

**THE IMPACT OF SCENE CONTEXT ON VISUAL OBJECT
RECOGNITION: COMPARING HUMANS AND MONKEYS**

ANAA SALIM ZAFER

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN BIOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2024

© ANAA SALIM ZAFER, 2024

Abstract

During natural vision, we rarely see objects on their own but rather within rich and complex environments. Understanding how the brain recognizes objects in these natural scenes by using contextual information remains a significant challenge. To explore neural mechanisms similar to human visual processing, it is crucial to have an animal model that behaves like humans. This way, the neural mechanisms we identify can offer insights relevant to the human brain. In this study, we investigated whether rhesus macaques could serve as a model for human context-driven object recognition. We did this by measuring their ability to identify visual objects with varying amounts, quality, and congruency of contextual cues. Our behavioral metrics showed remarkably similar context-dependent patterns between humans and monkeys. However, neural responses in the inferior temporal (IT) cortex of monkeys, who were not explicitly trained to recognize objects in context, could only partly explain this cross-species similarity. The behavioral similarities that were not explained by context-naïve neural data highlight significant gaps in our understanding. Our findings reveal a fascinating alignment between human and monkey visual object processing that cannot be fully explained by neural activity in a key visual region.

Acknowledgements

I am deeply grateful to Dr. Kohitij Kar for giving me the opportunity to work under his guidance and be a part of the ViTA Lab. His willingness to take a chance on me and his patient explanations have greatly enriched my understanding of neuroscience, ensuring clarity on concepts that were hard to grasp. I also extend my thanks to York University for providing practical experience as part of their master's program.

My heartfelt appreciation goes to our ViTA lab post-doc advisor, aka the Captain MD, Dr. Hamid Ramezanpour. His readiness to address my countless questions and his clear explanations of complex concepts have been invaluable. His constant motivation, boundless patience, and technical support have been crucial to my progress. I would also like to thank my labmates, Ms. Shirin Taghian Alamooti and Ms. E(liza)veta Yakubovskaya, for their academic support and for being wonderful colleagues. Their insightful comments and constructive criticisms at various stages of my research were thought-provoking and helped me refine my ideas and stay motivated. I am grateful to each of them for maintaining high research standards and teaching me the essence of effective research.

I extend my gratitude to all those who have provided immense support throughout my project. Mrs. Ekta Mehta Ramteke, who inspired me to pursue this master's degree, and Mr. Barkat Jasani, who encouraged it. Mr. Siddhant Nagpal, for helping me settle into a new country while balancing his own academic pursuits. Mr. Rajiv Badaloo, for being my confidant and constant pillar of strength, and my best friend, Ms. Safiya Ali, for being my unwavering support system through challenges. My lab mates, Ram Ahuja and Soroush Ziaee, for their camaraderie and for making the lab environment both enjoyable and supportive. Special thanks to Ms. Sara Djambazovska for her collaboration on this project.

Most importantly, I am profoundly grateful to my family, without whom none of this would have been possible. Special appreciation goes to my mother for her kindness, gentleness, love and unwavering support.

Lastly, but certainly not least, I thank God for His all-encompassing mercy and presence in every part of my life.

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents	v
List of Figures	ix
1 Introduction	1
1.1 Visual Object Recognition	1
1.2 Ventral Visual Pathway	3
1.3 Macaques as a model for studying the human brain	5
1.4 IT cortex, visual object recognition, and neural population decoding	6
1.5 Contextual processing in the brain	10
1.6 Structure of this thesis	15
2 Aim 1: Compare monkey and human behavior on effects of visual context during object discrimination tasks.	16

2.1	Background	16
2.2	Hypothesis	18
2.3	Visual stimuli	20
2.3.1	Image details	21
2.3.2	Contextual variations	21
2.4	Methods	24
2.4.1	Behavioral task in humans	24
2.4.2	Behavioral task in monkeys	25
2.5	Behavioral Metrics for evaluating performance	26
2.6	Internal Behavioral Consistency	29
2.7	Human-monkey shared behavioral variance	31
2.8	Statistical Analyses	31
2.9	Result	32
2.10	Potential Confounds and mitigation strategy	52
3	Aim 2: Analyze neuronal activity recorded across the IT cortex in context-naïve macaque monkeys to study contextual modulation effects during passive image observation task.	57
3.1	Background	57
3.2	Hypothesis	60
3.3	Visual stimuli	62
3.4	Methods	63

3.4.1	Experimental design - Passive Fixation Task	63
3.4.2	Eye Tracking	64
3.4.3	Macaque Neural Data	66
3.4.4	Partial correlation analysis	68
3.5	Results	69
3.5.1	Population Activity of Neurons: Context-Dependent Visual Processing, Dynamic Firing Rate Patterns, and Neural Decoding	69
3.5.2	Modulation in the pattern of neural prediction performance activity showing the impact of context	73
3.5.3	Neural Decoding Accuracy Across Increasing Numbers of Neurons . .	75
3.5.4	Population activity across the IT cortex in a context-naive monkey fully explains the shared behavioral variance between humans and monkeys at the overall context level	77
3.5.5	Population activity across the IT cortex in a context-naive monkey only partially explains the shared context-driven behavioral variance between humans and monkeys at the image level	79
3.5.6	Population activity across the IT cortex in a context-naive monkey showing a decline in explaining the shared context-driven behavioral variance between humans and monkeys at different behavioral metrics.	82
4	Discussion	85
4.1	Context modulates visual object recognition in humans and monkeys	87

4.2	Goodness of monkeys as a model of human contextual processing	88
4.3	Role of IT Cortex in Processing Scene Context	91
5	Appendix	95

List of Figures

1.1	A view of Ventral Visual Pathway	5
1.2	What is this object? Can you recognize it?	11
1.3	What is this object here? Look at Figure. 1.2 first.	11
2.1	Various Contextual variations illustrated on an image	21
2.2	Binary Object Identity Discrimination Task: Human Online Participation and Monkey Homecage Implementation	26
2.3	Multidimensional performance evaluation for object discrimination showing different behavioral metrics	30
2.4	Impact of Contextual Manipulations on Human Object Recognition Accuracy and Reliability	34
2.5	Impact of Contextual Variations on Object Recognition Performance in different images	38
2.6	Impact of Contextual Manipulations on Monkey Object Recognition Accuracy and Reliability	40

2.7	Impact of Contextual Variations on Monkey Object Recognition Performance	42
2.8	Correlative Analysis of Contextual and Image-Level Accuracy Between Two Monkeys	44
2.9	Correlative Analysis of Contextual and Image-Level Accuracy Between Humans and Monkeys	45
2.10	Monkeys and humans show similar (but not identical) context-driven behavioral changes	48
2.11	Comparative Analysis of Shared Behavioral Variance Between Humans and Monkeys at Different level of Analysis	51
2.12	Influence of Low-Level Image Features on Object Recognition	54
2.13	Low Predictive Value of Low-Level Image Features on Object Recognition at Context and Image Levels	55
3.1	Hypotheses on the contribution of IT neuron populations to the shared variance in human and monkey visual processing systems	63
3.2	Neural decoding of object categories from visual stimuli in monkeys using chronically implanted arrays and cross-validated classification decoders . . .	65
3.3	Normalized neuronal firing rate responses to visual context modulations . . .	70
3.4	Temporal Dynamics of IT Cortex Neural Response Decoding Accuracy . . .	72
3.5	Impact of Contextual Manipulations on Neural Prediction Accuracy	73
3.6	Decoding Accuracy as a Function of the Number of Neurons	76

3.7	Context-Level Human-Monkey Behavioral Variance Explained by Neural Data	
	(HM-EV)	78
3.8	Image-Level Human-Monkey Behavioral Variance Explained by Neural Data	
	(HM-EV)	81
3.9	Quantifying the IT Cortex’s Contribution to Explained Variance Across Be-	
	havioral Metrics in Primates	83
5.1	Learning and generalization of object recognition in context by macaque	
	monkeys: performance over time	96
5.2	Evaluation of object-level performance metrics in visual recognition task . .	98
5.3	Correlative Analysis of Object-Level Accuracy Between Primates	99
5.4	Confusion Matrix Visualization of Human Object Recognition Errors by Cate-	
	gory(1 - O2 Metric)	101
5.5	Confusion Matrix Visualization of Monkey Object Recognition Errors by	
	Category(1 - O2 Metric)	102

Chapter 1

Introduction

1.1 Visual Object Recognition

Recognizing a stop sign while driving, finding your favorite cereal in the store, or knowing your friend's face in a photo seems easy to us. But these actions are very complex. We quickly identify and categorize many different things in matter of seconds, even with variations in object composition and background (DiCarlo, 2012). From a survival perspective, being good at this is important for our lives. We need to quickly and accurately recognize things around us for everyday tasks like finding food, talking to others, using tools, and following directions. The aim is to understand how the visual system takes a retinal image, identifies the objects within it, and the underlying mechanisms involved. To understand how we recognize objects, Kar et al [1] propose establishing a foundational set of tasks for evaluating and understanding both biological performance and the performance of computational models. By defining precise and consistent tasks, researchers can accurately measure how well computational

models replicate biological processes, thereby deepening our insight into visual recognition systems. DiCarlo et al. (2012) [2] mention "core object recognition" as an effective start point to this evaluation. The author talks about "object recognition as the ability to assign labels (e.g., nouns) to particular objects, ranging from precise labels ("identification") to coarse labels ("categorization"). More specifically, we focus on the ability to complete such tasks over a range of identity-preserving transformations (e.g., changes in an object's position, size, pose, and background context), without any object-specific or location-specific pre-cuing". Previous studies have shown that macaques were able to identify objects in the central field of view at approximately 250ms [3] and reaction time for humans has been around under 200ms [4]. This performance is also available even if the presentation time of images sequentially at rates even less than 100ms [5]. Looking at the time needed to make a response in task behavior, it can be speculated that the central visual view is processed to help recognition in less than 200ms [4, 5, 6]. Primates usually explore their surroundings with quick eye movements and resulting short fixations of 200ms, during which they quickly identify one or more objects in their central vision [7, 8]. Dicarlo et al[9] refer to this extremely fast, and very accurate behavior in object recognition as "core recognition". Besides identifying the object category, humans also infer other details about objects like size, position, orientation, color, and material properties. These factors influence our perception of object identity and are also part of core object recognition [10].

1.2 Ventral Visual Pathway

Primates can identify and classify various objects in a remarkably rapid manner, often within just 200 milliseconds following the presentation of the stimulus despite the variation in object appearance [3, 11, 12, 13]. The intricate and complex details of primate vision offer a glimpse into a sophisticated mechanism honed through millions of years of evolutionary progress. One promising strategy for attempting to understand human visual recognition is to characterize the neural system that accomplishes it: the ventral visual pathway (VVP) which has been illustrated in Figure 1.1, which extends from the occipital lobe into inferior and lateral regions of the temporal lobe. At the core of the primate visual pathway lies the retina, a specialized layer of tissue responsible for light detection and conversion into electrical signals. These signals then traverse through the optic nerve to reach the Lateral Geniculate Nucleus (LGN) within the thalamus for initial processing. Subsequently, they proceed to the primary visual cortex (V1) located in the occipital lobe of the brain. Here, within an elaborate network of neurons, specialized neurons analyze the orientation and movement of light stimuli [14, 15, 16]. V1 primarily serves the function of fundamental feature extraction, particularly in tasks such as detecting orientations and edges [17]. Following V1, the primate visual system diverges into two distinct pathways known as the dorsal and ventral streams. The dorsal stream, often referred to as the "Where pathway," specializes in processing motion and spatial comprehension. In contrast, the ventral stream, known as the "What pathway," operates in a structured hierarchical manner with a unidirectional flow

of information, focusing on fundamental object recognition. The impressive capability of high-speed object recognition highlights the rapidity and efficiency with which computations are processed through the ventral visual pathway [18, 2, 19, 20]. The ventral stream, in particular, is subdivided into several areas including V1, V2, V4, and IT (inferior temporal cortex) [2]. V2 delves further into analyzing more intricate attributes, including textures and contours [21]. Ahead of V2, V4 which although remains engaged in the processing of orientation [22], it primarily associates with color analysis [23] contributing to interpreting hues, shades, and color saturations present in the visual field and is pivotal in upholding color constancy. This function ensures a steady perception of color despite changes in illumination [24], thereby contributing significantly to visual stability and accuracy. Subsequent to V4, the Inferior Temporal (IT) cortex, plays a pivotal role in object recognition as well as object identity discrimination. This region of the brain is adept at decoding complex visual stimuli, providing neural representation of objects [25, 26, 27]. IT cortex is further segmented into posterior, central, and anterior IT cortices (pIT, cIT, aIT). These areas progressively refine visual signals, transitioning from rudimentary feature detection in earlier stages such as V1, to more complex object representations in regions like IT. This progression is characterized by an expansion in the receptive fields (RFs) of neurons, indicating a broader scope of information integration as one moves from V1 towards IT [28]. Such intricate organization underscores the sophisticated nature of primate visual processing, finely tuned through evolutionary adaptations to discern and interpret the intricacies of the surrounding environment. IT cortex cells exhibit selective responsiveness to complex objects, maintaining a certain degree of

recognition accuracy despite variations in the object’s size and position [29, 30, 31]. This consistency enables primates to identify objects across different circumstances [25, 32] and it is seen that groups of neurons are frequently specialized to exhibit selectivity towards distinct categories, such as faces, locations, and fruits [33] etc.

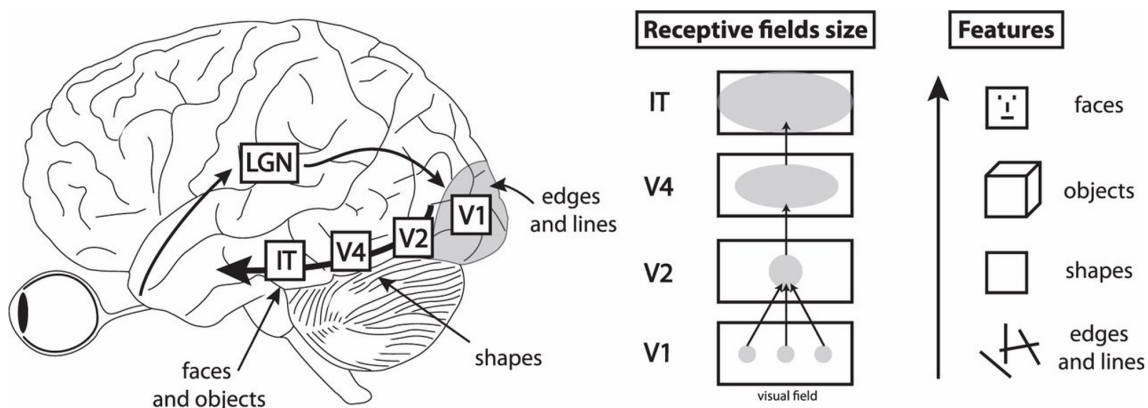


Figure 1.1: Left: The primate ventral visual stream processes information from the retina through the LGN, V1, V2, V4, to IT. Receptive fields expand from lower to higher visual areas, aggregating more of the visual field. Image obtained from Manassi et. al (2013)[34]. This illustration underscores the complexity of visual perception and the intricate pathways that enable the brain to interpret visual stimuli. Right: V1 codes basic features like edges, while higher-level neurons integrate these, representing more intricate features.

1.3 Macaques as a model for studying the human brain

To uncover the neural mechanism behind human visual processing, researchers have sought the aid of multiple animal models like Hubel and Wiesel studying cats’ striate cortex [15], Van Essen et al report on the mammalian visual cortex [28] and a rodent model being utilized by zoccolan et al [35] to study invariant object recognition. The most widely used animal

model for studying human visual perception has been the rhesus macaque monkeys [36, 27]. Fascinatingly, the ventral visual processing stream, which consists of a series of interconnected layers of the neocortex [28], is notably well-conserved across various primate species. This includes Old-World monkeys like rhesus macaques (*Macaca mulatta*) and baboons (*Papio papio*), which diverged from humans approximately 25 million years ago [37]. Years of research tell us that humans and monkeys share strong similar structures and functions in their visual cortical areas [38, 39, 40]. However, humans possess abilities not seen in monkeys, and there are significant differences in their brain structures [37]. Previously, we observed striking similarities in invariant visual object recognition behavior between these two primate species [13, 12]. Therefore, it has long been believed that monkeys and humans share similar high-level visual behaviors and the relative brain structures that support them. While there are important distinctions in capabilities and brain structures, the fundamental conservation of visual processing mechanisms across primates provides a robust basis for using animal models to advance our understanding of human visual cognition.

1.4 IT cortex, visual object recognition, and neural population decoding

Desimone et al[31] investigated the response selectivity of neurons in the inferior temporal (IT) cortex of macaques. The study systematically examines the responses of IT neurons to both simple stimuli, such as edges and bars, and complex stimuli, such as models of flowers, snakes,

hands, and faces. It was found that many responsive neurons reacted to virtually every tested stimulus. However, certain neurons exhibited stimulus selectivity based on shape, color, or texture. Another study on the IT cortex neurons in monkeys performing a delayed matching-to-sample task revealed that IT cells demonstrated size and location invariance in their sensory responses to object stimuli. These findings suggest IT neuronal populations mediate both invariant object recognition and perception of size/location transformations, supporting both size-dependent and size-independent visual processing [30, 41]. Researchers speculate that neurons in the IT do not respond exclusively to entire objects but rather to specific subsets of features within objects. A study was conducted on cortical areas along the ventral stream in anesthetized macaque monkeys. By employing a reduction method to pinpoint critical features that elicit maximal responses from neurons, this method involved simplifying complex object images step-by-step until only essential features remained, allowing them to identify the minimal set of features that trigger robust neuronal activity. Their findings suggest that neurons in these cortical regions do not solely respond to entire objects but also exhibit sensitivity to specific subsets of object features. Another study showed that cells posterior IT responded to various complex shapes, textures, and color combinations, while those in anterior IT showed selective responses to more abstract object features. Importantly, many neurons in V4 and posterior IT also displayed varying degrees of responsiveness to simpler features like bars and squares, indicating hierarchical processing of visual information. This hierarchical processing likely supports the brain's ability to recognize objects through the integration of simpler visual features into more complex representations. The organized

arrangement of neurons within columns in these cortical areas further suggests structured processing of visual information, where cells with similar feature selectivities are vertically aligned across the cortical sheet. Overall, these findings underscore the intricate nature of object representation in the ventral visual pathway, highlighting the roles of V2, V4, and IT in integrating and processing visual information from basic features to complex object representations [42, 43, 44, 45]. Some neurons respond to specific, detailed aspects of objects, like textures or colors, which helps in recognizing fine details whereas other neurons process the overall spatial arrangement of object parts, focusing on how these parts are organized relative to each other, which is crucial for understanding an object’s overall structure and functionality. This dual processing capability allows the brain to effectively handle both detailed and holistic aspects of object perception in three-dimensional space[46, 47]. Since single IT neurons cannot code complete object identities, this leads to the thinking that a population code could be utilized. Parts-Based encoding shown by the Recognition-by-Components model, breaks an object into separate, distinct parts. Neurons individually respond to these parts, aiding in object recognition by piecing together the different elements of the object. In contrast, holistic encoding methods like the eigenfaces approach for face recognition, decompose an object into overlapping features. Here, each neuron contributes to a composite view of the object, with multiple neurons together capturing the ensemble of features that make up the object as a whole [48, 49]. Also, neurons with similar selectivities towards objects tend to cluster together [44, 50]. Rather than individual neurons recognizing whole objects, each neuron responds to specific features of objects. The collective activity of

a group of these neurons encodes complete objects. Every object elicits a unique pattern of neural activity across this neuronal population, resulting in a distinct "response vector". This vector represents the object within the neural activity space [33, 51]. The neural response to a specific object can be conceptualized as a point within a high-dimensional space, which may encompass up to 10,000 dimensions based on the number of neurons involved. This expansive dimensional capacity enables the encoding of a diverse range of objects and their distinct features. Objects that are similar or represent different perspectives of the same object tend to activate similar patterns of neurons, leading these response vectors to cluster within particular regions of this space. These clusters are referred to as "manifolds," which are defined as areas where each point reflects shared transformations, such as changes in the object's viewpoint. Object recognition and categorization thus depend on the brain's ability to discriminate among these manifolds, each of which corresponds to a set of similar objects. Effective object recognition, therefore, hinges on distinguishing these separate manifolds within the neural representational space [9, 42, 52]. Hence, population-based decoding seems an effective way because it utilizes the collective responses of neuron populations to encode complex visual information within this high-dimensional neural space. By distinguishing between these manifolds, the brain achieves robust and nuanced object recognition. This approach showcases the brain's ability to dynamically interpret sensory information through a sophisticated neural architecture.

1.5 Contextual processing in the brain

As it is known, in the real world, objects rarely exist in isolation; they are typically encountered as part of a large scene, surrounded by other objects, and set within a broader contextual framework. This natural embedding of objects is a critical aspect of how we perceive and interact with the world. Viewing and interpreting an image, for instance, involves more than just identifying specific objects in a void, it also involves a nuanced understanding of how those objects within the image relate to each other in terms of their size, spatial positioning, and how they collectively contribute to the overall setting 1. This process is exemplified in Figure. 1.2 . There is no manipulation of the image, or any alteration present, yet recognizing the object remains challenging. The difficulty in identifying the object in this unaltered state underscores the intricate nature of visual processing. Then, look at Figure. 1.3; the box highlighted shows the same object information as Figure. 1.2. Here identifying the object is significantly easier [53]. This stark difference illustrates the profound impact of contextual cues on our perception and interpretation of visual information suggesting that surrounding context greatly affects the way object information is perceived.

Even though, we have seen through various studies how context can influence our perception of visual information including psychophysical demonstrations, which show how contextual elements can dramatically alter our interpretation and understanding of a scene. Bomatter et al. (2021) [54] investigate how unusual or out-of-context elements in both artificial and natural scenes challenge our standard contextual understanding, highlighting the brain's



Figure 1.2: What is this object? Can you recognize it?



Figure 1.3: What is this object here? Look at Figure. 1.2 first.

response to atypical configurations. Oliva and Torralba's (2007) [55] research underscores the crucial role of context in object recognition, illustrating how familiar surroundings aid in faster and more accurate identification of objects, and how this process is disrupted in unfamiliar settings. Bar (2004) [56] delves into the neuroscience aspect, examining how the brain integrates contextual information with specific object perception, shaping our overall interpretation of visual scenes. These studies collectively underscore that our perception of visual information is not merely isolated to the object observed but is deeply influenced by surrounding context and past experiences, a concept vital for fields ranging from cognitive

psychology to artificial intelligence. However, there is still a notable gap in our knowledge concerning the specific cognitive and neural mechanisms that allow us to integrate information about objects and their surroundings to understand the overall picture. There is a lack of understanding about the mechanisms that the brain uses to integrate previous knowledge with new input to enhance visual perception as well as speedy information processing.

Studies historically have focused on examining the “low-level” effects of context on visual perception in the brain area of primary visual cortex i.e., V1. The research by Henry and Kohn (2020) [57] and Coen-Cagli et al. (2015)[58] offers insightful perspectives into the varying impacts of context on visual perception, especially in the primary visual cortex (V1). The 2020 study by Henry and Kohn zeroes in on V1 limitations in distinguishing features within visually crowded environments, emphasizing V1’s challenges in segregating individual elements in densely populated visual fields. This highlights the constraints of early visual processing in spatial contexts. Conversely, Coen-Cagli et al.’s 2015 study explored the adaptability of surround suppression in the primary visual cortex (V1) using natural images. They discovered that the strength of surround suppression varied significantly across different images. This finding suggests that surround suppression in V1 is not just a fixed response but is modulated based on the brain’s inference about the visual environment’s statistical properties, revealing a sophisticated and flexible aspect of cortical processing. Together, these studies illuminate the intricate balance between early and advanced visual processing stages in interpreting complex visual scenes, revealing a critical area of ongoing research in understanding the full spectrum of visual perception. Despite the significant amount of

research that has been conducted, there is still crucial void in our understanding of how context affects visual recognition at more advanced stages- high-level processes, meaning in the later areas of the ventral visual stream. These high-level processes are critical for tasks such as differentiating between objects, also known as object-identity discrimination. This distinction is pivotal for a more profound comprehension of how the brain interprets complex visual scenes.

