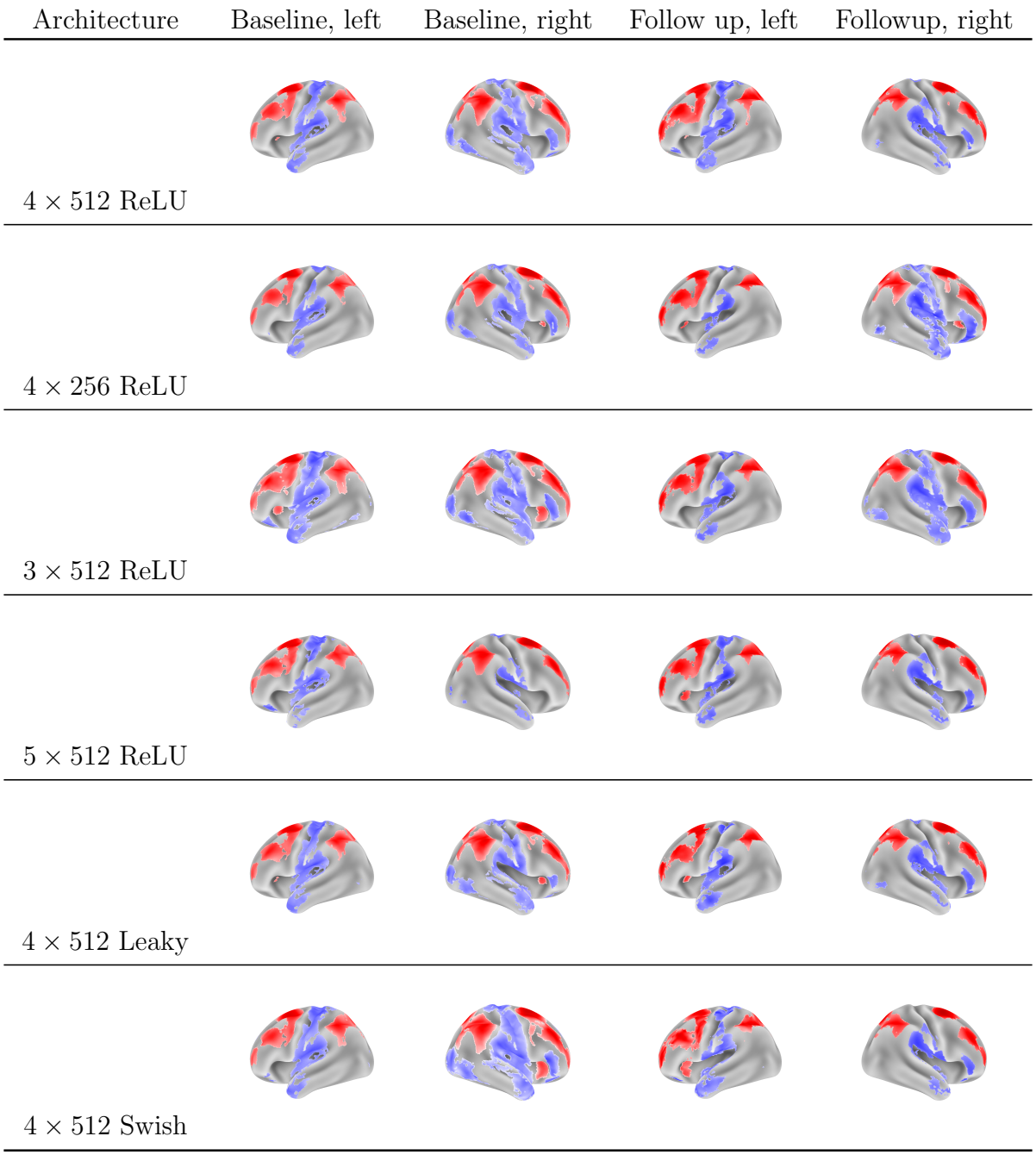


Figure B.5: Sensitivity analysis of different DNN architectures in terms of the number of layers, the number of nodes per layer, and the activation function applied to ABCD cortical surface data for the left and right hemispheres at baseline and follow-up time points. The color scheme represents values within the range [-1, 1].

B.4.2 EEG Brain-Computer Interface

Table B.2: Sensitivity analysis on BCI study for individual-level classifier for character-level prediction accuracy using different DNN configurations in terms of the number of layers, the number of nodes per layer, and the activation function.

