Learning Visual and Barrel Cortex Representations from Multimodal Convergence

by

Candace Ng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2026

© Candace Ng 2026

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Multisensory perception emerges from interactions between sensory pathways, but biologically inspired computational models often examine sensory modalities in isolation. This thesis presents a computational model of mouse visual and tactile processing through whiskers that incorporates anatomically motivated cortical pathways and modulation between sensory modalities. The visual pathway was based on MouseNet, a convolutional neural network derived from the organization of the mouse visual cortex. The whisker pathway was designed to reflect the hierarchical and temporal processing of signals through the somatosensory cortex and higher cortical regions. Representations produced from whisker input were used to modulate activity in the superficial layers of the primary visual cortex through a learned inhibitory gate motivated by experimentally observed somatosensory influences on visual processing. The visual and whisker encoders were trained jointly using a contrastive learning objective that aligned representations from paired simulated sensory observations. Training produced substantially greater similarity and retrieval performance for matched visual and whisker samples than for untrained representations, indicating that the model learned a shared embedding space across modalities. To evaluate how multimodal training shaped the individual sensory pathways, visual and whisker representations were separately compared with modality-specific neural recordings using representational similarity analysis (RSA). The learned visual representations exhibited measurable correspondence with recorded visual cortical activity. Enabling the inhibitory gate modestly increased neural alignment relative to disabling it within the same trained models, particularly in VISp layer 5, but gated models did not reliably outperform models trained without the gate. Orientation selectivity increased in most evaluated visual layers despite not being explicitly included in the training objective, and trained VISp values were broadly comparable to physiological measurements from mouse visual cortex. Whisker representations retained condition-dependent structure but did not show consistent positive correspondence with the neural geometry, with animal-level RSA varying in sign and magnitude. These results demonstrate a framework for integrating visual and tactile processing within a biologically grounded neural network and evaluating the resulting representations directly against experimental neural data. This work provides a foundation for investigating how interactions between sensory modalities and anatomical constraints shape cortical representations.

Acknowledgements

I would like to thank my supervisor, Dr. Bryan Tripp, for his guidance and invaluable feedback throughout my master's studies. I would also like to thank Dr. Mitra Hartmann, Kevin Kleczka, and Dr. Terry Stewart for their technical insight and collaborative support throughout this project. Their expertise greatly contributed to the development and evaluation of this work.

Finally, I would like to thank my parents for their encouragement and support throughout my graduate studies.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures ix

List of Tables x

1 Introduction 1

1.1 Background and Motivation 1

1.2 Problem Statement 2

1.3 Research Objectives 3

1.4 Thesis Organization 4

2 Background & Related Work 5

2.1 The Mouse Visual System 5

2.2 Biologically Inspired Visual Models 6

2.3 The Mouse Whisker System 8

2.3.1 Available Barrel Cortex Data 9

2.3.2 Computational Models of Barrel Cortex Processing 10

2.4 Multisensory Integration in the Mouse Cortex . . . 11
2.5 Self-Supervised Representation Learning . . . 12
2.6 Neural Representation Analysis . . . 14
2.7 Summary and Research Gap . . . 16

3 Methods **18**
3.1 System Overview . . . 18
3.2 Unity Simulation Environment . . . 19
3.2.1 Whisker Geometry and Kinematics . . . 20
3.2.2 Whisker Contact Detection . . . 20
3.2.3 Multimodal Training Trials . . . 22
3.2.4 Concave and Convex Evaluation Stimuli . . . 24
3.3 Whisker Representation . . . 25
3.3.1 Contact Feature Representation . . . 25
3.3.2 Temporal and Spatial Organization of Whisker Signals . . . 26
3.4 Whisker Encoder Architecture . . . 27
3.4.1 Biological Motivation . . . 27
3.4.2 Encoder Overview . . . 29
3.4.3 Temporal Integration in SSp-bfd2/3 . . . 29
3.4.4 Biologically Constrained Architectural Parameters . . . 30
3.4.5 Interareal Projections and Higher-Order Processing . . . 32
3.5 Visual Encoder . . . 33
3.5.1 Visual Processing Pathway . . . 33
3.5.2 Visual Input Preprocessing . . . 34
3.5.3 Visual Embedding Extraction . . . 36
3.5.4 Whisker-Conditioned Inhibitory Gating . . . 36
3.6 Multimodal Contrastive Learning . . . 37
3.6.1 Motivation for Cross-Modal Alignment . . . 37

3.6.2 Joint Embedding Space . . . 37
3.6.3 Contrastive Objective . . . 38
3.6.4 Training Procedure . . . 38
3.7 RSA Evaluation . . . 39
3.7.1 Motivation for RSA . . . 39
3.7.2 Whisker Encoder Evaluation . . . 40
3.7.3 Allen Brain Observatory Evaluation . . . 42
3.8 Orientation Selectivity Evaluation . . . 44
3.8.1 Oriented Grating Stimuli and Response Extraction . . . 44
3.8.2 Orientation Selectivity Index . . . 45

4 Results **46**
4.1 Experimental Overview . . . 46
4.2 Multimodal Representation Learning . . . 47
4.2.1 CLIP Training . . . 47
4.2.2 Embedding Alignment . . . 50
4.3 Whisker Representation Learning . . . 50
4.3.1 Six-Condition Neural Representational Similarity . . . 50
4.4 Visual Representation Learning . . . 53
4.4.1 Inhibitory Gating . . . 53
4.4.2 Orientation Selectivity . . . 56

5 Discussion **59**
5.1 Cross-Modal Alignment of Visual and Whisker Representations . . . 59
5.2 Visual Representations Following Multimodal Training . . . 60
5.3 Functional Effects of the Learned Inhibitory Gate in VISp . . . 61
5.4 Whisker Encoder Correspondence with Neural Data . . . 63
5.5 Limitations . . . 64
5.6 Future Work . . . 66

6 Conclusion **68**

6.1 Thesis Summary 68

6.2 Main Contributions 70

6.3 Final Remarks 71

References **72**

List of Figures

3.1 Overview of the proposed multimodal architecture 19

3.2 Representative Unity simulation environment and visual observations used for multimodal training 23

3.3 Unity implementation of the six stimulus conditions 24

3.4 Architecture of the whisker encoder 28

3.5 Modified MouseNet visual encoder architecture 35

3.6 General representational similarity analysis (RSA) workflow used to compare neural and model representations 41

4.1 Average CLIP contrastive training loss as a function of training epoch. . . 48

4.2 Retrieval performance of the learned multimodal embedding space before and after CLIP training. 49

4.3 Representational comparison of the whisker encoder and Rodgers neural responses 52

4.4 Representational similarity between model VISp2/3 activations and the Allen VISp L2/3 neural representation 55

4.5 Activity of model VISp2/3 units in response to Allen Brain Observatory images 56

4.6 Orientation selectivity across MouseNet visual cortical layers before and after CLIP training 58

List of Tables

3.1 Neuron-count estimates and implemented channel dimensions of the whisker encoder 31
3.2 Tensor shapes through the whisker encoder 33

4.1 Cosine similarity statistics between visual and whisker embeddings before and after CLIP training 51
4.2 Representational similarity between MouseNet and Allen Brain Observatory responses 54

Chapter 1

Introduction

1.1 Background and Motivation

Recent advances in deep learning have enabled computational models to achieve strong performance across a wide range of visual recognition tasks [26, 43]. Convolutional neural networks (CNNs) and vision transformers have demonstrated success in image classification, object detection, and scene understanding. While these models have achieved strong engineering performance, many remain only loosely inspired by biological neural systems and often lack the structural and functional characteristics observed in real brains.

Computational neuroscience seeks to bridge this gap by developing models that not only perform useful tasks but also provide insight into the organization and function of biological neural circuits. One example is MouseNet [48], a convolutional neural network architecture designed using anatomical connectivity data from the mouse visual cortex. Unlike conventional CNNs, MouseNet incorporates biologically informed cortical areas and interareal projections derived from experimental measurements, enabling comparisons between model representations and neural recordings obtained from living animals.

Although vision has been extensively studied in both neuroscience and machine learning, animals rarely rely on a single sensory modality when interacting with their environment. Rodents in particular make extensive use of their whiskers to perceive object shape, texture, and spatial location. Tactile information obtained through whisker contacts is processed within specialized somatosensory pathways and contributes significantly to navigation and object recognition. Experimental studies have further suggested that interactions between visual and somatosensory systems occur throughout the brain, indi-

cating that perception emerges from the integration of multiple sensory modalities rather than independent processing streams.

Previous computational models have integrated visual and tactile inputs for material-property prediction and cross-modal generation [55, 28]. However, these systems were developed primarily for machine perception and robotic sensing rather than as anatomically grounded accounts of biological visual–somatosensory pathways. In contrast, MouseNet models anatomically informed pathways in the mouse visual cortex but does not incorporate whisker input. It therefore cannot examine how tactile signals influence visual representations or how visual–somatosensory interactions contribute to recorded neural activity.

This thesis addresses this limitation by extending MouseNet with a biologically motivated whisker-processing pathway and a mechanism for integrating both tactile and visual information. By combining anatomical inspiration with modern representation learning techniques, the proposed framework aims to explore how multisensory information may be represented within cortical networks and to evaluate the resulting representations against experimental neural datasets.

1.2 Problem Statement

The absence of tactile pathways is particularly limiting for models of rodent perception because rodents rely heavily on active whisker sensing to interact with their environment [9]. Whisker contacts provide information about object shape and location, which is processed through somatosensory thalamocortical and barrel-cortex pathways [1, 9, 40, 46]. Experimental studies also show that visual and somatosensory activity interact within association cortex and through a direct interareal inhibitory circuit linking the two systems [50, 52]. These findings motivate models that examine visual and tactile processing within a unified biological framework.

Existing computational approaches address only parts of this problem, but do not combine anatomically motivated visual and whisker-processing pathways with learned cross-modal interactions within a unified framework. These limitations and the most closely related models are reviewed in detail in Chapter 2. The present thesis develops and evaluates a multisensory extension of MouseNet that integrates visual and whisker-derived signals and compares its internal representations with recordings from both visual and somatosensory cortex.

1.3 Research Objectives

The overall objective of this thesis is to develop and evaluate a biologically inspired multisensory neural network model that integrates visual and whisker-derived tactile information within a unified framework. Building upon MouseNet, the proposed model seeks to incorporate anatomically motivated somatosensory pathways and mechanisms for multisensory interaction while maintaining biological interpretability.

To achieve this objective, the following research goals are pursued:

1. Develop a biologically inspired whisker-processing pathway based on known anatomical organization within the rodent somatosensory system, including representations of the barrel cortex and higher-order somatosensory areas.
2. Integrate the visual and tactile pathways through a learned whisker-conditioned inhibitory gate that modulates intermediate VISp representations.
3. Train the resulting multimodal model using a self-supervised contrastive objective to encourage the emergence of shared visual and tactile representations.
4. Evaluate the learned tactile representations by comparing model activity with experimentally recorded neural responses obtained from whisker-based behavioural datasets.
5. Evaluate the learned visual representations by comparing model activity with Allen Brain Observatory recordings using representational similarity analysis and by characterizing orientation selectivity across visual layers.
6. Investigate the extent to which biologically inspired multisensory integration influences internal visual representations within the model.

The resulting framework combines biologically inspired visual and tactile pathways within a unified architecture and is evaluated against experimental neural recordings from both somatosensory and visual cortical systems. In addition, this work investigates self-supervised approaches for multimodal representation learning and examines the effects of biologically motivated multisensory interactions on internal model representations.

1.4 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 reviews relevant literature in biologically inspired neural networks, the organization of the rodent visual and somatosensory systems, multisensory integration, and self-supervised representation learning. Particular attention is given to prior work on MouseNet and experimental evidence supporting interactions between visual and tactile sensory pathways.

Chapter 3 describes the proposed multisensory modelling framework. This chapter presents the MouseNet visual architecture, the biologically inspired whisker processing pathway, the mechanism used to integrate tactile and visual information, the self-supervised training procedures, and the methodologies used for neural data comparison.

Chapter 4 presents the experimental results. Model representations are evaluated using representational similarity analysis against whisker and visual neural datasets from the Rodgers dataset and Allen Brain Observatory, and orientation selectivity is assessed across VISp and VISrl before and after multimodal contrastive training.

Chapter 5 discusses the implications of the experimental findings, examines the biological relevance of the learned representations, and considers the strengths and limitations of the proposed modelling approach.

Finally, Chapter 6 summarizes the main findings of this work and outlines directions for future research in biologically inspired multisensory modelling and neural representation learning.

Chapter 2

Background & Related Work

2.1 The Mouse Visual System

Large-scale physiological surveys and anatomical connectivity datasets have made the mouse visual system a valuable model for investigating cortical sensory processing [8, 18, 37, 32]. These studies show that the mouse visual cortex comprises a distributed hierarchy of interconnected areas with distinct functional properties.

Visual information is relayed from the retina through the lateral geniculate nucleus (dLGN) to the primary visual cortex (VISp). Neurons in VISp respond selectively to features including orientation, spatial frequency, and motion [8, 11]. VISp subsequently distributes visual information to multiple higher-order visual areas through structured interareal projections [18, 37]. Beyond VISp, the mouse visual cortex includes higher areas such as VISal, VISl, VISli, VISpl, VISpor, and VISrl. These areas exhibit distinct response properties, connectivity patterns, and functional specializations that support complementary processing streams across the visual cortex [8, 18, 32].

Anatomical tracing studies show that mouse cortical visual areas form a hierarchical network containing reciprocal corticocortical projections and thalamocortical connections [18]. This architecture is not a strictly sequential pathway, as information can be exchanged among primary and higher visual areas through extensive interareal connectivity [37, 18]. Analyses of projection strength and direction identify a hierarchy spanning visual cortical and thalamic regions [18]. Functional studies further show that these areas have distinct response properties, consistent with distributed and specialized processing across the visual cortical network [8, 32]. Together, these anatomical and functional findings motivate

treating visual representations as products of distributed cortical processing rather than of a single isolated stage.

The existence of extensive connectivity between cortical areas has important implications for computational modelling. Conventional convolutional neural networks are generally organized as sequential feedforward processing hierarchies and do not explicitly reproduce the regional organization or anatomically constrained interareal connectivity of biological visual systems [26, 35]. In contrast, biologically inspired visual models incorporate anatomical, physiological, or functional constraints and evaluate whether these constraints improve correspondence with neural activity [54, 48]. Representative approaches are discussed in the following section.

2.2 Biologically Inspired Visual Models

Deep convolutional neural networks have achieved remarkable success across a wide range of computer vision tasks [26] and have frequently been used as computational models of sensory processing. Early studies demonstrated that representations learned by convolutional networks exhibit similarities to neural activity observed within biological visual systems [54], motivating the use of deep learning models as tools for investigating cortical computation. However, most conventional architectures, such as AlexNet, VGG, and ResNet, were designed primarily to maximize task performance and do not explicitly incorporate anatomical constraints derived from biological neural circuits.

A growing body of work has developed biologically inspired models that incorporate architectural, functional, or anatomical properties of the visual system. Early hierarchical models, such as HMAX [45], used alternating layers of selectivity and invariance to approximate computations performed along the primate ventral visual pathway. Subsequent task-optimized deep neural networks demonstrated that training for object recognition can produce internal representations that predict neural responses in the primate visual cortex [54].

Other approaches have incorporated temporal processing and recurrent computation. PredNet [30] implements a hierarchical predictive coding architecture in which recurrent units propagate prediction errors during video processing, while task-driven convolutional recurrent networks and CORnet-S [27] use recurrent connections to better account for the temporal dynamics and hierarchical organization of cortical visual processing [35].

Models have also introduced more explicit biological constraints. VOneNet [7] places a fixed, VISp-inspired front end containing Gabor-like filters, nonlinearities, and neuronal

stochasticity before a conventional convolutional neural network. Topographic deep artificial neural networks (TDANNs) [31] incorporate spatial constraints that reproduce aspects of the functional organization observed across the ventral visual cortex, whereas self-supervised predictive models with parallel processing pathways can develop functional specializations resembling those of distinct cortical visual pathways [33].

More recently, models trained through self-supervised learning have been developed specifically to explain representations in the mouse visual cortex [36]. MouseNet [48] similarly incorporates anatomical constraints derived from the Allen Mouse Brain Connectivity Atlas [37], including the organization and relative connectivity of mouse visual areas, to provide a more biologically grounded model of mouse visual processing. Collectively, these models demonstrate that incorporating recurrence, predictive learning, cortical topography, parallel pathways, and anatomical constraints can improve the biological relevance of artificial visual systems.

Among these biologically grounded models, MouseNet [48] is particularly relevant to the present work because its architecture explicitly reflects the organization and connectivity of the mouse visual system. Unlike conventional sequential CNNs, MouseNet explicitly represents multiple cortical regions and anatomically constrained feedforward connections between them. Although biological visual cortex contains extensive reciprocal and recurrent connectivity, these connections are not represented in the feedforward MouseNet architecture. The architecture represents dLGN, VISp, and the higher visual areas VISal, VISl, VISli, VISpl, VISpor, and VISrl. Each visual area is divided into representations of layers 4, 2/3, and 5, and projections between regions are implemented using convolutional operations.

Structural parameters were constrained, where measurements were available, using estimates of neuron numbers, interareal connectivity, projection extent, and connection statistics between cortical layers [37, 48]. Where direct anatomical constraints were unavailable, the model used simplifying assumptions, including manually selected stride values for some projections. The resulting architecture therefore preserves selected organizational properties of the mouse visual system while remaining compatible with contemporary deep-learning frameworks.

On classification benchmarks, MouseNet achieved more than 90% validation accuracy on CIFAR-10 and approximately 37% top-1 validation accuracy on ImageNet, compared with approximately 60% for VGG16 using the same 64×64 input [48]. MouseNet used substantially fewer parameters than VGG16, so this comparison was not parameter matched. Using neural recordings from the Allen Brain Observatory Visual Coding dataset [8], the authors applied representational similarity analysis to compare MouseNet activations with

responses from mouse visual cortex [8, 25]. Because individual model components correspond to specified cortical regions and layers, the architecture supports direct comparisons between intermediate network representations and experimentally recorded neural activity.

MouseNet was developed exclusively as a visual model and does not incorporate tactile or other sensory pathways [48]. This limits its ability to investigate multisensory processing during natural rodent behaviour. Whisker-mediated sensing provides rodents with information about object identity and location [9], while experimental evidence demonstrates direct interactions between somatosensory and visual cortical circuits [52]. Extending MouseNet with a whisker-processing pathway and a mechanism for cross-modal interaction provides a means of investigating how visual and tactile signals jointly influence cortical representations.

2.3 The Mouse Whisker System

In addition to vision, mice rely heavily on their facial whiskers, or vibrissae, to acquire information about their surroundings [9]. Whiskers function as highly sensitive tactile sensors that allow animals to detect object location, shape, texture, and spatial relationships through active exploration. Unlike passive sensory systems that primarily receive external stimuli, whisker sensing is characterized by active movement, with rodents rhythmically sweeping their whiskers across the environment to generate sensory input [22]. This process enables rodents to gather detailed information about nearby objects even under conditions where visual information is limited or unavailable.

The vibrissal system is one of the most extensively studied sensory systems in neuroscience because of its experimental accessibility and clearly defined anatomical organization [9]. Whiskers are arranged in a regular grid on the rodent snout, and their mechanical deflection during object contact produces sensory signals that travel through the brainstem and thalamus to the somatosensory cortex [40]. Along this pathway, the spatial identity of individual whiskers is largely preserved.

Tactile information from the whiskers is primarily processed within the barrel cortex, located in the primary somatosensory cortex (SSp-bfd) [9]. The barrel cortex contains a distinctive anatomical organization in which groups of neurons are arranged into barrel-shaped structures corresponding to individual whiskers. This one-to-one mapping provides a highly structured representation of tactile information and has made the whisker system a valuable model for investigating sensory coding and neural computation [40].