To address this research gap, our study aims to define and measure high-level visual recognition behavior. We intend to employ a specific methodology, coupled with associated tasks, to probe the depth of how the brain discerns and differentiates between different objects within a scene. By focusing on the task of object-identity discrimination (Aim 1), we hope to uncover the intricate mechanisms by which the brain integrates previous high-level knowledge with incoming visual input, thereby enhancing the overall process of visual perception.

To improve our comprehension of the fundamental mechanisms of contextual reasoning, it is important to study how incoming sensory information is altered by top-down processes both in terms of behavior and neurophysiological activity. We can gain insight into this by examining and comparing how context influences visual recognition in both non-human primates (NHP) and humans through quantitative methods (Aim 1). Furthermore, by measuring the activity of neurons in certain regions of the brain, such as the IT cortex, during the passive image observation task, we can gain an understanding of the neural mechanisms that regulate contextual reasoning. (Aim 2).

To complement our study's focus on high-level visual recognition and contextual reasoning,

we have chosen to employ neural recording techniques, particularly in the IT Cortex of non-human primates. This methodological choice is informed by the distinct advantages neural recording offers over fMRI in this specific context. Neural recording provides unparalleled temporal resolution, capturing rapid neuronal activities that are crucial for understanding the dynamic processes of visual perception and recognition. Its direct measurement of neuronal activity enables a more precise exploration of how the brain processes visual stimuli down to milliseconds [26]. This approach aligns with our aims to deeply understand the neural mechanisms of contextual reasoning in visual recognition, bridging the gap between sensory input and cognitive processing. Through these investigations, we aspire to elucidate the nuanced interplay between contextual cues and object recognition, enhancing our understanding of visual perception.

A crucial initial step in understanding how context affects visual recognition is to study the effect of scene context on behavior. This can be achieved by defining the problem and creating experimental methods to investigate the role of visual context in humans and non-human primates. Behavioral experiments in humans will also be carried out to compare the core mechanisms in both species. The goal is to elucidate whether context improves or hinders object recognition, which context cues are crucial for recognition, and the dynamic aspects of contextual reasoning. Behavioral experiments were performed on monkeys and data were compared against human behavioral experiment results. The correlation, as expected, established a foundation for studying the neural mechanisms in the second objective, through neurophysiological experiments. In the footsteps of the behavioral experiments, we conducted

large-scale, chronic multi-electrode recordings in the macaque inferior temporal (IT) cortex, a region known to be involved in object recognition [25, 59]. By analyzing these recordings using neural decoders, we aim to predict human behavior with respect to object recognition. Research in the field of vision has often led to progress in other areas as well. Gaining knowledge about visual reasoning will also aid research in other senses where context plays an important role [60, 61].

1.6 Structure of this thesis

Following the introduction above, the next sections are divided as follows:

In Chapter 2, I conducted a systematic investigation into how different scene contexts influence object recognition behavior in humans and how much of this behavior is shared with monkeys.

In Chapter 3, I present the results of the research involving the recording and analysis of neuronal activity in the IT cortex of macaques during passive image observation tasks in context-naïve monkeys. We specifically ask — do the responses of the IT cortex in monkeys that cannot perform the contextual processing-based task already explain patterns in the observed behavior of humans and context-trained monkeys?

In Chapter 4, I discuss the major findings of the thesis, the implications of these findings for future neuroscience research, and reflect on the broader impacts of understanding object recognition across different species.

Chapter 2

Aim 1: Compare monkey and human behavior on effects of visual context during object discrimination tasks.

2.1 Background

It is essential to determine whether scene context influences behavior to evaluate the role of context in visual recognition. Interest in the statistical properties of natural images has been long-standing, accompanied by foundational studies examining the influence of context in visual perception [62, 63]. Studies have also investigated how sensory response in the visual cortex’s pyramidal cells is contextually modulated, specifically through surround suppression[64]. There are other studies that have qualitatively analyzed the role of context to show its impact on visual processing e.g Auckland, Cave, and Donnelly (2007) [65] investigate the impact of non-target objects on the recognition process, suggesting that extraneous elements within a scene can significantly modulate perceptual attention and accuracy. Biederman, Mezzanotte, and Rabinowitz (1982) explore the cognitive repercussions of encountering objects that deviate from expected relational norms, shedding light on the

mechanisms of scene perception when faced with incongruities [66]. Meanwhile, Hollingworth (1998) considers whether a consistent and expected scene context can facilitate the perceptual processing of objects, implicating the role of environmental familiarity in visual cognition [67]. Together, these studies underscore the complexity of how visual context informs and influences the recognition and interpretation of objects in our surroundings. Additionally, Zhang et al. (2020) [68] investigated the effects of context on object recognition in humans using a behavioral task and computational modeling. Participants were presented with objects in various contexts, and it demonstrated that context-dependent modulation influences human behavior performance. Rajalingham (2015) [13] examined and compared object recognition behaviors in rhesus monkeys and humans, finding that both share a highly similar pattern of object confusion, suggesting a common neural shape representation that supports object perception. This resemblance extends beyond simple visual processing to advanced computational models, implying parallel visual recognition competencies between the species. Again, Rajalingham et al. [12] conducted a large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. The results of the study showed the performance of monkeys was like that of humans, although humans performed slightly better overall. This suggests that humans and monkeys have similar characteristics concerning visual object recognition. However, to our best estimation, there is more information needed on whether monkeys utilize contextual cues. In addition, it is not clear whether humans and monkeys have comparable characteristics in integrating contextual factors, as there is a deficiency of information on this topic. To gain

more insight on this, we propose the following:

2.2 Hypothesis

Our main questions are:

1. Does context affect human and monkey behavior at all?

We speculate that contextual information (information that is present in the environment surrounding an object) can impact the behavior of both humans and monkeys. This means that the presence or absence of certain cues or stimuli in the environment may influence how an individual acts or makes decisions.

H0 (null hypothesis): There is no effect of context on human and monkey behavior. This would mean that contextual information does not play a significant role in shaping the behavioral accuracies of primate object discrimination behavior.

H1 (alternative hypothesis): There is a significant and reliable effect of context on human and monkey behavior.

If H1 is supported by the data, it would indicate that primates can use contextual information to guide their behavior and make decisions. To test these hypotheses, we will conduct experiments in which humans and monkeys are presented with images containing different contextual variations, and their behavior is measured and compared across image variations. If the data supports H0, there would be no significant differences in behavior between individuals in different contextual conditions, then we can examine

the disparities in the collected data. However, if the data supports H1, there would be significant differences in behavior that can be attributed to the presence or absence of contextual cues.

2. Does the pattern of behavior under contextual influences align between human and monkeys?

We want to test if the pattern of behavior under contextual influences aligns between human and monkeys. We assume that there are cognitive mechanisms involved in context-dependent behavior that are shared between the two species. We will compare the contextual effects at two levels of granularity. First, we will pool all images that belong to the same context type and estimate the behavioral accuracies across the different contexts (context-level analysis). Second, we will compare the behavioral accuracies at the image-by-image level (image-level analysis).

H0 (null hypothesis): There is no significant similarity in the pattern of behavior exhibited by humans and monkeys when performing object identity discrimination tasks under different contextual modulations.

H1 (alternative hypothesis): There is a significant similarity in the pattern of behavior exhibited by humans and monkeys when performing object identity discrimination tasks under different contextual influences.

This implies that there are similarities in the cognitive mechanisms involved in context-dependent behavior between macaques and humans. If supported by the data, it would suggest that both species process and use contextual information in similar ways. The

hypothesis will be tested by measuring and comparing the behavioral responses between humans and monkeys to object identity discrimination tasks. If the patterns of behavior exhibited by the two species are found to be similar, it would support the hypothesis. It would suggest that non-human primates, such as macaques, are a valuable model for studying context-dependent behavior in humans.

2.3 Visual stimuli

We generated an image dataset from ten object categories: bear, elephant, person, cars, dogs, apples, chair, plane, bird, zebra. The total number of images is 600 in gray-scale, with 60 images from each object category. For each object category, we selected 6 natural images that were downloaded from Microsoft Common Objects in Context (COCO) [69] dataset, varying in object size and location, which were center cropped (if needed), converted to gray-scale, and finally re-scaled to 512x512 pixels. We then generated 10 different contextual variations for each image, including congruent, incongruent, no context, masked object, blurred context, blurred object, blurred incongruent boundary, minimal, textured, and jigsaw context. Examples of these have been illustrated in Figure 2.1. The changes were made using the object segmentation for each image obtained from the COCO object annotation masks and (for some conditions) replacing the background with different backgrounds based on the contextual manipulation conditions. COCO is a large-scale object detection, segmentation, and captioning dataset. These images are manipulated programmatically to generate images with contextual information changed. These multiple variations of the contextual information

further our understanding of what aspects of the object’s surroundings impact recognition.

2.3.1 Image details

The stimulus resolution is 512×512 pixels and the objects are presented approximately about 8 degrees of visual angle.

2.3.2 Contextual variations

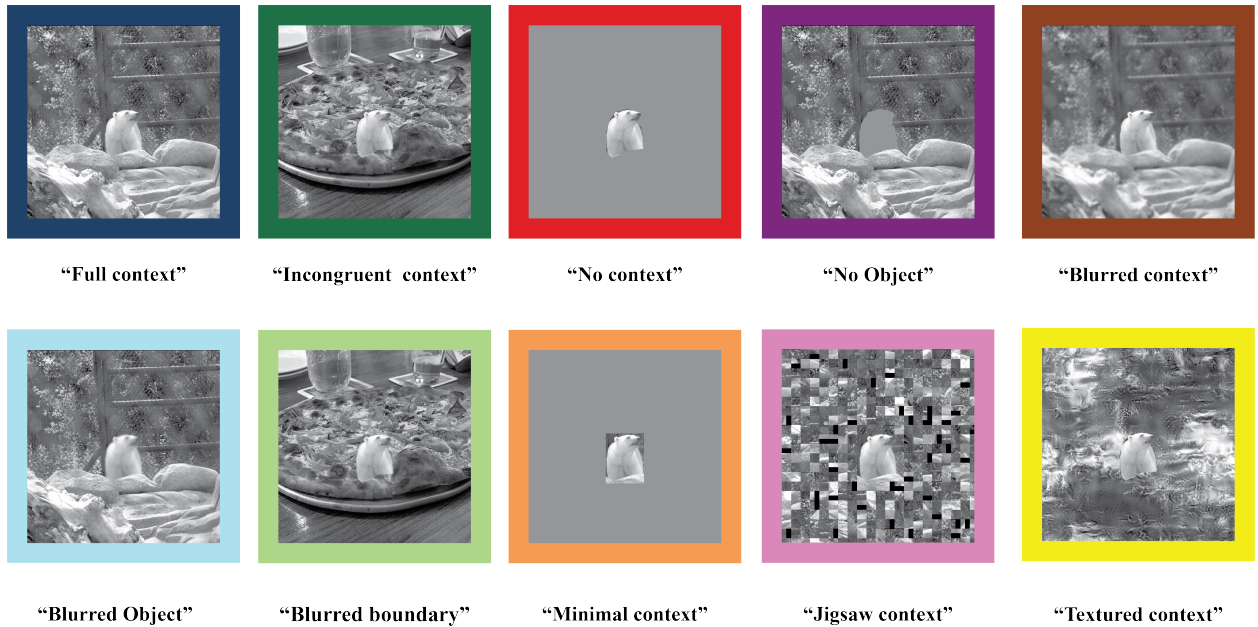


Figure 2.1: Figure showcases a series of images demonstrating the effects of different contextual manipulations on the perception of a single object used for the experiments. Through spatial transformations and texture alterations, this collection illustrates how visual context can dramatically alter the appearance and interpretation of an image. The frames around each image indicate the color associated with that context type (only used for reference in the thesis, not in the actual experiments).

Full context

This image type will have the object and its contextual information in full resolution with no manipulation to the scene context of the object (Natural image) serving as the reference image with the object in a congruent context (Color coded - Dark Blue).

Incongruent context

These images are where the target object is presented in a different (wrong) context (e.g., showing a car in the middle of the ocean) (Color coded - Dark Green).

No context

This variation includes no contextual information, as context is swapped with gray pixels and only “segmented object” information remains (Color coded - Red).

No object

This variation includes no object information, as the object is swapped with gray pixels and only contextual information remains (Color coded - Purple).

Blurred context

This variation has the object has the full resolution, whereas Gaussian blurred with kernel size 2 applied on the context (Color coded - Brown).

Blurred object

This variation has the context has the full resolution, whereas Gaussian blurred with kernel size 2 applied on the target object (Color coded - Light Blue).

Blurred object boundary(with incongruent context)

This variation has the Gaussian blur with kernel size 2 applied on the object context boundary and presented in an incongruent context (Color coded - Light Green).

Minimal context

This variation will have all context apart from the smallest bounding box around the object removed (Color-coded - Orange).

Jigsaw context

Contextual information (but not the target object) is chopped into a grid with dimensions of 25x25 pixels. The piece containing the target object remained in the same position as in the original image, and the other pieces were randomly shuffled around the object (Color coded - Pink).

Textured context

Context swapped with texture generated with Portilla and Simoncelli method [70] (5 iterations) on the baseline image. This contains the object with a background that matches the

surrounding information in terms of low-level texture statistics but lacks any high-level object information (Color coded - Yellow).

2.4 Methods

2.4.1 Behavioral task in humans

A total of 309 human subjects participated in the binary object discrimination tasks. The behavioral task in humans has been conducted using the online Amazon Mechanical Turk (MTurk) platform. MTurk is a web-based platform that enables researchers to collect large-scale psychophysical data from crowd-sourced 'human intelligence tasks' (HITs) in a cost-effective and efficient manner. The reliability of the online MTurk platform has been validated by comparing results obtained from online and in-lab psychophysical experiments. [13, 26]. Conducting the study online through the MTurk platform provides several other advantages. It eliminates the need for researchers to recruit and manage participants manually, which can be time-consuming and costly. The platform also provides a standardized interface for presenting the behavioral task to participants, ensuring consistency across participants. Furthermore, online experiments using MTurk can also allow for rapid and efficient data collection, reducing the time needed to conduct the study. This can be especially beneficial when conducting large-scale studies that require a significant amount of data to be collected.

Experimental design

The techniques and procedures employed by Kar et. al [71] are utilized for our experiments. Experiments are created using jsPsych. jsPsych is a JavaScript framework for creating behavioral experiments that run in a web browser.

Each trial starts with a blank gray screen with a fixation cross (0.2°) for 100 ms. Next, subjects are presented with a Test image (with varying context) including the target object for 100 ms, then the stimulus is followed by a blank gray screen for 100 ms, subjects are presented with two choices containing a standard image of the target object (the correct choice) and a standard image of the distractor object and they must indicate which one matches the object perceived in the Test image (<1500 ms). The human subjects indicate their choice by touching the screen or clicking the mouse over the target object while the monkey subjects will indicate their choice by touching the screen over the target object. Each subject will see an image only once. This is repeated for multiple test images with naturalistic context as well as images with contextual modifications. A pictorial description of it is shown in Figure 2.2 (information regarding images- section 2.3). (no. of images – 600)

2.4.2 Behavioral task in monkeys

A total of 2 male Macaque monkeys performed the behavioral tasks. The experiment was conducted in their home cages [72]. The experimental design is the same for monkeys as that for humans. We obtained a minimum of 32 trials per image from the pooled monkey responses.

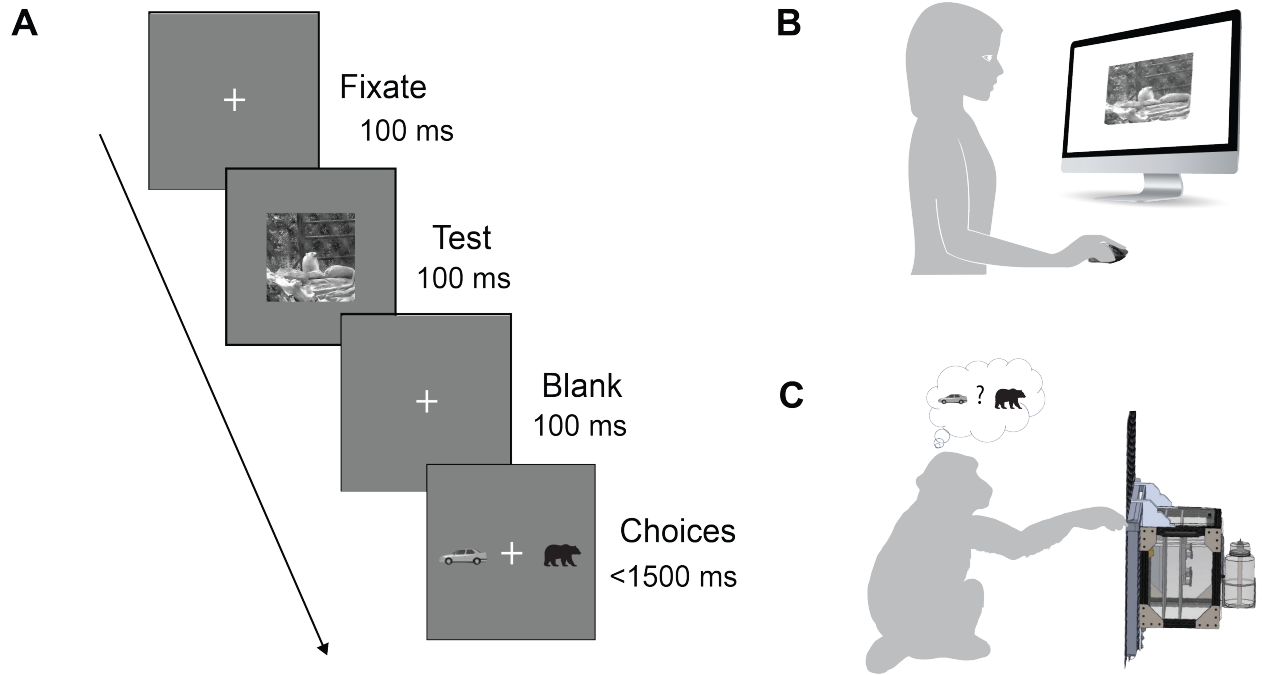


Figure 2.2: A. depicts the experimental setup for a binary object discrimination task, showing the sequence of events a subject undergoes in each trial. Subjects fixate on a cross, then the test image containing one of ten possible objects and contextual manipulations is shown at the center of the visual field (subtending 8 degrees of visual angle) for 100 ms. After a 100-ms delay, a canonical view of the target object (the same category as, but not a template match to, the test image) and a distractor object (one of the other nine objects) appears. B. Schematic illustrates a human participant engaged in the task via MTurk, for 309 participants recruited from Amazon MTurk. C. Schematic shows a monkey in a home cage environment [72], representing two context-trained adult macaques.

2.5 Behavioral Metrics for evaluating performance

To capture the multifaceted nature of object recognition within varied contexts, a comprehensive matrix has been developed, encapsulating the performance data across all experimental variables. This matrix, as depicted in Figure 2.3, serves as a foundational structure from which the six behavioral metrics are derived.

The matrix’s vertical axis categorizes the target object types, while the horizontal axis

represents the distractor object categories. Each cell within the matrix corresponds to a unique combination of target and distractor, filled with aggregated data from numerous trials per image per object category. To assess the impact of scene context on object recognition more comprehensively, we've developed six behavioral metrics that go beyond simple accuracy measurements across all images.

The first metric termed the behavioral signature at the context level (B.C1), focuses on human performance in recognizing objects within specific context categories (C). This approach offers an in-depth look at how context influences recognition performance in a categorical sense. The one-versus-all context-level performance metric (B.C1) estimates the discriminability of all images of context category c , essentially pooling the accuracies across all images of context type C and all object/distractor pairs within. Because we tested 10 context categories, the resulting B.C1 signature has 10 independent values.

The color gradient within the cells reflects the accuracy rate, with lighter shades indicating higher rates of correct recognition. This nuanced visualization allows for an immediate assessment of recognition patterns and anomalies at a glance.

The second metric, at the context level (B.C2), examines how well objects are recognized across different contexts, considering the performance through various distractor object categories. C2 thus captures the consistency of recognition across diverse contextual backdrops, revealing objects' context-dependent recognizability and robustness against confounding factors presented by distractors.

The third metric, the Behavioral Image-Level Metric (I1), zeroes in on the distinctiveness

of individual images. It measures the system’s ability to accurately identify each object (‘o’) in different contextual setups on a per-image basis. By doing so, I1 sheds light on the variability of performance at the individual image level, providing a more granular perspective on how well objects are recognized in various contexts in each image individually. The one-versus-all image-level performance metric (B.I1) estimates the discriminability of each image, pooling across all distractors. Because we have an image test set of 600 images (60 per object), the resulting B.I1 signature has 600 independent values. Given an image i of object category o , and all nine distractor objects (do), we computed the average performance image as:

$$I1_i^o = \left(\sum_{d=1}^{10} PC_i^{o,d \neq o} \right) \div 9$$

, where PC (percent correct) is the fraction of correct responses for the binary task between object categories o and d . Considering every image of context type c , the B.C1 performance for each context type is the mean across the performance of all images (60 per context type):

$$C1_c = \left(\sum_{i_c=1}^{60} I1_{i_c} \right) \div 60$$

Expanding our suite of metrics, the Image-Level Metric (I2) captures image-level performance across different object-distractor pairings. I2 quantifies the recognition difficulty within each image, considering the complexity and influence of the distractor categories present. This metric is pivotal in understanding how the composition of an image and the interplay of objects and distractors contribute to the challenges of object recognition.

To access these four vectors, we first compile data regarding all the tasks, having image performance from all individuals and all trials for all target object categories against every distractor category, the pictorial representation of this matrix is shown below.

2.6 Internal Behavioral Consistency

To evaluate the alignment between human and monkey visual systems to specific behavioral metrics, we adopted a 'consistency' metric similar to the 'human consistency' established in the literature [12]. This metric gauges the agreement between behavioral patterns by calculating a noise-corrected correlation between behavioral signatures derived from two distinct metrics.

To establish a benchmark for behavioral performance, we divided all accuracy trials per image into two halves, each representing independent samples. We then applied the behavioral metrics to these subsets to yield two separate portrayals of the system's performance, taking the mean across each split half and this is repeated 20 times with different trial splits. The degree of agreement between these portrayals is determined by calculating the Pearson correlation between them. The result is refined using the Spearman-Brown correlation formula to correct for any potential bias due to the split of data. This figure represents the self-consistency of the system, essentially assessing the trustworthiness of the behavioral signature in light of the volume of data collected. The ceiling estimates are the pooled human internal consistency, showing the full range of values (min-max).