Group	ID	4×512 ReLU	4×256 ReLU	3×512 ReLU	5×512 ReLU	4×512 Leaky	4×512 Swish
Control	106	86	83	86	88	84	86
	108	93	92	92	91	91	91
	111	83	82	82	82	83	81
	114	95	94	94	95	95	93
	115	91	90	89	89	90	90
	117	97	95	95	97	95	94
	119	65	68	64	63	64	69
	120	87	86	87	88	85	86
	121	92	92	92	92	91	90
	122	87	86	82	86	83	87
	123	98	98	98	98	98	98
	171	92	91	91	91	92	92
	172	81	79	81	82	80	80
	177	92	91	92	93	90	94
	183	92	90	92	94	92	92
ALS	143	88	85	87	86	84	85
	145	76	76	75	76	78	76
	146	90	90	91	91	89	90
	147	93	92	92	93	92	92
	151	94	93	93	93	93	93
	152	69	60	69	67	67	67
	156	82	81	83	83	85	84
	158	73	71	71	73	72	75
	159	78	77	75	76	76	73

APPENDIX C

Supplementary Materials for Chapter IV

Here, we include detailed specifications of our U-Net architecture for training the A-DDPM generation framework. Additionally, we provide further image generation results for the same target generation under various seeds.

C.1 Architecture details

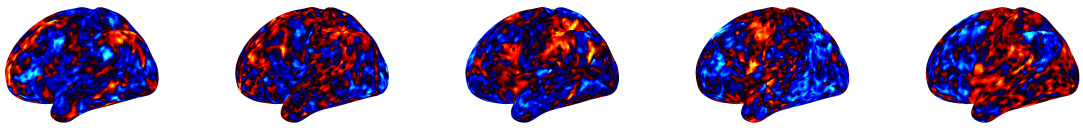
The denoising backbone used in the diffusion framework is a U-Net backbone [Ronneberger et al., 2015] adapted from Cardoso et al. [2022], optimized for spatially structured neuroimaging data. For both MNIST and fMRI data, the U-Net model accepts a single-channel image as input and outputs a one-channel reconstruction for the predicted noise at a given timestep. We employ sinusoidal embeddings for the input timestep, as outlined in Ho et al. [2020] with label inputs, which are integrated into all residual blocks throughout the network. Each encoder stage applies group normalization, SiLU activation, and residual convolutional transformations, with self-attention introduced at the deeper layers to capture long-range spatial dependencies across the cortical surface. The bottleneck incorporates both residual and attention blocks, forming a high-level latent representation that guides spatially coherent reconstruction in the decoder. Skip connections between corresponding encoder and decoder layers preserve fine spatial details, while learned upsampling restores the full-resolution activation maps. The output layer projects the reconstructed features back to a single-channel image via a 3×3 convolution.

For the MNIST dataset, we train the BACNet for 80K steps in U-Net and 160K steps in ControlNet with a learning rate of 0.0001. For ABCD dataset, we train the BACNet for 250K steps in both U-Net and ControlNet with a learning rate of 0.0001. All the computation are conducted on a 2x 2.9 GHz Intel processor with 378G RAM on a single NVIDIA A100 GPU.

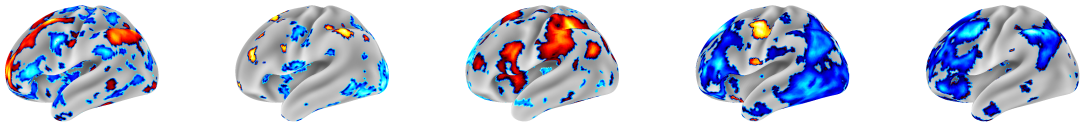
C.2 Additional results

In this section, we include multiple image generation results given the same spatial activation map condition under different initial noises. Here, we show that for each generation, the BACNet can consistently preserve the spatial conditioning while generate plausible fMRI contrast map with variability and fidelity to fMRI data.