Neural activity within the whisker system encodes a variety of behaviourally relevant

variables. Experimental studies have demonstrated that cortical responses are influenced not only by whisker contact events but also by variables such as whisker position, movement, object location, and sensorimotor context [9, 1]. As a result, tactile perception emerges from the integration of sensory signals with information related to ongoing whisker motion. These findings suggest that temporal dynamics play an important role in whisker-based perception and motivate the inclusion of temporal processing mechanisms in computational models of tactile sensing.

Beyond the primary somatosensory cortex, whisker information is further processed in higher-order cortical regions, including secondary somatosensory cortex (SSs) [40, 29]. These higher-order areas contribute to object discrimination, sensorimotor processing, and interactions with other sensory modalities. The existence of these hierarchical processing pathways parallels organizational principles observed within the visual system and suggests that tactile perception emerges through distributed processing across multiple cortical regions rather than within a single cortical area.

2.3.1 Available Barrel Cortex Data

Experimental data from the barrel cortex span several recording methods and behavioural conditions. Studies using passive, experimenter-controlled whisker deflections provide precise control over stimulus timing and have characterized temporal integration across cortical layers [42]. These data are useful for isolating stimulus-response relationships, although they do not capture the sensorimotor interactions associated with active whisking.

Experiments involving active touch provide complementary data collected during behaviour. Peron et al. [39] used volumetric two-photon calcium imaging to sample approximately 75% of excitatory layer 2/3 neurons within the relevant imaged regions of barrel cortex while mice performed an object localization task. Sofroniew et al. [49] combined two-photon imaging and extracellular electrophysiology to measure responses across cortical layers while mice used their whiskers to follow walls in a virtual environment. These datasets characterize population responses to object location, whisker movement, and touch during active sensing.

The Rodgers dataset [46] used in the present work provides high-speed videography of multiple individually tracked whiskers together with aligned behavioural events and extracellular recordings spanning the depth of barrel cortex. Recordings were collected while mice performed shape discrimination and shape detection tasks, permitting comparisons between tactile inputs and neural activity. Compared with calcium imaging datasets,

these electrophysiological recordings provide greater temporal resolution and direct measurements of spiking activity, although they sample fewer neurons. The combination of tracked whisker contacts and simultaneously recorded cortical activity makes this dataset particularly suitable for evaluating temporally structured representations learned by the whisker-processing pathway developed in this thesis.

2.3.2 Computational Models of Barrel Cortex Processing

Computational models of the whisker system have been developed at several levels of biological detail. Early circuit models focused on transformations between the thalamus and layer 4 of barrel cortex. Pinto et al. [41] showed that strong cortical inhibition could produce a damping regime that accounts for the sensitivity of cortical responses to thalamic synchrony and the sharpening of receptive fields between the thalamus and cortex. Sharp et al. [47] simulated multiple barrel columns using leaky integrate-and-fire neurons and reproduced experimentally observed thalamocortical response transformations and the propagation of activity between columns.

Other models have investigated computations occurring during active touch. Peron et al. [38] combined behavioural recordings, targeted perturbations, and a recurrent network model of layer 2/3 to show that enhanced connectivity among similarly tuned neurons can selectively amplify touch-related signals. These circuit-level models provide mechanistic accounts of particular neural response properties but are generally not trained to perform complete recognition tasks.

At the sensor level, WHISKiT [57] provides a dynamical model of the rodent head and vibrissal array that can generate physically grounded tactile signals during interactions with objects. More recently, task-optimized convolutional recurrent networks have been trained on realistic whisker-derived input sequences and compared quantitatively with recordings from rodent somatosensory cortex [6]. These models demonstrate that temporal recurrence and task optimization can improve correspondence with biological tactile representations. However, existing models primarily treat whisker processing in isolation. The model developed in this thesis instead integrates a temporal whisker pathway with an anatomically organized visual network, allowing the influence of tactile information on visual representations to be investigated within a unified multisensory architecture.

The present work focuses specifically on the tactile-to-visual direction of this interaction. This choice was motivated by Weiler et al. [52], who identified a direct SSp-bfd-to-VISp pathway through which whisker stimulation suppresses visual activity via feedforward inhibition in VISp layer 2/3. Under the conditions examined in that study, visual stimulation

did not produce a corresponding population-level modulation of SSp-bfd activity, suggesting an asymmetric effect. This experimentally characterized pathway therefore provided a specific anatomical and functional constraint for the cross-modal interaction implemented in the present model.

2.4 Multisensory Integration in the Mouse Cortex

Natural perception relies on the integration of information from multiple sensory modalities. Rather than processing sensory inputs in complete isolation, the brain combines visual, auditory, somatosensory, and other sensory signals to generate coherent representations of the external environment [51, 14, 12]. Multisensory integration has been shown to improve perceptual accuracy, reduce uncertainty, and support robust behaviour in complex environments [51]. Understanding how information from different sensory systems is combined has become an important objective in both experimental and computational neuroscience.

Historically, sensory processing was often viewed as a largely modular process in which primary sensory cortices operated independently before information was integrated within higher-order association areas. However, accumulating experimental evidence has challenged this view by demonstrating that sensory activity in one modality can influence neural responses within cortical regions traditionally associated with another modality [14]. These findings suggest that cross-modal interactions occur throughout the cortex and may play an important role in shaping sensory representations.

In rodents, interactions between visual and somatosensory systems are particularly relevant because both modalities contribute to navigation, object localization, and environmental exploration [52]. During natural behaviour, whisker contacts frequently occur alongside visual observations of nearby objects, providing complementary sources of information regarding object identity and spatial location. The integration of these signals enables animals to form richer representations of their surroundings than would be possible using either modality alone [12].

Recent experimental work has provided direct evidence for visual-somatosensory interactions within the mouse cortex [52, 12]. In particular, studies have demonstrated that somatosensory activity can influence neural responses within visual cortical regions through long-range corticocortical projections [12]. In particular, Weiler et al. [52] demonstrated that whisker stimulation can suppress visually evoked activity in VISp through a direct corticocortical pathway from SSp-bfd. This finding indicates that tactile information is

capable of modulating visual representations even at relatively early stages of cortical processing.

Among cortical regions anatomically classified as higher visual areas, the rostrolateral visual area (VISrl) was selected because single-cell imaging has identified bimodal neurons in this area that respond to both visual stimuli and whisker deflections [17]. Despite its classification as a visual area, these multisensory response properties suggest that VISrl also participates in cross-modal processing. Many of these neurons prefer corresponding locations in visual and tactile space, providing a shared representation of peri-personal space.

VISrl is situated between primary visual cortex (VISp) and primary somatosensory cortex (SSp-bfd) [17]. Experimental evidence indicates that VISrl contains multimodal neurons and exhibits whisker-related activity, motivating its use as a higher-order region for visual and tactile representation learning [17, 50]. Developmental work further suggests that correlated activity between VISp and SSp-bfd helps establish co-aligned visual and somatosensory projections within VISrl [12]. Separately, the direct SSp-bfd-to-VISp inhibitory circuit identified by Weiler et al. [52] motivated the whisker-conditioned gate applied within VISp.

While biologically inspired visual models such as MouseNet [48] provide useful frameworks for investigating visual computation, they do not explicitly represent tactile sensory pathways or mechanisms for cross-modal interaction. Incorporating multisensory processing into such models offers an opportunity to investigate how tactile information may influence visual representations and to explore the computational consequences of biologically motivated sensory integration.

2.5 Self-Supervised Representation Learning

Self-supervised representation learning uses the structure of unlabelled data to generate training targets, removing the need for manually assigned category labels [3, 20]. In computer vision, prominent approaches include contrastive learning, non-contrastive bootstrap methods, and joint-embedding predictive methods [5, 19, 16, 2, 43]. These approaches differ primarily in how they define related examples, prevent representational collapse, and determine which aspects of an input should be preserved in the learned representation.

Contrastive methods learn representations by increasing the similarity between positive pairs while reducing similarity to negative examples. SimCLR [5] constructs positive pairs

from two augmented views of the same image and treats views of other images in the minibatch as negatives. Its effectiveness depends strongly on the choice of augmentations, the use of a nonlinear projection head, and sufficiently large minibatches. Momentum Contrast (MoCo) [19] addresses the dependence on large batches by maintaining a queue of negative representations produced by a slowly updated momentum encoder. The queue provides a large and relatively consistent set of negatives while allowing training with smaller minibatches. Both methods learn invariances determined by their data augmentations, but they require mechanisms for selecting and maintaining negative examples.

Non-contrastive approaches demonstrate that explicit negatives are not always required. Bootstrap Your Own Latent (BYOL) [16] uses an online network to predict the representation produced by a target network for a different augmented view of the same image. The target network is updated using an exponential moving average of the online network, while gradients are applied only through the online pathway. This design enables BYOL to avoid collapsed representations without explicitly separating different images in the embedding space.

Joint-embedding predictive architectures provide another alternative. The Joint-Embedding Predictive Architecture (I-JEPA) [2] learns by using the representation of a visible context region to predict the latent representations of masked target regions within the same image. Unlike pixel reconstruction methods, I-JEPA predicts abstract representations rather than the exact appearance of missing pixels. It is therefore intended to preserve semantic and spatial information while reducing reliance on hand-designed transformations and low-level reconstruction objectives.

Multimodal contrastive learning extends these principles to paired observations from different modalities. Contrastive Language-Image Pre-training (CLIP) [43] uses separate visual and language encoders and trains them so that matched image-text pairs have similar embeddings while mismatched pairs are separated. The present thesis adopts an analogous CLIP-style objective for paired visual and whisker-derived inputs. Unlike SimCLR, MoCo, BYOL, and I-JEPA, which primarily construct related views within one modality, the multimodal objective uses correspondence between sensory modalities as its supervisory signal. This makes it directly suited to investigating whether visual and tactile pathways can learn a shared representation without category labels.

Self-supervised objectives are relevant to computational neuroscience because they provide testable accounts of how structured representations may emerge from sensory experience without explicit category supervision. They should not, however, be interpreted as direct models of biological learning mechanisms. Their value in the present context is that they allow the learned representations to be evaluated against neural data and compared

across alternative training objectives.

2.6 Neural Representation Analysis

A central challenge in computational neuroscience is determining whether artificial neural networks and biological neural systems represent the same stimuli in similar ways. Direct comparisons between individual biological neurons and artificial network units are often difficult to interpret because the two systems differ in dimensionality, population size, tuning properties, and recording modality [24]. Consequently, several methods have been developed to compare representations without assuming a one-to-one correspondence. These methods formalize different notions of similarity and can produce different conclusions about the same pair of representations [53].

Representational Similarity Analysis (RSA) compares the geometry induced by a common set of stimuli [25]. Responses to each stimulus are used to construct a representational dissimilarity matrix (RDM), whose entries describe the pairwise dissimilarities between stimulus and response patterns. The upper triangular entries of RDMs obtained from two systems can then be correlated. Because RSA compares relationships among stimuli rather than individual units, it can be applied when representations have different dimensionalities and no meaningful neuron-to-unit correspondence. A high RSA score indicates similar stimulus geometry, but does not demonstrate that individual units have equivalent tuning or that one representation can directly predict the other.

Centered Kernel Alignment (CKA) [23] is another geometry-based method that compares centered similarity matrices constructed from two representations. Linear CKA is invariant to orthogonal transformations and isotropic scaling and can compare representations containing different numbers of units. It has been used extensively to compare layers of neural networks trained from different initializations. Although CKA and RSA are mathematically related under particular choices of kernels, centering, and similarity measures, they need not produce identical results in practice because their preprocessing and normalization assumptions can differ.

Canonical-correlation-based methods instead search for linear combinations of units that are maximally correlated between two representations. Singular Vector Canonical Correlation Analysis (SVCCA) [44] first uses singular value decomposition to retain dominant representational dimensions and then applies canonical correlation analysis. Projection-Weighted Canonical Correlation Analysis (PWCCA) [34] extends this approach by weighting canonical directions according to their contribution to the original representation.

These methods are useful when two systems encode similar information in different linear bases. Because it permits each representation to be remixed through learned linear combinations, it can assign high similarity to populations that encode the same information even when their individual units have different response profiles.

Predictive methods evaluate whether activity in one system can be used to predict activity in another. For example, neural predictivity fits a regularized linear mapping from model features to recorded neural responses and evaluates the mapping on held-out stimuli [54]. This approach directly tests whether neural responses are linearly accessible from model representations, but it is directional and depends on the amount of available data, the dimensionality of the representations, the regularization procedure, and the cross-validation design. Flexible mappings can also remove representational differences that would remain visible to geometry-based comparisons.

More restrictive alignment methods include orthogonal Procrustes analysis and unit-matching approaches. Orthogonal Procrustes analysis finds the rotation or reflection that best aligns two representation matrices while preserving their internal Euclidean geometry [10]. Unit-matching methods instead compare individual response profiles: hard matching searches for a one-to-one permutation of units, while soft matching uses continuous assignments so that representations with different numbers of units can be compared [21]. These methods ask whether representations share geometric structure or individual-unit tuning under constrained alignments, rather than whether one representation can be transformed into the other by an unrestricted linear mapping. Recent systematic comparisons show that these stricter and more flexible metrics capture complementary notions of representational correspondence and can rank the same model pairs differently [53].

RSA was selected as the primary method in this thesis because the model and neural datasets contain responses to corresponding stimulus conditions but differ substantially in the number and identity of their units. This is especially important for the analysis of the whisker encoder with the Rodgers dataset [46], which contains only six experimental conditions. RSA permits direct comparison of the pairwise condition structure without estimating a neuron-to-unit transformation. RSA scores quantify whether model and neural populations organize the evaluated stimuli similarly. They should therefore be interpreted as evidence of shared representational geometry rather than as proof of identical neural mechanisms or individual-unit correspondence.

2.7 Summary and Research Gap

The literature reviewed in this chapter highlights significant advances in both computational neuroscience and machine learning. Experimental studies have established the mouse visual system as a valuable model for investigating sensory processing, while large-scale anatomical datasets have enabled the development of biologically inspired neural network architectures such as MouseNet. These models provide a framework for studying the relationship between neural structure, learning, and cortical representations while supporting direct comparisons with experimentally recorded neural activity.

At the same time, extensive research on the rodent vibrissal system has demonstrated the importance of whisker-mediated tactile sensing for perception and behaviour. The hierarchical organization of somatosensory pathways, together with the availability of neural recordings from whisker-related cortical regions, provides a strong foundation for the development and evaluation of biologically inspired models of tactile processing.

Recent experimental work establishes direct and associative routes for visual-tactile interaction in mouse cortex, including somatosensory-to-visual feedforward inhibition and co-aligned visual and tactile representations in higher visual area VISrl [52, 12, 17]. Existing computational studies, however, address distinct parts of this problem. Goal-driven and task-optimized whisker models learn tactile object representations from simulated whisker mechanics and can reproduce aspects of rodent somatosensory activity, but do not include a visual pathway [56, 6]. Conversely, MouseNet models the anatomically constrained mouse visual hierarchy without whisker input [48]. Visual-tactile systems have also learned cross-modal material and object representations, but are not organized according to mouse cortical anatomy or evaluated against mouse neural recordings [55, 28].

The closest biologically grounded multisensory models explicitly connect VISp, SSp, and VISrl. Dwulet et al. [12] model how correlated spontaneous activity can refine co-aligned VISp-VISrl and SSp-VISrl projections during development. Guyoton et al. [17] use recurrent excitatory-inhibitory rate networks representing VISp, SSp, VISrl, and VISal to explain the transfer of a learned spatial sensorimotor rule between touch and vision. These studies focus on developmental map formation and cross-modal rule generalization, respectively. They do not combine a visual hierarchy with a spatiotemporal whisker encoder, train the resulting system on paired visual and simulated whisker observations, or test how tactile modulation alters the correspondence between learned visual representations and visual cortical activity.

The contribution of this thesis is the combination of an anatomically constrained MouseNet visual pathway and a spatially and temporally structured whisker pathway. The

principal architectural extension to MouseNet is a learned whisker-conditioned inhibitory gate applied at the VISp4-to-VISp2/3 projection, providing a direct route through which tactile representations can modulate the visual hierarchy. The model is trained using a self-supervised contrastive objective and is evaluated against visual and somatosensory neural recordings. Multimodal behaviour is examined through cross-modal embedding alignment, the functional effects of inhibitory gating, and orientation selectivity following multimodal training. This framework tests how biologically motivated tactile modulation alters visual representations and neural alignment within a unified system inspired by the mouse cortex.

Chapter 3

Methods

3.1 System Overview

This thesis investigates multimodal sensory integration in biologically inspired neural networks through the incorporation of whisker-based tactile sensing. Rodents rely heavily on both vision and active whisker sensing to interact with their environment, particularly during object localization and shape discrimination tasks. Although previous biologically inspired convolutional models such as MouseNet primarily focus on visual processing pathways, fewer models incorporate tactile sensory representations and cross-modal interactions between whisker and visual systems.

To study these interactions, a multimodal framework was developed consisting of a simulated whisker environment, a biologically structured whisker encoder, and a visual encoder based on MouseNet. A simulation environment developed with the Unity game engine (Unity Technologies; https://unity.com) was used to generate trial-paired visual observations and whisker-contact data during controlled object-interaction tasks. Visual observations from the environment were processed through the MouseNet visual pathway, while whisker contact signals were processed using a custom whisker encoder organized according to anatomical structures inspired by the rodent barrel cortex and associated somatosensory pathways. The resulting visual and whisker embeddings were jointly trained using a contrastive learning objective to encourage alignment between the two sensory modalities.

In addition to multimodal representation learning, this work also investigates cross-modal inhibitory interactions between tactile and visual pathways. Motivated by exper-

imentally observed somatosensory-to-visual inhibition discussed in Section 2.4, whisker-derived embeddings were used to modulate intermediate visual representations through a learned feature-wise inhibitory gating mechanism. Representational similarity analysis (RSA) was used to compare learned representations against neural recordings from mouse sensory datasets, including an object classification dataset by Rodgers and the Allen Brain Observatory. Mouse cortical representations can likely be better explained by a combination of mechanisms, including unimodal self-supervised learning and reinforcement learning. However, this thesis focuses on the effects of cross-modal self-supervised learning, which have not been previously examined in the mouse cortical models reviewed in Chapter 2. The overall multimodal framework developed in this thesis is illustrated in Figure 3.1.

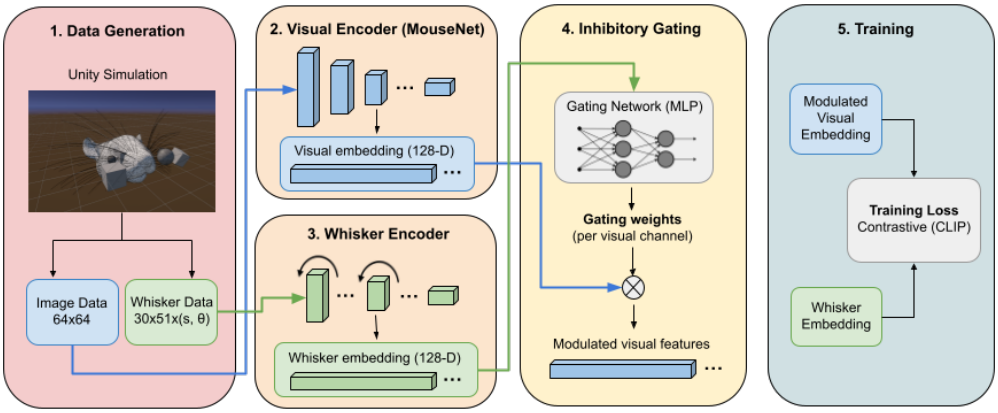


Figure 3.1: **Overview of the proposed multimodal architecture.** Visual and whisker data are generated in a Unity simulation environment and processed by biologically inspired visual and whisker encoders. Whisker-derived gating signals modulate visual representations during multimodal training using a contrastive objective.