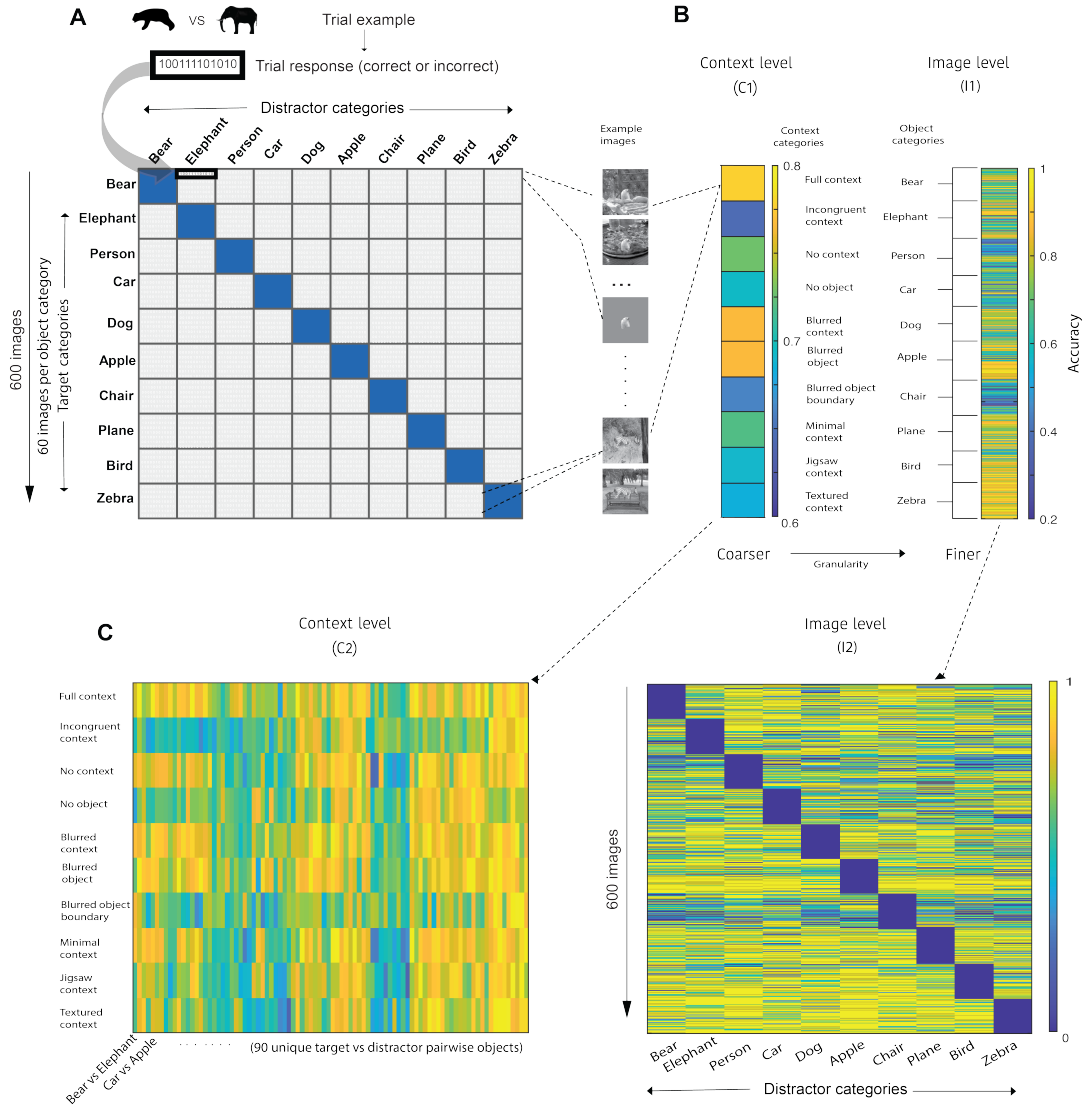


Figure 2.3: A. Schematic showing matrix containing all the image task responses for all the object categories against all distractor categories with an example of how trial responses are considered. Each matrix cell compiles trial data for specific target-distractor pairings, with the blue color diagonally indicating a non-applicable combination (car vs car). B. Schematic explaining how human performance is evaluated from task matrix, at context and image level. Metric C1 examines accuracy within particular context categories, collapsing data across target-distractor combinations to evaluate the influence of context on recognition. Image-Level Metric (I1) focuses on individual image performance, assessing the recognizability of objects in distinct contexts on a per-image basis showing granularity of evaluation from coarser context to finer image level metric. C. Metric C2 explores object recognition consistency across varying contexts, revealing how different distractors impact recognition. Lastly, Metric I2 quantifies the challenge of recognizing objects amidst various distractor influences, enhancing understanding of image composition's role in recognition difficulty.

2.7 Human-monkey shared behavioral variance

To quantify the behavioral pattern similarity at a context and image level across humans and monkeys, we calculated the percent of shared behavioral variance (SV) for both signatures. The SV is obtained as the square of the correlation (Pearson’s R) of the pooled humans and pooled monkeys’ behavioral signatures, corrected by the human and monkey signature internal consistency. This was repeated 20 times choosing 600 images with repetition(bootstrap).

2.8 Statistical Analyses

For each statistical analysis, we first tested the normality of the data. We used the Lilliefors test assuming a normal distribution, with a threshold 5% (normal distribution: $p > 0.05$). To test for statistical significance with a normal distribution of the data, a paired (monkey-human comparison) independent (comparing contextual variations) T-test was performed. This is a test for the null hypothesis that two samples have identical average (expected) values. The $t(\text{DOF})$ - statistic value quantifies the difference between the arithmetic means of the two samples. It is calculated as the mean of the difference of the two variables, divided by the standard error. The p-value quantifies the probability of observing as or more extreme values assuming the null hypothesis, that the samples are drawn from populations with the same population means, is true. A Wilcoxon signed rank (paired variables) or rank-sum (independent variables) was performed in case of a non-normal data distribution. The null hypothesis is that two (paired or independent respectively) samples come from the same

distribution. In particular, it tests whether the distribution of the differences $x - y$ is symmetric about zero. It is a non-parametric version of the T-test. We chose the threshold 5% ($p < 0.05$) to reject the null hypothesis for all tests.

2.9 Result

Building on these metrics, we aim to determine the extent of behavioral overlap between humans and monkeys. Our comparative analysis might reveal one of three scenarios. Firstly, due to species-specific differences, monkeys may process visual context differently from humans, resulting in no behavioral overlap. Alternatively, monkeys might share only a portion of behavioral variance with humans. Lastly, it's possible that in our chosen tasks, images, and contextual variations, monkey and human behaviors completely align. We can evaluate these scenarios using each behavioral metric, expecting that more granular metrics will better delineate the shared behavioral characteristics of these two systems.

Our initial analysis involved assessing the overall performance of both species in the binary discrimination task. This broad metric provides a baseline understanding of the proficiency with which each species can interpret and recognize objects overall.

The collective performance of our human subjects, represented by a composite figure derived from multiple trials across different images, stands at 71.35 %. This level of accuracy reflects humans' ability to integrate contextual information and discern objects within their environments, laying the groundwork for more detailed comparisons.

In parallel, The overall performance of monkeys was slightly lower, with an accuracy

rate of 69.16 %. This comparison sets a preliminary stage for understanding the relative capabilities of monkeys in similar object recognition tasks and provides a benchmark for interspecies comparisons.

Behavioral modulation showing significant and reliable impact of context on human behavior

Having established a foundational understanding of the overall performance in the binary discrimination task, we now turn to a slightly more detailed examination of the data. The initial findings indicate a strong ability in both species to utilize image information effectively to produce high performance. The next step is to dissect how information is perceived, at context and image level. By venturing beyond the surface-level performance metrics and into the depths of behavioral analysis, we stand to enhance our understanding of primate vision cognition significantly.

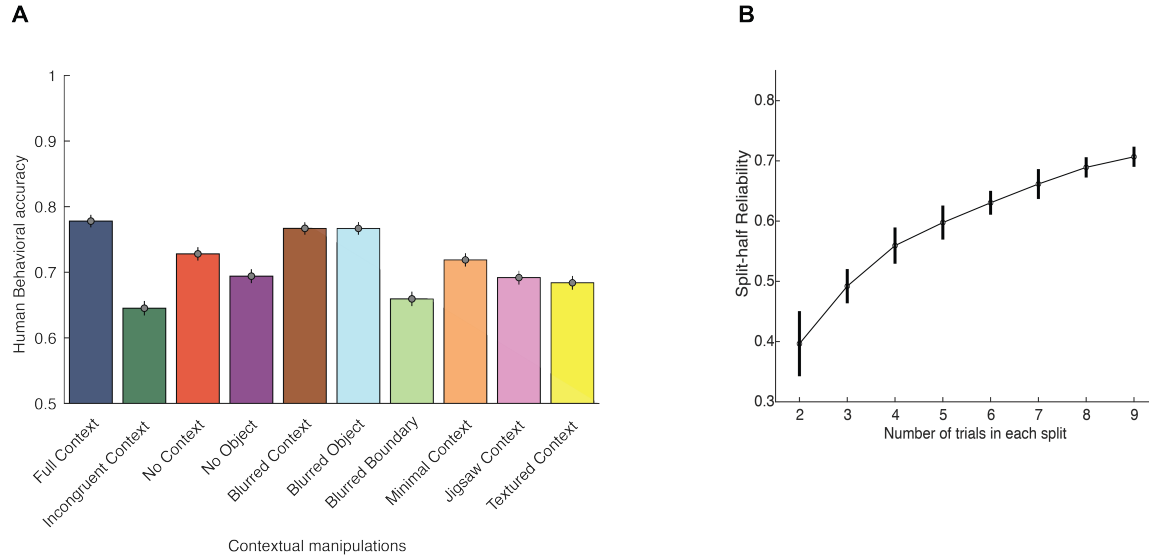


Figure 2.4: A. This figure illustrates the mean accuracy rates of human participants under varying contextual conditions. Each bar represents the average performance in recognizing objects with the contextual manipulation indicated by the x-axis labels. Error bars denote the standard error of the mean, providing a visual representation of the variability within each condition. B. This graph shows the split-half reliability scores for human participants across an increasing number of trials. The reliability scores are plotted on the y-axis, and the number of trials in each split is plotted on the x-axis. The error bars indicate the standard error of the reliability estimate for each number of trials, showing the precision of the estimate. The upward trend demonstrates that the reliability of the behavioral data increases with the number of trials, indicating high internal consistency within the data set.

But before dividing into inter-species comparisons, we must ascertain if contextual changes influence the object recognition capabilities of human participants. If behavioral changes are observed, this investigation will stage for subsequent comparative analyses with non-human primates, our foundational premise of AIM 1, primarily to ascertain the degree to which context-dependent visual processing is a shared trait across species.

To address this initial aspect of our hypothesis, we rigorously tested the behavioral accuracy of human subjects across a spectrum of contextual conditions. The alternate hypothesis posits that context plays a pivotal role in object recognition, and thus, any

alterations to the usual environmental backdrop of an object would significantly modulate the recognition accuracy of individuals.

The empirical data, as depicted in Figure 2.4, substantiates our hypothesis. Our results show that varying the context of the image changes the performance of the human participants. For instance, consistent with previous research [68, 54] humans show a significant reduction in accuracy for incongruent compared to congruent contexts (Accuracy = 0.13 ± 0.21 ; Lilliefors test: full context $p=0.004$, incongruent context $p=0.371$, non-normal distribution, $p>0.005$; Wilcoxon rank-sum test statistic=4.4, $p=0.0001$; Figure 2.4A: blue vs green bars). The images with 'Full Context' conditions correlate with the highest mean accuracy rates, thereby underscoring the hypothesis that humans leverage contextual information to facilitate object recognition. Conversely, the introduction of 'Incongruent Context' conditions – where objects are juxtaposed against non-typical environmental cues – shows a marked decline in performance. Even when the context/object boundary was blurred (termed blurred boundary), we observed the same effect. Predictably, removing the object, retaining only its silhouette, also led to reduced accuracy; however, performance remained well above chance, indicating that the context alone (with the object outline) provided enough information for accurate object discrimination. Moreover, when the context was removed or minimized, there was again a decrease in performance, confirming that humans also rely on the surroundings for object recognition. The blurring process itself seemed to have minimal influence on human responses. Using a synthesized texture (textured context), which retained the visual attributes of the original context, also adversely affected human behavior. This not only illustrates the

significant effect of context but also refutes the null hypothesis (H0) that context does not influence human behavior. The consistency of variability within each condition, as evidenced by the error bars, further reinforces the robustness of these findings, showing that the observed effect was not due to outliers or irregular data distribution.

These findings falsify the null hypothesis (H0) that context does not affect human behavior, providing strong evidence in support of the alternative hypothesis (H1) that context significantly influences human object recognition performance.

The reliability of these behavioral findings is further bolstered by the data's internal consistency, as demonstrated by the split-half reliability analysis shown in Figure 2.4. The analysis reveals a clear trend of increasing reliability with a greater number of trials. Such a positive relationship is indicative of the robust internal consistency within the dataset, as the number of trials in each split grows, so too does the consistency of the participants' responses.

This trend is especially notable when considering the lowest values of reliability associated with a smaller number of trials. As expected, a limited data set can result in greater variability due to chance. However, the graph illustrates that with an increasing number of trials, the reliability measures steadily climb, indicating a decrease in variability and an increase in the precision of measured behavioral responses. This point is further clarified by the narrowing of standard error bars as the number of trials rises, emphasizing the reliability of the data collected.

Variable Impact of Contextual Congruence on Human Object Recognition Behavior

In exploring the dynamics of object recognition, Figure 2.5 offers a nuanced exploration that illustrates the complex influence of contextual congruence on object recognition accuracy. While congruent contexts generally facilitate better recognition by providing supportive environmental cues aligned with the object's typical setting as seen above, our data reveal more nuanced interactions. Dual scatter plot analysis comparing recognition accuracies under full versus incongruent contexts, and full versus no context conditions, demonstrates that although full contexts typically enhance recognition, there are notable instances where incongruent or absent contexts lead to better performance. This suggests that under certain conditions, incongruent backgrounds or the absence of any background may highlight distinctive features of the object more effectively, focusing attention more sharply on the object itself or challenging typical perceptual expectations.

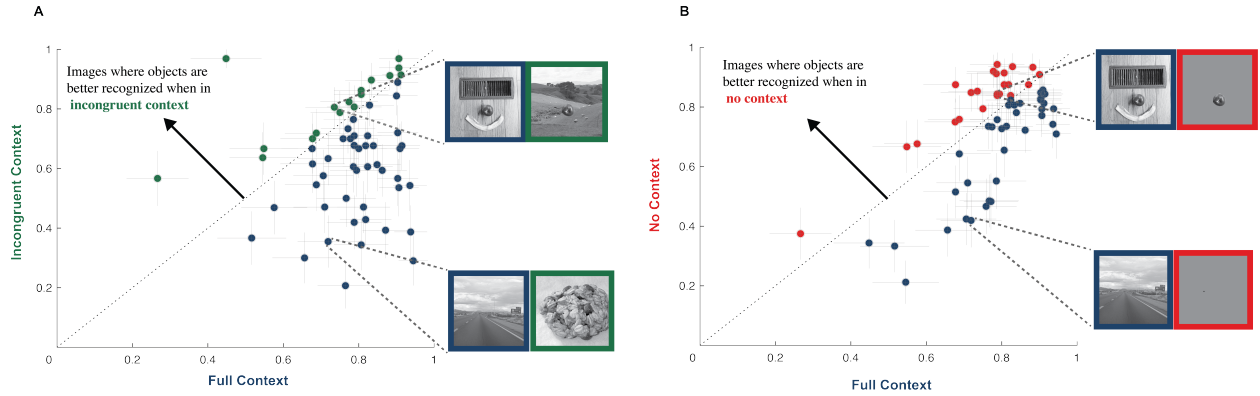


Figure 2.5: This scatter plot vividly illustrates image-level accuracies in object recognition tasks, comparing different contextual manipulations and their impact on human behavioral accuracy. Left: Each dot plots the accuracy of recognizing the same object when embedded in full versus incongruent contexts. The x-axis shows accuracies in full context, and the y-axis in incongruent context, with error bars representing the standard error across image trials to highlight variability in performance. Specific examples are included where objects demonstrate superior recognition in incongruent contexts compared to full, illustrating unexpected cognitive processing advantages under atypical contextual settings. Right: In a similar format, this graph compares accuracies in full context versus no context, revealing a generally lower performance when context is removed, yet some objects, like the very small car (<1 degree of visual angle), show significant challenges in recognition without contextual cues. Together, these comparisons shed light on the nuanced effects of context—both congruent and absent—on visual perception, challenging the blanket assumption that more context unilaterally enhances recognition accuracy.

This variable impact underscores the vital role of context as a dynamic component of visual perception—one that can either support or challenge the recognition process depending on the specific attributes and familiarity of the objects and their settings. Particularly, the difficulty in recognizing smaller objects like cars when deprived of contextual cues, underscores the complex and variable nature of context in visual perception, challenging the assumption that more context straightforwardly enhances recognition.

Behavioral modulation showing significant and reliable impact of context on monkey behavior

Having established that contextual changes do indeed affect human object recognition, we can now extend this line of inquiry to a comparative analysis with monkeys. By evaluating whether similar patterns emerge across species, we aim to uncover the extent of behavioral overlap between humans and monkeys. Subsequent experiments were conducted with monkeys to investigate the influence of context on non-human primates. The experimental setup mirrored that of the human trials, utilizing the same range of images with contextual variations. To ensure that macaques are familiar with the scene context per object category, they were first explicitly trained them with images in full context (from the Microsoft COCO dataset, 160 images per object for ten objects). Macaques showed robust cross-validated accuracy during such training (Figure 5.1). Once the monkeys ($n=2$) were fully trained (i.e., reached 80 % performance) in their home cages (see learning curve Figure 5.1), we measured their object discrimination performances with the same contextually manipulated images as humans.

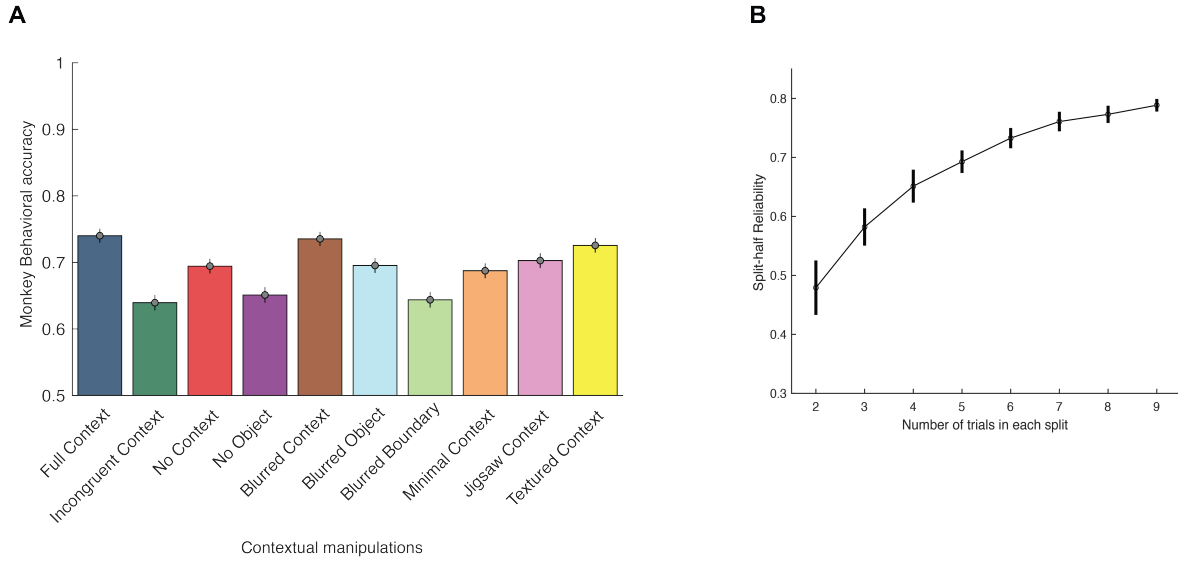


Figure 2.6: A. This figure illustrates the mean accuracy rates of monkeys under varying contextual conditions. Each bar represents the average performance in recognizing objects with the contextual manipulation indicated by the x-axis labels. Error bars denote the standard error of the mean, providing a visual representation of the variability within each condition. B. This graph shows the split-half reliability scores for monkeys across an increasing number of trials. The reliability scores are plotted on the y-axis, and the number of trials in each split is plotted on the x-axis. The error bars indicate the standard error of the reliability estimate for each number of trials, showing the precision of the estimate. The upward trend demonstrates that the reliability of the behavioral data increases with the number of trials, indicating high internal consistency within the data set.

Mirroring the human study, the accuracy of monkeys in object recognition tasks also exhibited variability contingent upon the contextual presentation of the images. The mean accuracies, presented in Figure 2.6, indicate a discernible impact of context on performance. Similar to humans, monkeys demonstrated heightened accuracy in 'Full Context' scenarios, which suggests a shared reliance on environmental cues between species for object recognition.

The 'Incongruent Context' presented a greater challenge, with a noticeable dip in accuracy. This pattern aligns with humans and the hypothesis that context exerts a significant influence on the perceptual processes involved in object discrimination tasks in primates. Similar to

humans, monkeys also showed a significant reduction in accuracy for incongruent compared to congruent contexts (Accuracy = 0.104 ± 0.18 ; Lilliefors test: full context $p=0.173$, incongruent context $p=0.58$, normal distribution, $p>0.005$; independent t-test, $t(59) = 3.305$, $p=0.001$, Figure 2.6A: blue vs green bars).

The split-half reliability analysis extended to the monkey data sets reinforces the behavioral trends observed. As depicted in Figure 2.6, a consistent increase in reliability with the number of trials affirms the robustness of the results. Despite the inter-species differences, the parallel in data consistency between humans and monkeys underscores a fundamental aspect of primate cognition concerning the utilization of contextual information.

The error bars on the reliability graph, denoting standard error, confirm the precision of these measurements across repeated trials. The implication is clear: the cognitive mechanisms engaged during the tasks are reliably accessed by monkeys, much like in humans, leading to consistent performance outcomes.

Variable Impact of Contextual Congruence on Monkey Object Recognition Behavior

In examining the dynamics of object recognition, Figure 2.7 illustrates the complex influence of contextual congruence on monkey object recognition accuracy. While congruent contexts are generally understood to facilitate better recognition as we see in our earlier analysis, this could be by providing supportive environmental cues that align with an object's typical setting, this analysis reveals more nuanced interactions. The scatter plot analysis comparing

recognition accuracies under full versus no context conditions demonstrates that although full contexts typically enhance recognition, there are notable instances where absent contexts lead to better performance. This observation suggests that under certain conditions, the absence of any background may highlight distinctive features of the object more effectively, focusing attention more sharply on the object itself or challenging typical perceptual expectations.

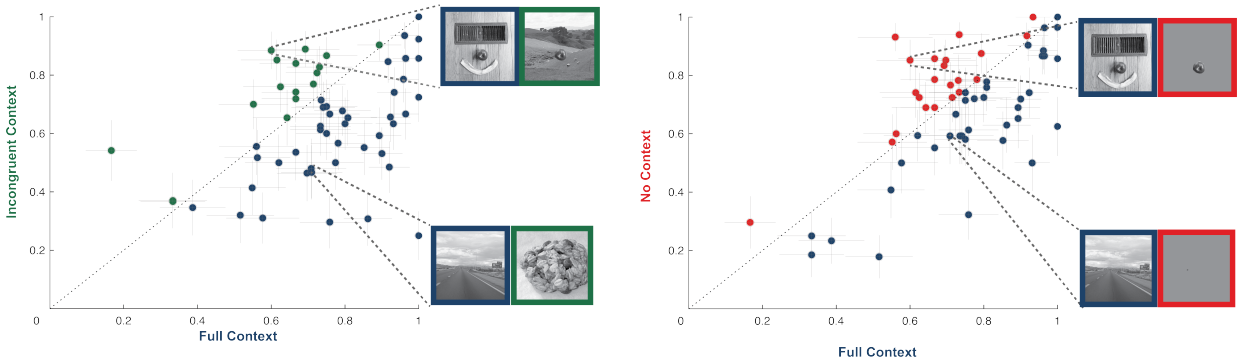


Figure 2.7: A. This scatter plot vividly illustrates image-level accuracies in object recognition tasks, comparing different contextual manipulations and their impact on monkey behavioral accuracy. Each dot plots the accuracy of recognizing the same object when embedded in full versus no contexts. The x-axis shows accuracies in full context, and the y-axis in incongruent context, with error bars representing the standard error across image trials to highlight variability in performance. Specific examples are included where objects demonstrate superior recognition in incongruent context compared to full, illustrating unexpected cognitive processing advantages under absurd contextual settings. B. In a similar format, this graph compares accuracies in full context versus no context, revealing a generally lower performance when context is removed, yet some objects show significant recognition performance without contextual cues.

This nuanced impact underscores the crucial role of context as a dynamic component of visual perception—one that can either support or challenge the recognition process depending on the specific attributes and familiarity of the objects and their settings. Particularly, the difficulty in recognizing smaller or less distinct objects when deprived of contextual cues underscores the complex and variable nature of context in visual perception, challenging the

assumption that more context straightforwardly enhances recognition.

High behavioral alignment between individual monkeys with respect to contextual modulations

The analysis of behavioral patterns within species is as crucial as cross-species comparisons for understanding the intricacies of cognitive function. To this end, we examined the alignment in behavior between two individual monkeys, referred to as Bento and Magneto, under varying contextual conditions.

At the context level (C1), the correlation between Monkey M and Monkey B was found to be high ($p = 0.98$), suggesting an almost perfect alignment in their responses to contextual cues. This remarkable congruence is illustrated in the left panel of Figure 2.8, where each data point corresponds to the average accuracy rate for each context category tested, plotted for both monkeys. The proximity of these points to the line of unity reflects their similar processing abilities within these contexts.