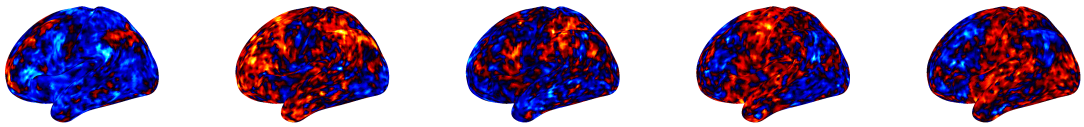
(a) True images



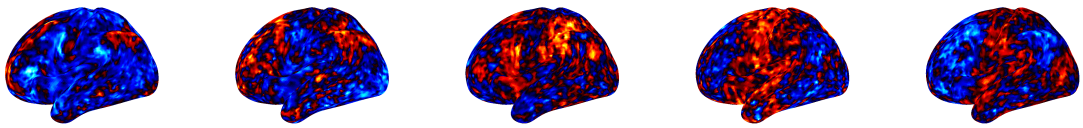
(b) Learned activation



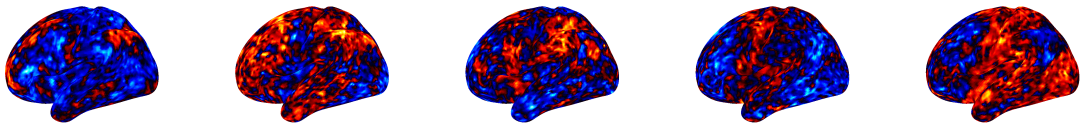
(c) Generation 1



(d) Generation 2



(e) Generation 3



(f) Generation 4

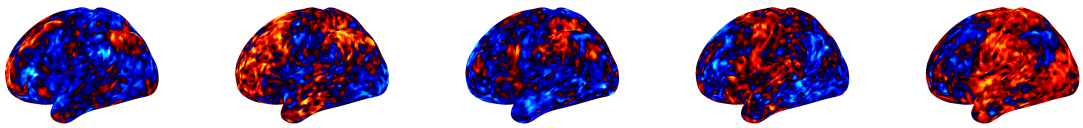


Figure C.1: BACNet generation results under different initial noise values.

BIBLIOGRAPHY

Meenu Ajith and Vince D Calhoun. Denoising diffusion probabilistic models for high-fidelity fMRI intrinsic connectivity network data generation. In *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2024.

Elena A. Allen, Erik B. Erhardt, and Vince D. Calhoun. Data visualization in the neurosciences: Overcoming the curse of dimensionality. *Neuron*, 74(4):603–608, 2012. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2012.05.001. URL https://www.sciencedirect.com/science/article/pii/S089662731200428X.

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://docs.pytorch.org/assets/pytorch2-2.pdf.

John Ashburner. SPM: a history. *Neuroimage*, 62(2):791–800, 2012.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Gregory Benton, Wesley J Maddox, Jayson Salkey, Julio Albinati, and Andrew Gordon Wilson. Function-space distributions over kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Ad-

cock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76:1–32, 2017.

Betty Jo Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018.

Gang Chen, Paul A Taylor, and Robert W Cox. Is the statistic value all we should care about in neuroimaging? *Neuroimage*, 147:952–959, February 2017.

Gang Chen, Yaqiong Xiao, Paul A Taylor, Justin K Rajendra, Tracy Riggins, Fengji Geng, Elizabeth Redcay, and Robert W Cox. Handling multiplicity in neuroimaging through Bayesian lenses with multilevel modeling. *Neuroinformatics*, 17:515–545, 2019.

Kathia Chenane, Youcef Touati, Larbi Boubchir, and Boubaker Daachi. Neural net-based approach to EEG signal acquisition and classification in BCI applications. *Computers*, 8 (4):87, 2019.

Robert W Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

Andreas Damianou and Neil D Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215. PMLR, 2013.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

Xavier Descombes, Frithjof Kruggel, and D Yves Von Cramon. Spatio-temporal fMRI analysis using markov random fields. *IEEE transactions on medical imaging*, 17(6):1028–1039, 1998.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Emanuel Donchin, Kevin M Spencer, and Ranjith Wijesinghe. The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering*, 8(2):174–179, 2000.

David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013.

Lawrence Ashley Farwell and Emanuel Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

Bruce Fischl and Anders M Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97 (20):11050–11055, 2000.

Karl J Friston. Statistical parametric mapping. *Neuroscience Databases: A Practical Guide*, pages 237–250, 2003.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2):389–402, 2019.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. 2007.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Jeff Goldsmith, Lei Huang, and Ciprian M Crainiceanu. Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64, 2014.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Evan M Gordon, Timothy O Laumann, Babatunde Adeyemo, Jeremy F Huckins, William M Kelley, and Steven E Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, 26(1):288–303, 2016.

Cui Guo, Jian Kang, and Timothy D Johnson. A spatial Bayesian latent factor model for image-on-image regression. *Biometrics*, 78(1):72–84, 2022.