3.2 Unity Simulation Environment

A synthetic data-generation environment was developed in Unity 6 (version 6000.0.35f1) to produce paired visual and tactile observations. The environment consists of a head-fixed mouse with simulated whisker kinematics. Two related datasets were generated using the same head model, whisker kinematics, and contact detection system. The first contained

paired visual and whisker observations for multimodal contrastive training. The second contained whisker observations from fixed concave and convex surfaces for whisker-only evaluation.

3.2.1 Whisker Geometry and Kinematics

The simulated whisker array was developed using a model of the mouse head provided by Kevin Kleczka of Northwestern University. The model used WHISKiT Physics [57], a three-dimensional biomechanical simulator of the complete rodent whisker array, to generate the corresponding whisker configurations. The resulting kinematic data described the mouse whisker array at protraction angles from -10° to 40° relative to rest, in one-degree increments. These data were converted into a representation compatible with Unity and were used to prescribe whisker motion directly.

The array contained 30 right-side whiskers. At each of the 51 protraction angles, every whisker was represented by 100 ordered three-dimensional points describing its centreline. For use in Unity, the coordinates were organized by side, pose, whisker identity, and centreline-point index. During initialization, Unity loaded these coordinates, exchanged the y- and z-axes to match the Unity coordinate convention, and transformed the resulting centrelines into world coordinates.

The 51 poses formed an ordered sequence spanning the whisking trajectory. Unity could reconstruct any pose directly from its centreline coordinates or linearly interpolate corresponding points in adjacent poses to produce a smooth whisking animation. The head remained stationary, and whisker geometry followed the predefined trajectory rather than deforming in response to contact.

3.2.2 Whisker Contact Detection

The 100-point centreline was used for rendering and precise contact localization. Because consecutive centreline points define 99 intervals, assigning a separate collider to every interval would have required 99 moving colliders for each whisker, or 2970 colliders across the complete whisker array. To reduce this computational cost, each whisker was instead divided into five broad regions, each represented by a kinematic capsule collider. These capsules moved with the prescribed whisker trajectory and provided an initial, approximate test for overlap with a stimulus. Whisker-to-whisker collisions were disabled, and stimulus objects remained static so that they could not be displaced by the kinematically controlled whiskers.

When a capsule overlapped a stimulus collider, the overlap identified the approximate portion of the whisker containing the contact. The corresponding section of the complete 100-point centreline was then sampled from base to tip. Successive samples were tested until the algorithm found the first transition from outside to inside the configured contact-distance threshold. This interval was refined using eight iterations of bisection: at each iteration, the midpoint was tested and the half containing the outside-to-inside transition was retained. The resulting position approximated the first location at which the whisker centreline entered the contact region. This point was then projected onto the centreline so that its cumulative arc length could be calculated. Its cumulative arc length was used to calculate the normalized contact coordinate

$$s = 1 - \frac{\ell_{\text{contact}}}{\ell_{\text{total}}},$$

where ℓ_{contact} is the centreline arc length from the whisker base to the estimated contact point and ℓ_{total} is the total whisker length. Thus, larger values of s indicate contact closer to the whisker base, with $s = 1$ at the base and $s = 0$ at the tip. A value of zero was also used to represent no detected contact. If multiple stimulus colliders contacted the same whisker during a sampling interval, the base-most contact was retained.

Contact was evaluated repeatedly throughout the whisking sequence rather than being latched at the first contact event. At each recorded sample, the current contact value was written to the corresponding whisker CSV and the contact state was cleared. Continued overlap generated another contact through Unity's trigger-stay callback during the following physics update, producing consecutive nonzero samples for sustained interactions. Because the whiskers followed prescribed kinematic trajectories and did not deform in response to contact, this signal represents the location and duration of geometric whisker and stimulus overlap rather than contact force or mechanically induced whisker bending.

The contact representation used here is deliberately simpler than that of WHISKiT Physics. WHISKiT incorporates realistic whisker-array morphology and experimentally constrained whisker dynamics to predict time-varying mechanical signals, including three-dimensional forces and moments at each whisker base. In contrast, the whisker centrelines in the present simulation followed prescribed, non-deforming trajectories, and tactile input was reduced to the normalized contact location s at each sample. This simplification supported efficient generation of trial-paired visual observations and tactile sequences across 100,000 Unity trials. Unity was chosen because the present study investigates cross-modal representation learning rather than whisker biomechanics. Consequently, the simulated tactile signals should be interpreted as patterns of contact location and duration, not as estimates of bending moment, contact force, or vibration responses.

3.2.3 Multimodal Training Trials

The multimodal training dataset contained $100,000$ independently generated Unity trials. Each trial presented a randomized field of five primitive objects surrounding the head-fixed mouse. Every object was independently selected as a cube or sphere with equal probability. Object diameter was sampled uniformly between 3 and 8 mm. Object centres were sampled within scene bounds of -20 to 20 mm laterally, -0.5 to 0.5 mm vertically, and -10 to 10 mm along the depth axis. Candidate placements were rejected if they entered a 5 mm exclusion region surrounding the head or overlapped a previously placed object. Up to 100 placement attempts were allowed for each object.

The dataset size of 100,000 trials was selected as a compromise between environmental diversity and computational feasibility. Generating each example required rendering an image, simulating contact for 30 whiskers over 51 temporal samples, and storing the resulting paired observations. Object identity, size, position, illumination, and ground colour were sampled independently, so each trial represented a distinct scene configuration rather than a repeated observation. Nevertheless, 100,000 examples constitute a relatively small dataset for self-supervised learning. The resulting model is therefore intended as a proof of concept for cross-modal learning in a biologically structured architecture, rather than a demonstration that representation quality had saturated at this dataset scale.

The accepted objects used Unity primitive colliders and remained static throughout the trial. No rigid body was attached to the stimuli, preventing the kinematically controlled whiskers from displacing them. Thus, contact altered the recorded value of s but did not alter either the prescribed whisker trajectory or the position of the object. Environmental appearance was varied through changes in sky illumination and ground colour, introducing visual variation that was unrelated to tactile contact.

Visual observations were obtained using a camera placed at the right eye of the mouse. The camera rendered an intermediate 256×256 image, which was downsampled to a 64×64 RGB image to align with the input to MouseNet. The final training configuration used a horizontal field of view of approximately 90°. A single image was captured at the first recorded tactile sample of each trial. Representative right-eye observations and an overview of the corresponding Unity simulation environment are shown in Figure 3.2.

Following image capture, the continuous interpolated whisking animation remained active while 51 tactile samples were recorded on consecutive Unity updates. For each of the 30 whiskers, a CSV file stored the normalized contact position, s, and the nominal interpolated protraction angle, θ, at every sample. Each multimodal trial consisted of an image taken from the right eye of the mouse and a trial-level paired tactile sequence with dimensions $30 \times 51 \times 2$.

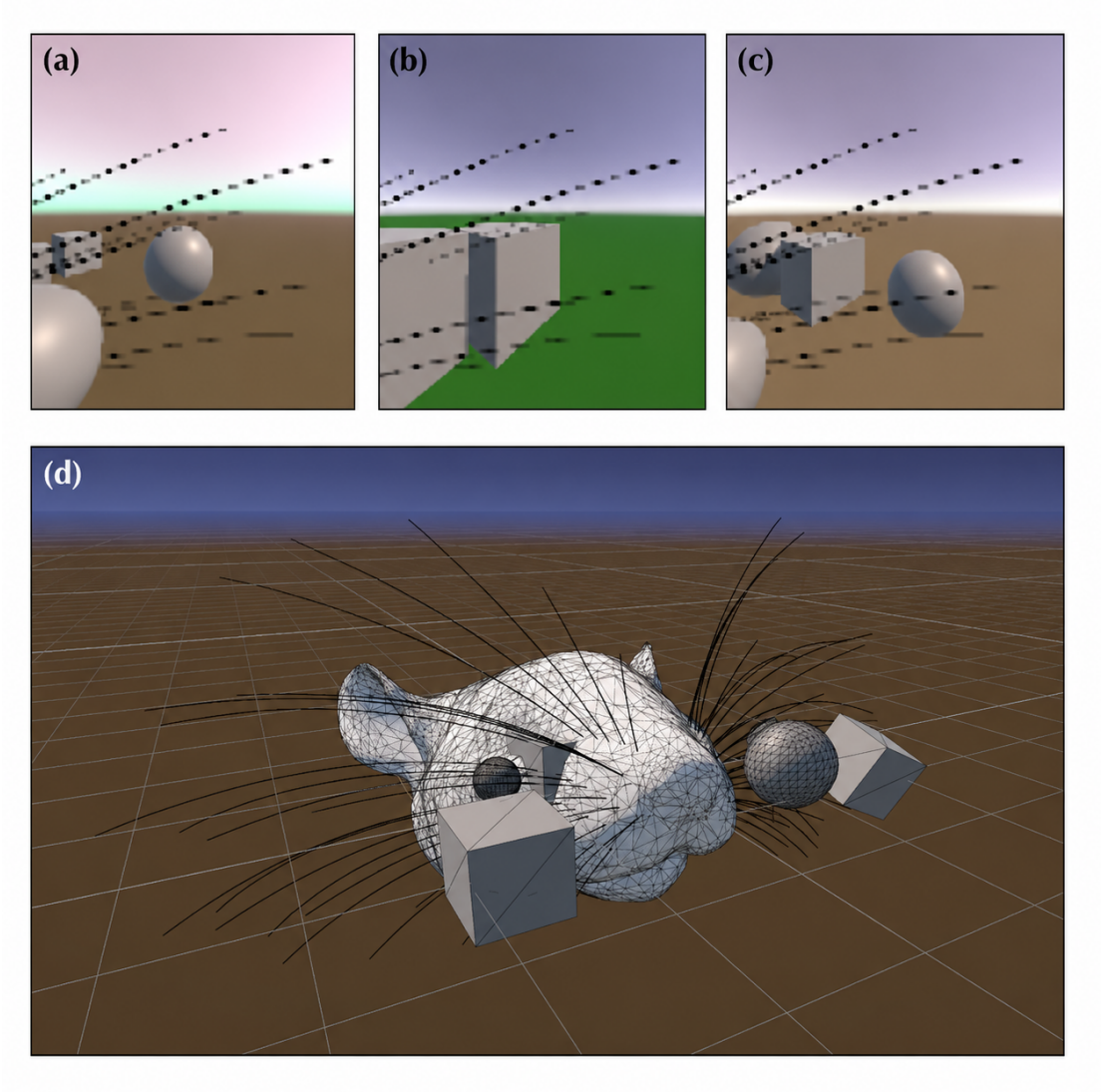


Figure 3.2: **Representative Unity simulation environment and visual observations used for multimodal training.** (a–c) Example images captured from the simulated mouse's right-eye camera across independently generated trials, illustrating variation in object geometry, position, illumination, and ground colour. (d) Unity scene view showing the head-fixed mouse, simulated whisker array, and surrounding primitive objects. Each trial paired one visual observation with the corresponding 51-frame whisker-contact sequence.

3.2.4 Concave and Convex Evaluation Stimuli

A separate Unity simulation generated tactile responses for six conditions defined by the combination of surface curvature (concave or convex) and stimulus position (near, medium, or far). These conditions corresponded to the experimental design of Rodgers [46], in which concave and convex stimuli were presented to mice at three positions while whisker behaviour and neural activity were recorded. Chung et al. [6] used a modified implementation of WHISKiT Physics to generate simulated tactile and mechanical inputs that are subsequently compared with neural responses from the Rodgers dataset. The present Unity implementation used the simulation resources released by Chung et al. to reconstruct stimulus conditions corresponding to the Rodgers experiment. The six reconstructed stimulus configurations are shown in Figure 3.3.

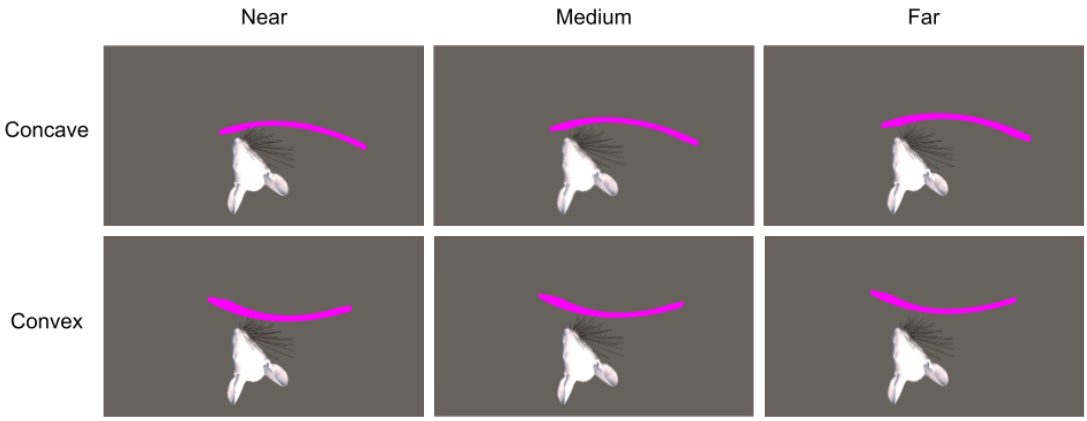


Figure 3.3: **Unity implementation of the six stimulus conditions used for tactile evaluation.** Columns show the near, medium, and far stimulus positions from left to right, while rows show the concave and convex surface geometries, respectively. The magenta surface indicates the curved stimulus presented to the right-side whisker array.

Specifically, the dimensions and orientation of the curved stimulus were derived from the released `convex-shape.obj` mesh, while its poses were reconstructed from the values in `rodgers_stimuli.sh`. The source configuration used the same curved mesh for both curvature conditions, with object rotations of $-25°$ and $155°$ for the concave and convex conditions, respectively. It also specified a head position of $(-5, -10, 0)$ mm, a head orientation of $(0°, 15°, 0°)$, and an object height of 3 mm. In Unity, the contact surface was represented as an extruded circular-arc mesh fitted to the source geometry. Collision de-

tection was approximated using 16 overlapping static box colliders positioned tangent to the arc.

The stimulus poses were transformed from the source simulator's coordinate frame into Unity's coordinate frame, accounting for differences in head pose, origin, and axis convention. For each curvature, the transformed near position was used as the reference placement. The medium and far stimuli were positioned by adding one and two copies, respectively, of the converted source displacement vector to the near position. Stimulus geometry and orientation were held constant across the three positions.

Each condition was evaluated using the same deterministic sequence of 51 whisker poses, corresponding to whisking angles from $-10°$ to $40°$ in $1°$ increments. At each pose, all whiskers were placed directly into the corresponding stored geometry and the Unity transforms were synchronized. Contact was then evaluated using an explicit geometric overlap query between the whisker collision capsules and the compound stimulus colliders. Candidate overlaps were refined against the sampled whisker centreline to estimate the most proximal point of contact.

Following the Chung et al. configuration, the stimulus was positioned on the right side of the whisker array and tactile contact only occurred on the 30 right-side whiskers. For every whisker and pose, the recorded representation contained the normalized contact location s and whisking angle θ. Contact position was expressed along the whisker centreline, with $s = 1$ at the base and $s = 0$ at the tip; non-contacting whiskers were assigned $s = 0$.

Each condition therefore produced an array with dimensions

$$30 \text{ whiskers} \times 51 \text{ poses} \times 2 \text{ features}.$$

The corresponding CSV file contained 30 consecutive rows for each pose and $51 \times 30 = 1530$ rows in total. Because the stimulus geometry, whisker poses, and contact detection procedure were deterministic, a single recording was retained for each condition rather than treating repeated identical simulations as independent observations.

3.3 Whisker Representation

3.3.1 Contact Feature Representation

Whisker-object interactions generated within the simulation environment were converted into structured tactile feature representations for input into the whisker encoder architecture. Initial whisker encoding experiments represented tactile interactions at individual

time points using contact position and protraction angle measurements generated within the Unity simulation environment. Each frame was processed independently before the resulting features were aggregated into a fixed-dimensional embedding. Because these frame-based representations did not explicitly preserve temporal order, they could not directly capture how whisker contacts evolved as the whiskers moved along stimulus surfaces.

This limitation was particularly relevant because neurons in mouse barrel cortex can integrate patterns of whisker stimulation over hundreds of milliseconds, with particularly strong temporal integration observed in a subset of upper-layer neurons [42]. This finding motivated the incorporation of temporal processing at the SSp-bfd2/3 stage of the whisker encoder. The encoder therefore preserved the 51-frame sequence through SSp-bfd4 and into SSp-bfd2/3, where a gated recurrent unit (GRU) integrated the ordered activity across the whisking trajectory before the representation was propagated to higher somatosensory regions. The implementation of this temporal integration is described in Section 3.4.3.

For each whisker at each temporal frame, a normalized contact position variable s was used to describe the radial distance of contact along the whisker shaft, as described in Section 3.2.2. Whisker protraction angle θ was additionally recorded for each temporal frame to capture whisker motion throughout the whisking cycle: $\theta_t \in [-10°, 40°]$. Before being provided to the whisker encoder, the angular component was normalized using

$$\tilde{\theta} = \frac{\theta - 15}{25}.$$

Because the simulated sweep ranged from $-10°$ to $40°$, this transformation mapped the angular component exactly to $[-1, 1]$. The resulting input for each whisker and time point therefore consisted of the normalized contact coordinate s and normalized angular position $\tilde{\theta}$. Together, these features enabled whisker interactions to be represented using compact spatiotemporal descriptors aligned with whisker-centered tactile sensing rather than global scene coordinates.

3.3.2 Temporal and Spatial Organization of Whisker Signals

Whisker interactions were represented as temporally ordered sequences spanning the duration of an active whisking cycle. For each trial, tactile features were recorded across 51 temporal frames corresponding to successive whisker protraction states during whisking behaviour. Rather than treating whiskers as independent unordered sensors, whisker signals were organized according to anatomical topology in the barrel cortex. Individual whiskers retained consistent spatial identities corresponding to canonical mouse whisker

labels arranged across rows A-E [40]. This organization preserved neighbourhood relationships between whiskers and enabled tactile information to be represented in a manner consistent with the spatial organization of the barrel cortex. Each of the 30 whiskers was assigned to its corresponding barrel-grid position, with unused grid locations left empty.

For each temporal frame, whisker features were arranged into structured spatial tensors corresponding to the barrel field layout prior to processing by the whisker encoder. This representation enabled convolutional operations within the whisker encoder to capture local spatial relationships between neighbouring whiskers while simultaneously preserving temporal dynamics across the whisking sequence.

The resulting tactile representation therefore encoded both spatial and temporal structure within whisker interactions. Spatial organization preserved anatomically meaningful relationships between whiskers, and temporal organization enabled downstream layers to model the evolution of tactile contact patterns across the prescribed whisking trajectory.

3.4 Whisker Encoder Architecture

3.4.1 Biological Motivation

The whisker encoder architecture, summarized in Figure 3.4, was designed to approximate hierarchical processing within the mouse somatosensory cortex while preserving anatomical organization and temporal dynamics. In rodents, whisker-based tactile information is processed through the barrel cortex, a structured region of primary somatosensory cortex in which individual whiskers correspond to spatially organized cortical "barrels" [40]. This organization preserves spatial relationships between whiskers and supports distributed processing of tactile sensory information during active whisking behaviour.

Motivated by this organization, the whisker encoder was constructed using processing stages corresponding to regions within the mouse barrel cortex and associated somatosensory pathways. Rather than treating whisker contacts as independent feature vectors, tactile representations were processed using spatially structured operations aligned with barrel field topology. This enabled neighbouring whiskers to interact locally through convolutional operations while preserving anatomically meaningful whisker organization throughout the network.