The observed disparity between context-level and image-level alignment may be attributed to individual differences in memory, and perceptual strategies that affect image processing. This intra-species variability is essential for understanding the range of cognitive abilities within a single species and provides insights into the adaptive nature of cognitive traits.

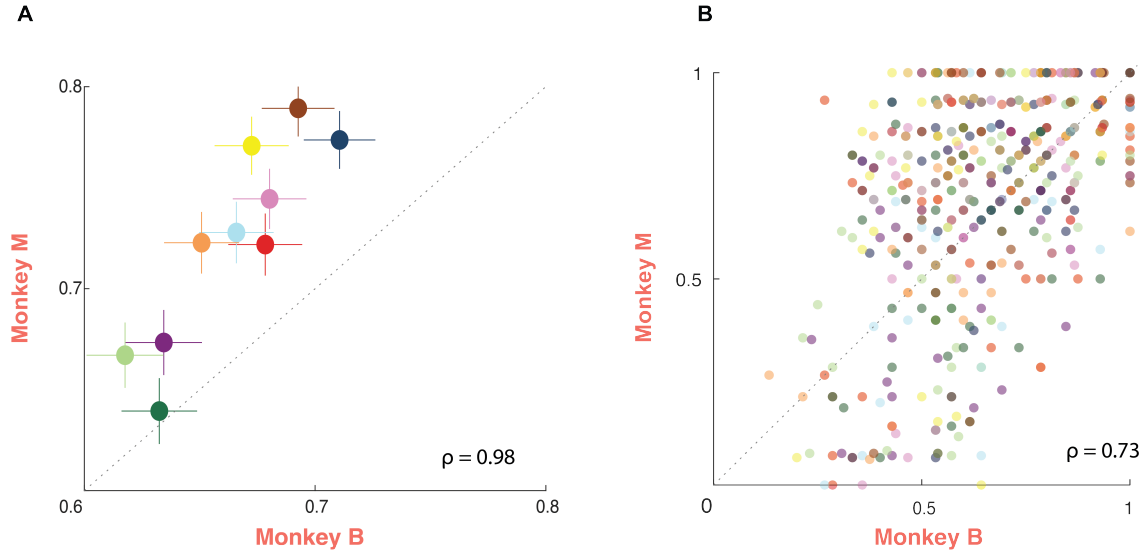


Figure 2.8: A. Scatter plot showing the correlation of accuracy rates at the context level between Monkey B and Monkey M, indicating a high alignment in their contextual processing abilities, where each point represents an average accuracy score within a specific context, with error bars denoting the standard error. The near-perfect correlation coefficient ($\rho = 0.98$) suggests a highly similar contextual processing ability between the monkeys. B. Scatter plot showing the correlation ($\rho = 0.73$) of accuracy rates at the image level between the two monkeys, with a moderate yet significant correlation, signifying individual differences in processing specific images.

Significant and reliable alignment in behavior between human and monkey under varying context

With the evidence demonstrating that both human and monkey subjects exhibit significant behavioral changes under varying contextual conditions, we advance to the next pivotal phase of our research – the comparative analysis to see if these behaviors are aligned between the two species.

We begin by taking the coarsest grain of comparison, which is at the C1 level. Upon comparing data at the context level(C1) between primates, we uncovered a strikingly high

correlation in their behavioral responses.

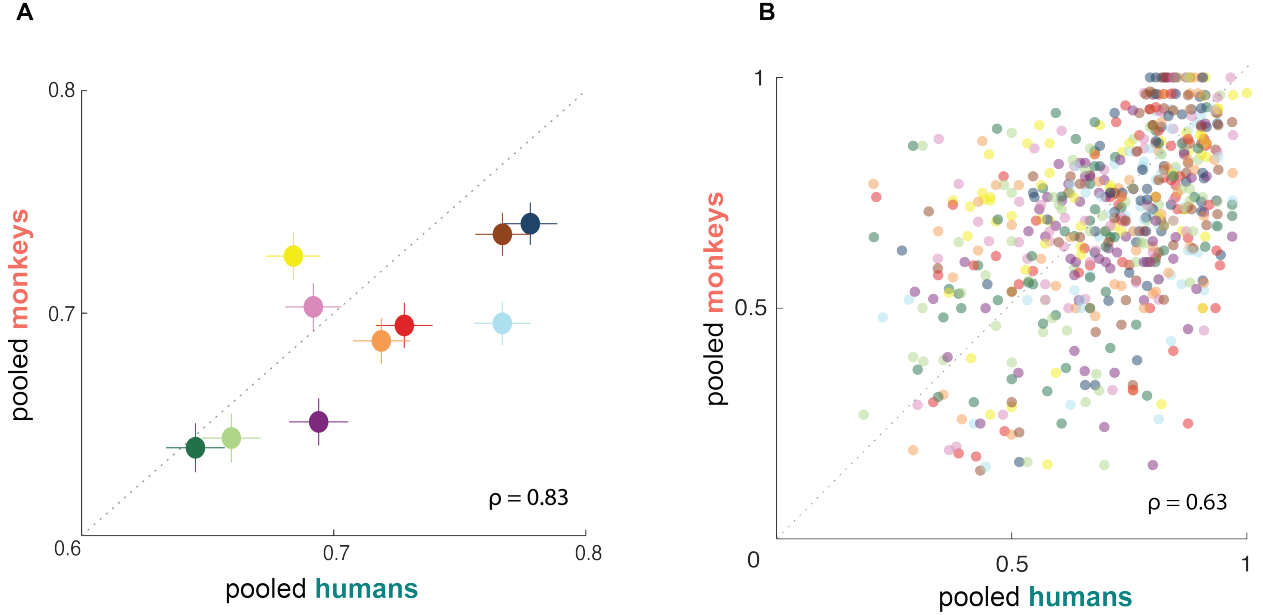


Figure 2.9: A. Scatter plot comparing the pooled mean accuracy rates at the context level between humans and monkeys. Each data point's color represents a different context category, with error bars indicating the standard error of the mean. The dotted line represents a perfect one-to-one correlation. The observed correlation coefficient of 0.83 underscores a high degree of behavioral concordance between the two species in context-dependent recognition tasks. B. Scatter plot illustrating the correlation of pooled mean accuracy rates at the image level between humans and monkeys. The data points represent individual images assessed during the study. The correlation coefficient of 0.63 indicates a moderate yet significant alignment in object recognition performance between humans and monkeys at the image level.

As illustrated in Figure 2.9, each point represents the mean accuracy rates pooled from humans and monkeys across the varying contexts we tested. Both monkeys and humans performed best in the full context condition (blue point) and worst in the incongruent context condition (green point). The proximity of these data points to the line of unity (dotted line) indicates a strong linear relationship between the two species' recognition capabilities within similar contextual frameworks. However, the majority of the points fall below the diagonal (human - monkey) = 0.02 ± 0.03 , Lilliefors $p = (0.517, 0.487)$: normal distribution; paired

t-test, $t(9) = 1.86$, $p=0.1$) indicating humans outperformed monkeys in most contextual conditions. The two exceptions were the jigsaw and textured context conditions, where monkeys slightly outperformed humans. The context level alignment (C1 corrected Pearson $R=0.83$) is indicative of a substantial alignment in the way both species process visual context.

Building upon the substantial correlation observed at the context level, we proceeded to examine the relationship between human and monkey behavioral responses at the image level. This granular analysis allowed us to explore the intricacies of object recognition across a myriad of specific images, providing a more detailed perspective on the perceptual alignment between the two species.

At this image-specific tier, our findings revealed a moderate correlation between the two species (I1 corrected Pearson $R=0.63$). The slope of the regression line in Figure 2.8B suggests that humans outperformed monkeys on average, but this difference was not as pronounced as at the context level ((human - monkey) $=0.02\pm0.18$, Lilliefors $p=(0.001, 0.001)$: non-normal distribution; Wilcoxon test: statistic = 79913.5, $p=0.02$). While this value denotes a significant positive relationship, it also highlights a greater degree of variability when compared to the context level. This distinction suggests that while the broader contextual processing shows a high degree of evolutionary conservation, the recognition of individual images may be influenced by more individual factors. The divergence in correlation coefficients between the context and image levels may be attributable to the specificity of the images themselves. Individual images contain unique features and combinations of elements that may be differentially salient or familiar to each species, thus affecting recognition accuracy.

Humans and monkeys share significant variance in context-driven changes in object recognition

To understand the variability in object recognition among humans, we calculated what we call the "human ceiling" by comparing the shared variance between two separate groups of human subjects. This is visually represented by the teal band in Figure 2.10. We then compared this human ceiling to the shared variance between humans and monkeys. Since we are comparing a large group of 309 humans to a smaller group of just 2 monkeys, we also considered how the size of the monkey group affects its consistency with human data.

The "pseudo" human consistency function is a method to estimate how shared variance between humans and monkeys would change with an infinitely large group of monkeys. Starting with 309 human subjects, we create multiple smaller subsamples, calculating the shared variance in object recognition tasks for each. These values are used to fit a sigmoid function, modeling how shared variance increases with larger human subsamples. Extrapolating this function to an infinite group provides an estimate of the maximum shared variance, termed the "human ceiling." Assuming monkeys follow a similar pattern, we use this function to predict the shared variance between humans and monkeys as the monkey group size increases.

When the number of monkeys increased from one to two, the shared variance between humans and monkeys increased by 4.3% (illustrated by the gray bars in Figure 2.10A). By extrapolating this trend to an infinitely large group of monkeys using a "pseudo" human consistency function (a sigmoid function derived from subsampling the human group), the behavioral results show that humans and macaques share a significant proportion of the

variance (62.43%) that is induced by context variations during object discrimination tasks. We also estimated that the shared variance between humans and monkeys would eventually reach approximately 80% of the human ceiling.

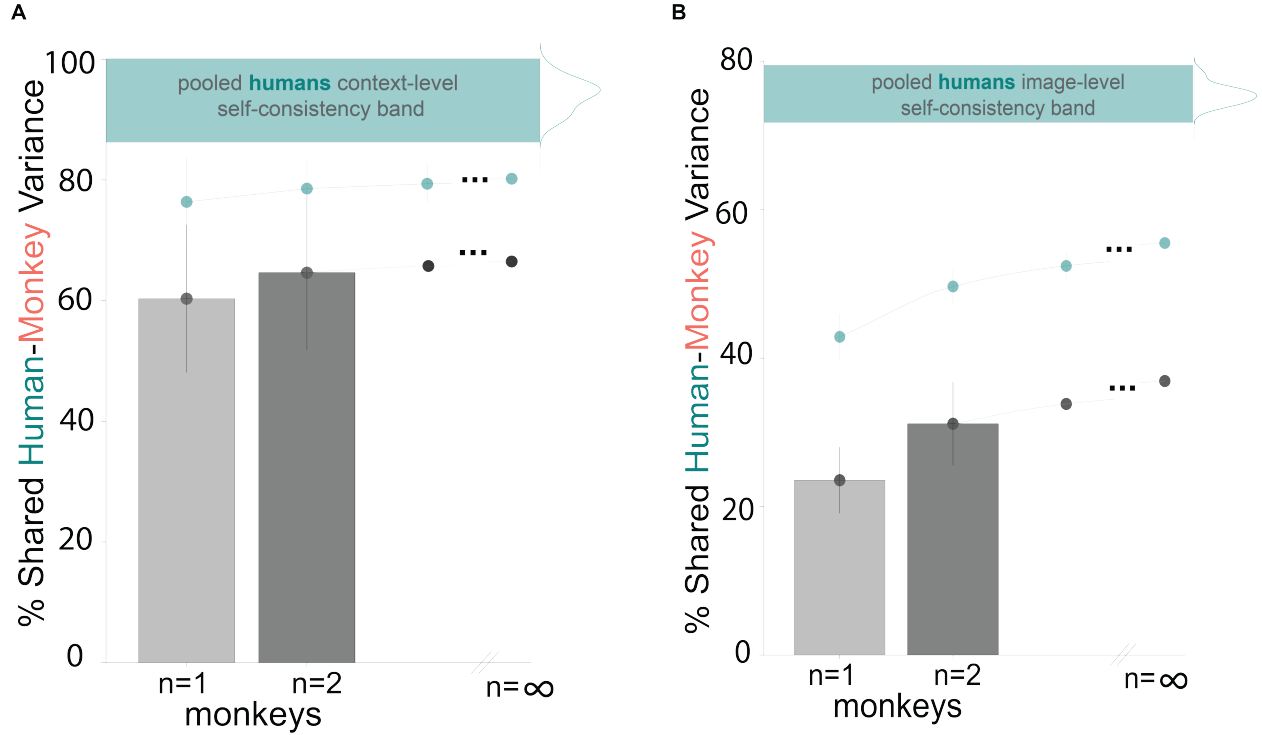


Figure 2.10: A. Shared human-monkey explained variance at a context level (mean with standard deviation across context types), as a function of the number of monkeys used for pooling. The asymptotic value for an infinite pool of monkeys is obtained by extrapolating the “pseudo” human consistency function. The human self-consistency ceiling is shown as a teal band. B. Similar to A but mean shared variance at an image level with standard deviation across image subsamples.

In addition to context-level analysis, we also examined the shared variance at the image level, as shown in Figure 2.10B. This analysis revealed that humans were less consistent with each other at the image level compared to the context level, which is expected due to the increased detail and variability of individual images. This effect was even more pronounced for monkeys, with a larger drop in shared variance at the image level. Increasing the number

of monkeys from one to two improved the shared variance between humans and monkeys by 8.2% at the image level (Figure 2.10B). Extrapolating to an infinite pool of monkeys, we estimated that the shared variance would eventually reach approximately 70% of the human ceiling at the image level. Our behavioral results demonstrate that humans and macaques share a proportion of variance (31%) induced by context variations during object discrimination.

Decline in Shared Variance between Humans and Monkeys from coarse to fine behavioral analysis

We next check the amount of behavioral variance shared between the two species at the 4 different behavioral metrics we defined earlier (Figure 2.3). The bar graph shown in Figure 2.10 visually quantifies the extent of shared behavioral variance between humans and monkeys, segmented into four specific conditions labeled C1, I1, C2, and I2. Each bar corresponds to a different condition and represents the mean percentage of shared variance between the two species, with error bars reflecting the variability of these measurements. The C1 and I1 conditions demonstrate a significant degree of shared variance, particularly in the C1 condition, where the shared variance peaks at 71.75%, indicating a potentially similar perceptual processing framework being engaged by both humans and monkeys in this context. In the I1 condition, the shared variance is 44.12%, still showing substantial overlap. At the C2 and I2 levels, the shared variance remains moderate but shows a decline, with values of 29.50% and 21.26%, respectively. This suggests that while there remains an alignment in

responses, the C2 and I2 conditions, which represent a more granular level of analysis, may engage different aspects of cognitive or perceptual processing that are less conserved between species. As our examination becomes more detailed, we begin to observe subtle differences, indicating that finer distinctions in task or stimulus processing may not be as conserved and shared across species.

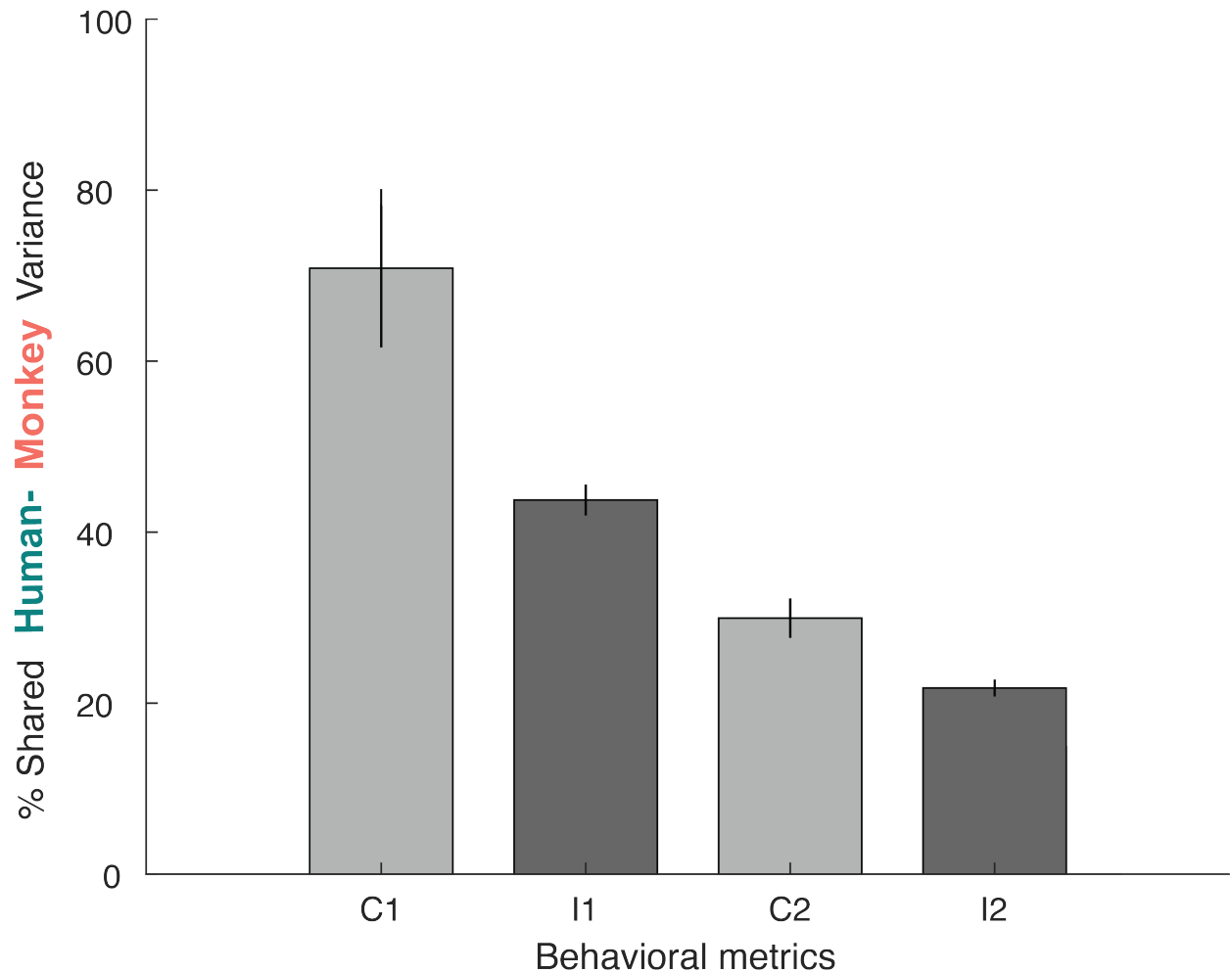


Figure 2.11: This bar graph illustrates the percentage of shared variance in behavioral metrics between humans and monkeys across four levels of analysis: C1, I1, C2, and I2. Each bar represents the mean percentage of shared variance, with error bars indicating the standard deviation across repetitions. Notably, condition C1 shows the highest shared variance, suggesting a strong similarity in behavioral responses between the two species at this level of analysis.

2.10 Potential Confounds and mitigation strategy

Other properties of images which are not specific to context can modulate behavior. These properties include the contrast, brightness of the image, and the sharpness of the image. All these properties can influence the way that people perceive, interpret, and respond to images, and can thus have an impact on behavior. In our study, each image underwent a thorough analysis to extract a range of fundamental image features. These features encompass several aspects, each offering unique insights into the image’s composition and characteristics. The following elements were meticulously evaluated: object size, object position (x and y position), contrast mean, contrast standard deviation, spectral mean, spectral standard deviation, pixel, object category, and context category.

Leveraging the COCO object annotations, we extracted the object size in degrees of visual angle. This measurement was determined as the fraction of the full image size (considering the entire image presented at 8 degrees) occupied by the smallest bounding square encompassing the object. Additionally, the object’s central position within the image was quantified through its x and y coordinates relative to the image boundaries. To understand the spectral properties of the images, we applied the Fast Fourier Transform (FFT). This transformation enabled us to delve into the spectral domain of each image, from which we calculated the spectral mean and standard deviation. This approach provides a deeper understanding of the image’s frequency components. For assessing image contrast, we employed the standard contrast metric tailored for grayscale images. This metric was derived from the highest and

lowest pixel values in the image. To further our analysis, we also computed the contrast standard deviation, which was obtained from the pixel-wise standard deviation across the grayscale image (Figure 2.11). This metric offers a quantitative measure of the variability in brightness and helps in understanding the visual complexity of the images. These low-level features were selected to capture essential properties of the images that might influence object recognition performance, such as an object’s saliency (contrast) and its placement within the scene (location). Our analysis showed that these low-level features do not account for the context-level shared behavioral variance apart from pixels.

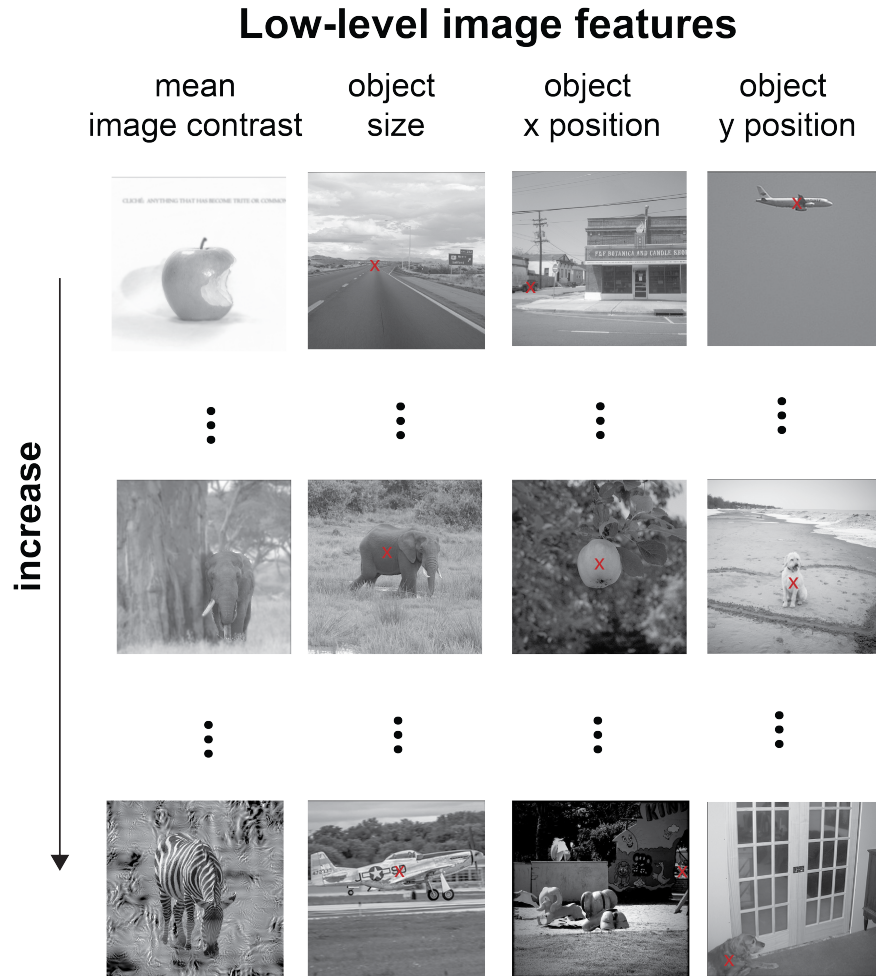


Figure 2.12: this visual array presents a systematic example of low-level image features examined for their effect on object recognition. The top row specifies the features examined, ranging from mean image contrast, object size, and object positions on the x and y axes. Vertically, the images demonstrate increasing complexity in these features. The first column shows a progression in mean image contrast, from high to low, affecting the visibility and perceptual salience of the objects. The second column alters the object size, illustrating how the object scale within an image impacts recognizability. The third and fourth columns manipulate the object's position horizontally (x-axis) and vertically (y-axis), showcasing the influence of spatial location on detection and identification. Red crosses mark the focal point of each feature's manipulation, serving as a reference for the changes applied. This collection underlines the significant role of basic visual properties in the cognitive process of recognizing and interpreting objects within a scene.