Niels Vaever Hartvig and Jens Ledet Jensen. Spatial mixture modeling of fMRI data. *Human brain mapping*, 11(4):233–248, 2000.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Yufan Hu, Wuyang Li, Yixuan Yuan, et al. Synthesizing realistic fMRI: A physiological dynamics-driven hierarchical diffusion model for efficient fMRI acquisition. In *The Thirteenth International Conference on Learning Representations*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Phillip A Jang, Andrew Loeb, Matthew Davidow, and Andrew G Wilson. Scalable Levy process priors for spectral kernel learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Saad Jbabdi, Mark William Woolrich, and Timothy Edward John Behrens. Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models. *NeuroImage*, 44(2):373–384, 2009.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Terry L Jernigan, Sandra A Brown, and Gayathri J Dowling. The adolescent brain cognitive development study. *Journal of research on adolescence: the official journal of the Society for Research on Adolescence*, 28(1):154, 2018.

Jian Kang, Brian J Reich, and Ana-Maria Staicu. Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184, 2018.

Matthias Kaper, Peter Meinicke, Ulf Grossekathoefer, Thomas Lingner, and Helge Ritter. BCI competition 2003-data set IIb: support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, 51(6):1073–1076, 2004.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational Bayes, 2013.

Insung Kong and Yongdai Kim. Posterior concentrations of fully-connected Bayesian neural networks with general priors on the weights. *arXiv preprint arXiv:2403.14225*, 2024.

Brian Kulis, Mátyás Sustik, and Inderjit Dhillon. Learning low-rank kernel matrices. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 505–512, 2006.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

Changwoo Lee, Zhao Tang Luo, and Huiyan Sang. T-LoHo: A Bayesian regularization model for structured sparsity and smoothness on graphs. *Advances in Neural Information Processing Systems*, 34:598–609, 2021.

Herbert KH Lee. Consistency of posterior distributions for neural networks. *Neural Networks*, 13(6):629–642, 2000.

June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

Jessica Leoni, Silvia Carla Strada, Mara Tanelli, Kaijun Jiang, Alessandra Brusa, and Alice Mado Proverbio. Automatic stimuli classification from ERP data for augmented communication via Brain-Computer Interfaces. *Expert Systems with Applications*, 184:115572, 2021.

Binxu Li, Wei Peng, Mingjie Li, Ehsan Adeli, and Kilian M Pohl. Integrating anatomical priors into a causal diffusion model. *arXiv preprint arXiv:2509.09054*, 2025.

Xinyi Li, Li Wang, Huixia Judy Wang, and Alzheimer's Disease Neuroimaging Initiative. Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression. *Journal of the American Statistical Association*, 116(536):1994–2008, 2021.

Zikai Lin, Yajuan Si, and Jian Kang. Latent subgroup identification in image-on-scalar regression. *The annals of applied statistics*, 18(1):468, 2024.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, 2011.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Guoxuan Ma, Bangyao Zhao, Hasan Abu-Amara, and Jian Kang. Bayesian image-on-image regression via deep kernel learning based Gaussian processes. *arXiv preprint arXiv:2311.05649*, 2023.

Jonathan Marchini and Anne Presanis. Comparing methods of analyzing fMRI statistical parametric maps. *Neuroimage*, 22(3):1203–1213, 2004.

Amanda F Mejia, Yu Yue, David Bolin, Finn Lindgren, and Martin A Lindquist. A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, 115(530):501–520, 2020.

Anna Menacher, Thomas E Nichols, Timothy D Johnson, and Jian Kang. Scalable scalar-on-image cortical surface regression with a relaxed-thresholded Gaussian process prior. *arXiv preprint arXiv:2403.13628*, 2024.

Vinod Menon. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in cognitive sciences*, 15(10):483–506, 2011.

Jeffrey S Morris, Philip J Brown, Richard C Herrick, Keith A Baggerly, and Kevin R Coombes. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–489, 2008.

Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Nilearn contributors. Nilearn, 2025. URL https://doi.org/10.5281/zenodo.8397156.

Evert J Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. 1930.

Hatice Okumuş and Önder Aydemır. Random forest classification for brain computer interface applications. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2017.

Junier B Oliva, Avinava Dubey, Andrew G Wilson, Barnabás Póczos, Jeff Schneider, and Eric P Xing. Bayesian nonparametric kernel-learning. In *Artificial Intelligence and Statistics*, pages 1078–1086. PMLR, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M Pohl. Generating realistic brain MRIs via a conditional diffusion probabilistic model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2023.

William D Penny, Nelson J Trujillo-Barreto, and Karl J Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

Nicholas G Polson and Veronika Ročková. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108 (504):1339–1349, 2013.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2005.