In addition to spatial organization, the whisker encoder incorporates recurrent temporal processing. Each simulated trial is provided to the encoder as a complete sequence of 51 prescribed whisker poses. SSp-bfd4 processes each frame separately while preserving

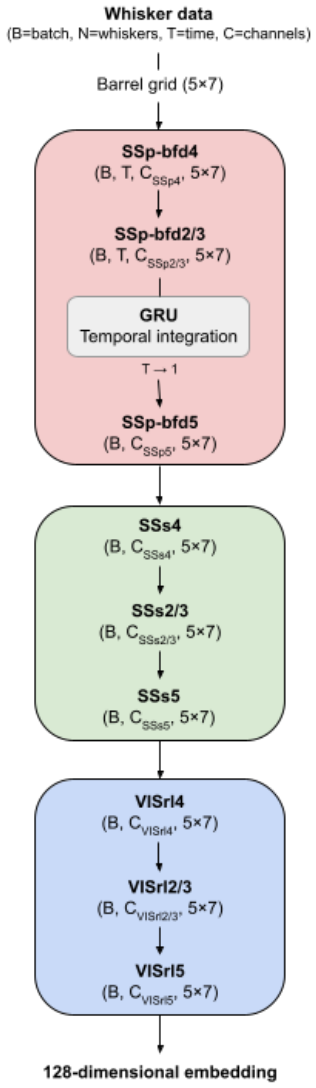


Figure 3.4: **Architecture of the whisker encoder.** The figure presents a simplified area-level overview of the implemented architecture. For readability, each interareal arrow summarizes a set of layer-specific, anatomically masked projections rather than depicting every individual connection. All three SSp-bfd representations project to SSs4, and all three SSp-bfd and SSs representations project to VISrl4. Direct projections originating from SSp-bfd4 use its temporally averaged representation, whereas ordered temporal integration is performed by the GRU in SSp-bfd2/3. The VISrl5 representation is spatially averaged and projected to the 128-dimensional whisker embedding.

the temporal dimension, after which a gated recurrent unit (GRU) in SSp-bfd2/3 processes the ordered sequence recurrently. At each barrel-grid location, the GRU hidden state is updated sequentially across the 51 frames, allowing information from earlier whisker poses to influence the representation of later poses. The final hidden state is then used as the temporally integrated SSp-bfd2/3 representation for subsequent processing. Thus, temporal recurrence occurs within SSp-bfd2/3 even though the simulated sensory trajectory itself is predetermined rather than generated online through closed-loop interaction.

3.4.2 Encoder Overview

The whisker encoder consisted of a directed hierarchy of processing stages with interareal projections organized according to the anatomical pathway of the somatosensory cortex. Tactile feature representations generated from the Unity simulation environment were first arranged into structured barrel field tensors and processed through layers corresponding to regions of the primary barrel cortex (SSp-bfd). Subsequent processing stages propagated tactile information toward the secondary somatosensory cortex and higher-order multimodal regions.

The encoder receives each trial as a tensor containing all 51 temporally ordered whisker poses. These frames remain separate during the initial SSp-bfd4 processing stage. A framewise projection from SSp-bfd4 produces an SSp-bfd2/3 feature map for each time point, yielding a 51-frame sequence of spatial feature maps. At each of the 35 barrel-grid locations, this sequence is then passed through a shared GRU one frame at a time. The recurrent hidden state therefore accumulates information across the ordered whisking sequence, and the hidden state after the 51st frame is projected to form the temporally integrated SSp-bfd2/3 representation. Subsequent processing through SSp-bfd5 and the recurrent-path contributions to higher regions uses this integrated representation.

3.4.3 Temporal Integration in SSp-bfd2/3

Temporal integration was performed within the SSp-bfd2/3 representation using a gated recurrent unit (GRU). Early architectural explorations considered treating whiskers as independent temporal signals processed using separate recurrent mechanisms. However, this approach reduced opportunities for spatial interactions between neighbouring whiskers and did not align well with the distributed spatial organization observed within barrel cortex processing [40, 39]. In the current model, temporal processing was incorporated after whisker features had first been spatially organized into barrel field representations.

At each temporal frame, a shared multilayer perceptron mapped the two tactile input features to the 194 channels assigned to SSp-bfd4. After placement on the 5×7 barrel grid, a shared anatomically masked 3×3 convolution produced a 250-channel SSp-bfd2/3 map at every frame. The resulting sequence was rearranged so that a single GRU with input size 250 and hidden size 64 was applied independently at each of the 35 grid locations. For each grid location, the GRU processed the 51 feature vectors in temporal order, updating a 64-dimensional hidden state at each step as a function of both the current frame and the hidden state from the preceding frame. The final hidden state at each location was projected back to 250 channels and reassembled into a spatial SSp-bfd2/3 representation. This preserved ordered temporal information without constructing a recurrent state from the complete flattened feature map. The exact tensor dimensions throughout the encoder are summarized in Table 3.2.

Not every anatomical pathway used this recurrently integrated representation. For the direct SSp-bfd4-to-SSs4 and SSp-bfd4-to-VISrl4 projections, the 51 SSp-bfd4 feature maps were averaged across the temporal dimension, producing a layer-4 representation of shape $(B, 194, 5, 7)$. This temporally averaged representation was passed through the corresponding anatomically masked 3×3 projections, producing SSs4 and VISrl4 representations with shapes $(B, 69, 5, 7)$ and $(B, 7, 5, 7)$, respectively. These direct pathways therefore conveyed a time-averaged summary of layer-4 activity, whereas ordered temporal information was retained through the recurrent SSp-bfd2/3 pathway.

The GRU summarized the evolution of whisker contacts across the complete 51-pose whisking cycle, reducing the temporal sequence to a single spatial representation. Convolutional processing captured interactions between neighbouring barrel-grid locations at each time point, while recurrent processing retained the order-dependent temporal information required to represent changes throughout the whisking sequence.

3.4.4 Biologically Constrained Architectural Parameters

The dimensionality and convolutional parameters of the whisker encoder were selected using an adaptation of the biologically constrained procedure used to construct MouseNet [48]. Rather than selecting layer widths independently, the number of feature channels was related to estimates of the neuronal population in each represented cortical area and layer. Layer-specific neuron-count estimates for SSp-bfd and SSs were obtained from the Blue Brain Cell Atlas [13]. Since equivalent estimates were not available for VISrl, the population of each VISrl layer was approximated from the corresponding VISl estimate by scaling according to the ratio of their estimated cortical surface areas.

All cortical representations retained a spatial feature map of height $H = 5$ and width $W = 7$. Because a literal one-model-unit-per-neuron representation would require several thousand channels on this comparatively small spatial grid, a uniform biological-to-model compression factor was applied. The channel count of cortical layer l was calculated as

$$C_l = \max\left(1, \left\lfloor \frac{N_l}{HWm} \right\rfloor\right),$$

where N_l is the estimated neuronal population, $HW = 35$ is the number of allocated spatial feature-map positions, and $m = 20$ is the number of estimated biological neurons represented by each model activation. The same value of m was used for every area and layer, preserving relative differences in estimated cortical population sizes while keeping the network computationally tractable. The resulting channel dimensions are shown in Table 3.1.

Table 3.1: **Neuron-count estimates and implemented channel dimensions of the whisker encoder.** All cortical representations retained a 5×7 spatial feature map. Channel counts were calculated as $C_l = \lfloor N_l/(35 \times 20) \rfloor$.

Cortical area	Layer	Estimated neurons	Channels
SSp-bfd	4	135,943	194
SSp-bfd	2/3	175,481	250
SSp-bfd	5	117,751	168
SSs	4	48,397	69
SSs	2/3	95,958	137
SSs	5	65,758	93
VISrl	4	4,927	7
VISrl	2/3	11,831	16
VISrl	5	8,161	11

Temporal integration was performed after framewise spatial processing from SSp-bfd4 to SSp-bfd2/3. The resulting 250-channel SSp-bfd2/3 sequence was rearranged from $(B, T, 250, 5, 7)$ to $(35B, T, 250)$ so that the same recurrent mechanism was applied independently at every barrel-grid location. A single-layer gated recurrent unit used an input dimensionality of 250 and a hidden dimensionality of 64. Its final hidden state was projected from 64 back to 250 channels and restored to a $(B, 250, 5, 7)$ spatial representation. Thus, temporal information from all 51 frames was integrated without spatial averaging before recurrence and without using a recurrent state comparable to the estimated biological population size.

Spatial projection parameters followed the anatomically constrained procedure introduced in MouseNet. Because the whisker encoder used rectangular 5×7 feature maps, $\sqrt{HW}$ was used as the effective linear spatial resolution when converting anatomical projection widths into feature-map units. Under the resulting parameterization, all anatomical projections used 3×3 convolutions with stride one and padding one. As in MouseNet, each convolution used a fixed sparse connectivity mask sampled from a Gaussian profile parameterized by the estimated width and peak probability of the corresponding anatomical projection, while the retained convolutional weights were learned. For interareal projections, the estimated biological in-degree was reduced using the same global compression factor applied to the neuronal populations and distributed among projections converging on the same target. Converging projections were summed before target-specific batch normalization and rectified linear activation.

3.4.5 Interareal Projections and Higher-Order Processing

The connectivity pattern followed the graph-based construction used in MouseNet. Each modelled area contained representations corresponding to layers 4, 2/3, and 5, with the within-area pathway proceeding from layer 4 to layer 2/3 and then layer 5. As in MouseNet, multiple layer-specific projections from lower-order areas converged on layer 4 of a target area, and their outputs were summed before target-specific normalization and nonlinear activation. In the whisker pathway, all three SSp-bfd representations projected to SSs4, while all three SSp-bfd and all three SSs representations projected to VISrl4. This retained MouseNet's principle of convergent, anatomically motivated connectivity, while the 5×7 barrel topology, neuron-count compression, and recurrent temporal integration were adaptations introduced for whisker processing.

The architecture additionally incorporated downstream projections toward visually associated cortical regions, including representations corresponding to the rostrolateral visual area (VISrl). This pathway was motivated by evidence that whisker stimulation evokes activity in VISrl [50] and that VISrl is engaged during whisker-based texture discrimination [15]. By propagating tactile representations toward visually associated regions, the architecture enabled later multimodal interactions between whisker and visually derived representations.

Interareal projections within the whisker encoder were implemented using convolutional operations preserving spatial organization across layers. This allowed tactile information to maintain structured spatial topology throughout successive processing stages rather than collapsing whisker representations during early processing.

Table 3.2: **Tensor shapes through the whisker encoder.** B denotes batch size, and each whisking sequence contains 51 frames. All cortical representations retain a 5×7 spatial feature map until the final spatial average.

Stage	Representation or operation	Tensor shape
Input	Raw whisker sequence	$(B, 30, 51, 2)$
Frame encoding	Shared frame features	$(B, 30, 51, 194)$
SSp-bfd4	Framewise sequence	$(B, 51, 194, 5, 7)$
	Temporal mean	$(B, 194, 5, 7)$
SSp-bfd2/3	Framewise sequence	$(B, 51, 250, 5, 7)$
	Per-site GRU input	$(35B, 51, 250)$
	Per-site GRU output	$(35B, 51, 64)$
	Final GRU hidden state	$(1, 35B, 64)$
	Integrated representation	$(B, 250, 5, 7)$
SSp-bfd5	Layer output	$(B, 168, 5, 7)$
SSs hierarchy	SSs4	$(B, 69, 5, 7)$
	SSs2/3	$(B, 137, 5, 7)$
	SSs5	$(B, 93, 5, 7)$
VISrl hierarchy	VISrl4	$(B, 7, 5, 7)$
	VISrl2/3	$(B, 16, 5, 7)$
	VISrl5	$(B, 11, 5, 7)$
Whisker output	Spatial mean	$(B, 11)$
	Projected embedding	$(B, 128)$

3.5 Visual Encoder

3.5.1 Visual Processing Pathway

Visual processing within the multimodal model was performed using a modified MouseNet architecture based on known anatomical connectivity patterns within the mouse visual cortex. MouseNet is a convolutional neural network designed to approximate hierarchical processing across multiple mouse visual cortical regions while preserving interareal connectivity observed in experimental tracing studies. Within the present work, MouseNet served as a visual backbone for extracting hierarchical visual representations from images generated within the Unity simulation environment.

Static input images were first processed by the dorsal lateral geniculate nucleus (dLGN) before being propagated through the primary visual cortex (VISp) and downstream visual areas. Visual embeddings used for multimodal learning were extracted directly from representations corresponding to the rostrolateral visual area (VISrl). VISrl was therefore selected as a candidate region for multimodal convergence because of its position within higher-order cortical pathways and its experimentally observed involvement in whisker-related processing [50, 15].

The visual pathway used the anatomically motivated MouseNet architecture but did not load pretrained MouseNet weights. Its parameters were randomly initialized and optimized jointly with the whisker encoder, inhibitory gate, and multimodal projection layers using the contrastive learning objective. MouseNet therefore provided the structural organization of the visual pathway rather than pretrained visual representations. The simplified MouseNet architecture, including the whisker-conditioned inhibitory pathway and visual embedding extraction used throughout this work, is illustrated in Figure 3.5.

3.5.2 Visual Input Preprocessing

Input images generated within the simulation environment were converted into grayscale representations prior to processing within MouseNet, which required a two-channel visual input. Although this design was originally motivated by the dichromatic nature of mouse vision (UV- and green-sensitive pathways), explicit UV information was unavailable in both the Unity-generated and Allen Brain Observatory image datasets.

To create a biologically plausible two-channel representation, RGB images were first converted to a grayscale intensity map. This intensity map was then decomposed into ON and OFF channels using a contrast-based preprocessing step. Let x denote the grayscale image and μ the global mean image intensity. The ON and OFF channels were computed as:

$$ON(x) = max(0, x - \mu)$$
$$OFF(x) = max(0, \mu - x),$$

where the ON channel encoded positive deviations from the mean intensity and the OFF channel encoded negative deviations. This decomposition produces separate neural activations for local brightness increments and decrements relative to the overall scene illumination, consistent with mechanisms observed in early visual processing. The resulting two-channel representation was then provided as input to the initial visual layers corresponding to projections from the lateral geniculate nucleus (dLGN) into primary visual cortex (VISp).

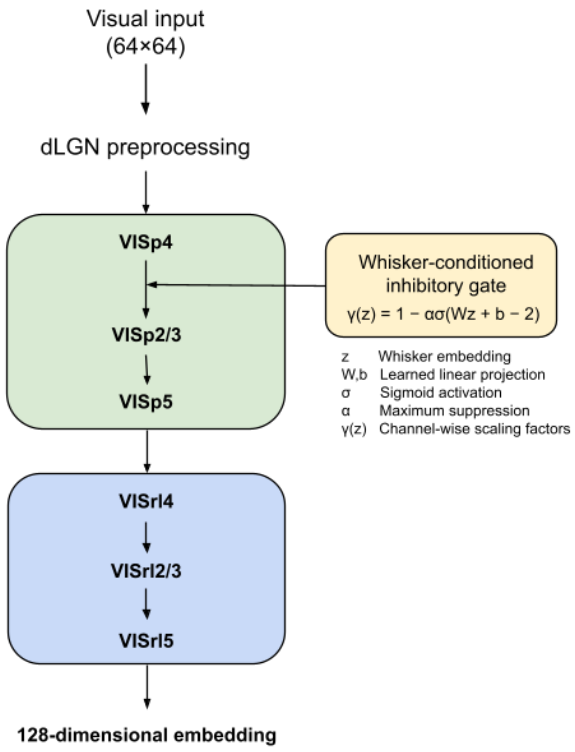


Figure 3.5: **Modified MouseNet visual encoder architecture.** Visual inputs are processed through the dorsal lateral geniculate nucleus (dLGN) for preprocessing before propagating through the primary visual cortex (VISp) and higher visual area VISrl using the anatomically derived MouseNet architecture. A whisker-derived inhibitory gate, motivated by experimentally observed somatosensory modulation of visual cortex, is applied to the VISp2/3 stage. The resulting VISrl5 representation is projected to a 128-dimensional embedding for multimodal contrastive alignment, cross-modal retrieval, and embedding similarity analysis.

3.5.3 Visual Embedding Extraction

For multimodal learning, the VISrl5 activation map produced by the configured retinotopic MouseNet was flattened and projected into a 128-dimensional embedding space for multimodal learning. These visual embeddings were trained jointly with whisker embeddings using a contrastive learning objective designed to align trial-paired visual and tactile observations.

The VISrl5 representation was used as the source of the visual embedding because VISrl is a higher-order cortical region with experimentally observed involvement in whisker-related and cross-modal processing [50, 15]. The resulting visual embeddings encoded hierarchical visual features extracted from the Unity simulation while remaining compatible with downstream multimodal alignment and representational similarity analysis experiments.

3.5.4 Whisker-Conditioned Inhibitory Gating

In addition to contrastive alignment at the embedding level, the model incorporated a direct whisker-to-visual interaction motivated by the somatosensory-to-visual inhibitory pathway discussed in Section 2.4. The interaction is implemented as a learned whisker-conditioned inhibitory gate applied to the projection from VISp layer 4 to layer 2/3.

Cross-modal modulation was applied to the output of the convolutional projection from VISp layer 4 to VISp layer 2/3. Let $F \in \mathbb{R}^{B\times C\times H\times W}$ denote the output of this projection for a batch of B samples, and let $z \in \mathbb{R}^{128}$ denote the corresponding whisker embedding. A learned affine transformation produced one scaling coefficient for each of the C output channels:

$$\gamma(z) = 1 - \alpha\sigma(Wz + b - 2),$$

where $W \in \mathbb{R}^{C\times 128}$ and $b \in \mathbb{R}^{C}$ were learned parameters, and $\sigma(\cdot)$ denotes the sigmoid function. The suppression parameter α was fixed at 0.25 throughout training rather than learned. The resulting coefficients were therefore bounded between 0.75 and 1.0, limiting the attenuation of any channel to a maximum of approximately 25% and preventing complete elimination of visual activity. The offset of -2, together with zero initialization of W and b, initialized the coefficients near the identity transformation, with $\gamma(z) \approx 0.97$ at the beginning of training.

The scaling coefficients were broadcast across the spatial dimensions of the corresponding feature maps and applied multiplicatively:

$$F'_{b,c,h,w} = \gamma_{b,c}(z) F_{b,c,h,w}.$$

Thus, each whisker observation could produce a different pattern of attenuation across VISp layer 2/3 channels. Because a single coefficient was applied at every spatial position within a channel, the gate preserved that channel's spatial activation pattern while changing its magnitude relative to the other channels.

The modulated projection was then passed through the existing VISp layer 2/3 normalization and nonlinear activation operations. Its output propagated through the unchanged MouseNet connections to VISp layer 5 and higher visual regions, including VISrl. Whisker input could therefore affect downstream visual processing by changing the relative gains of the VISp layer 2/3 channels supplying these later regions.

3.6 Multimodal Contrastive Learning

3.6.1 Motivation for Cross-Modal Alignment

During active exploration, mice simultaneously receive visual and tactile sensory information describing the same external environment. Visual observations and whisker interactions contain complementary information regarding object identity, geometry, and spatial structure. Motivated by this relationship, the present work used multimodal contrastive learning to encourage alignment between visual and whisker-derived representations corresponding to the same interaction trial.

Rather than directly supervising object labels or behavioural outputs, the model was trained to associate trial-paired visual and tactile observations within a shared embedding space. This allowed the network to learn relationships between modalities using paired sensory experiences generated within the simulation environment. Each positive pair consisted of the single visual image captured at the first tactile sample and the complete $30 \times 51 \times 2$ whisker sequence from the same Unity trial. The modalities were therefore paired at the trial level rather than aligned as two temporally sampled sequences.

3.6.2 Joint Embedding Space

Visual and tactile representations were independently processed through the MouseNet visual encoder and whisker encoder before being projected into a shared latent embedding

space. Visual embeddings were extracted from VISrl representations generated by the MouseNet backbone, while tactile embeddings were generated from the whisker encoder described in the previous section.

Both modalities were projected into a shared embedding space of fixed-dimensionality, enabling similarity comparisons between visual and tactile representations corresponding to the same interaction trial. Prior to similarity computation, embeddings from both modalities were normalized to unit magnitude. The resulting embedding space was intended to encourage paired tactile and visual observations to occupy nearby regions of the latent representation space while separating unrelated sensory observations.