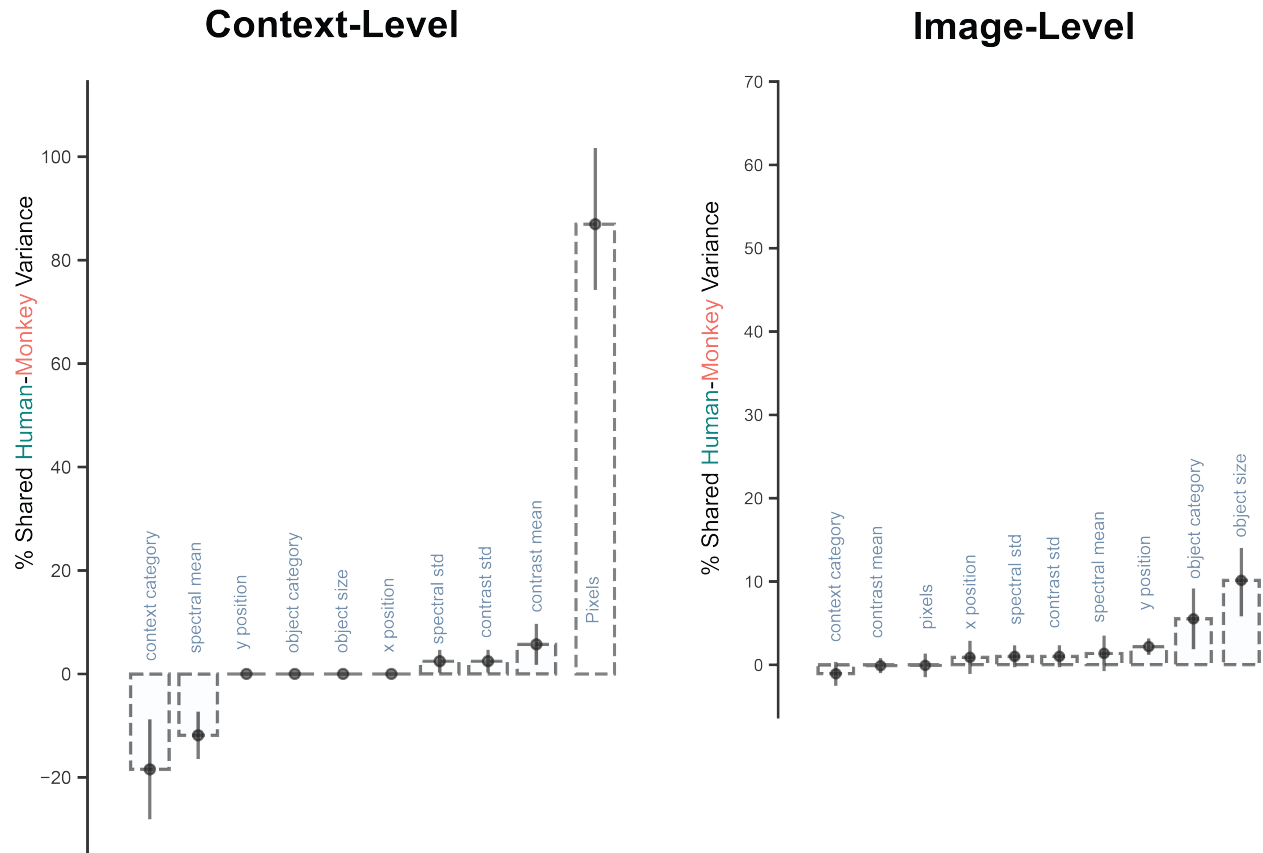


Figure 2.13: The graphs illustrate the predictive impact of various low-level image features on object recognition accuracy, separated into context-level and image-level analyses. On the left, context-level predictors such as category, spectral information, and spatial positioning are measured for their influence on recognition across a broad context. On the right, the image-level graph details the contribution of similar features to recognition accuracy within individual images. Bar heights represent the percentage of variance in recognition performance that each feature can explain, with higher bars indicating greater predictive value.

We utilized individual low-level features to make predictions for image-level performance using a linear regression model. We found predictions from each feature and obtained their correlation with primate behavioral accuracy on all images. The bar plot below reveals that the low-level features of images were not very effective in predicting behavioral performance. Specifically, we found that the correlation between low-level features and primate accuracy was low.

The paired graphs compare the ability of low-level image features to predict recognition accuracy at two distinct levels: context-level and image-level. The left graph (Context-Level) indicates that the 'pixels' feature has the highest predictive value for recognition accuracy across various contexts, which means that the pixel composition of an image strongly determines recognition performance in a broad context.

The right graph (Image-Level) reflects the predictive value of these features for recognition within individual images. Unlike the context level, 'object size' is seen with a prominent positive value, indicating its strong predictive power on the image level, suggesting that the size of the object within a specific image is a key factor in recognition accuracy by both monkeys and humans.

These findings suggest that while the low-level features may be informative to some extent, they are not reliable predictors of behavioral performance. This indicates the necessity for considering more complex, higher-order processing to comprehensively understand context-dependent object recognition performance.

Chapter 3

Aim 2: Analyze neuronal activity recorded across the IT cortex in context-naïve macaque monkeys to study contextual modulation effects during passive image observation task.

3.1 Background

In our first aim, we derived that contextual influences significantly impact object recognition behavioral outcomes observed in both humans and non-human primates. Over the last several decades, significant advancements have been made in the field of cognitive neuroscience, particularly in understanding the neural basis of object recognition. Central to this progress has been the delineation of the primate ventral visual pathway, which has been identified as the critical component for the neural circuits that underpin object recognition abilities in both humans and non-human primates [25, 26, 3]. The strategic use of rhesus macaques as a model system has been pivotal, not only because their object recognition behaviors

closely mirror those of humans [13, 12], but also due to the practical feasibility of conducting invasive studies that allow for detailed exploration of neural mechanisms [71, 73]. To further explore the underlying neural mechanisms, we propose employing a primate model, focusing on the inferior temporal (IT) cortex—a well-documented region in the brain integral to visual object recognition [26, 51]. Researchers have leveraged sophisticated neuroimaging and electrophysiological techniques to map the activity within the inferior temporal (IT) cortex, situated at the apex of the macaque ventral visual pathway. Findings from authors like Logothetis and colleagues (1995) [59] demonstrate through single-unit recordings that neurons within the IT cortex show selective responses to varied object categories. This suggests a role for the IT cortex in high-level visual processing, where it may be involved in distinguishing between different types of visual stimuli. Further research by Hung et al. in 2005 [25] expanded on this by showing that the IT cortex harbors neurons that respond to specific feature combinations, indicating its capacity for integrating diverse visual features to construct detailed object representations. Additional investigations have revealed that neurons in the IT cortex also specialize in processing images of faces, hands, and body parts, pointing to its versatility in handling complex visual information ([45, 74, 31, 75]. Another study investigated how neurons in the IT cortex of macaques encode visual stimuli during a discrimination task, analyzing the roles of firing rate and spike timing. They found that 94% to 99% of the stimulus information was encoded by the firing rate alone. The results suggest that visual information in this brain region is predominantly encoded through the rate of neuron firing [76]. A study examined how neurons in the macaque IT cortex respond to

images of various materials under different lighting and shapes. It found that these neurons are selective to material textures, largely independent of shape, but their responses vary significantly with illumination direction suggesting that texture recognition involves specific phase information as well as texture patterns [77]. Studies employing techniques such as single-unit recording and functional MRI have revealed that a linear combination of image-driven population activity in the IT cortex can effectively predict human object recognition behaviors across various tasks [56, 20]. Intriguingly, these neural responses are typically captured while monkeys passively view images, without active engagement or direct learning of the tasks. This passive viewing paradigm underscores the predominantly bottom-up nature of these neural representations, suggesting that the encoded information is inherently task-independent [56, 20, 31]. In a related line of research, studies comparing the processing capabilities of the human visual cortex and deep convolutional neural networks (DCNNs) during object categorization tasks have demonstrated similar findings. Both early stages of EEG activity in humans and early layers in DCNNs were predominantly responsive to the backgrounds of objects rather than the objects themselves. This suggests a fundamental aspect of visual processing where effectively categorizing objects is heavily dependent on the ability to distinguish them from their surrounding environments. This capability is crucial both for biological visual systems and artificial intelligence models designed to mimic these processes [78]. Despite these advancements, there remains a gap in our understanding of how contextual changes influence the responses of the IT cortex and whether there is a direct correlation between the patterns of activity in IT neuron populations and the behavioral

outputs reported in primates. Our proposed study aims to fill these gaps by leveraging neural recording and decoding techniques to provide deeper insights into the complex dynamics of visual perception and object recognition in primate models.

3.2 Hypothesis

Our main question is:

1. Is there a significant relationship between the neural activity patterns in the IT cortex of context-naive monkeys and the shared behavioral variance observed between humans and monkeys when responding to test images under various contextual influences?

Our goal is to determine the degree to which image-driven responses in the Inferotemporal (IT) cortex of context-naive macaques can explain the observed shared variance in behavior between humans and monkeys. We aim to explore the correlation between the population activity in the IT cortex of context-naive monkeys and the behavioral patterns shared variance of human- monkeys in object recognition tasks influenced by contextual cues. Our hypothesis suggests that the neural responses in the IT cortex when decoded, will reflect the extent of shared behavioral variance seen between humans and monkeys. To test this, we will expose context-naive monkeys to the same set of images as in previous experiments, recording their IT cortical activity. This neural data will then be used to make predictions for the images which will then be compared to

the shared primate variance. Ultimately, our analysis will focus on quantifying how much of the variance between human and monkey behavior is accounted for by the image-driven neural responses in the IT cortex of context-naive macaques.

H0 (null hypothesis): The population activity in the Inferotemporal (IT) cortex of context-naive monkeys explains none of the behavioral variance observed between humans and monkeys, possibly suggesting that prior contextual cues have a significant influence on primate object discrimination behavior.

Our null hypothesis (H0) posits that the population-level neural activity in the Inferotemporal Cortex (ITC) of context-naive monkeys will not contribute to explaining the shared behavioral variance observed in humans and monkeys during object recognition tasks that involve contextual cues. This hypothesis suggests that the neural mechanisms responsible for context-dependent visual processing may not be entirely discernable or active in the ITC of monkeys without prior exposure to contextual information. Alternatively, it may imply the involvement of additional brain regions, distinct from the IT cortex, that are responsible for processing contextual information.

H1 (alternative hypothesis): The population activity in the Inferotemporal (IT) cortex of context-naive monkeys explains some of the behavioral variance observed between humans and monkeys.

This hypothesis suggests that the population activity in the Inferotemporal (IT) cortex of context-naive monkeys contributes to explaining a portion of the behavioral variance observed between humans and monkeys. It may suggest that while contextual cues might

have some influence, they are not the primary drivers of primate object discrimination behavior. This would suggest that inherent capabilities of the IT cortex might enable effective object discrimination, independent of contextual learning. It would suggest that other factors, possibly including innate neural processing, or another part of the brain could also contribute to the observed behavioral patterns.

H2 (alternative hypothesis): The population activity in the Inferotemporal (IT) cortex of context-naïve monkeys explains all the behavioral variance observed between humans and monkeys.

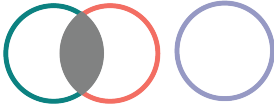
This hypothesis asserts that the population activity in the IT cortex of context-naïve monkeys accounts for all the shared behavioral variance observed between humans and monkeys. This hypothesis would imply that the neural response in the native monkeys' IT cortex determines how primates, both human and non-human, react to object recognition tasks without any relevance to prior contextual influences. It would suggest that the innate neural mechanisms in the IT cortex are entirely responsible for the observed behaviors, overriding any potential impact of contextual information or learning.

3.3 Visual stimuli

The images presented during the behavioral task are the ones employed in this part of the study.

How much of the shared **human-monkey** variance is explained by the **IT population**?

H0: **None** of the shared variance is explained



H1: **Some** of the shared variance is explained



H2: **All** of the shared variance is explained



Figure 3.1: This figure illustrates three hypotheses (H0, H1, H2) regarding the overlap in shared variance accounted for by human and monkey visual processing systems with that explained by the naive IT neuron population. H0 posits that the IT neuron population does not explain any of the shared variance, H1 suggests that the IT neuron population accounts for some of the shared variance, and H2 proposes that the IT neuron population explains all of the shared variance between humans and monkeys visual processing.

3.4 Methods

3.4.1 Experimental design - Passive Fixation Task

Large-scale recording of neural activity from IT cortex of two male contextually naive macaque monkeys using multi-electrodes arrays (Utah arrays; Blackrock microsystems Inc.) while they passively view test images from Aim1 (Figure 3.2). For this study, each neural array was equipped with 96 electrodes, configured in a grid layout with 400 μm between electrodes, covering a 4mm x 4mm area. The placement of these arrays enabled the sampling of neural activity across various regions of the inferotemporal (IT) cortex, particularly along its posterior-anterior axis. In our analysis, we did not consider the specific spatial location of the sites, considering each as an independent sample from the aggregate IT cortex population. The experimental design involved a passive viewing task, wherein the subjects (monkeys) were not required to engage in any active task but simply to view visual stimuli, specifically

test images, while their neural responses were recorded. To initiate a trial, the subject needed to maintain focus on a fixation point, represented by a cross spanning 0.2° of visual angle, for 300 milliseconds. Subsequently, a series of 5 to 10 images was displayed sequentially; each image was ON for 100 milliseconds and was interspersed with 100 milliseconds of a gray (background, 'OFF') screen. Following this sequence, a water reward was administered, succeeded by a 500-millisecond inter-trial interval before commencing the next image sequence. The animals ($n = 2$, male rhesus macaques) used in the passive fixation experiments study can be classified as “categorization task naive” since they have not been explicitly trained to perform any object categorization tasks. We combined the most reliable neural sites ($n=122$, out of the 122 sites - 30 sites from monkey 1 and 92 sites from monkey 2) across the two monkeys to generate a pooled neural population for further analysis. All procedures involving surgical and animal care adhered strictly to the guidelines set forth by the Canadian Council on Animal Care (CCAC).

3.4.2 Eye Tracking

Eye calibration and eye tracking of macaque subjects were conducted utilizing established operant conditioning techniques, incorporating fluid rewards, head fixation, and the employment of real-time video-based eye tracking for eye movement monitoring. Eye movements were precisely tracked using the TRACKPixx system by VPixx Technologies for monkey 1 and SR Research EyeLink 1000 for monkey 2. The two macaques underwent training to maintain their gaze on a centrally positioned fixation cross, measuring 0.2° in visual

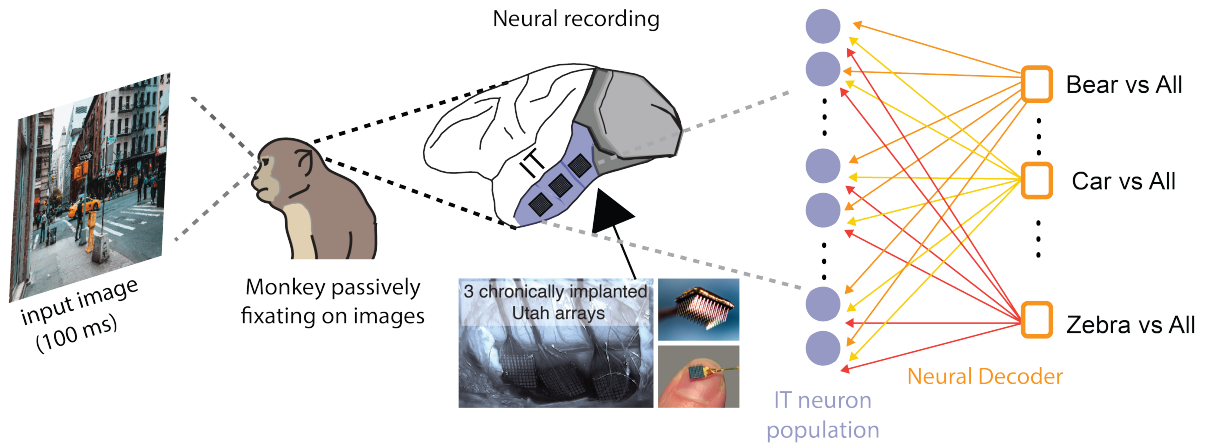


Figure 3.2: The image illustrates the process of recording and decoding neural responses to visual stimuli. In the depicted experiment, a monkey passively views a series of images, each presented for 100 milliseconds. The images are centrally located within the monkey's gaze. Neural activity is captured via electrodes in chronically implanted Utah arrays within the monkey's brain. The recorded neural data is analyzed using a machine learning approach—a cross-validated one-vs-all classifier that distinguishes between different object categories. This classifier operates on a "one vs. all" strategy for each object category and is rigorously cross-validated. The classification is based on neural activity patterns across varying time intervals from the onset of the image presentation. The mean evoked firing rate is an entry in a response vector (green vertical vector, green saturation is proportional to response magnitude) that summarizes the population response per image. For the decoding process, the classifier predicts the probability that a certain object category is present in the image. It identifies the category with the highest probability, distinguishing it from all other categories. Importantly, the behavioral predictions made by the neural decoder are based on images that are completely novel to the model, ensuring that no image of the same object—irrespective of contextual changes was used in both the training and testing phases.

angle. Each neural recording session commenced with an eye-tracking calibration exercise, wherein the subjects were required to execute saccades towards various spatial targets and sustain their gaze for a duration of 500 milliseconds. Calibration procedures were periodically revisited throughout the session to correct any detected drift. The administration of stimuli and the regulation of rewards were adeptly handled via MATLAB’s PsychToolbox Software. Real-time eye-tracking was employed to ensure that eye jitter did not exceed $\pm 2^\circ$, otherwise, the trial was aborted. Stimulus display and reward control were managed using the MWorks Software (<https://mworks.github.io>).

3.4.3 Macaque Neural Data

The neural data was gathered while the macaque was engaged in a passive fixation exercise, utilizing a dataset of 600 images that were contextually varied and displayed 31 times each from 122 neural sites. The neural activity, band-pass filtered between 0.1 Hz and 10 kHz, was continuously captured at a 20 kHz sampling rate using Intan Recording Controllers (Intan Technologies, LLC). The data discussed herein reflect multiunit activity measurements. The recorded raw data was preprocessed to extract firing rates, which were then organized into a matrix where data was organized with trial information for all neural sites, images, and timebin activity. Each neural site’s response to an image was taken as the mean rate during a time window of 100-170ms following image onset, a window that has been chosen based on the neuronal population’s self-consistency and aligns with the visually-driven latency of IT neurons [26]. We chose the most stable responses (self-consistency greater than 0.1) from

the combined arrays data matrix, which was used as a threshold for reliability. This process yielded a selection of 122 of the most reliable sites within the IT cortex population. The neural data were collected over multiple days, which is essential for accumulating sufficient data to achieve reliable conclusions. There is challenge of collecting data over multiple days recording arrays are fixed in the brain tissue, they do not guarantee that the same neurons are recorded across all days. However, the absolute stability of recording the same neurons was not necessary because the analyses assumed randomly selected samples of IT neurons. We applied a specific normalization procedure to address the day-to-day variability in the neural recordings. This procedure involved two main steps. First, we subtracted the background neural response—the average response to blank images—from the recorded response for each neural site. This subtraction was crucial to eliminate baseline neural activity that was not related to the stimulus, ensuring that subsequent analyses would reflect true stimulus-driven activity. Second, we normalized these adjusted responses by dividing them by the standard deviation of all responses recorded at that site on that day. This normalization by sample standard deviation helped to standardize the response data across different recording sessions, making them less sensitive to daily variations in recording conditions such as electrode placement or neural responsiveness. This methodological approach was designed to minimize the impact of day-to-day fluctuations on the analysis, enabling more reliable comparisons of neural activity over time [26]. To the reliable sites, we applied an l_2 -regularized decoder, which allowed us to generate test prediction probabilities for each image. We implemented an l_2 -regularized logistic regression model to decode patterns in the neural activity. The trained

decoder was applied to the test set to generate prediction probabilities for each image. These prediction probabilities indicate the likelihood that the observed neural activity corresponds to each object class. The object with the highest probability is then chosen as the predicted class for that image. These test prediction probabilities were used to compute performance metrics, such as accuracy, which quantify the decoder’s ability to correctly identify the images based on neural activity.

3.4.4 Partial correlation analysis

To estimate the fraction of shared human-monkey variance that is explained by the neural sites, we calculated the partial correlation for the pooled human and monkey population - while controlling individually for neurons. The partial correlation gives us the fraction of the primate shared variance that is independent of the neural variance. The percentage of shared human-monkey variance explained by neurons is then given by the formula:

$$\frac{R^2 - R_p^2}{R^2},$$

where R^2 is the human-monkey Pearson correlation, R_p is the human-monkey partial correlation, while controlling for neurons. The neural correction of the partial correlation is done by fitting a sigmoid extrapolation to an infinite number of neural trials. Both R^2 and R_p are corrected by the (Spearman-Brown corrected) human and monkey split-half reliabilities $\sqrt{\rho_h \cdot \rho_m}$, where ρ_h is the human split-half reliability and ρ_m is the monkey split-half reliability. However, due to the normalization by R^2 , we do not need to account for the

human or monkey noise.

3.5 Results

3.5.1 Population Activity of Neurons: Context-Dependent Visual Processing, Dynamic Firing Rate Patterns, and Neural Decoding

This study investigates the dynamic neural responses to various visual contexts to understand how the brain processes diverse stimuli. We aim to explore how different contextual presentations—ranging from full and incongruent contexts to various blurred and minimalistic scenes— affect the firing rates of neurons. Figure 3.3 illustrates the variability in neuronal firing rates across different visual contexts. Each colored line in the graph represents the mean neuronal response to a specific type of visual stimulus, showing how the brain’s activity patterns shift over milliseconds after an image is presented. The lines reveal distinct peaks and troughs, suggesting rapid adaptation and processing by the visual cortex in response to changes in visual context.

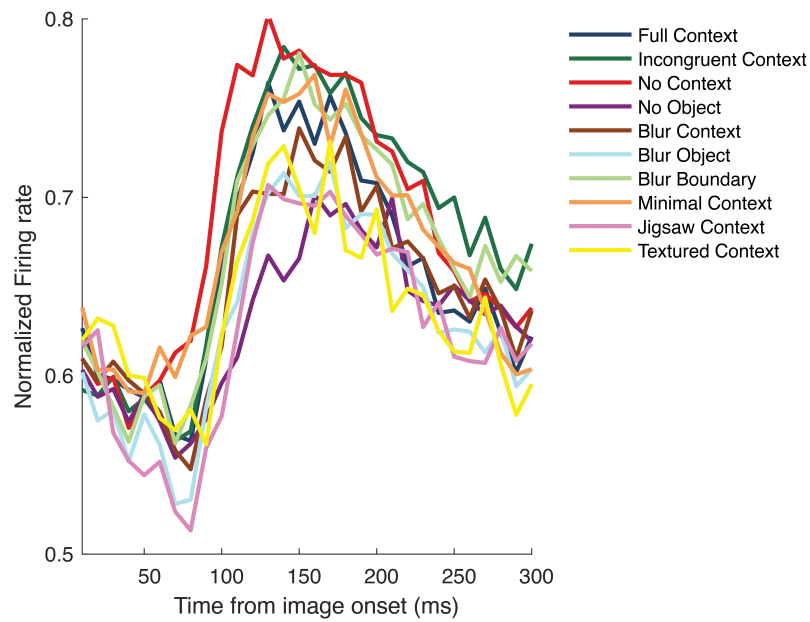


Figure 3.3: This graph depicts the normalized firing rates of neurons over time (milliseconds) from the onset of the image presentation. Each colored line represents the mean neuronal response to a different visual context, showcasing the dynamic patterns of neural activity associated with the processing of diverse visual stimuli.

However, examining these firing rates alone, it is challenging to decipher the underlying mechanisms of cognition and the specific context influencing these neural responses. This necessitates the use of decoding methods to interpret the data more accurately.

Our analysis employs a logistic regression model with ridge regularization to decode patterns in neural activity linked to image recognition. We used 5-fold cross-validation on a dataset of 600 images, which ensures our model’s robustness and its capability to generalize beyond the training data. Feature standardization was achieved through z-score normalization of neural firing rates, providing a uniform scale for model input and enabling an impartial evaluation of the model’s predictive accuracy.