Richard C. Reynolds, Daniel R. Glen, Gang Chen, Ziad S. Saad, Robert W. Cox, and Paul A. Taylor. Processing, evaluating, and understanding fMRI data with afni_proc.py. *Imaging Neuroscience*, 2:imag–2–00347, 11 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00347. URL https://doi.org/10.1162/imag_a_00347.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

Alec Sargood, Lemuel Puglisi, James H Cole, Neil P Oxtoby, Daniele Ravì, and Daniel C Alexander. CoCoLIT:Controlnet-conditioned latent image translation for MRI to amyloid PET synthesis. *arXiv preprint arXiv:2508.01292*, 2025.

Franco Scarselli and Ah Chung Tsoi. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Networks*, 11 (1):15–37, 1998.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

Per Sidén, Anders Eklund, David Bolin, and Mattias Villani. Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. *NeuroImage*, 146:211–225, 2017.

Per Sidén, Finn Lindgren, David Bolin, Anders Eklund, and Mattias Villani. Spatial 3d Matérn priors for fast whole-brain fMRI analysis. *Bayesian Analysis*, 16(4):1251–1278, 2021.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Chandra Sripada, Saige Rutherford, Mike Angstadt, Wesley K Thompson, Monica Luciana, Alexander Weigard, Luke H Hyde, and Mary Heitzeg. Prediction of neurocognition in youth from resting state fMRI. *Molecular Psychiatry*, 25(12):3413–3421, 2020.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Paul A Taylor, Richard C Reynolds, Vince Calhoun, Javier Gonzalez-Castillo, Daniel A Handwerker, Peter A Bandettini, Amanda F Mejia, and Gang Chen. Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *Neuroimage*, 274:120138, 2023.

Paul A. Taylor, Daniel R. Glen, Gang Chen, Robert W. Cox, Taylor Hanayik, Chris Rorden, Dylan M. Nielson, Justin K. Rajendra, and Richard C. Reynolds. A set of fMRI quality control tools in afni: Systematic, in-depth, and interactive qc with afni_proc.py and more. *Imaging Neuroscience*, 2:imag–2–00246, 08 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00246. URL https://doi.org/10.1162/imag_a_00246.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ryan J Tibshirani. *The solution path of the generalized lasso*. Stanford University, 2011.

Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435 – 1463, 2008. doi: 10.1214/009053607000000613. URL https://doi.org/10.1214/009053607000000613.

Sadraque S Viana, Danilo M Batista, and Danilo B Melges. Logistic regression models: Feature selection for P300 detection improvement. In *XXIV Brazilian Congress on Biomedical Engineering–CBEB*, volume 2014, pages 979–982, 2014.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Andrew S Whiteman, Timothy D Johnson, and Jian Kang. Bayesian inference for group-level cortical surface image-on-scalar regression with Gaussian process priors. *Biometrics*, 80(4):ujae116, 2024.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2016.

Mark W Woolrich, Timothy EJ Behrens, and Stephen M Smith. Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage*, 21(4):1748–1761, 2004.

Yuliang Xu and Jian Kang. Bayesian image mediation analysis. *arXiv preprint arXiv:2310.16284*, 2023.

Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 922–930, 2012.

Zijian Zeng, Meng Li, and Marina Vannucci. Bayesian image-on-scalar regression with a spatial global-local spike-and-slab prior. *Bayesian Analysis*, 1(1):1–26, 2022.

Daiwei Zhang, Lexin Li, Chandra Sripada, and Jian Kang. Image response regression via deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1589–1614, 2023a.

Linlin Zhang, Michele Guindani, Francesco Versace, Jeffrey M Engelmann, and Marina Vannucci. A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. 2016.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023b.

Bangyao Zhao, Jane E Huggins, and Jian Kang. Bayesian inference on brain-computer interfaces via glass. *arXiv preprint arXiv:2304.07401*, 2023.

Haokai Zhao, Haowei Lou, Lina Yao, Wei Peng, Ehsan Adeli, Kilian M Pohl, and Yu Zhang. Diffusion models for computational neuroimaging: A survey. *arXiv preprint arXiv:2502.06552*, 2025.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

Hongtu Zhu, Jianqing Fan, and Linglong Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.

Jerry Zhu, Jaz Kandola, Zoubin Ghahramani, and John Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. *Advances in Neural Information Processing Systems*, 17, 2004.

Jinfeng Zhuang, Ivor W Tsang, and Steven CH Hoi. A family of simple non-parametric kernel learning algorithms. *The Journal of Machine Learning Research*, 12:1313–1347, 2011.