3.6.3 Contrastive Objective

Multimodal alignment was performed using a symmetric contrastive learning objective similar to those used in CLIP multimodal representation learning frameworks [43]. For a batch of paired visual and tactile embeddings, pairwise cosine similarities were computed between all visual and whisker representations within the batch.

Matching visual-tactile pairs originating from the same simulation trial were treated as positive examples, while embeddings originating from different trials were treated as negative examples. A temperature scaled cross-entropy objective was then applied symmetrically across both modalities:

$$\mathcal{L} = \frac{1}{2}\left(CE\left(\frac{Z_v Z_w^T}{\tau}\right) + CE\left(\frac{Z_w Z_v^T}{\tau}\right)\right),$$

where Z_v and Z_w denote normalized visual and whisker embeddings respectively, and τ denotes the temperature parameter controlling similarity scaling. This objective encouraged embeddings corresponding to paired sensory observations to exhibit high similarity while discouraging similarity between unrelated visual and tactile samples.

3.6.4 Training Procedure

The standard multimodal model was trained by jointly optimizing the MouseNet visual pathway, whisker encoder, inhibitory gate, and modality-specific projection layers using the symmetric contrastive objective described above. The MouseNet pathway was randomly initialized and did not use pretrained weights. Consequently, the visual and tactile representations, inhibitory modulation, and cross-modal alignment developed jointly during training. The training dataset contained 100,000 trial-paired visual snapshots and

tactile sequences generated from the Unity simulation environment. Each batch therefore contained visual and tactile observations from the same object-interaction trials.

To evaluate whether the effects of inhibitory gating were robust to model initialization, training was repeated independently using five random seeds, numbered 42 through 46. For each seed, one standard gated model and one no-gate model were trained from scratch, producing ten trained checkpoints. The random seed was applied to Python, NumPy, and PyTorch random-number generation, including CUDA initialization and data-loader shuffling. Gated and no-gate runs carrying the same seed label therefore used seed-matched initialization of their shared components and seed-matched sample ordering, although their parameters were optimized independently. The five independently trained seeds were used collectively only for the Allen Brain Observatory RSA analysis described in Section 3.7.3. Unless otherwise stated, all other analyses used the final-epoch gated checkpoint trained with seed 42.

In the gated models, the whisker-dependent inhibitory gate remained active throughout training. In the no-gate models, both the inhibitory-gate module and its VISp layer 2/3 forward hook were removed before the first optimization step. The no-gate condition therefore represented an independently trained architectural ablation, rather than a gated model whose gate was disabled only during evaluation.

The seeded models were trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-5} and a batch size of 16. Training samples were shuffled at the beginning of each epoch. The contrastive temperature parameter, τ, was learnable and initialized to 0.1. To ensure positivity, the temperature was parameterized in log space as $\tau = \exp(\eta)$, with η initialized to $\log(0.1)$, and the value supplied to the loss was constrained to the interval $[0.05, 0.30]$. The checkpoint from the final training epoch was retained for each seed and architecture.

3.7 RSA Evaluation

3.7.1 Motivation for RSA

Representational similarity analysis (RSA) was used to compare internal model representations against neural recordings obtained from mouse sensory cortex datasets. Direct comparison between neural activity and deep neural network activations is challenging due to differences in dimensionality, neuron counts, and representational structure. RSA

provides a framework for comparing representations at the level of pairwise stimulus relationships rather than requiring direct neuron-to-neuron correspondence.

Within RSA, neural and model responses are transformed into representational dissimilarity matrices (RDMs) describing the similarity structure of responses across stimuli. These matrices can then be compared statistically to evaluate whether two systems encode stimuli using similar representational geometry. In the present work, RSA was used to evaluate the extent to which visual and tactile representations generated by the model aligned with neural response structure observed in mouse sensory cortex. The overall RSA evaluation procedure used throughout this thesis is illustrated in Figure 3.6.

3.7.2 Whisker Encoder Evaluation

RSA was used to evaluate the whisker encoder against the Rodgers dataset [46]. Concave and convex surfaces were each presented at near, medium, and far positions, producing the six stimulus conditions. Because the stimulus geometry, placement, and sequence of 51 whisker poses were deterministic, one simulated recording was retained for each condition.

Each recording was processed independently by the frozen whisker encoder. Although the encoder ultimately produces a 128-dimensional embedding for multimodal contrastive training, that final embedding is generated only after activity has propagated through the complete whisker pathway to VISrl5. For comparison with the Rodgers barrel-cortex recordings, the intermediate SSp-bfd2/3 representation was used instead because recurrent integration of the complete whisking sequence occurs at this stage. Because the Rodgers neural population included neurons spanning cortical layers L2/3–L6, this selection was not intended as a depth-matched comparison, but was treated as a representative processing stage.

At SSp-bfd2/3, the gated recurrent unit processed the 51-frame sequence independently at each of the 35 barrel-grid locations. The final recurrent hidden state at each location is projected back to the 250 channels assigned to SSp-bfd2/3, producing a temporally integrated feature map with dimensions $250 \times 5 \times 7$. This feature map is spatially averaged across the barrel grid to produce one 250-dimensional model response vector,

$$\mathbf{z}_i \in \mathbb{R}^{250},$$

for each condition i. The response is extracted before activity is propagated to SSp-bfd5, SSs, or VISrl and is used to construct the six-condition model RDM. The model representational dissimilarity matrix is calculated using Pearson correlation distance:

$$D_{ij}^{\text{model}} = 1 - r(\mathbf{z}_i, \mathbf{z}_j),$$

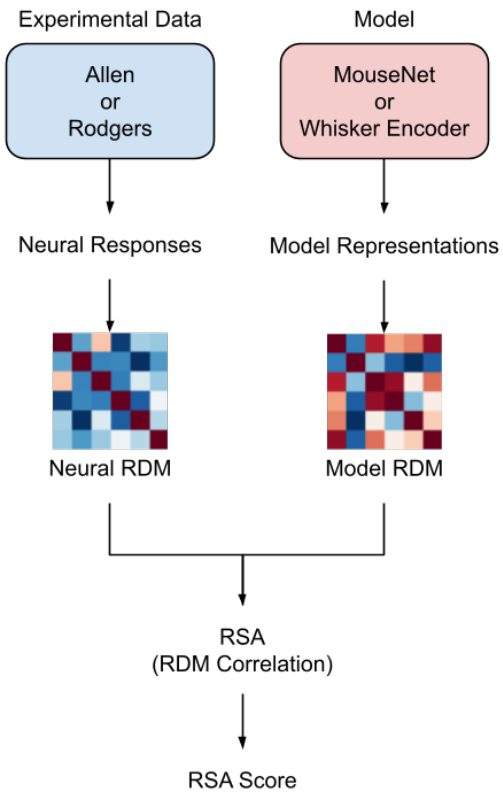


Figure 3.6: **General representational similarity analysis (RSA) workflow used to compare neural and model representations.** Experimental neural responses from the Rodgers or Allen Brain Observatory datasets and model representations extracted from the proposed network are independently transformed into representational dissimilarity matrices (RDMs). The RSA score is calculated by comparing the upper-triangular entries of the neural and model RDMs. Pearson correlation is used for the Rodgers analysis, while Spearman rank correlation is used for the Allen Brain Observatory analysis.

where r denotes the Pearson correlation coefficient.

Model-to-neural correspondence was evaluated by adapting the RSA reliability-filtering and noise-correction procedure used by Chung et al. [6] for the Rodgers dataset. The preprocessed dataset contained neural responses for the same six combinations of stimulus shape and position. Animals were analyzed separately. To estimate the reliability of each animal's six-condition neural representational geometry, neural trials were randomly divided into two halves over 1000 bootstrap iterations. Condition-mean neural response vectors and their RDMs were calculated independently for the two trial halves, and the correlation between their upper-triangular entries provided the raw split-half reliability, r_{yy}. The Spearman-Brown formula was applied to this split-half correlation to estimate the reliability expected from the full set of trials:

$$r_{yy}^{\mathrm{SB}} = \frac{2r_{yy}}{1 + r_{yy}}.$$

Following the reliability-filtering procedure of Chung et al., bootstrap iterations with raw split-half reliability $r_{yy} < 0.5$ were excluded before noise correction. This threshold removes iterations in which the neural representational geometry is insufficiently consistent across trial halves to provide a stable estimate of model-to-neural correspondence. Retained noise-corrected RSA values were averaged within each animal. The resulting animal-level estimates were summarized using their mean, median, and standard error of the mean. For visualization only, units were concatenated across Rodgers recording sessions and condition-mean responses were used to construct a pooled neural RDM.

3.7.3 Allen Brain Observatory Evaluation

The Allen Brain Observatory dataset [8] was used to evaluate visual cortical representations generated by MouseNet and to investigate the effect of cross-modal inhibitory modulation on alignment with neural activity in mouse visual cortex. Unlike the Rodgers experiments, which focused primarily on tactile representations, Allen Brain Observatory experiments focused on visual cortical activity recorded during repeated presentation of natural scene stimuli. For each experiment, neural responses were extracted for all stimulus presentations corresponding to the 118 natural scene images included within the dataset. Neural activity was then aggregated across repeated presentations of each image to obtain a single response vector per stimulus.

Several preprocessing configurations were investigated during development of the RSA pipeline. Initial experiments used calcium fluorescence traces, whereas the final analysis used event-based activity estimates, which provided a sparser representation of neural

activity. To reduce the influence of unreliable neurons, split-half reliability analysis was performed on repeated stimulus presentations. Reliability scores were computed independently for each neuron, and thresholding was applied to retain only neurons exhibiting sufficiently consistent responses across stimulus repetitions. Multiple reliability thresholds were evaluated to examine their effect on RSA performance.

Following reliability filtering, neural responses were aggregated across experiments belonging to the same cortical region. Image similarity matrices were computed using Pearson correlation, and corresponding representational dissimilarity matrices were generated for both neural and model representations. RSA scores were then calculated as Spearman correlations between the upper triangular entries of the neural and model RDMs.

Additional analyses were performed using experiment subsets corresponding to different cortical depth ranges in VISp, allowing comparisons across cortical layers. These layer-specific evaluations were used to investigate whether particular model representations exhibited stronger correspondence with specific cortical populations. Because the Allen natural-scene images did not have corresponding tactile measurements, each image was evaluated using the same no-contact whisker input with dimensions $30 \times 51 \times 2$, representing a fixed no-contact condition. In the gate-on condition, this input was processed by the whisker encoder to produce the representation supplied to the inhibitory gate. The natural-scene image was simultaneously propagated through the MouseNet visual pathway. Visual responses were extracted from the selected model layer and used to construct the model representational dissimilarity matrix.

Four model conditions were evaluated for each of the five seeds. The first was a seed-matched, randomly initialized MouseNet baseline evaluated with the gate disabled. The second was a trained gated checkpoint evaluated with its gate disabled. The third used the same trained gated checkpoint with its gate enabled and supplied with the no-contact whisker sequence. The fourth was an independently trained no-gate checkpoint for which the inhibitory-gate module and its forward hook had been absent throughout training. The gate-off and gate-on conditions therefore formed a paired inference-time comparison in which model weights were held constant. The comparison between gate-on and no-gate models instead tested whether including the gate during training provided an advantage over a seed-matched architectural ablation.

RSA was calculated separately for every model seed before aggregation. For each condition and reliability threshold, the five seed-specific RSA scores were summarized using their arithmetic mean, sample standard deviation, and sample variance, with the variance calculated using $n-1$ in the denominator. Paired differences were also calculated within each seed. The gating effect was defined as the gate-on RSA minus the gate-off

RSA from the same gated checkpoint, whereas the gate-inclusion contrast was defined as the gate-on RSA minus the RSA of the independently trained no-gate model carrying the same seed label.

Inferential comparisons were performed at a split-half reliability threshold of $r_{\text{split}} \geq 0$. Two-sided one-sample t-tests were applied to the five within-seed differences, which is equivalent to a paired t-test between the corresponding model conditions. These exploratory tests used $n = 5$ and 4 degrees of freedom, and the reported p-values were not corrected for multiple comparisons. Reliability thresholds from -0.20 to 0.90, evaluated in increments of 0.05, were additionally examined as a descriptive sensitivity analysis. Because these thresholds selected nested subsets of the same neural population, they were not treated as independent observations or additional experimental replicates.

3.8 Orientation Selectivity Evaluation

3.8.1 Oriented Grating Stimuli and Response Extraction

Orientation tuning was evaluated using procedurally generated sinusoidal grating stimuli. Each grating was defined by its orientation, spatial frequency, and phase and was generated at the same spatial resolution as the visual inputs used during model training. The resulting image was converted into the input representation required by the visual encoder using the preprocessing procedure described in Section 3.5.2.

Responses were extracted from layers 4, 2/3, and 5 of both VISp and VISrl. Forward hooks were used to record activations from the selected layer after the nonlinear activation function. This ensured that the recorded responses corresponded to the non-negative output of each analyzed model unit. Because the response of a visual unit may depend on spatial frequency and phase in addition to orientation, these stimulus parameters were first optimized separately for each analyzed unit. An automated parameter search was used to identify the spatial frequency and phase combination that maximized the unit's response.

After the spatial frequency and phase had been selected, both parameters were held constant while orientation was varied. Thirty orientations evenly spanning 0 to π radians were presented to the model. This range was used because grating orientations separated by 180° are equivalent. The activation recorded at each orientation was used to construct an orientation response curve for the corresponding model unit.

The procedure was applied independently to the randomly initialized and CLIP-trained models. The same stimulus generation, parameter selection, response extraction, and

orientation sampling procedures were used for both conditions, allowing differences in orientation tuning associated with training to be evaluated directly.

3.8.2 Orientation Selectivity Index

Orientation selectivity was quantified using a circular-variance-based orientation selectivity index. For a model channel with response $R(\theta_k)$ to orientation θ_k, the index was calculated as

$$\frac{|\sum_{k=1}^{K} R(\theta_k) e^{2i\theta_k}|}{\sum_{k=1}^{K} R(\theta_k)},$$

where K denotes the number of evaluated orientations. The factor of two accounts for the equivalence of orientations separated by 180°.

The resulting index ranges from 0 to 1. An OSI near 0 indicates similar responses across orientations, whereas an OSI near 1 indicates that the channel responds primarily to a restricted range of orientations. Channels that produced zero response across all evaluated orientations had an undefined denominator and were assigned an undefined OSI value. These channels were excluded from the calculation of layer-wise means and standard deviations but were retained when reporting the number of responsive and non-responsive channels.

Layer-level results were summarized using the mean and population standard deviation of the finite channel-wise OSI values. Individual channel values were also retained to illustrate the variability of orientation tuning within each model layer.

Chapter 4

Results

4.1 Experimental Overview

This chapter evaluates the proposed model across three complementary branches: multimodal training evaluation, whisker encoder evaluation, and visual encoder evaluation. Together, these experiments examine whether the model preserves meaningful tactile structure, corresponds with recorded mouse cortical activity, and learns associations between paired visual and whisker observations.

The first branch evaluates multimodal representation learning. Paired visual and whisker observations generated in Unity were used to train the model with a contrastive objective. Training behaviour was assessed using loss convergence, retrieval performance, and cosine similarity between matched and mismatched visual and whisker embeddings. The trained model showed substantially greater similarity and retrieval performance for matched observations than the untrained model, demonstrating that it learned a shared representation capable of associating corresponding sensory inputs.

The second branch evaluates the whisker encoder. Simulated responses to six combinations of surface curvature and distance were examined to determine whether the tactile inputs contained distinguishable condition structure. Representational similarity analysis was then used to compare frozen encoder representations with neural recordings from the Rodgers dataset using the reliability filtering and noise correction procedure described by Chung et al. Although the simulated contact patterns differed systematically across conditions, the encoder produced highly similar latent representations.

The third branch evaluates the visual pathway. MouseNet representations were compared with neural activity from the Allen Brain Observatory using representational similar-

ity analysis. These experiments also examined whether inhibition derived from the whisker pathway altered visual correspondence and whether the visual representations exhibited orientation selectivity. The visual pathway demonstrated measurable correspondence with recorded cortical activity, while the inhibitory gate produced a modest improvement across most of the evaluated reliability thresholds. Orientation selectivity was additionally examined before and after multimodal training to characterize changes in visual response properties.

Unless otherwise stated, experiments were implemented in PyTorch and executed using NVIDIA GPU hardware on the Compute Canada Graham and Nibi clusters.

4.2 Multimodal Representation Learning

The following analyses examined the behaviour of the multimodal embedding space learned through contrastive training. These experiments were intended to provide insight into the degree of alignment achieved between visual and whisker representations and to identify potential limitations of the multimodal learning framework.

4.2.1 CLIP Training

Multimodal representation learning was performed using the CLIP-style symmetric contrastive objective described in Chapter 3. During training, paired visual and whisker observations generated from the Unity simulation environment were projected into a shared 128-dimensional embedding space. The objective encouraged embeddings originating from the same simulated trial to become more similar while separating embeddings from unrelated trials.

Training convergence was first assessed using the average contrastive loss. As shown in Figure 4.1, the loss decreased rapidly during the initial epochs and continued to decline more gradually throughout the remainder of training. By the final epochs, the loss had approached a low and approximately stable value, indicating that the model successfully optimized the multimodal contrastive objective on the training dataset.

The quality of the learned embedding space was evaluated using cross-modal retrieval accuracy. A reproducible random subsample of 20,000 whisker embeddings was used as the query set, with each query compared against the complete training dataset of 100,000 visual embeddings. Recall@K measured the proportion of queries for which the corresponding

visual embedding was ranked among the top K candidates. Figure 4.2 summarizes performance before and after training. The untrained model performed at approximately chance level for all evaluated values of K, whereas the trained model achieved Recall@1, Recall@5, and Recall@10 values of 0.33, 0.67, and 0.79, respectively. These results show that contrastive training produced strong pair-specific alignment between visual and whisker representations within the training dataset.

Together, the decrease in training loss and the substantially above-chance retrieval performance demonstrate that the reformulated multimodal architecture learned pair-specific correspondence between visual and whisker observations within the training dataset. Retrieval was not perfect, but the large improvement relative to the randomly initialized model indicates that contrastive training established meaningful cross-modal alignment.

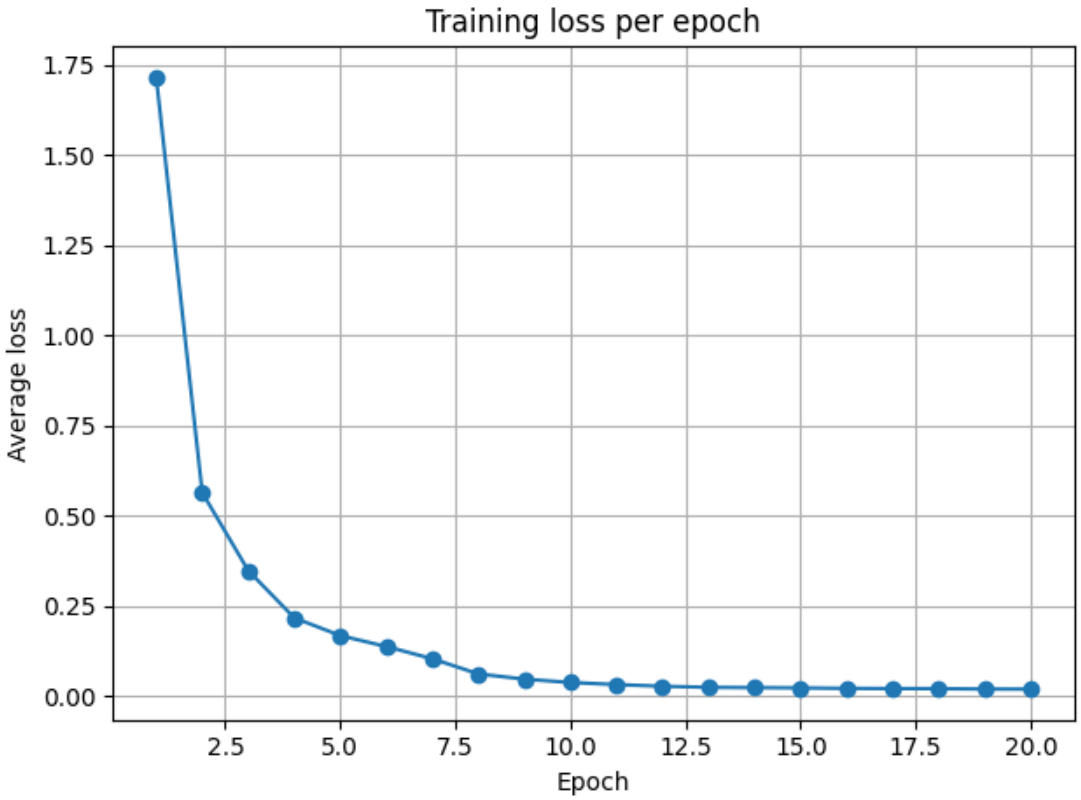


Figure 4.1: **Average CLIP contrastive training loss as a function of training epoch.** The loss decreased rapidly throughout training before approaching a stable value, indicating successful optimization of the multimodal contrastive objective.