Advancing beyond the scope of conventional correlation studies between human and monkey behavioral metrics, we introduced a neural image metric (Neural I1). This metric is pivotal in determining its modulatory influence on the correlational relationship between human and monkey responses over a spectrum of time frames. By calculating partial correlations for an array of starting times and bin durations, our analysis provides a granular perspective on the interaction between Neural I1 and the behavioral indices of both species.

The scatter plot in Figure 3.4 illustrates the visually encoded complex data using color and size gradients to represent various analytical and temporal dimensions. This visualization strategy not only highlights the temporal specificity of neural influence on behavior. Our refined approach successfully elucidates the portion of shared variance in primate behavior that can be attributed to neural activity, offering enhanced insights into the cognitive and neural congruences between humans and monkeys. From this analysis, we have honed in on

the meantime window of 100-170ms post-image onset, identified as the most reliable time-bin for capturing neural responses as within this critical window, we found neural sites exhibiting a self-consistency greater than 0.1, which led to the identification of 122 reliable IT cortex sites from the pooled neural array that will form the foundation for our further calculations and inferences.

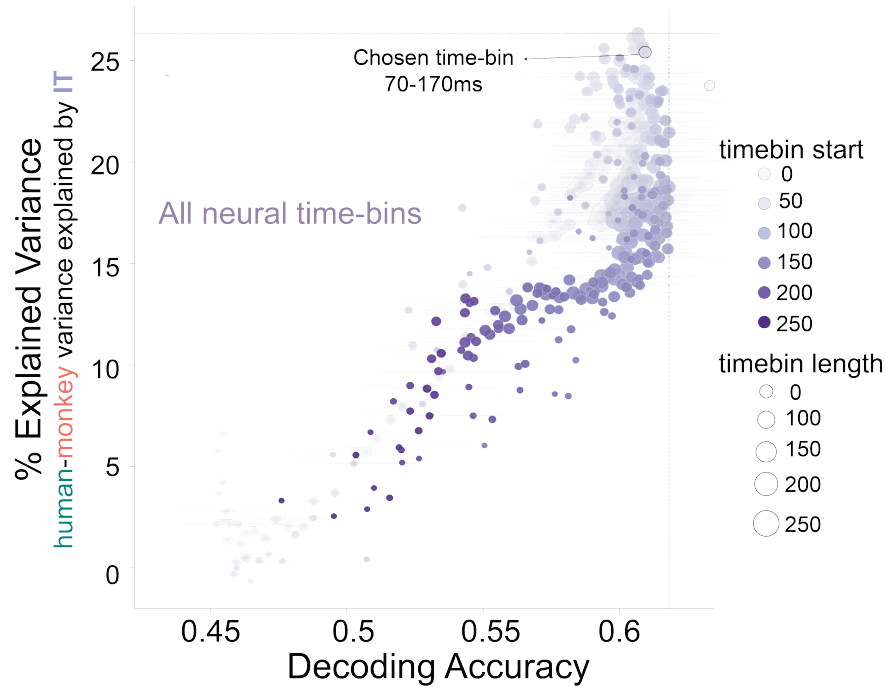


Figure 3.4: This scatter plot maps the mean decoding accuracy of all time-bins((filtered with self-consistency >0.1) from the neural data on the x-axis, against the percentage of human-monkey shared variance explained by IT cortex neural responses at image level in a macaque during a passive fixation task on the y axis with standard error across images. Each point represents a neural time-bin, with the size of the point correlating to the time-bin length and the color gradient indicating the time-bin starts from the onset of the image stimulus. The most reliable responses are marked and concentrated around the average interval of 100-170 milliseconds post-image onset, identifying the most dependable time bin for analysis.

3.5.2 Modulation in the pattern of neural prediction performance activity showing the impact of context

We generated a neural C1 metric (similar to the behavioral C1 metric; see Chapter 2, section 2.3) from the test prediction probabilities. The graph provides a clear visualization of how contextual manipulations impact neural prediction accuracy.

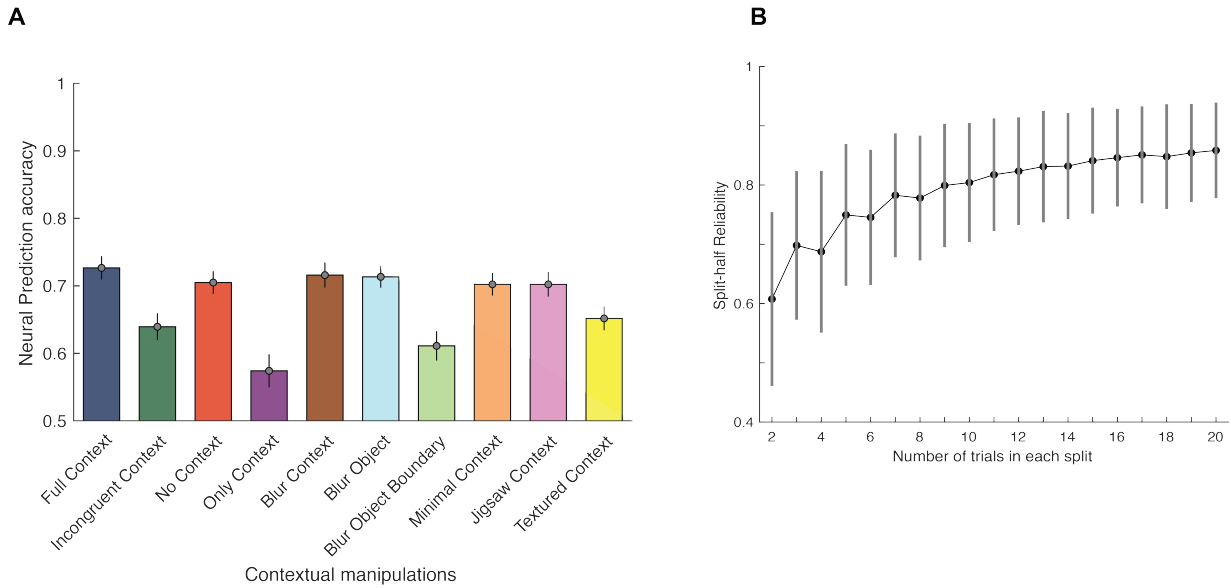


Figure 3.5: A: Each bar represents the average performance in recognizing objects with the contextual manipulation indicated on the x-axis. Error bars denote the standard error of the mean, providing a visual representation of the variability within each condition. B: This graph shows the split-half reliability for neural sites on an increasing number of trials. The reliability is plotted on the y-axis, and the number of trials in each split is plotted on the x-axis. The error bars indicate the standard error of the reliability estimate for each number of trials, showing the precision of the estimate. The upward trend demonstrates that the reliability of the neural data increases with the number of trials, indicating high internal consistency within the data set.

As illustrated, the highest prediction accuracy is observed with full context, indicating optimal neural decoding when complete contextual information is available. A slight decrease

in accuracy is noted with incongruent context, suggesting that mismatched context slightly impairs performance. (Accuracy = 0.09 ± 0.14 ; Lilliefors test: full context $p=0.002$, incongruent context $p=0.032$, non-normal distribution, $p>0.005$; Wilcoxon rank-sum test statistic=3078, $p=0.0038$; Figure 3.5A: blue vs green bars). Notably, the "No Context" condition maintains relative accuracy, indicating that the IT cortex can still predict image identities well without contextual information. The "Only Context" condition, however, shows the lowest performance, indicating that context alone, without the object, is insufficient for accurate neural prediction. Other manipulations, such as "Blur Context," "Blur Object," "Minimal Context," and "Jigsaw Context," show varying levels of accuracy, generally decreasing as context becomes less informative or more distorted. Overall, the trend indicates that neural prediction accuracy varies with the type and quality of contextual information, emphasizing the importance of both context and object information for accurate neural decoding in the IT cortex. The error bars represent the standard error of the mean, with smaller bars suggesting less variability in the measured accuracy for each condition. The robustness of these findings is reinforced by the data's internal consistency, highlighted in the split-half reliability analysis presented in Figure 3.5B. This analysis demonstrates a discernible pattern: as the number of trials per split increases, so does the reliability of the results. The correlation underscores the internal consistency across the dataset, reflecting greater stability in neural responses with more trials.

3.5.3 Neural Decoding Accuracy Across Increasing Numbers of Neurons

To understand the capacity of the IT cortex for object recognition, it is essential to determine how the number of neurons sampled impacts the accuracy of neural decoding and to reach the accuracy of primates. This analysis provides insights into the limits of neural decoding accuracy and helps identify the required scale of neural recordings to achieve optimal performance.

Figure 3.5 explores the relationship between neural decoding accuracy and the number of neurons sampled. Each point on the graph represents the mean decoding accuracy for a subsample of neurons ranging from 1 to 122 neurons within the 70-170 ms time-bin, with error bars indicating the standard deviation across ten different sub-samples. As the number of neurons increases, there is an initial rapid improvement in decoding accuracy, demonstrating that even a small number of IT neurons can provide substantial information about visual stimuli. Beyond approximately 50 neurons, the rate of accuracy improvement plateaus, suggesting diminishing returns as additional neurons contribute less to the overall accuracy. The extrapolation line (dashed red curve) indicates that recording from about 4537 neurons would be necessary to achieve the pooled monkey accuracy of approximately 0.69, highlighting the potential of large-scale neural recordings to reach high decoding performance. Additionally, the decreasing size of the error bars with more neurons indicates reduced variability and more stable decoding performance, emphasizing the reliability of larger neural populations.

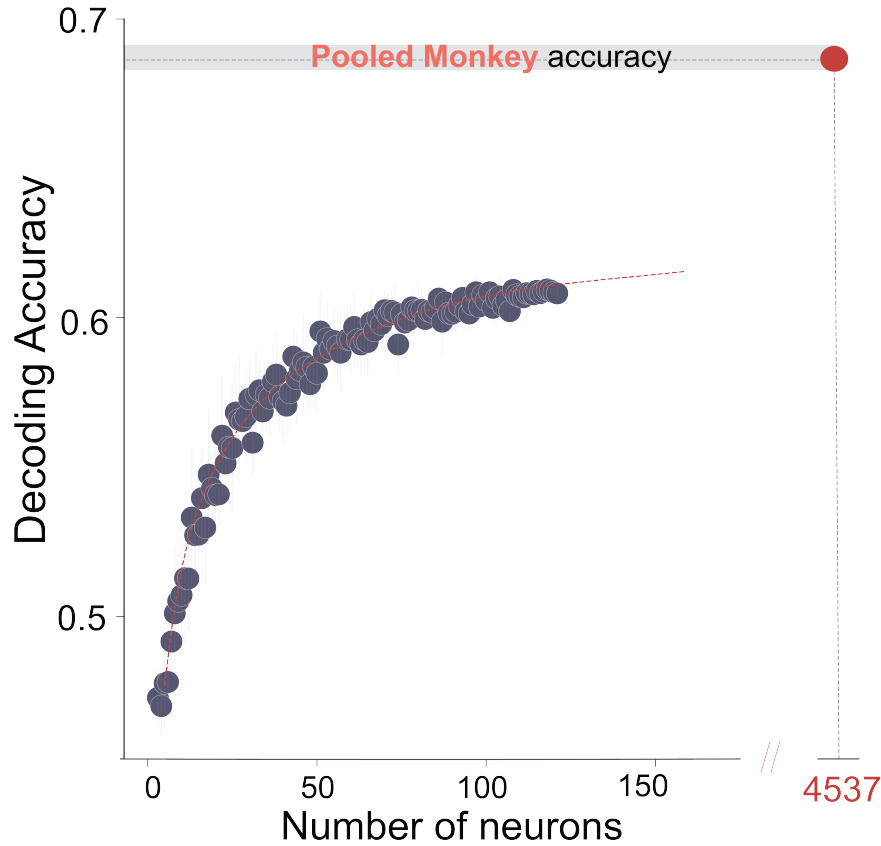


Figure 3.6: This graph shows the decoding accuracy with standard deviation (one-vs-all accuracy, chance level = 0.5) across neuron subsamples for the 70-170 ms time-bin, as a function of the number of neurons. The extrapolation (dashed red curve) estimates that decoding accuracy from a neural population of 4537 recorded neural sites would reach the overall pooled monkey accuracy (0.69). The gray band indicates the monkey accuracy mean with standard error across the 600 images.

The gray band representing pooled monkey accuracy serves as a benchmark, illustrating that the current neuron samples' accuracy approaches this reference point, showcasing the effectiveness of IT cortex neurons decoding objects.

The analysis highlights a key insight: increasing the number of IT cortex neurons significantly enhances the stability and accuracy of neural decoding. This improvement continues as more neurons are sampled, showing that even small increments in neuron count

can contribute to better decoding outcomes. Although the rate of accuracy improvement shows signs of leveling off after a certain point, this reflects an efficient utilization of neural information. The ability to approach primate-level accuracy with a larger sampling pool demonstrates the remarkable potential of large-scale neural recordings.

3.5.4 Population activity across the IT cortex in a context-naive monkey fully explains the shared behavioral variance between humans and monkeys at the overall context level

To evaluate the ability of the IT cortex to account for shared variance in human-monkey behavioral patterns, we investigate how context-level explained variance (EV) changes with the number of neurons sampled. This analysis helps determine the extent to which larger neural populations can capture context-driven behavioral variance, providing insights into the neural mechanisms underlying context-dependent behavior. Figure 3.6 explores how well macaque neural responses can account for the shared variance between human and macaque behavioral patterns at the context level (B.C1). We performed a partial correlation analysis between human and macaque C1 behavioral patterns while controlling for the IT population activity-based predictions of B.C1. Interestingly, with data from 122 neural sites, we explained 75 % of the context-level shared variance between humans and monkeys. To address limitations due to the limited number of neural recordings, we extrapolated the neural decoding accuracy to match monkey accuracy using a logarithmic function. This extrapolation led to an estimate that 4537 neural sites would be needed to reach monkey

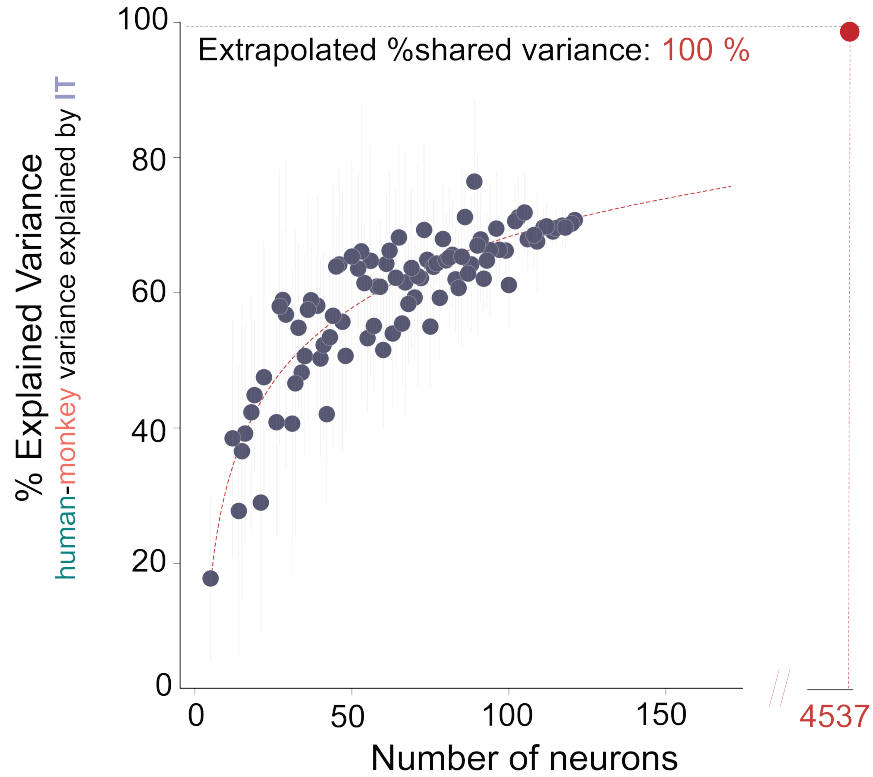


Figure 3.7: This graph depicts the context-level human-monkey shared behavioral variance explained by the neural data. The explained variance (EV) is shown as a function of the number of neurons used for decoding. Each point represents the average EV for a subsample of neurons, with error bars indicating the standard deviation across ten subsamples. As the number of neurons increases, the EV also increases, suggesting that larger neural populations can better account for the shared variance in human-monkey behavioral patterns at the context level. The extrapolated curve predicts that recording from approximately 4537 neurons would fully explain the human-monkey explained variance (HM-EV).

accuracy. The graph shows that the explained variance (EV) increases with the number of neurons, with each point representing the average EV for a subsample of neurons and error bars indicating the standard deviation across ten different subsamples. The extrapolated curve predicts that 4537 neurons would fully explain the human-monkey explained variance (HM-EV), suggesting that the IT cortex can capture context-driven behavioral variance when sufficient neurons are recorded.

This analysis effectively demonstrates that even without prior contextual conditioning, the IT cortex in monkeys exhibits a remarkable capacity to account for a significant portion of the shared behavioral variance observed between humans and monkeys. By analyzing the relationship between the number of sampled neurons and explained variance, we infer that the naive IT cortex can capture essential aspects of context-driven behaviors. The ability of context-naive neurons to substantially explain inter-species behavioral congruences underscores the robustness of the IT cortex in decoding complex visual information, independent of specific contextual learning.

3.5.5 Population activity across the IT cortex in a context-naive monkey only partially explains the shared context-driven behavioral variance between humans and monkeys at the image level

This section delves into the image-level variance explained by neural data. It demonstrates how the IT cortex, even when the monkey was deprived of contextual training, responds to visual stimuli on an image-by-image basis. We assessed how such neural responses predict shared variance between human and macaque behavioral patterns (as estimated in Aim 1) at the image level by performing a partial correlation analysis between human and macaque I1 behavioral patterns while controlling for the IT population activity-based predictions. Neural Data from 122 neural sites indicated that 25% of the image-level shared variance could be

explained. To overcome the limitation of the small neural sample, we extrapolated neural decoding accuracy using a logarithmic function, estimating that approximately 4537 neural sites would be necessary to match the pooled monkey accuracy. Each point on the graph represents the average explained variance (EV) for a subsample of neurons, with standard deviation error bars showing variability across ten subsamples. The graph suggests that while larger neural populations increase EV, they only partially explain the shared behavioral variance at the image level, reaching a plateau at approximately 48.3%, as indicated by the extrapolated curve.

This analysis underscores that the IT cortex of a context-naïve monkey only partially explains the shared behavioral variance between humans and monkeys at the image level. Despite the lack of contextual training, these neurons demonstrate a limited capacity to correlate with shared human-macaque behavioral responses based on visual stimuli alone. The data reveals that with increasing numbers of neurons sampled, the explained variance (EV) in behavior increases, but plateaus at approximately 48.3%. This plateau in explained variance (EV) highlights fundamental limitations in the predictive capabilities of neurons that have not been contextually trained, particularly at the detailed image-specific level. Although the IT cortex of a context-naïve monkey provides significant insights into general visual processing, its ability to accurately categorise objects based on specific images is inherently restricted. This limitation suggests that the neural mechanisms involved in image recognition and response are complex and potentially influenced by a variety of factors beyond innate neural processing, such as learned associations and contextual familiarity.

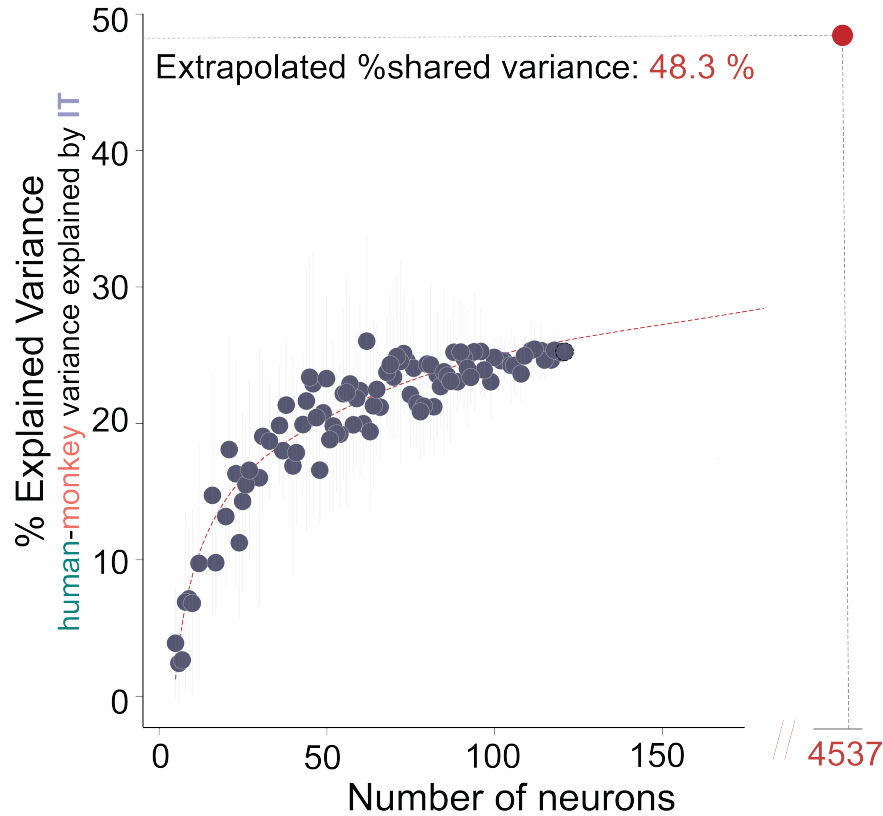


Figure 3.8: This graph displays the image-level variance explained by neural data from the IT cortex of a context-naïve monkey. The explained variance (EV) increases with the number of neurons, but only to a certain extent, plateauing at about 48.3%. Each point represents the average EV for a subsample of neurons, and error bars indicate the standard deviation across ten different subsamples. The extrapolated curve predicts that even with a substantial increase to 4537 neurons, the neural data would only partially explain the shared human-monkey behavioral variance at the image level. This finding highlights the challenges and limitations of using context-naïve neural responses to predict detailed, image-specific behavioral patterns and underscores the need for a larger and more contextually enriched neural dataset to fully understand the neural basis for such variance.

3.5.6 Population activity across the IT cortex in a context-naive monkey showing a decline in explaining the shared context-driven behavioral variance between humans and monkeys at different behavioral metrics.

The bar graph visually quantifies the extent of shared human monkey shared variance is explained by the naive monkeys IT neural population, segmented into four specific conditions labeled C1, I1, C2, and I2. Each bar corresponds to a different condition and represents the mean percentage of percent explained variance between the two species by IT, with error bars reflecting the variability of these measurements. The C1 and I1 conditions demonstrate a significant degree of explained variance, particularly in the C1 condition, where the explained variance peaks, indicating that at a coarser grain IT can explain the shared variance to an extent. At the C2 and I2 levels, the shared variance drops decline from the C1 level. This suggests that while there remains an alignment in responses at a coarser matrix, the C2 and I2 conditions, which represent a finer level of analysis, may engage different aspects of cognitive or perceptual processing that are less conserved between species. As our examination becomes more detailed, we begin to observe subtle differences, indicating that finer distinctions in task or stimulus processing may not be as universally shared across species.

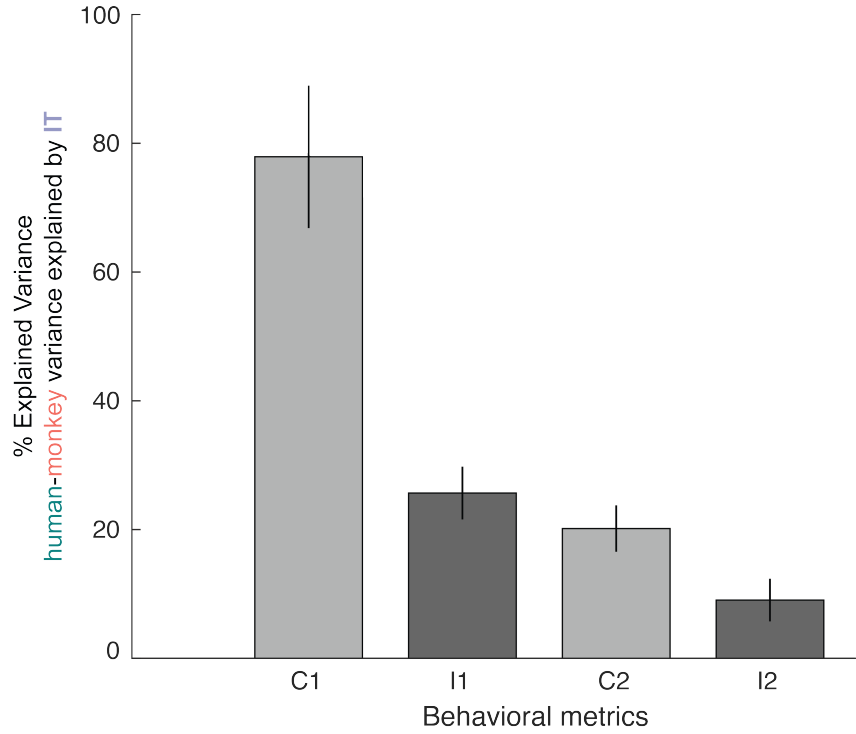


Figure 3.9: The bar graph quantifies the percentage of shared explained variance in primates which is explained by IT cortex activity across four different behavioral metrics introduced in Aim1(C1, I1, C2, I2)- (Figure 2.3). The bars represent the mean with the error bars representing the standard deviation indicating the variability of this influence. The pronounced peak in the 'C1' condition suggests a heightened role of IT cortex activity in explaining variance within that specific context.

The bar graph illustrates the percentage of explained variance in human-monkey behavior attributed to the IT cortex across four different conditions: C1, I1, C2, and I2. The y-axis represents the percentage of explained variance, ranging from 0 to 100 percent, while the x-axis categorizes the conditions. Each bar signifies the percentage of variance explained for its respective condition, error bars are standard deviations across iterations, with the height indicating the amount of variance.

The graph shows that C1 has the highest explained variance at approximately 77.91 %, with an error bar reflecting some variability. In contrast, I1 exhibits a lower explained

variance of around 25.67 %, but with less variability as indicated by a smaller error bar. C2 follows with an explained variance of about 20.16 %, and moderate variability. I2, with the lowest explained variance of approximately 9.04 %, showing some variability.

Overall, the graph suggests that the C1 condition accounts for the highest percentage of explained variance in the human-monkey variance by IT, while the I2 condition accounts for the least. The error bars provide insights into the reliability of these estimates, indicating that as the granularity of the metric of analysis increases, the explainability decreases.

Chapter 4

Discussion

This thesis aimed to elucidate the role of context in visual object recognition, focusing on both humans and macaque monkeys. Our approach combined behavioral analysis with neurophysiological data and decoding models to quantify how context influences visual recognition. A key finding from our experiments is that both humans and monkeys significantly leverage scene context to enhance object recognition capabilities. The visual object recognition abilities of monkeys trained to categorize objects in their natural context were strongly modulated when we deliberately varied the contextual cues, which went beyond low-level image attributes. This was evidenced by both species exhibiting higher accuracy in object recognition tasks when presented with objects in natural contexts compared to incongruent or no-context scenarios. Experimental results demonstrated that humans showed a recognition accuracy of 71.35 % across different contexts, while monkeys showed slightly lower overall accuracy at 69.16 %. The performance was notably better in full context conditions for both species, indicating the importance of contextual cues in facilitating the recognition process. One of the most compelling findings from our study was the high degree of behavioral

concordance observed between humans and monkeys, suggesting a fundamental similarity in how both species process contextual information in visual tasks. This concordance was quantitatively supported by a strong correlation coefficient of 0.83 at the context level and a moderately significant correlation of 0.63 at the image level. This high behavioral alignment in context-dependent recognition tasks is significant as it not only supports the hypothesis that primate vision—both human and non-human—relies heavily on contextual information but also suggests that studies on monkeys can provide valuable insights into human visual cognition. We recognized that context plays a critical role in this process and propose that monkeys serve as an effective model for exploring the neural mechanisms involved. This cross-species similarity highlights the potential of using non-human primates as effective models for studying human perception, particularly in understanding how context influences visual processing and decision-making in complex environments. These findings are particularly surprising given the detailed level of alignment across multiple contexts and images, suggesting a more profound and foundational role of context in visual perception than previously documented. However, our analysis also revealed that, at the image level, monkeys do not fully replicate human behavioral patterns. This discrepancy may be attributed to potential limitations in the depth and duration of their training or inherent species-specific differences in sensory processing and cognition. Furthermore, we observed that the population activity distributed across the IT cortex of naïve monkeys, which were not explicitly trained with objects in context, does not fully account for the context-driven behavioral patterns observed in context-trained monkeys.

4.1 Context modulates visual object recognition in humans and monkeys

Our results underscore the critical role of context in primate object recognition, aligning with a substantial body of literature on visual cognition [68, 54, 56, 55, 73]. This body of research has established that contextual cues significantly influence human visual object recognition capabilities. These cues are informed by our understanding of how frequently certain objects appear together, which shapes our perceptions of what is congruent or incongruent within a given scene. Interestingly, our findings reveal that monkeys exhibit a similar sensitivity to these contextual statistical cues, much like humans. When we presented objects in varying contexts, we observed notable decreases in object discrimination accuracy in both species when the context was incongruent compared to fully congruent scenes. Specifically, humans showed up to a 13% decrease in accuracy, while monkeys exhibited a 10.4% decrease under incongruent conditions. This demonstrates that both humans and monkeys rely on contextual information to recognize objects accurately. These striking parallels between the two primate species highlight the potential of macaques as a valuable model system for investigating the neural mechanisms underlying context processing. The fact that monkeys show similar context sensitivity suggests that they can provide important insights into the cognitive and neural processes that humans use. Our study extends beyond previous research by demonstrating context sensitivity across a wide range of manipulations and employing rigorous, multi-faceted behavioral metrics designed to quantify performance changes induced by contextual cues.

Earlier studies might have focused on more limited aspects of context, but our comprehensive approach shows that both species are affected by a broad spectrum of contextual variations.

The tight similarity between human and monkey behavior points to potential shared cognitive mechanisms. These mechanisms likely include knowledge of usual object co-presence, which helps in understanding which objects typically appear together, the relative sizes of objects, and positional regularities within a scene. For example, knowing that a cup is often found on a table rather than floating in the air helps both humans and monkeys recognize objects more accurately when they are in expected positions.

These shared mechanisms could account for the context-facilitation effects observed in both species. In essence, the ability to use context to improve object recognition is not unique to humans but is also present in monkeys, suggesting that these cognitive processes could be shared between the two species. Understanding these mechanisms better can provide insights into how our brains process complex visual information and adapt to different contexts, which can have implications for fields ranging from neuroscience to artificial intelligence.

4.2 Goodness of monkeys as a model of human contextual processing

While there is a largely consistent effect at the broader, context-level validate macaques as a model for studying visual processing, some differences were still noted between the two species. Our analysis showed that, at the individual image level, monkeys did not fully

imitate human behavioral patterns. To familiarize them with various objects in their natural surrounding context—such as cars on roads, bears in the jungle, and chairs in rooms — the monkeys were trained extensively using natural images from the MS COCO image dataset. The training continued until their performance leveled off, as shown in Figure 5.1.

Despite achieving high performance on the object discrimination task, the mismatch between human and monkey image responses might suggest that the training might not have been sufficient for the number of objects they learn to generalize as well as the duration of the time of their training. Enhancing the training program could potentially lead to better alignment with human behavior at the context level, reducing the differences observed at the image level. For instance, extending the duration of training and exposing monkeys to an even wider variety of images with varied natural contextual scenarios might help them develop a more nuanced understanding of context, closer to that of humans.

Furthermore, this discrepancy suggests inherent species-level differences in sensory processing and cognition between the two primates. One possible explanation for these differences is the varying extent and quality of training received by the two species. While humans typically have extensive and diverse visual experiences throughout their lives, which can enhance their ability to recognize and discriminate between images, monkeys in experimental settings might have a more constrained exposure. The training protocols for monkeys, although rigorous, might not fully encompass the broad range of visual stimuli that humans encounter, potentially leading to differences in performance on visual tasks.

Additionally, factors, like the limited number of stimuli and specific task requirements,

might also contribute to these discrepancies in performance between humans and monkeys. The current set of stimuli might not capture the full complexity and breadth of real-world environments, and the tasks used might not fully encompass the range of cognitive processes humans employ. Expanding the stimulus set to include more diverse and representative images could provide a more accurate assessment of the monkeys' capabilities. Moreover, designing tasks that better reflect everyday visual challenges could help bridge the gap between human and monkey performance. Interestingly, when we used context manipulations like blurring, we saw that performance in both species was only slightly affected. This suggests that certain types of contextual changes do not significantly impact object recognition, indicating which aspects of context are most critical for accurate object recognition.

Another important factor to consider is the natural differences between humans and monkeys, such as unique species traits and brain structure variations [79]. These biological disparities may restrict the extent to which monkeys can mimic human contextual processing. Consequently, while improved training methods and greater exposure to varied visual contexts might enhance the performance of monkeys, it's vital to acknowledge that certain differences may inevitably remain. These differences stem from the distinct evolutionary trajectories, sensory capacities, and cognitive abilities of humans and monkeys. Recognizing these limitations is essential for the accurate interpretation of comparative studies and the effective use of monkeys as models in exploring human visual processing.

Future research should aim to explore these fundamental differences more deeply, potentially through comparative studies and behavioral experiments that target specific aspects

of sensory and cognitive processing. By doing so, we can gain a more comprehensive understanding of the shared and unique elements of visual perception and context recognition in primates, ultimately improving the validity and applicability of findings derived from monkey models to human contexts.

Understanding and acknowledging these limitations is essential as we continue to refine the use of monkeys as models for human visual processing. While macaques provide a valuable model for studying visual processing, it is crucial to consider the limitations and potential discrepancies between species. By enhancing training protocols, expanding stimulus sets, and understanding biological differences, we can improve the alignment between monkey and human behavioral patterns. This will not only enhance the validity of macaques as a model system but also contribute to a deeper understanding of the neural and cognitive mechanisms underlying context processing in primates.

4.3 Role of IT Cortex in Processing Scene Context

The inferior temporal (IT) cortex plays a vital role in visual object processing, and this has been studied extensively [25, 71, 26]. However, our findings indicate that the IT cortex responses from monkeys that have not been trained to discriminate objects within specific contexts (e.g., cars on the road, planes in the sky, bear in a forest) do not fully predict scene context behavior as measured in humans or in monkeys that have been explicitly context-trained. Our findings reveal that the neural data from context-naïve monkeys only partially explain the shared behavioral variance with humans. This partial explanation

underscores the complexity of the neural processes at play and the scope to probe further the neural circuitry underlying this behavior. This discrepancy points to the necessity for a more detailed approach in future research to truly understand the role of the IT cortex (in conjunction with other brain areas) in processing scene context.

Given this partial explanation by context-naïve monkeys' neural data, a future research avenue involves training these monkeys in object recognition tasks that incorporate contextual cues. The goal would be to test whether their behavioral performance and IT neural responses align more closely with context-trained monkeys over time. This training would involve exposing the monkeys to visual stimuli featuring both the target object and relevant contextual information and then tracking their recognition abilities and neural activity.

Given that the IT responses from contextually naïve monkeys could not account for the shared behavioral patterns observed between humans and monkeys, several potential explanations for this shortcoming can be considered. One possible reason for this limitation is that our data might be too restricted to draw comprehensive conclusions. The range of our IT neural data sampling, the variety of images used, the number of trials conducted, and the different objects and context scenarios might not be extensive enough to fully showcase the IT cortex's capabilities in a processing context. To tackle these limitations, we conducted extrapolation analyses (Figure 3.5) to estimate how increasing our sample size could impact our findings. These analyses suggest that the current data from the IT cortex of context-naïve monkeys are insufficient to fully explain human-like behavior.

Another reason might be that the IT cortex in context-naïve monkeys lacks the refined

ability to represent scene context due to insufficient exposure to varied contextual cues as compared to context-trained monkeys. Providing monkeys with more diverse and extensive contextual training could potentially enhance the IT cortex's ability to process and represent scene context more effectively.

Furthermore, exploring how the IT cortex interacts with other brain regions is essential. This includes areas within the ventral visual stream, such as V4, as well as areas outside the ventral stream, like the ventrolateral prefrontal cortex (vlPFC). Investigating these interactions and their correlation with behavior in both trained and untrained monkeys could reveal novel neural mechanisms. This could help determine if other brain regions compensate for or enhance the IT cortex's role in context processing, offering a more holistic view of the neural networks involved in this complex task. By integrating behavioral and neural analyses, we can develop comprehensive explanations that connect the cognitive effects of context with their neural underpinnings. Overall, this multi-faceted approach paves the way for a deeper understanding of how context facilitates robust object perception across primates.

In summary, to fully understand the IT cortex's role in context processing, future research should: increase the variety and quantity of data to better capture the full capabilities of the IT cortex in processing context; provide more diverse and extensive contextual training to monkeys to improve their IT cortex's representational capacity; investigate the extent of neural plasticity in the IT cortex with increased contextual exposure and training. Additional investigation of how the IT cortex interacts with other brain regions to understand the broader neural network involved in context processing and combine findings from behavioral,

and neural studies to develop a comprehensive understanding of visual processing.

In conclusion, this thesis highlights the importance of context in visual recognition and supports using macaques as models for human visual processing. Optimizing training and exploring neural interactions can further align monkey and human behaviors, advancing our understanding of visual cognition.

Chapter 5

Appendix

Monkey learning to do task

To verify macaques as a suitable animal model for investigating the neural underpinnings of human context processing, it's essential to first determine if macaques exhibit similar behavioral responses to contextual effects as humans. To ensure the macaques were accustomed to the scene context associated with each object category, they were systematically trained using images featuring full context. Specifically, images were sourced from the Microsoft COCO dataset, encompassing 160 images for each of ten distinct objects. Throughout this training phase, the macaques demonstrated strong cross-validated accuracy, indicating their ability to recognize and respond to these contextual cues effectively. This preliminary behavioral similarity suggests that macaques can indeed serve as a valuable model for studying context-processing mechanisms comparable to those in humans.

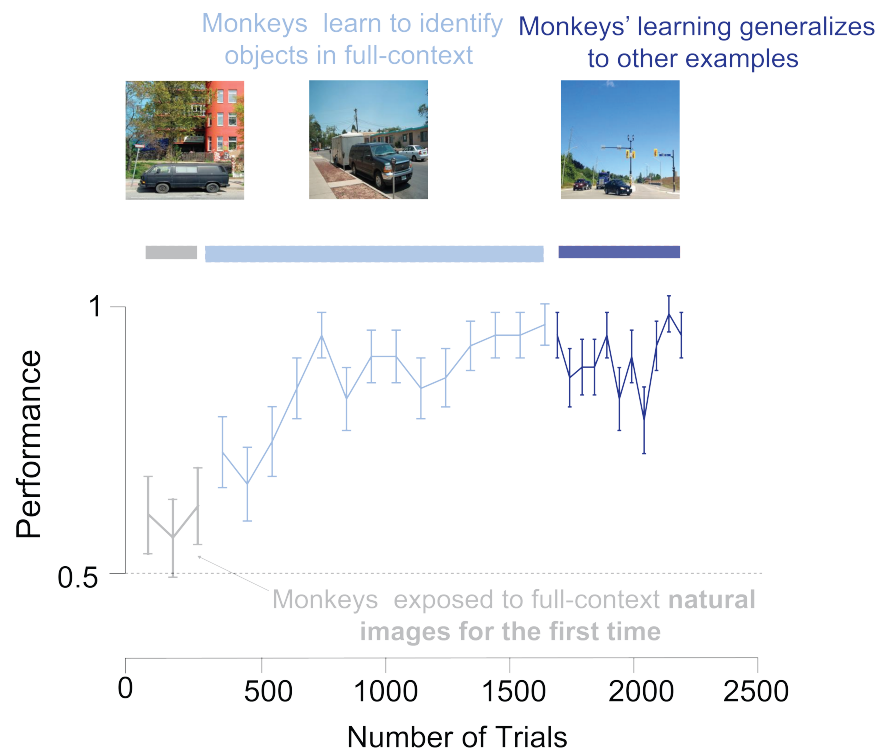


Figure 5.1: Training process for one macaque in the task shown in Figure 2.2A (chance = 0.5). The monkey, who was previously not exposed to images with objects in context, has low starting performance for full-context images (light blue curve). However, the monkey quickly learns to recognize images in full context (blue curve). Furthermore, this ability generalizes to new images (dark blue curve). Error bars show the standard deviation across object categories.

Behavioral metric- object level (O1) and object level (O2)

Object Level metric (O1) which steps in to quantify the average recognition accuracy for each target object category across all presented contexts and images. This measure cuts across the contextual and image-level data, providing a high-level overview of object recognizability. O1 is particularly useful in identifying which objects have a higher or lower recognition rate in general, offering a broad-stroke understanding of object recognition trends.

Secondly, the Object Level metric(O2) investigates how the recognition of each object is influenced by the distractor object category it appears with. By measuring recognition performance fluctuations due to different distractor objects, O2 provides insights into which objects are highly context-dependent, enhancing our understanding of context's role in visual perception and cognition.

Significant and reliable alignment in behavior between primates under different object categories

The comparative study at the object level, focusing on how two individual primates, Monkey B and Monkey M, process and recognize specific object categories reveals a robust alignment in recognition capabilities between these two subjects. Each point on the plot corresponds to the average accuracy rate for a particular object, with error bars reflecting the standard error. The presence of these data points near the line of unity (dotted line) and a high correlation coefficient of 0.76 suggests a strong linear relationship in their object recognition processes. This high degree of correlation indicates that, despite potential individual variations, the

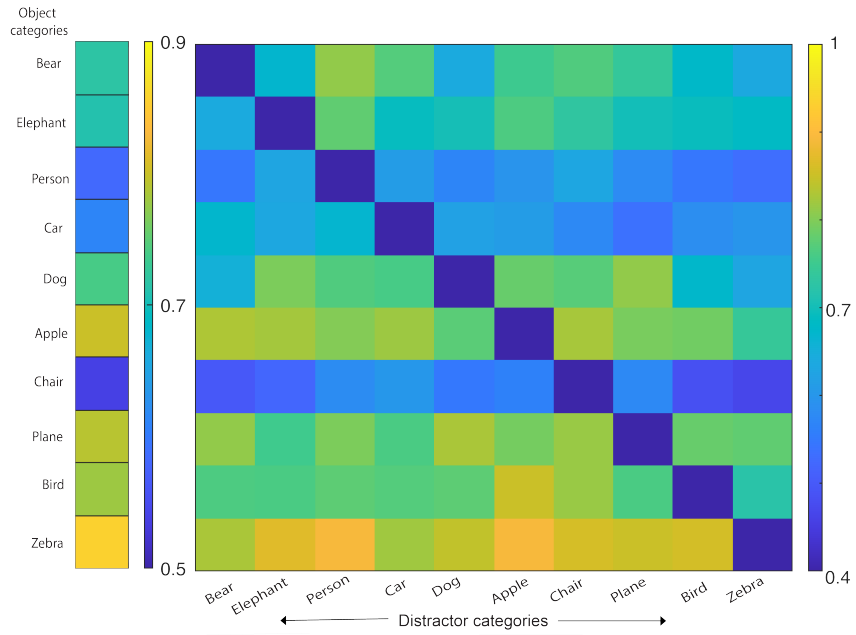


Figure 5.2: Schematic explaining how performance is evaluated from the task matrix at the object level. Metric O1 examines accuracy within particular object categories, collapsing data across target-distractor combinations to evaluate the influence of objects on recognition. Object-level metric (O2) explores object recognition consistency across distractor categories, revealing how different distractors impact recognition.

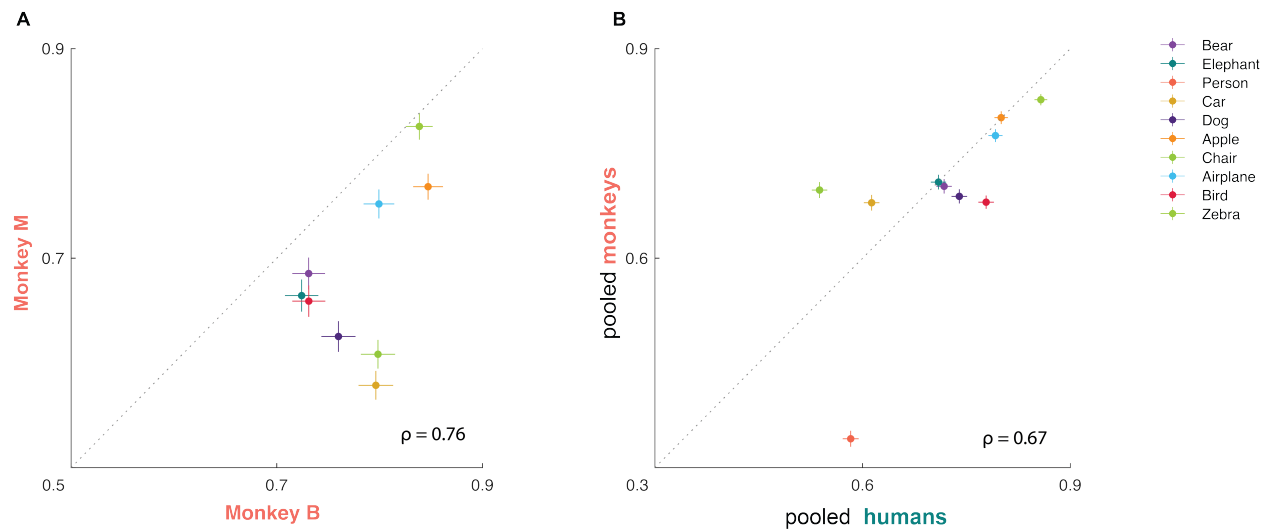


Figure 5.3: A. Scatter plot illustrating the correlation of object recognition accuracy between Monkey B and Monkey M, showing a strong alignment in their perceptual capabilities. Each point on the plot represents the average accuracy for a specific object category, marked with error bars indicating the standard error. The correlation coefficient ($\rho = 0.76$) underscores a significant agreement in how these two monkeys recognize and process various objects. B. Scatter plot demonstrating the correlation of object recognition accuracy between pooled monkeys and humans, marked by a correlation coefficient of 0.67. Each data point represents a distinct object category. Despite a lower correlation compared to the inter-monkey comparison, the correlation remains moderate and significant, suggesting substantial but nuanced similarities and differences in how monkeys and humans perceive the same object categories.

fundamental aspects of object processing are remarkably consistent between these two monkeys.

Expanding upon the insights gained from the inter-monkey comparison, we further explore the relationship in object recognition between a collective group of monkeys and humans as shown in Figure B. Each data point represents the pooled mean accuracy rates for both species per object category. With a correlation coefficient of 0.67, the plot highlights a moderate yet significant positive correlation, underscoring substantial similarities in how both species perceive and interpret the same objects.

Object Misidentification Patterns: Detailed Confusion Matrices Compare Human and Monkey Recognition Errors

The object confusion matrix represents a measure of the influence that various distractor object categories have on the recognition of different target objects, and it is calculated based on the o2 metric. The matrix emphasizes the degree of interference or difficulty caused by distractors during object recognition tasks.

Each row and column of the matrix corresponds to a specific object category. The intersection of a row and column represents the performance impact when the object of the row is the target and the object of the column is the distractor. The color gradient reflects this impact, with warmer shades indicating higher levels of interference (lower recognition performance), and cooler shades indicating less interference (higher recognition performance).

Black cells indicate the instances where the data is not available (NaN – Not a Number).