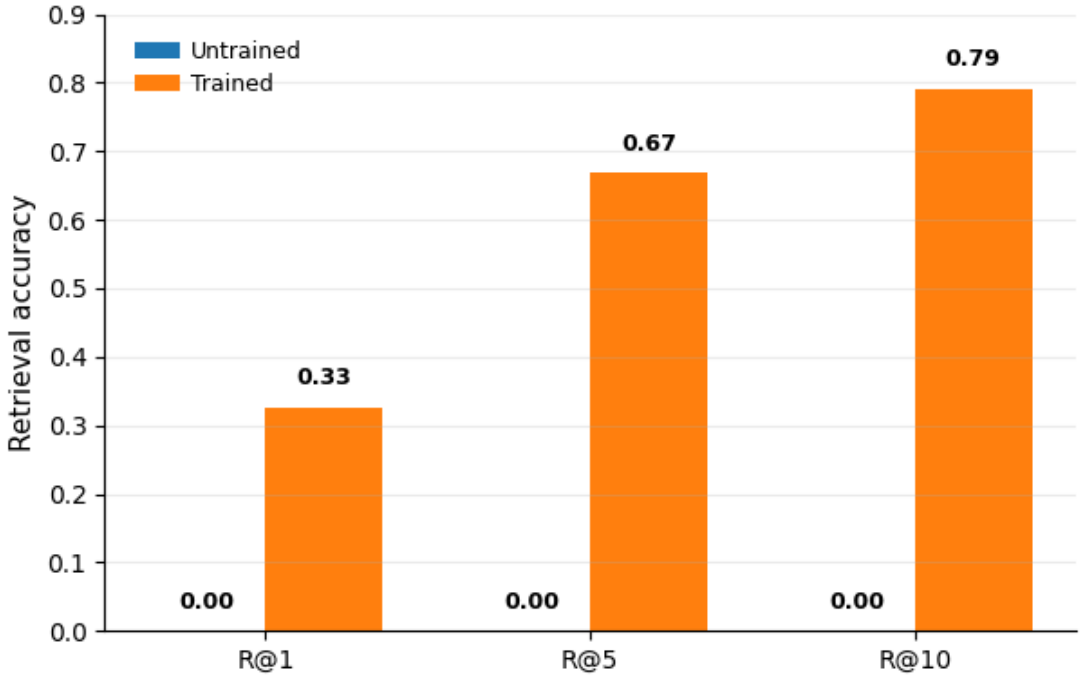


Figure 4.2: **Retrieval performance of the multimodal embedding space before and after CLIP training.** Recall@K measures the proportion of whisker embeddings for which the corresponding visual embedding was ranked among the top K candidates. Performance was estimated using a reproducible random subsample of 20,000 whisker queries, each compared against the complete dataset of 100,000 visual embeddings. The same query indices were used for the trained and untrained models. The trained model achieved Recall@1, Recall@5, and Recall@10 values of 0.33, 0.67, and 0.79, respectively, while untrained performance remained approximately at chance level.

4.2.2 Embedding Alignment

To further examine the structure of the learned multimodal representation, cosine similarity statistics were computed between visual and whisker embeddings before and after contrastive training. For each sample, a matched pair consisted of the visual and whisker embeddings originating from the same simulated trial. A corresponding mismatched pair was generated by randomly pairing each whisker embedding with a visual embedding from a different trial. Cosine similarity statistics were then computed across all matched and randomly generated mismatched pairs.

Table 4.1 summarizes the resulting cosine similarity statistics for trained and untrained encoder combinations. Prior to training, matched and randomly mismatched embedding pairs exhibited nearly identical cosine similarity values regardless of the encoder combination. Consequently, there was essentially no separation between corresponding and non-corresponding multimodal observations within the embedding space.

Following CLIP training, the mean cosine similarity between matched visual and whisker embeddings increased to 0.972 ± 0.015, whereas mismatched pairs had a mean similarity of 0.007 ± 0.383. This produced a difference of approximately 0.966 between the matched and mismatched means, indicating that corresponding multimodal observations were mapped to highly similar regions of the latent space while unrelated observations remained well separated.

All control combinations containing at least one randomly initialized encoder produced effectively identical matched and mismatched similarity statistics, with gaps below 0.001. The strong pair-specific separation therefore emerged only when the visual and whisker encoders had been trained jointly using the multimodal contrastive objective. Together, these results demonstrate that contrastive training successfully established correspondence between trial-paired visual and whisker observations rather than relying on similarities inherent to the encoder architectures or untrained representations.

4.3 Whisker Representation Learning

4.3.1 Six-Condition Neural Representational Similarity

The simulated whisker responses varied across the six evaluation conditions. Each condition consisted of contact measurements from 30 right-side whiskers over 51 prescribed whisking poses. The near surfaces produced contact across a larger portion of the whisking cycle,

Table 4.1: **Cosine similarity statistics between visual and whisker embeddings before and after CLIP training.** Matched pairs correspond to embeddings originating from the same simulated trial, while mismatched pairs correspond to embeddings originating from different trials. Values report the mean and standard deviation across pairs. The gap denotes the difference between the matched and mismatched means; gaps with an absolute magnitude below 0.001 are reported as <0.001.

Encoder Pair	**Matched** ($\mu \pm \sigma$)	**Mismatched** ($\mu \pm \sigma$)	**Gap**
V_{trained} vs. W_{trained}	0.972 ± 0.015	0.007 ± 0.383	0.966
V_{trained} vs. W_{init}	0.017 ± 0.105	0.017 ± 0.105	<0.001
V_{init} vs. W_{trained}	0.001 ± 0.062	0.001 ± 0.062	<0.001
V_{init} vs. W_{init}	$-0.027 \pm <0.001$	$-0.027 \pm <0.001$	<0.001

whereas contact became progressively sparser as the stimulus was moved to the medium and far positions.

The six simulated recordings were subsequently processed by the frozen whisker encoder, and the temporally integrated representation from SSp-bfd2/3 was extracted for analysis. Following temporal integration, the SSp-bfd2/3 feature map was spatially pooled to produce one 250-dimensional response vector for each condition. Pairwise correlation distances between these six condition responses ranged from 0.003 to 0.484, with a mean of 0.144 (Figure 4.3A). The intermediate barrel-cortex representation therefore retained some condition-dependent structure, although several condition pairs were represented very similarly. The organization of these dissimilarities also differed from the descriptive Rodgers neural RDM (Figure 4.3B).

Across the 11 animals, the mean uncorrected Pearson RSA was -0.202, with a median of -0.202 and individual scores ranging from -0.470 to 0.184. Following split-half reliability filtering and noise correction, the mean RSA was -0.230 ± 0.066 SEM and the median was -0.234. Individual noise-corrected scores ranged from -0.550 to 0.214 (Figure 4.3D), with nine of the eleven animals producing negative values. Although the SSp-bfd2/3 representation distinguished several of the six tactile conditions, its pattern of pairwise condition relationships differed from that observed in the Rodgers neural data. Nine of the eleven animals produced negative RSA values, indicating an overall tendency for condition relationships in the model to vary in the opposite direction from those in the neural representation rather than exhibiting positive correspondence.

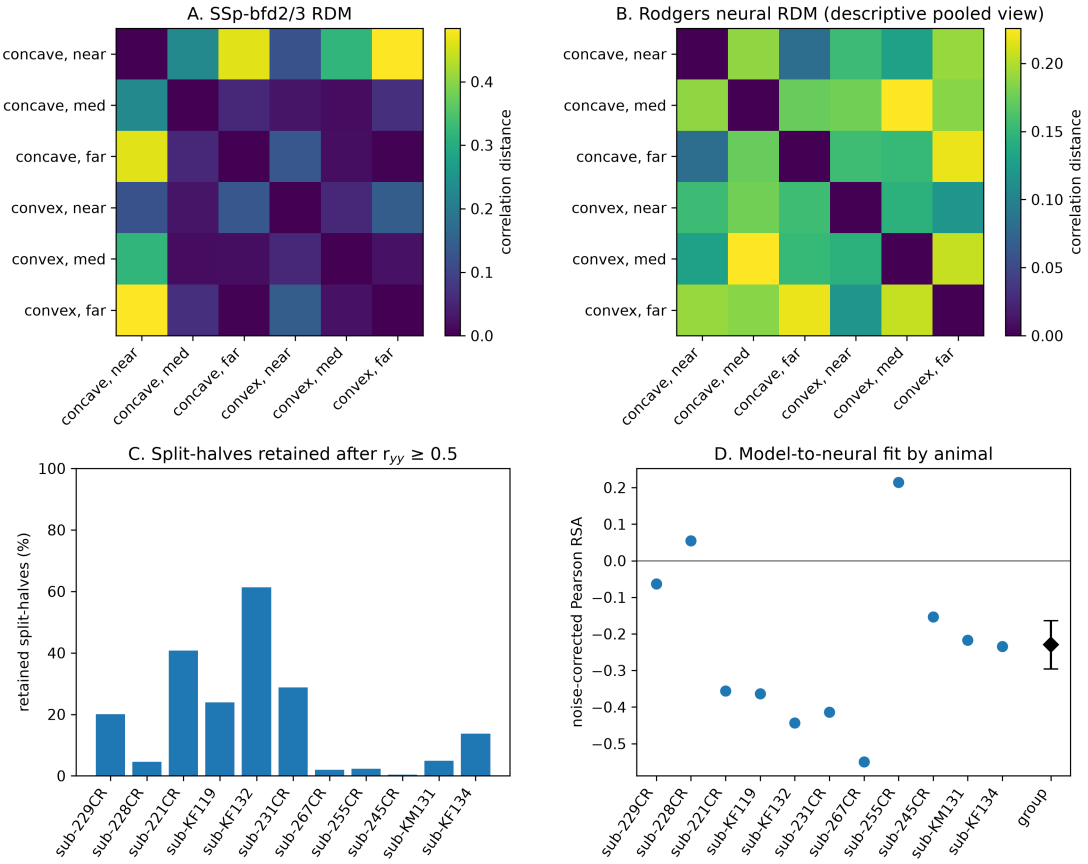


Figure 4.3: **Representational comparison of the whisker encoder and Rodgers neural responses.** **A**, Correlation-distance RDM calculated from the spatially pooled, temporally integrated 250-dimensional SSp-bfd2/3 representation of the frozen whisker encoder. **B**, Descriptive neural RDM obtained after pooling units across Rodgers recording sessions. Separate colour scales are used for the model and neural RDMs. **C**, Percentage of the 1000 bootstrap split-halves retained for each animal after applying the neural reliability criterion $r_{yy} \geq 0.5$. **D**, Noise-corrected Pearson RSA for each animal. The group marker denotes the mean across animals, and the error bar denotes the standard error of the mean.

4.4 Visual Representation Learning

Representational similarity analysis (RSA) was used to evaluate the correspondence between model representations and the representational geometry derived from neural activity recorded by the Allen Brain Observatory. Consistent with the methodology in MouseNet, neural and model representations were converted into representational dissimilarity matrices (RDMs), and similarity between these matrices was quantified using Spearman correlation.

4.4.1 Inhibitory Gating

To investigate whether tactile information could influence visual representations, the whisker-conditioned inhibitory gating mechanism described in Chapter 3 was incorporated into the MouseNet visual pathway. The gate was applied to the projection from VISp layer 4 to VISp layer 2/3 and produced channel-specific suppression based on the whisker embedding.

The effect of the inhibitory gate was evaluated using representational similarity analysis (RSA) on natural-scene responses from the Allen Brain Observatory. Because these experiments did not include paired tactile measurements, the same all-zero right-whisker input sequence was supplied to the whisker encoder for every natural scene. The analysis was repeated across five independently initialized training seeds (42-46). Four conditions were compared: an untrained MouseNet baseline, the multimodally trained gated models evaluated with the gate disabled, the same models evaluated with the gate enabled, and independently trained models from which the inhibitory gate was removed before training.

The layer-specific RSA results at a neuronal split-half reliability threshold of $r_{\text{split}} \geq 0$ are summarized in Table 4.2. Values are means across the five seeds, with sample variances shown in parentheses. In VISp4, mean RSA was 0.0938 for the untrained baseline, 0.1336 for both gate-off and gate-on evaluation, and 0.1339 for the independently trained no-gate model. Gate-off and gate-on values were identical for every seed because VISp4 precedes the gated projection.

Within the gated models, enabling the gate increased mean RSA in VISp2/3 from 0.1093 to 0.1178 ($\Delta_{\text{on}-\text{off}} = +0.0085$) and in VISp5 from 0.0736 to 0.1046 ($\Delta_{\text{on}-\text{off}} = +0.0311$). The VISp5 gate-on minus gate-off difference was positive for all five seeds across the evaluated reliability thresholds. An exploratory paired test at $r_{\text{split}} \geq 0$ found a nominally significant VISp5 difference, $t(4) = 3.02$, $p = 0.039$, although this value was not corrected for comparisons across layers or reliability thresholds.

The architectural ablation produced a different conclusion. In VISp2/3, the independently trained no-gate models achieved a higher mean RSA than the gated models evaluated with the gate enabled (0.1230 versus 0.1178; $\Delta_{\text{on}-\text{no gate}} = -0.0053$). In VISp5, gate-on RSA was higher on average than no-gate RSA (0.1046 versus 0.0958; $\Delta_{\text{on}-\text{no gate}} = +0.0088$), but this direction occurred in only three of the five seeds and was not significant ($t(4) = 0.62$, $p = 0.570$). Thus, applying the learned gate produced a reproducible acute change in downstream representational geometry, particularly in VISp5, but including the gate during multimodal training did not reliably improve alignment with the Allen neural representations.

Table 4.2: **Representational similarity between MouseNet activations and Allen Brain Observatory natural-scene responses across five seeds.** Values are mean Spearman correlations across seeds 42–46, with sample variances in parentheses. Neurons satisfied $r_{\text{split}} \geq 0$. The gate-on condition used an all-zero right-whisker input.

A. RSA by model condition

Population	Experiments	Neurons	Base	Gate off	Gate on	No gate
VISp4	42	6533	0.0938 (0.000808)	0.1336 (0.001097)	0.1336 (0.001097)	0.1339 (0.001279)
VISp2/3	20	3989	0.0947 (0.000646)	0.1093 (0.000022)	0.1178 (0.000183)	0.1230 (0.000152)
VISp5	26	2892	0.0742 (0.000123)	0.0736 (0.000117)	0.1046 (0.000808)	0.0958 (0.000042)

B. Paired RSA contrasts

Population	$\Delta_{\text{on}-\text{off}}$	Variance	$\Delta_{\text{on}-\text{no gate}}$	Variance
VISp4	0.0000	0.000000	−0.0003	0.000327
VISp2/3	+0.0085	0.000101	−0.0053	0.000514
VISp5	+0.0311	0.000528	+0.0088	0.001021

Note. Base denotes an independently initialized untrained MouseNet. Gate off and Gate on use the same gated checkpoints. No gate denotes independently trained models from which the inhibitory gate was removed before training. $\Delta_{\text{on}-\text{off}}$ = Gate on − Gate off and $\Delta_{\text{on}-\text{no gate}}$ = Gate on − No gate.

The reliability of neurons was evaluated using the split half reliability procedure in MouseNet. As shown in Figure 4.4, the gate-on RSA remained consistently higher than the gate-off RSA. Across the full range of reliability thresholds, the mean gate-on RSA remained higher than the mean gate-off RSA. The mean difference gradually decreased

from 0.0085 at $r_{\text{split}} \geq 0$ to 0.0051 at 0.8, 0.0049 at 0.85, and 0.0018 at 0.9. At these higher thresholds, progressively fewer neurons were retained: 563 neurons at 0.8, 355 at 0.85, and 162 at 0.9. Thus, the direction of the gating effect was preserved in the five-seed mean across the complete threshold range, although its magnitude decreased and its uncertainty increased as the retained neural population became smaller.

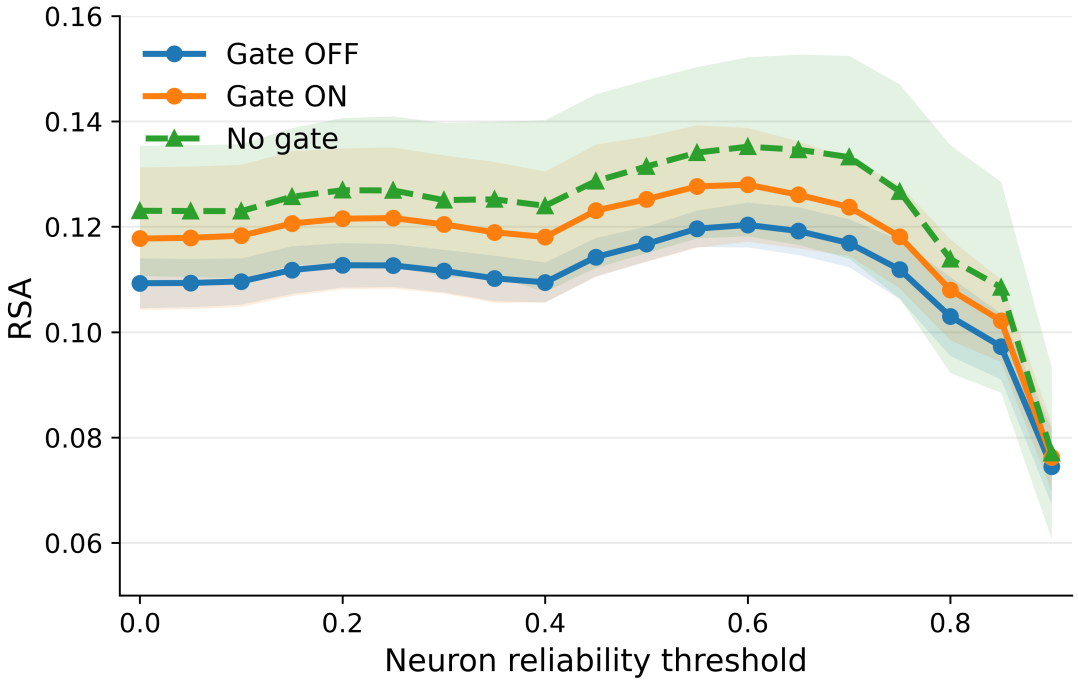


Figure 4.4: **Representational similarity between model VISp2/3 activations and the Allen VISp L2/3 neural representation as a function of minimum neuronal split-half reliability.** Lines show mean RSA across five training seeds (42–46), and shaded regions show ±1 sample standard deviation. Gate ON denotes evaluation using an all-zero right-whisker input. No gate denotes independently trained models from which the inhibitory gate was removed before training. Mean gate-on RSA remained above mean gate-off RSA across all evaluated thresholds, although the difference decreased at the highest thresholds as fewer neurons were retained.

VISp2/3 activations were also compared with the gate disabled and enabled. As shown in Figure 4.5, enabling the gate reduced the magnitude of VISp2/3 responses while preserving their stimulus-dependent activation structure. This behaviour was consistent with the intended suppressive action of the inhibitory mechanism.

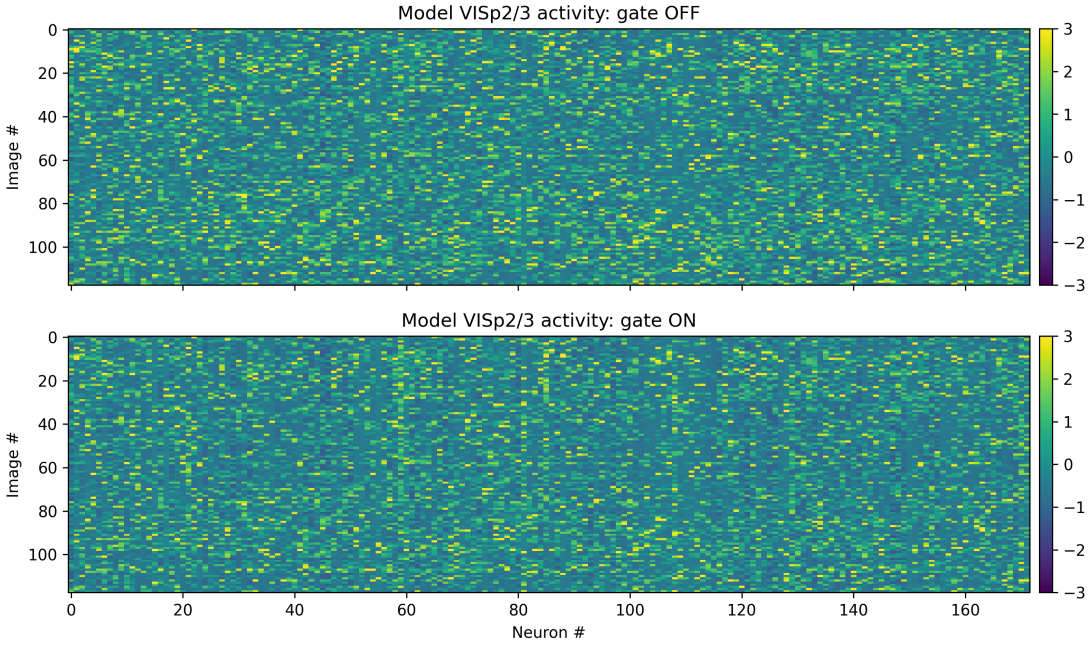


Figure 4.5: **Activity of model VISp2/3 units in response to the 118 Allen Brain Observatory natural-scene images with the inhibitory gate disabled (top) and enabled using an all-zero right-whisker input (bottom)**. Rows correspond to images and columns show the same representative subset of VISp2/3 neurons in both conditions. Activations were standardized using the gate-off mean and standard deviation and are displayed using identical colour limits. Enabling the gate reduced activation magnitude while retaining the overall stimulus-dependent response structure.

4.4.2 Orientation Selectivity

To evaluate whether the learned visual representations exhibited functional properties associated with biological visual cortex, orientation selectivity was measured across layers of VISp and VISrl before and after multimodal contrastive training. Model responses were recorded during presentation of oriented grating stimuli, and an orientation selectivity index (OSI) was calculated for each channel using the procedure described in Chapter 3. This analysis provided a complementary evaluation to RSA by examining the tuning of individual model channels rather than population-level representational geometry.

Figure 4.6 compares the layer-wise OSI distributions before and after CLIP training. Bars indicate the mean OSI across channels with defined responses, error bars indicate the

population standard deviation, and individual points show the OSI of each valid channel. Channels that produced no response across the evaluated orientations yielded undefined OSI values and were excluded from the plotted summary statistics.

Before training, mean orientation selectivity was generally low, particularly in VISp layer 5 and throughout VISrl. Mean OSI was approximately 0.29 in VISp layer 4, 0.38 in VISp layer 2/3, and 0.05 in VISp layer 5. Orientation selectivity was similarly limited in VISrl, with mean values of approximately 0.10 or lower across the evaluated layers.

Following CLIP training, mean orientation selectivity increased substantially in five of the six evaluated layers. Mean OSI values reached approximately 0.53, 0.38, and 0.52 in VISp layers 4, 2/3, and 5, respectively. In VISrl, mean OSI increased to approximately 0.63 in both layers 4 and 2/3 and to 0.49 in layer 5. Orientation selectivity was therefore higher in VISrl than in the corresponding VISp layers 4 and 2/3, whereas VISp layer 5 exhibited slightly greater selectivity than VISrl layer 5. All trained channels produced finite OSI values under the evaluated grating conditions, in contrast to the large proportion of undefined responses observed before training.

The trained VISp values were comparable to physiological measurements obtained using the same circular-variance-based OSI formulation. Durand et al. [11] reported a mean OSI of 0.42 ± 0.02 across awake mouse VISp and 0.48 ± 0.05 specifically in layer 4, while Camillo et al. [4] reported population means between 0.37 and 0.41 in VISp. The model value of approximately 0.53 in VISp layer 4 was therefore close to the experimentally reported layer 4 value. The lower value in model VISp layer 2/3 and higher value in layer 5 indicate that orientation selectivity varied considerably across the model hierarchy.

Physiological measurements from higher visual cortex have also indicated greater orientation selectivity in VISrl than in VISp. Marshel et al. [32] reported approximate mean values of 0.50 in VISp and 0.69 in VISrl. However, the study calculated OSI from responses to the preferred and orthogonal orientations rather than using circular variance, so its numerical values cannot be compared directly with those of the present model. Qualitatively, the elevated selectivity observed in model VISrl layers 4 and 2/3 is consistent with the reported tendency toward stronger orientation tuning in VISrl, although this relationship was not observed in layer 5.

Overall, these results show that substantial orientation tuning emerged across much of the visual hierarchy following multimodal contrastive training, despite orientation selectivity not being included explicitly in the training objective. The close correspondence between model and physiological OSI in VISp layer 4 provides evidence that the learned visual representations captured a functional property of mouse visual cortex.

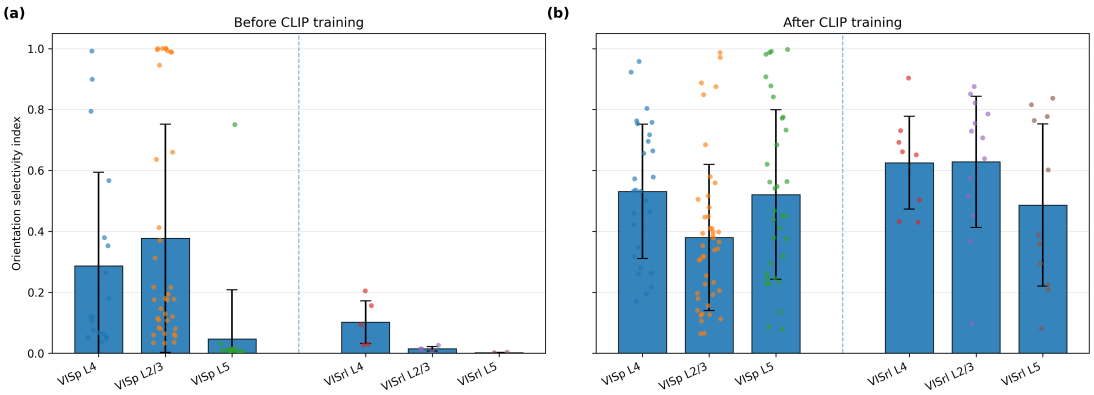


Figure 4.6: **Orientation selectivity across MouseNet visual cortical layers before and after CLIP training.** Bars show the mean orientation selectivity index (OSI) across channels with finite responses, error bars indicate the population standard deviation, and points represent individual channel values. (a) Before training, orientation selectivity was generally low and variable, particularly in VISp layer 5 and across VISrl, where many channels produced no response across the evaluated orientations. (b) Following CLIP training, mean OSI increased substantially in five of the six evaluated layers, while VISp layer 2/3 remained approximately unchanged. All trained channels produced finite OSI values. The dashed line separates VISp and VISrl layers.

Chapter 5

Discussion

5.1 Cross-Modal Alignment of Visual and Whisker Representations

The multimodal contrastive objective successfully aligned representations produced from paired visual and whisker observations. Following training, the mean cosine similarity between matched visual and whisker embeddings was approximately 0.972, whereas randomly mismatched embeddings had a mean similarity close to zero. This produced a matched-mismatched gap of approximately 0.966. Retrieval analysis similarly showed that the model could identify the corresponding observation from the other modality at rates substantially above chance. These findings demonstrate that the visual and whisker encoders learned a shared embedding space in which paired sensory observations were strongly associated.

Comparisons with randomly initialized encoders indicate that this correspondence was produced by joint multimodal training rather than by the architecture or embedding dimensionality alone. Pairings containing at least one untrained encoder did not show a meaningful difference between matched and mismatched trials. The large separation observed when both encoders were trained therefore reflects information learned from the correspondence between the two simulated sensory streams.

However, successful retrieval does not necessarily imply that the embedding space recovered a complete or biologically realistic representation of the external objects. Contrastive learning rewards features that distinguish one paired trial from other trials in the batch. The encoders may therefore rely on any consistent information shared between the image and whisker observations, including object geometry, contact configuration, object

position, or other correlated aspects of the simulation. In the absence of explicit object metadata, the contribution of each factor could not be separated directly.

Overall, the multimodal results demonstrate that anatomically motivated visual and whisker pathways can be trained jointly to establish a strong correspondence between paired sensory signals. The model was able to learn strong cross-modal associations within the training dataset while preserving interpretable properties within the visual pathway. Together with the VISp gating results, this provides evidence that both representational alignment and intermediate cross-modal modulation can be incorporated into a unified model of mouse sensory processing.

5.2 Visual Representations Following Multimodal Training

The relationship between the trained visual pathway and the seed-matched randomly initialized reference was layer dependent. Across five training seeds, the trained gate-off models produced higher mean RSA than the untrained baseline in VISp4 and VISp2/3, but not in VISp5. In contrast, the independently trained no-gate models produced higher mean RSA than the baseline in all three layers. These comparisons suggest that multimodal contrastive training can preserve or improve correspondence with visual cortical activity, but that this effect cannot be attributed specifically to the inhibitory gate. Nayebi et al. [36] similarly found that shallow networks trained with contrastive objectives generally achieved greater model-to-neural correspondence than untrained, supervised, and alternative self-supervised models, including under an RSA-based evaluation. Absolute RSA values cannot be compared directly because Nayebi et al. used animal-level aggregation and selection of the maximally predictive model layer, whereas the present analysis used reliability-thresholded pooled populations and anatomically matched layers.

The orientation selectivity analysis provided evidence that the visual pathway developed structured responses following training. Relative to the randomly initialized model, mean OSI increased in VISp layers 4 and 5 and across all evaluated VISrl layers, with particularly large increases in VISp layer 5 and VISrl. VISp layer 2/3 was an exception, remaining approximately unchanged in mean OSI after training. The trained VISp values were comparable to physiological measurements obtained using the same circular-variance-based OSI formulation. Durand et al. [11] reported a mean OSI of 0.42 ± 0.02 across awake mouse VISp and 0.48 ± 0.05 specifically in layer 4, while Camillo et al. [4] reported population means between 0.37 and 0.41 in VISp. VISp layer 5 exhibited a some-

what higher mean value of approximately 0.52. The emergence of orientation-selective responses is notable because the network was not trained using orientation labels or an explicit orientation-discrimination objective. Instead, these responses developed while the model was optimized to align paired visual and whisker observations through multimodal contrastive learning.

Orientation selectivity was higher in VISrl than in the corresponding VISp layers 4 and 2/3, suggesting that parts of the higher visual pathway developed sharper orientation tuning. This relationship was not uniform across the hierarchy, however, because VISp5 exhibited greater mean selectivity than VISrl5. The results therefore support a layer-dependent organization of orientation tuning rather than a consistent increase at every successive stage. The VISrl values should also not be interpreted as evidence that the model precisely reproduces biological VISrl responses. Available physiological studies of higher visual areas may use different definitions of orientation selectivity, making direct comparison with the formulation used in this thesis difficult. In addition, the optimization procedure identified stimuli that maximized individual model-unit responses rather than reproducing an in vivo grating experiment exactly. The result therefore primarily demonstrates that multimodal training produced units with systematic and sharply tuned orientation preferences.

Taken together, the neural similarity and orientation selectivity results suggest that the visual encoder did not merely learn an arbitrary code for solving the contrastive objective. Its representations preserved a degree of correspondence with mouse visual cortical activity while also developing interpretable visual tuning. This supports the use of an anatomically constrained architecture when examining how multimodal learning affects sensory representations. At the same time, these findings do not establish that the representations are fully biologically equivalent to those of the mouse visual system. They instead show that biologically relevant response properties can coexist with a representation optimized for cross-modal alignment.

5.3 Functional Effects of the Learned Inhibitory Gate in VISp

The Allen Brain Observatory comparison was used to determine how the learned inhibitory gate affected the correspondence between model and biological visual representations. The analysis was repeated across five independently trained seeds. Two related comparisons were considered. First, enabling and disabling the gate within the same gated checkpoint

isolated the immediate effect of applying the learned modulation. Second, comparison with independently trained no-gate models tested whether including the gate during training provided an advantage over an architecture in which the gate was completely absent.

The gate was applied to the projection from VISp4 to VISp2/3, and its most direct effect was therefore expected in VISp2/3. At the primary reliability threshold of $r_{\text{split}} \geq 0$, enabling the gate increased VISp2/3 RSA by an average of 0.0085 relative to disabling it. The five-seed mean difference remained positive across the evaluated neuronal reliability thresholds, although its magnitude varied among seeds and the paired difference was not statistically significant ($p = 0.132$). The VISp2/3 result therefore provides evidence of a small and directionally consistent mean change, but not strong evidence of an improvement that generalizes across independently trained models.

A larger downstream effect was observed in VISp5. Enabling the gate increased VISp5 RSA by an average of 0.0311, and the gate-on value exceeded the gate-off value in all five seeds across the principal range of reliability thresholds. At $r_{\text{split}} \geq 0$, this paired difference was nominally significant, $t(4) = 3.02$, $p = 0.039$, with a 95% confidence interval of approximately $[0.0025, 0.0596]$. This result suggests that the representational change introduced at the VISp4-to-VISp2/3 projection propagated to the downstream VISp5 representation. Because the analysis involved only five seeds and multiple layer comparisons without correction, the statistical result should nevertheless be considered exploratory. VISp4 RSA was identical with the gate enabled and disabled. This was expected because VISp4 precedes the gated projection and therefore served as an internal check that the manipulation affected only the gate and downstream representations. The contrast between the null VISp4 result and the changes in VISp2/3 and VISp5 supports the interpretation that the observed differences originated from the location of the inhibitory mechanism.

Importantly, the independently trained no-gate models did not show that including the gate produced a general advantage in neural alignment. In VISp2/3, mean RSA was 0.1230 for the no-gate models and 0.1178 for the gate-on condition. In VISp5, gate-on RSA was higher than no-gate RSA, 0.1046 compared with 0.0958, but the difference varied substantially among seeds and was not significant ($p = 0.570$). No reliable gate-on advantage over no-gate training was observed in any evaluated layer. The results therefore distinguish an acute functional effect of enabling the learned gate from an architectural benefit of including the gate during training. The gate reproducibly altered downstream representational geometry, particularly in VISp5, but it did not reliably improve correspondence relative to models trained without the gate.

The activation analysis provides a complementary, but illustrative, interpretation of this effect. In the representative checkpoint shown in Figure 4.5, enabling the gate reduced the

magnitude of the VISp2/3 feature vectors while broadly preserving the stimulus-dependent activation pattern. The larger change in activation magnitude relative to the RSA change is consistent with the gate acting partly as a form of gain control, because RSA is comparatively insensitive to uniform scaling of activity. However, this activation visualization represents one checkpoint rather than an aggregate analysis across the five seeds.

Overall, the five-seed results show that applying the learned inhibitory gate changes the geometry of downstream visual representations, with the strongest acute effect occurring in VISp5. They do not show that gated training generally produces representations that are more biologically aligned than those learned by a no-gate architecture. Since both gate conditions used an all-zero whisker sequence, the experiment characterizes the gate's learned no-contact modulation and does not establish how meaningful tactile activity would influence visual representations.

5.4 Whisker Encoder Correspondence with Neural Data

The whisker-only evaluation did not provide evidence of consistent positive correspondence between the frozen encoder and the Rodgers neural representational geometry. This weak correspondence did not result from the conditions being mapped to exactly identical representations. Pairwise correlation distances among the SSp-bfd2/3 responses ranged from 0.003 to 0.484, with a mean of 0.144, indicating clear condition-dependent structure in the intermediate barrel-cortex representation. However, several condition pairs were represented very similarly, indicating that much of the variation present in the simulated whisker inputs had been compressed by this stage of the encoder. Several condition pairs remained highly similar, but others were strongly separated. The negative model-to-neural RSA therefore primarily reflects a difference in how the model and neural populations organized the six conditions, rather than an absence of condition information or complete representational collapse.

The present result was substantially weaker than the correspondence reported by Chung et al. [6] using the same six-condition Rodgers dataset. Chung et al. reported a mean animal-to-animal consistency of $r = 0.175 \pm 0.161$ SEM for their raw simulated force-and-torque inputs. By comparison, the present SSp-bfd2/3 representation produced a mean noise-corrected RSA of -0.230 ± 0.066 SEM. A possible explanation is a mismatch between the representations encouraged by the different training procedures. The present whisker encoder was optimized through cross-modal contrastive learning to associate tactile

features with corresponding visual observations. In contrast, Chung et al. evaluated models optimized for tactile object categorization and tactile-specific self-supervised objectives on biomechanically richer force-and-torque inputs [6]. These objectives may encourage tactile representations that are more closely related to the structure captured by the Rodgers neural recordings.

5.5 Limitations

Several limitations should be considered when interpreting the results of this thesis. These limitations concern the sensory simulation, the biological detail of the model, the available neural datasets, and the scope of the evaluation procedures.

First, the Unity environment provides a controlled but simplified representation of active whisker sensing. Whisker movements followed predetermined kinematic trajectories while the head and stimulus objects remained stationary. The whiskers did not deform in response to contact, and the recorded tactile signal represented geometric overlap rather than contact force, bending moment, vibration, or mechanoreceptor activity. Real mice continuously adjust whisker and head movements in response to sensory feedback. The absence of this closed loop behaviour limits the extent to which the simulated inputs reproduce natural tactile exploration. Environmental and object variability were also restricted relative to the range of conditions encountered by behaving animals.

The whisker encoder represented the 30 right-side whiskers within a unilateral barrel-field topology and used anatomically motivated pathways through somatosensory and higher cortical regions. These pathways were implemented as computational abstractions rather than detailed simulations of the underlying circuitry. Consequently, the findings concern the representational effects of the modelled topology and information flow. They do not address contributions from mechanisms not explicitly included, such as thalamic processing, laminar cortical circuitry, motor signals, or biophysical synaptic dynamics.

The whisker evaluation was limited by both the size and reliability of the neural comparison. The Rodgers benchmark contained only six combinations of stimulus curvature and distance, producing 15 unique pairwise dissimilarities. Correlations calculated from such a small representational space can be sensitive to individual condition pairs. Each model condition was also represented by one deterministic simulation, whereas the neural data contained trial-to-trial variation arising from behaviour, movement, and biological noise.

The preprocessed Rodgers responses additionally combined neurons recorded across

multiple barrel-cortex depths. Consequently, the SSp-bfd2/3 model representation was compared with a neural population that was not restricted to biological layer 2/3. Neural reliability also varied substantially among animals: the median proportion of retained bootstrap split-halves was 13.7%, and five animals retained fewer than 5% of their splits. Although all animals produced finite noise-corrected estimates, estimates based on very few retained splits may be unstable.

The visual evaluation presents a different set of limitations. The Allen Brain Observatory recordings were collected during passive presentation of visual stimuli without corresponding tactile measurements. Both gating conditions used the same all-zero whisker input sequence and differed only in whether the resulting inhibitory scaling was applied to the visual features. Enabling the gate changed RSA relative to disabling it within the same trained checkpoints, particularly in VISp5. However, the independently trained no-gate models were not reliably worse than the gated models. The results therefore demonstrate that the gate's learned no-contact modulation alters visual representational geometry, but they do not establish that including the gate generally improves biological correspondence. They also do not establish that meaningful whisker signals would produce the same effect or that the learned modulation reproduces the mechanism or magnitude of tactile influences in behaving mice.

The multimodal training results demonstrate strong alignment between paired visual and whisker embeddings, but this alignment should not be interpreted as evidence of biological correspondence on its own. Contrastive training directly rewards the model for distinguishing matched from mismatched observations. Substantially above-chance retrieval performance and strong similarity between matched embeddings therefore show that the training objective was successfully optimized, but they do not establish that the shared representation captures object identity, perceptual similarity, or the organization of neural activity. Additional evaluation on new environments, object categories, and behavioural tasks would be required to determine how broadly the learned representation generalizes.

The visual and whisker pathways were evaluated against separate neural datasets collected from different animals, sensory conditions, and behavioural paradigms. At the time of writing, no publicly available dataset contains the corresponding visual and somatosensory population recordings required to evaluate both pathways within a common RSA framework. Consequently, the present experiments cannot directly test whether the learned interaction between the two pathways resembles multisensory processing in the same animal. The study that motivated the inhibitory gate [52] demonstrated that contralateral whisker stimulation suppressed visually evoked responses in an anterior region of VISp. These findings provide a qualitative biological motivation for the modelled inter-

action rather than direct validation of its implementation.

Finally, the inhibitory-gating RSA analysis was repeated across five independent training seeds, allowing its mean effect and seed-to-seed variability to be evaluated. This analysis showed that the immediate gate-on versus gate-off effect was most consistent in VISp5, whereas comparisons with independently trained no-gate models were more variable. Nevertheless, five seeds provide limited statistical power, and the reported tests should be interpreted as exploratory. Other results, including retrieval performance, orientation selectivity, activation visualizations, and whisker RSA, primarily describe one trained model configuration unless otherwise stated. Repeating those analyses across additional training runs would be necessary to determine their stability.

5.6 Future Work

Several extensions could build on the framework developed in this thesis. A primary direction would be to improve the realism and diversity of the simulated whisker environment. Future simulations could incorporate whisker deformation, contact forces, bending moments, and vibration rather than representing contact only through geometric intersection. Allowing the mouse head and whiskers to respond dynamically to sensory feedback would further support the study of active sensing. A closed-loop environment in which whisking and head movement are adapted according to previous contacts would provide a more realistic test of whether the model can learn representations that support tactile exploration and object recognition. Greater variation in object geometry, position, material, and motion would also help determine whether the learned representations generalize beyond the conditions used during training.

The whisker pathway could also be extended to incorporate additional biological structure, such as thalamic processing, recurrent cortical interactions, feedback from higher cortical regions, and motor-related signals. These additions would make it possible to examine how sensory and motor information jointly shape whisker representations, although they would need to be introduced selectively to avoid increasing model complexity without a corresponding improvement in explanatory value.

A particularly important next step would be to evaluate the model using richer neural and behavioural datasets. The whisker RSA analysis was restricted to six experimental conditions, limiting the resolution with which model and neural representational structures could be compared. Neural recordings obtained across a larger set of object shapes, distances, contact patterns, and whisking behaviours would provide a more informative

benchmark. Ideally, future experiments would collect synchronized visual stimuli, whisker kinematics, contact measurements, behaviour, and neural activity from visual and somatosensory cortical regions in the same animals. Such data would permit a direct test of whether the learned inhibitory interaction resembles neural modulation during genuine multisensory behaviour.

Finally, the learned multimodal representation should also be evaluated using alternative learning objectives. The contrastive objective used in this thesis directly encouraged embeddings from paired visual and whisker observations to become similar while separating them from mismatched observations drawn from the training batch. Although effective computationally, this dependence on explicit negative pairs limits biological plausibility. Future work could investigate other approaches such as Bootstrap Your Own Latent (BYOL) [16]. Although BYOL is not a mechanistic model of biological learning, it may provide a less strongly contrastive account of how representations emerge from paired sensory experience. Comparing these objectives would also help determine whether the alignment, orientation selectivity, and neural similarity observed here depend specifically on contrastive training.

Together, these extensions would allow the present framework to progress from a proof of concept toward a more complete computational account of visual and tactile interactions in the mouse cortex.

Chapter 6

Conclusion

6.1 Thesis Summary

This thesis investigated how visual and whisker-derived tactile information can be incorporated into a unified computational model of the mouse cortex. Existing biologically inspired models such as MouseNet represent visual processing using anatomically motivated cortical pathways, but they do not explicitly include the whisker system or interactions between visual and somatosensory pathways. To address this limitation, MouseNet was extended with a custom whisker encoder, a mechanism for tactile modulation of visual activity, and a multimodal training framework based on paired sensory observations generated in simulation.

The whisker encoder was designed to preserve the spatial and temporal organization of tactile signals. Contact position and whisker angle were recorded across a complete whisking sequence and arranged according to the anatomical organization of the barrel field. Temporal integration was introduced within the pathway corresponding to superficial barrel cortex before representations were propagated through higher somatosensory regions and VISrl. This architecture provided a biologically motivated method for transforming simulated whisker dynamics into fixed-dimensional tactile representations.

A Unity simulation environment was developed to generate the sensory observations required for training and evaluation. The multimodal dataset contained image and whisker contact sequences produced during interactions with randomized objects. A separate evaluation dataset represented concave and convex surfaces at three distances, reproducing the six stimulus conditions used in the Rodgers neural dataset. These simulations made it

possible to evaluate the model under controlled sensory conditions while preserving whisker identity, contact timing, and object geometry.

The frozen whisker encoder retained substantial differences among several of the six conditions, demonstrating that condition-dependent tactile structure remained present at SSp-bfd2/3. However, this representational geometry did not align consistently with what was observed in the Rodgers neural recordings. The mean noise-corrected Pearson RSA across animals was negative on average, and individual animal scores varied in both sign and magnitude. The current whisker representations therefore did not exhibit consistent correspondence with the neural representational geometry observed in the Rodgers recordings. This result identifies a limitation of the learned tactile representation rather than an absence of information in the simulated inputs.

The visual pathway exhibited measurable correspondence with neural activity from the Allen Brain Observatory. Enabling the learned gate increased mean RSA relative to disabling it within the same gated checkpoints, particularly in VISp5. However, independently trained no-gate models performed comparably, indicating that the gate modulated downstream representations without providing a reliable overall advantage in neural alignment. Visual representations evaluated following multimodal training also exhibited increased orientation selectivity throughout all evaluated layers of VISp and VISrl, except VISp2/3 where the OSI value remained approximately unchanged. Although some higher visual representations were more selective than typically observed physiologically, the analysis showed that the model captured functional properties associated with visual cortical processing.

Multimodal training successfully aligned paired visual and whisker observations within a shared embedding space. The contrastive loss decreased during training, retrieval performance improved substantially above chance, and matched sensory observations became clearly distinguishable from mismatched observations. These results demonstrate that the model learned strong associations between trial-paired visual and tactile inputs within the training dataset, even though the whisker encoder alone did not reproduce the neural geometry found in the Rodgers dataset.

Taken together, the findings establish a framework for integrating visual and tactile processing within an anatomically motivated neural network. The visual and multimodal results demonstrate the potential of this approach, while the whisker evaluation provides a clear account of the limitations of the current tactile representation.

6.2 Main Contributions

A central contribution of this thesis is the development of a biologically motivated whisker encoder. The architecture preserves the spatial arrangement of the whisker array, incorporates temporal integration across the whisking sequence, and propagates tactile representations through cortical regions inspired by the mouse somatosensory hierarchy. This provides an initial computational model for investigating how structured whisker dynamics may be transformed across cortical processing stages.

The thesis also contributes a simulation framework for generating trial-paired visual and tactile sensory observations. The Unity environment models whisker geometry, prescribed whisking trajectories, object contact, and visual input under controlled conditions. In addition to supporting multimodal training, the simulation reproduces the six shape and distance conditions required for comparison with the Rodgers neural recordings.

Another contribution is the implementation of a biologically motivated mechanism for direct interaction between the sensory pathways. Rather than combining the modalities only through concatenation, the model uses representations from the whisker pathway to produce feature-dependent inhibition within superficial VISp. The inhibitory mechanism reproducibly altered downstream representational geometry across five training seeds, with its strongest within-model effect in VISp5, although it did not reliably improve correspondence relative to an architecture with the gate removed.

The work further demonstrates that contrastive learning can align representations produced by anatomically structured visual and tactile encoders. Training produced a shared embedding space in which paired observations could be retrieved substantially above chance from 100,000 candidates drawn from the training dataset. This establishes a practical method for learning associations between sensory pathways without requiring manual labels for object identity or category.

Finally, the thesis evaluates the model directly against neural recordings from both visual and somatosensory experiments. The Allen analysis identified measurable correspondence between model and visual cortical representations, while the Rodgers analysis revealed weak correspondence in the whisker pathway. Reporting both outcomes provides a more complete evaluation of the model and establishes a reproducible benchmark for improving future tactile encoders.

These contributions combine sensory simulation, anatomical constraints, multimodal learning, and neural evaluation within a single modelling framework. The resulting system can therefore support future investigations into how cortical organization and interactions between sensory pathways influence learned representations.

6.3 Final Remarks

The results of this thesis illustrate both the potential and difficulty of developing biologically grounded models of multisensory processing. Anatomical structure and trial-paired sensory experience were sufficient to produce measurable visual correspondence and strong multimodal alignment. They were not, however, sufficient to ensure that the whisker encoder reproduced the representational structure observed in somatosensory neural recordings.

Interpretation of the whisker evaluation was limited by the availability of suitable neural data, as whisker-related neural recordings are often collected during highly specific behavioural tasks. The Rodgers recordings were obtained during a particular object discrimination task, whereas the whisker encoder was trained through multimodal contrastive learning and was not optimized to perform the same behaviour. Consequently, weak neural correspondence may reflect both limitations of the current encoder and a mismatch between the experiences and objectives of the model and those of the animals.

The evaluation of multisensory representations was similarly constrained by the absence of a unified neural dataset containing synchronized visual input, whisker activity, behaviour, and recordings from visual and somatosensory cortical regions. Separate datasets were therefore used to evaluate the visual and tactile pathways. Although this approach allowed each pathway to be compared with experimental neural activity, it did not provide a direct biological test of the learned multimodal representations or the inhibitory interaction between the pathways. Neural recordings collected during controlled visual and tactile behaviour would provide a stronger basis for evaluating these components.

The present model should therefore be understood as a foundation for studying multisensory cortical computation rather than a complete account of mouse perception. Its main value lies in bringing visual and whisker processing into one anatomically motivated framework and making the resulting representations testable against available experimental data. By identifying both successful components and limitations in the model and available neural benchmarks, this work provides a basis for developing more accurate computational models of how sensory pathways interact within the brain.

References

[1] Ehud Ahissar, Ronen Sosnik, and Sebastian Haidarliu. Transformation from temporal to rate coding in a somatosensory thalamocortical pathway. *Nature*, 406:302–306, 2000.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[4] Daniela Camillo, Christiaan N. Levelt, and J. Alexander Heimel. Lack of functional specialization of neurons in the mouse primary visual cortex that have expressed calretinin. *Frontiers in Neuroanatomy*, 8:89, 2014.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020.

[6] Trinity Chung, Yuchen Shen, Nathan C. L. Kong, and Aran Nayebi. Task-optimized convolutional recurrent networks align with tactile processing in the rodent brain. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

[7] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of CNNs improves

robustness to image perturbations. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[8] Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel K. Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020.

[9] Mathew E. Diamond, Moritz von Heimendahl, Per M. Knutsen, David Kleinfeld, and Ehud Ahissar. 'Where' and 'what' in the whisker sensorimotor system. *Nature Reviews Neuroscience*, 9(8):601–612, 2008.

[10] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[11] Séverine Durand, Ramakrishnan Iyer, Kenji Mizuseki, Saskia de Vries, Stefan Mihalas, and R. Clay Reid. A comparison of visual response properties in the lateral geniculate nucleus and primary visual cortex of awake and anesthetized mice. *The Journal of Neuroscience*, 36(48):12144–12156, 2016.

[12] JaeAnn M. Dwulet, Nawal Zabouri, Jan H. Kirchner, Marina E. Wosniack, Alessandra Raspanti, Deyue Kong, Gerrit J. Houwen, Paloma P. Maldonado, Christian Lohmann, and Julijana Gjorgjieva. Correlated spontaneous activity sets up multi-sensory integration in the developing higher-order cortex. *eLife*, 2024. Reviewed Preprint.

[13] Csaba Erö, Marc-Oliver Gewaltig, Daniel Keller, and Henry Markram. A cell atlas for the mouse brain. *Frontiers in Neuroinformatics*, 12:84, 2018.

[14] Asif A. Ghazanfar and Charles E. Schroeder. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285, 2006.

[15] Ariel Gilad and Fritjof Helmchen. Spatiotemporal refinement of signal flow through association cortex during learning. *Nature Communications*, 11(1):1744, 2020.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[17] Maëlle Guyoton, Giulio Matteucci, Charlie G. Foucher, Matthew P. Getz, Julijana Gjorgjieva, and Sami El-Boustani. Cortical circuits for cross-modal generalization. *Nature Communications*, 16(1):4230, 2025.

[18] Julie A. Harris, Stefan Mihalas, Karla E. Hirokawa, Jennifer D. Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, Aaron Feiner, David Feng, Nathalie Gaudreault, Charles R. Gerfen, Nile Graddis, Peter A. Groblewski, Alex M. Henry, Anh Ho, Robert Howard, Joseph E. Knox, Leonard Kuan, Xiuli Kuang, Jerome Lecoq, Phil Lesnar, et al. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575:195–202, 2019.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[20] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021.

[21] Meenakshi Khosla and Alex H. Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 326–341, 2024.

[22] David Kleinfeld, Ehud Ahissar, and Mathew E. Diamond. Active sensation: insights from the rodent vibrissa sensorimotor system. *Current Opinion in Neurobiology*, 16(4):435–444, 2006.

[23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529, 2019.

[24] Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013.

[25] Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[27] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J. Majaj, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[28] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, June 2019.

[29] Yanqi Liu, Pol Bech, Keita Tamura, Lucas T. Délez, Sylvain Crochet, and Carl C. H. Petersen. Cell class-specific long-range axonal projections of neurons in mouse whisker-related somatosensory cortices. *eLife*, 13:RP97602, 2024.

[30] William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017.

[31] Eshed Margalit, Hyodong Lee, Dawn Finzi, James J. DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14):2435–2451.e7, 2024.

[32] James H. Marshel, Marina E. Garrett, Ian Nauhaus, and Edward M. Callaway. Functional specialization of seven mouse visual cortical areas. *Neuron*, 72(6):1040–1054, 2011.

[33] Patrick Mineault, Shahab Bakhtiari, Timothy Lillicrap, Christopher C. Pack, and Blake A. Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[34] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[35] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J. DiCarlo, and Daniel L. K. Yamins. Task-driven convolutional recurrent

models of the visual system. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[36] Aran Nayebi, Nathan C. L. Kong, Chengxu Zhuang, Justin L. Gardner, Anthony M. Norcia, and Daniel L. K. Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, 19(10):e1011506, 2023.

[37] Seung Wook Oh, Julie A. Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M. Henry, Marty T. Mortrud, Benjamin Ouellette, Thuc Nghi Nguyen, Staci A. Sorensen, Clifford R. Slaughterbeck, Wayne Wakeman, Yang Li, David Feng, Anh Ho, Eric Nicholas, Karla E. Hirokawa, Phillip Bohn, Kevin M. Joines, Hanchuan Peng, Michael J. Hawrylycz, John W. Phillips, John G. Hohmann, Paul Wohnoutka, Charles R. Gerfen, Christof Koch, Amy Bernard, Chinh Dang, Allan R. Jones, and Hongkui Zeng. A mesoscale connectome of the mouse brain. *Nature*, 508:207–214, 2014.

[38] Simon Peron, Ravi Pancholi, Bettina Voelcker, Jason D. Wittenbach, H. Freyja Ólafsdóttir, Jeremy Freeman, and Karel Svoboda. Recurrent interactions in local cortical circuits. *Nature*, 579:256–259, 2020.

[39] Simon P. Peron, Jeremy Freeman, Vijay Iyer, Caiying Guo, and Karel Svoboda. A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron*, 86(3):783–799, 2015.

[40] Carl C. H. Petersen. The functional organization of the barrel cortex. *Neuron*, 56(2):339–355, 2007.

[41] David J. Pinto, Jed A. Hartings, Joshua C. Brumberg, and Daniel J. Simons. Cortical damping: Analysis of thalamocortical response transformations in rodent barrel cortex. *Cerebral Cortex*, 13(1):33–44, 2003.

[42] Anna Pitas, Ana Lía Albarracín, Manuel Molano-Mazón, and Miguel Maravall. Variable temporal integration of stimulus patterns in the mouse barrel cortex. *Cerebral Cortex*, 27(3):1758–1764, March 2017.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[44] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[45] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[46] Chris C. Rodgers. A detailed behavioral, videographic, and neural dataset on object recognition in mice. *Scientific Data*, 9:620, 2022.

[47] Thomas Sharp, Rasmus S. Petersen, and Steve B. Furber. Real-time million-synapse simulation of rat barrel cortex. *Frontiers in Neuroscience*, 8:131, 2014.

[48] Jianghong Shi, Bryan Tripp, Eric Shea-Brown, Stefan Mihalas, and Michael A. Buice. MouseNet: A biologically constrained convolutional neural network model for the mouse visual cortex. *PLOS Computational Biology*, 18(9):e1010427, 2022.

[49] Nicholas J. Sofroniew, Yurii A. Vlasov, Samuel A. Hires, Jeremy Freeman, and Karel Svoboda. Neural coding in barrel cortex during whisker-guided locomotion. *eLife*, 4:e12559, 2015.

[50] Chenchen Song, Denise M. Piscopo, Cristopher M. Niell, and Thomas Knöpfel. Cortical signatures of wakeful somatosensory processing. *Scientific Reports*, 8(1):11977, 2018.

[51] Barry E. Stein and Terrence R. Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4):255–266, 2008.

[52] Simon Weiler, Vahid Rahmati, Marcel Isstas, Johann Wutke, Andreas Walter Stark, Christian Franke, Jürgen Graf, Christian Geis, Otto W. Witte, Mark Hübener, Jürgen Bolz, Troy W. Margrie, Knut Holthoff, and Manuel Teichert. A primary sensory cortical interareal feedforward inhibitory circuit for tacto-visual integration. *Nature Communications*, 15(1):3081, 2024.

[53] Jialin Wu, Shreya Saha, Yiqing Bo, and Meenakshi Khosla. Comparing and integrating different notions of representational correspondence in neural systems. *arXiv preprint arXiv:2509.21628*, 2026.

[54] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural

responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[55] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, July 2017.

[56] Chengxu Zhuang, Jonas Kubilius, Mitra J. Z. Hartmann, and Daniel L. Yamins. Toward goal-driven neural network models for the rodent whisker-trigeminal system. In *Advances in Neural Information Processing Systems*, volume 30, pages 2552–2562, 2017.

[57] Nadina O. Zweifel, Nicholas E. Bush, Ian Abraham, Todd D. Murphey, and Mitra J. Z. Hartmann. A dynamical model for generating synthetic data to quantify active tactile sensing behavior in the rat. *Proceedings of the National Academy of Sciences*, 118(27):e2011905118, 2021.