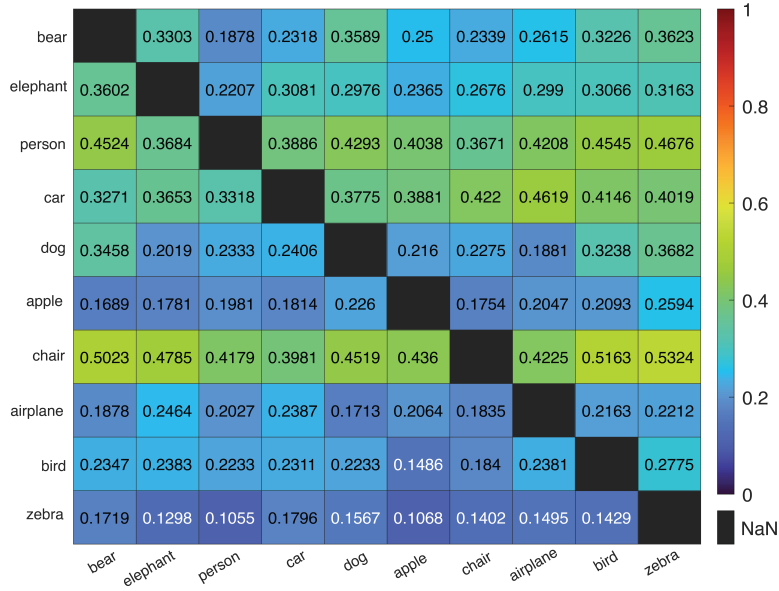


Figure 5.4: The confusion matrix displays the performance of human subjects in object recognition tasks. Each cell in the matrix represents the probability that a given object (rows) is misidentified as another object (columns). The color gradient from blue to yellow illustrates increasing error rates, where darker blue indicates lower confusion and yellow represents higher confusion. Notable confusions include misidentifying 'chair' as 'table' and 'bird' as 'airplane,' which may indicate visual or semantic similarities perceived by humans. The 'NaN' (black cells) denotes categories that were not applicable or not tested in this setup.

This means that no data was collected since the target and distractor are the same, and hence no interference measurement is applicable.

The human confusion matrix provides a rich visualization of how accurately or inaccurately humans categorize various objects in recognition tasks. The matrix reveals intriguing patterns where certain objects are consistently confused with others, suggesting semantic or visual similarities that might be leading to these errors. For example, animals like 'bear' and 'dog' show mutual confusion, which could be due to similar body shapes or sizes perceived by the participants. Some misclassifications highlight how visual similarities can lead to errors. For

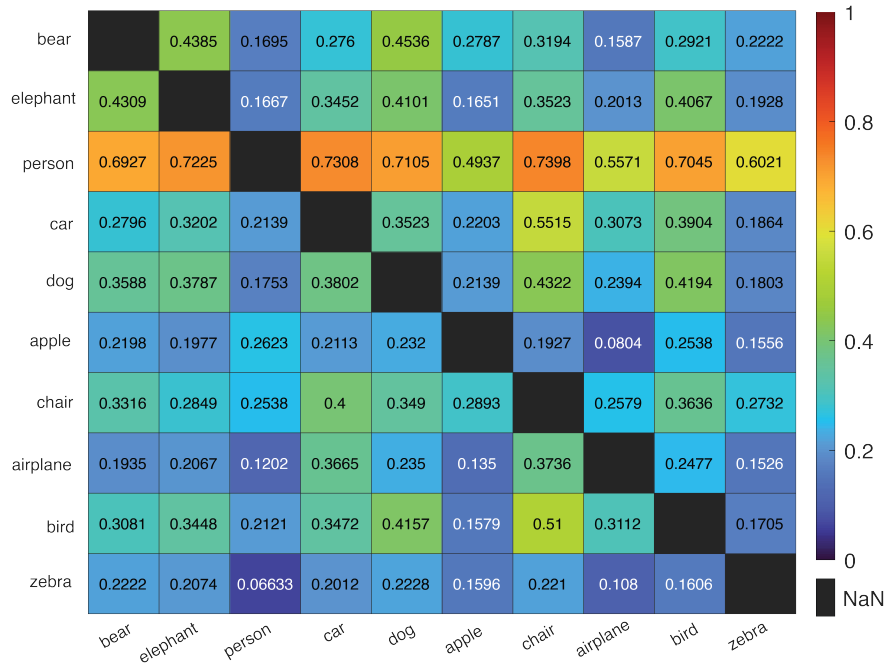


Figure 5.5: This matrix illustrates the object recognition performance of monkeys, highlighting differences in error rates compared to human performance. Each element indicates how frequently one object (row) is mistaken for another (column) by the subjects. The color scale from blue to yellow represents an increasing likelihood of confusion, with deeper blue showing lesser confusion and yellows indicating higher misidentification rates. Significant confusions, such as 'person' being confused with 'chair', suggest perceptual or associative challenges unique to non-human primates.

instance, 'airplane' and 'bird' might be confused due to similar silhouettes when viewed from certain angles. Some categories might show higher overall error rates, evidenced by more colorful (towards yellow) non-diagonal cells in their rows. This could indicate that these objects are inherently more difficult to recognize, possibly due to less distinctive features or greater variability in appearance across different instances of the category.

The matrix highlights several notable misidentifications, such as monkeys frequently confusing 'person' with other object categories. This might indicate perceptual similarities perceived by monkeys or a lack of distinction in the monkeys' cognitive categorization between

human forms and common objects within human environments. There's a noticeable tendency for monkeys to confuse different animal categories, such as 'bear' and 'elephant'. Some errors, such as the confusion between 'bird' and 'airplane', which is less pronounced in human data, suggest unique aspects of how monkeys process objects that fly.

The correlation analysis yielded a correlation coefficient (r) of 0.57, which indicates a moderate positive relationship between the two confusion matrices. This correlation reflects a notable but not strong association, suggesting that while there is a link between the two, other factors may also play a significant role in the observed correlation.

Bibliography

- [1] Kohitij Kar and James J DiCarlo. *The Quest for an Integrated Set of Neural Mechanisms Underlying Object Recognition in Primates*. 2023. DOI: 10.48550/ARXIV.2312.05956. URL: <https://arxiv.org/abs/2312.05956>.
- [2] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. “How Does the Brain Solve Visual Object Recognition?” In: *Neuron* 73.3 (Feb. 2012), pp. 415–434. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.01.010. URL: <http://dx.doi.org/10.1016/j.neuron.2012.01.010>.
- [3] Michèle Fabre-Thorpe, Ghislaine Richard, and Simon J. Thorpe. “Rapid categorization of natural images by rhesus monkeys”. In: *NeuroReport* 9.2 (Jan. 1998), pp. 303–308. ISSN: 0959-4965. DOI: 10.1097/00001756-199801260-00023. URL: <http://dx.doi.org/10.1097/00001756-199801260-00023>.
- [4] Simon Thorpe, Denis Fize, and Catherine Marlot. “Speed of processing in the human visual system”. In: *Nature* 381.6582 (June 1996), pp. 520–522. ISSN: 1476-4687. DOI: 10.1038/381520a0. URL: <http://dx.doi.org/10.1038/381520a0>.

- [5] C. Keysers et al. “The Speed of Sight”. In: *Journal of Cognitive Neuroscience* 13.1 (Jan. 2001), pp. 90–101. ISSN: 1530-8898. DOI: 10.1162/089892901564199. URL: <http://dx.doi.org/10.1162/089892901564199>.
- [6] B. Fischer and H. Weber. “Express saccades and visual attention”. In: *Behavioral and Brain Sciences* 16.3 (Sept. 1993), pp. 553–567. ISSN: 1469-1825. DOI: 10.1017/S0140525X00031575. URL: <http://dx.doi.org/10.1017/S0140525X00031575>.
- [7] James J. DiCarlo and John H. R. Maunsell. “Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing”. In: *Nature Neuroscience* 3.8 (Aug. 2000), pp. 814–821. ISSN: 1546-1726. DOI: 10.1038/77722. URL: <http://dx.doi.org/10.1038/77722>.
- [8] Antje Nuthmann. “Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task”. In: *Psychonomic Bulletin and Review* 24.2 (Aug. 2016), pp. 370–392. ISSN: 1531-5320. DOI: 10.3758/s13423-016-1124-4. URL: <http://dx.doi.org/10.3758/s13423-016-1124-4>.
- [9] James J. DiCarlo and David D. Cox. “Untangling invariant object recognition”. In: *Trends in Cognitive Sciences* 11.8 (Aug. 2007), pp. 333–341. ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.06.010. URL: <http://dx.doi.org/10.1016/j.tics.2007.06.010>.
- [10] Stefania Bracci and Hans P. Op de Beeck. “Understanding Human Object Vision: A Picture Is Worth a Thousand Representations”. In: *Annual Review of Psychology* 74.1

- (Jan. 2023), pp. 113–135. ISSN: 1545-2085. DOI: 10.1146/annurev-psych-032720-041031. URL: <http://dx.doi.org/10.1146/annurev-psych-032720-041031>.
- [11] A Delorme, G Richard, and M Fabre-Thorpe. “Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans”. In: *Vision Research* 40.16 (July 2000), pp. 2187–2200. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(00)00083-3. URL: [http://dx.doi.org/10.1016/S0042-6989\(00\)00083-3](http://dx.doi.org/10.1016/S0042-6989(00)00083-3).
- [12] Rishi Rajalingham et al. “Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks”. In: *The Journal of Neuroscience* 38.33 (July 2018), pp. 7255–7269. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.0388-18.2018. URL: <http://dx.doi.org/10.1523/JNEUROSCI.0388-18.2018>.
- [13] Rishi Rajalingham, Kailyn Schmidt, and James J. DiCarlo. “Comparison of Object Recognition Behavior in Human and Monkey”. In: *The Journal of Neuroscience* 35.35 (Sept. 2015), pp. 12127–12136. ISSN: 1529-2401. DOI: 10.1523/JNEUROSCI.0573-15.2015. URL: <http://dx.doi.org/10.1523/JNEUROSCI.0573-15.2015>.
- [14] D. H. Hubel and T. N. Wiesel. “Receptive fields of single neurones in the cat’s striate cortex”. In: *The Journal of Physiology* 148.3 (Oct. 1959), pp. 574–591. ISSN: 1469-7793. DOI: 10.1113/jphysiol.1959.sp006308. URL: <http://dx.doi.org/10.1113/jphysiol.1959.sp006308>.
- [15] D. H. Hubel and T. N. Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of Physiology* 160.1 (Jan.

- 1962), pp. 106–154. ISSN: 1469-7793. DOI: 10.1113/jphysiol.1962.sp006837. URL: <http://dx.doi.org/10.1113/jphysiol.1962.sp006837>.
- [16] D. H. Hubel and T. N. Wiesel. “Shape and arrangement of columns in cat’s striate cortex”. In: *The Journal of Physiology* 165.3 (Mar. 1963), pp. 559–568. ISSN: 1469-7793. DOI: 10.1113/jphysiol.1963.sp007079. URL: <http://dx.doi.org/10.1113/jphysiol.1963.sp007079>.
- [17] Charles E Connor, Scott L Brincat, and Anitha Pasupathy. “Transformation of shape information in the ventral pathway”. In: *Current Opinion in Neurobiology* 17.2 (Apr. 2007), pp. 140–147. ISSN: 0959-4388. DOI: 10.1016/j.conb.2007.03.002. URL: <http://dx.doi.org/10.1016/j.conb.2007.03.002>.
- [18] Edmund T Rolls. “Functions of the Primate Temporal Lobe Cortical Visual Areas in Invariant Visual Object and Face Recognition”. In: *Neuron* 27.2 (Aug. 2000), pp. 205–218. ISSN: 0896-6273. DOI: 10.1016/s0896-6273(00)00030-1. URL: [http://dx.doi.org/10.1016/s0896-6273\(00\)00030-1](http://dx.doi.org/10.1016/s0896-6273(00)00030-1).
- [19] Dwight J. Kravitz et al. “The ventral visual pathway: an expanded neural framework for the processing of object quality”. In: *Trends in Cognitive Sciences* 17.1 (Jan. 2013), pp. 26–49. ISSN: 1364-6613. DOI: 10.1016/j.tics.2012.10.011. URL: <http://dx.doi.org/10.1016/j.tics.2012.10.011>.
- [20] Alomit Ishai et al. “Distributed representation of objects in the human ventral visual pathway”. In: *Proceedings of the National Academy of Sciences* 96.16 (Aug. 1999),

- pp. 9379–9384. ISSN: 1091-6490. DOI: 10.1073/pnas.96.16.9379. URL: <http://dx.doi.org/10.1073/pnas.96.16.9379>.
- [21] E Peterhans and R von der Heydt. “Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps”. In: *The Journal of Neuroscience* 9.5 (May 1989), pp. 1749–1763. ISSN: 1529-2401. DOI: 10.1523/jneurosci.09-05-01749.1989. URL: <http://dx.doi.org/10.1523/JNEUROSCI.09-05-01749.1989>.
- [22] Anitha Pasupathy and Charles E. Connor. “Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation”. In: *Journal of Neurophysiology* 86.5 (Nov. 2001), pp. 2505–2519. ISSN: 1522-1598. DOI: 10.1152/jn.2001.86.5.2505. URL: <http://dx.doi.org/10.1152/jn.2001.86.5.2505>.
- [23] SJ Schein and R Desimone. “Spectral properties of V4 neurons in the macaque”. In: *The Journal of Neuroscience* 10.10 (Oct. 1990), pp. 3369–3389. ISSN: 1529-2401. DOI: 10.1523/jneurosci.10-10-03369.1990. URL: <http://dx.doi.org/10.1523/JNEUROSCI.10-10-03369.1990>.
- [24] BC Motter. “Neural correlates of attentive selection for color or luminance in extrastriate area V4”. In: *The Journal of Neuroscience* 14.4 (Apr. 1994), pp. 2178–2189. ISSN: 1529-2401. DOI: 10.1523/jneurosci.14-04-02178.1994. URL: <http://dx.doi.org/10.1523/JNEUROSCI.14-04-02178.1994>.
- [25] Chou P. Hung et al. “Fast Readout of Object Identity from Macaque Inferior Temporal Cortex”. In: *Science* 310.5749 (Nov. 2005), pp. 863–866. ISSN: 1095-9203. DOI: 10.1126/science.1117593. URL: <http://dx.doi.org/10.1126/science.1117593>.

- [26] Najib J. Majaj et al. “Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance”. In: *The Journal of Neuroscience* 35.39 (Sept. 2015), pp. 13402–13418. ISSN: 1529-2401. DOI: 10.1523/jneurosci.5181-14.2015. URL: <http://dx.doi.org/10.1523/JNEUROSCI.5181-14.2015>.
- [27] Keiji Tanaka. “Inferotemporal Cortex and Object Vision”. In: *Annual Review of Neuroscience* 19.1 (Mar. 1996), pp. 109–139. ISSN: 1545-4126. DOI: 10.1146/annurev.ne.19.030196.000545. URL: <http://dx.doi.org/10.1146/annurev.ne.19.030196.000545>.
- [28] D. J. Felleman and D. C. Van Essen. “Distributed Hierarchical Processing in the Primate Cerebral Cortex”. In: *Cerebral Cortex* 1.1 (Jan. 1991), pp. 1–47. ISSN: 1460-2199. DOI: 10.1093/cercor/1.1.1. URL: <http://dx.doi.org/10.1093/cercor/1.1.1>.
- [29] David J. Freedman et al. “Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex”. In: *Science* 291.5502 (Jan. 2001), pp. 312–316. ISSN: 1095-9203. DOI: 10.1126/science.291.5502.312. URL: <http://dx.doi.org/10.1126/science.291.5502.312>.
- [30] Andreas Lueschow, Earl K. Miller, and Robert Desimone. “Inferior Temporal Mechanisms for Invariant Object Recognition”. In: *Cerebral Cortex* 4.5 (1994), pp. 523–531. ISSN: 1460-2199. DOI: 10.1093/cercor/4.5.523. URL: <http://dx.doi.org/10.1093/cercor/4.5.523>.

- [31] R Desimone et al. “Stimulus-selective properties of inferior temporal neurons in the macaque”. In: *The Journal of Neuroscience* 4.8 (Aug. 1984), pp. 2051–2062. ISSN: 1529-2401. DOI: 10.1523/jneurosci.04-08-02051.1984. URL: <http://dx.doi.org/10.1523/JNEUROSCI.04-08-02051.1984>.
- [32] Nuo Li et al. “What Response Properties Do Individual Neurons Need to Underlie Position and Clutter “Invariant” Object Recognition?” In: *Journal of Neurophysiology* 102.1 (July 2009), pp. 360–376. ISSN: 1522-1598. DOI: 10.1152/jn.90745.2008. URL: <http://dx.doi.org/10.1152/jn.90745.2008>.
- [33] Roozbeh Kiani et al. “Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex”. In: *Journal of Neurophysiology* 97.6 (June 2007), pp. 4296–4309. ISSN: 1522-1598. DOI: 10.1152/jn.00024.2007. URL: <http://dx.doi.org/10.1152/jn.00024.2007>.
- [34] M. Manassi, B. Sayim, and M. H. Herzog. “When crowding of crowding leads to uncrowding”. In: *Journal of Vision* 13.13 (Nov. 2013), pp. 10–10. ISSN: 1534-7362. DOI: 10.1167/13.13.10. URL: <http://dx.doi.org/10.1167/13.13.10>.
- [35] Davide Zoccolan et al. “A rodent model for the study of invariant visual object recognition”. In: *Proceedings of the National Academy of Sciences* 106.21 (May 2009), pp. 8748–8753. ISSN: 1091-6490. DOI: 10.1073/pnas.0811583106. URL: <http://dx.doi.org/10.1073/pnas.0811583106>.
- [36] Takafumi Minamimoto, Richard C. Saunders, and Barry J. Richmond. “Monkeys Quickly Learn and Generalize Visual Categories without Lateral Prefrontal Cortex”. In: *Neuron*

- 66.4 (May 2010), pp. 501–507. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2010.04.010. URL: <http://dx.doi.org/10.1016/j.neuron.2010.04.010>.
- [37] Richard Passingham. “How good is the macaque monkey model of the human brain?” In: *Current Opinion in Neurobiology* 19.1 (Feb. 2009), pp. 6–11. ISSN: 0959-4388. DOI: 10.1016/j.conb.2009.01.002. URL: <http://dx.doi.org/10.1016/j.conb.2009.01.002>.
- [38] Guy A. Orban, David Van Essen, and Wim Vanduffel. “Comparative mapping of higher visual areas in monkeys and humans”. In: *Trends in Cognitive Sciences* 8.7 (July 2004), pp. 315–324. ISSN: 1364-6613. DOI: 10.1016/j.tics.2004.05.009. URL: <http://dx.doi.org/10.1016/j.tics.2004.05.009>.
- [39] Dante Mantini et al. “Interspecies activity correlations reveal functional correspondence between monkey and human brain areas”. In: *Nature Methods* 9.3 (Feb. 2012), pp. 277–282. ISSN: 1548-7105. DOI: 10.1038/nmeth.1868. URL: <http://dx.doi.org/10.1038/nmeth.1868>.
- [40] Nikolaus Kriegeskorte et al. “Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey”. In: *Neuron* 60.6 (Dec. 2008), pp. 1126–1141. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2008.10.043. URL: <http://dx.doi.org/10.1016/j.neuron.2008.10.043>.
- [41] M. Ito et al. “Size and position invariance of neuronal responses in monkey inferotemporal cortex”. In: *Journal of Neurophysiology* 73.1 (Jan. 1995), pp. 218–226. ISSN: 1522-1598.

- DOI: 10.1152/jn.1995.73.1.218. URL: <http://dx.doi.org/10.1152/jn.1995.73.1.218>.
- [42] Sidney R Lehky and Keiji Tanaka. “Neural representation for object recognition in inferotemporal cortex”. In: *Current Opinion in Neurobiology* 37 (Apr. 2016), pp. 23–35. ISSN: 0959-4388. DOI: 10.1016/j.conb.2015.12.001. URL: <http://dx.doi.org/10.1016/j.conb.2015.12.001>.
- [43] E. Kobatake and K. Tanaka. “Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex”. In: *Journal of Neurophysiology* 71.3 (Mar. 1994), pp. 856–867. ISSN: 1522-1598. DOI: 10.1152/jn.1994.71.3.856. URL: <http://dx.doi.org/10.1152/jn.1994.71.3.856>.
- [44] Ichiro Fujita et al. “Columns for visual features of objects in monkey inferotemporal cortex”. In: *Nature* 360.6402 (Nov. 1992), pp. 343–346. ISSN: 1476-4687. DOI: 10.1038/360343a0. URL: <http://dx.doi.org/10.1038/360343a0>.
- [45] S. Yamane, S. Kaji, and K. Kawano. “What facial features activate face neurons in the inferotemporal cortex of the monkey?” In: *Experimental Brain Research* 73.1 (1988), pp. 209–214. ISSN: 1432-1106. DOI: 10.1007/bf00279674. URL: <http://dx.doi.org/10.1007/BF00279674>.
- [46] Yukako Yamane et al. “Representation of the Spatial Relationship Among Object Parts by Neurons in Macaque Inferotemporal Cortex”. In: *Journal of Neurophysiology* 96.6 (Dec. 2006), pp. 3147–3156. ISSN: 1522-1598. DOI: 10.1152/jn.01224.2005. URL: <http://dx.doi.org/10.1152/jn.01224.2005>.

- [47] Yukako Yamane et al. “A neural code for three-dimensional object shape in macaque inferotemporal cortex”. In: *Nature Neuroscience* 11.11 (Oct. 2008), pp. 1352–1360. ISSN: 1546-1726. DOI: 10.1038/nn.2202. URL: <http://dx.doi.org/10.1038/nn.2202>.
- [48] Irving Biederman. “Recognition-by-components: A theory of human image understanding.” In: *Psychological Review* 94.2 (Apr. 1987), pp. 115–147. ISSN: 0033-295X. DOI: 10.1037/0033-295x.94.2.115. URL: <http://dx.doi.org/10.1037/0033-295X.94.2.115>.
- [49] Matthew Turk and Alex Pentland. “Eigenfaces for Recognition”. In: *Journal of Cognitive Neuroscience* 3.1 (Jan. 1991), pp. 71–86. ISSN: 1530-8898. DOI: 10.1162/jocn.1991.3.1.71. URL: <http://dx.doi.org/10.1162/jocn.1991.3.1.71>.
- [50] K. Tanaka. “Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities”. In: *Cerebral Cortex* 13.1 (Jan. 2003), pp. 90–99. ISSN: 1460-2199. DOI: 10.1093/cercor/13.1.90. URL: <http://dx.doi.org/10.1093/cercor/13.1.90>.
- [51] Sidney R. Lehky and Anne B. Sereno. “Comparison of Shape Encoding in Primate Dorsal and Ventral Visual Pathways”. In: *Journal of Neurophysiology* 97.1 (Jan. 2007), pp. 307–319. ISSN: 1522-1598. DOI: 10.1152/jn.00168.2006. URL: <http://dx.doi.org/10.1152/jn.00168.2006>.
- [52] Shimon Edelman. “Representation is representation of similarities”. In: *Behavioral and Brain Sciences* 21.4 (Aug. 1998), pp. 449–467. ISSN: 1469-1825. DOI: 10.1017/s0140525x98001253. URL: <http://dx.doi.org/10.1017/s0140525x98001253>.

- [53] Gabriel Kreiman. *Biological and Computer Vision*. Cambridge University Press, 2021.
- [54] Philipp Bomatter et al. “When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes”. In: *CoRR* abs/2104.02215 (2021). arXiv: 2104.02215. URL: <https://arxiv.org/abs/2104.02215>.
- [55] Aude Oliva, Antonio Torralba, et al. “The role of context in object recognition”. In: *Trends in cognitive sciences* 11.12 (2007), pp. 520–527.
- [56] Moshe Bar. “VISUAL OBJECTS IN CONTEXT”. In: *Neuroscience* 5 (2004), pp. 617–629.
- [57] Christopher A. Henry and Adam Kohn. “Spatial contextual effects in primary visual cortex limit feature representation under crowding”. In: *Nature Communications* 11.1 (Apr. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-15386-7. URL: <http://dx.doi.org/10.1038/s41467-020-15386-7>.
- [58] Ruben Coen-Cagli, Adam Kohn, and Odelia Schwartz. “Flexible gating of contextual influences in natural vision”. In: *Nature Neuroscience* 18.11 (Oct. 2015), pp. 1648–1655. ISSN: 1546-1726. DOI: 10.1038/nn.4128. URL: <http://dx.doi.org/10.1038/nn.4128>.
- [59] Nikos K Logothetis, Jon Pauls, and Tomaso Poggio. “Shape representation in the inferior temporal cortex of monkeys”. In: *Current Biology* 5.5 (1995), pp. 552–563.

- [60] Claire Chambers et al. “Prior context in audition informs binding and shapes simple features”. In: *Nature Communications* 8.1 (Apr. 2017). ISSN: 2041-1723. DOI: 10.1038/ncomms15027. URL: <http://dx.doi.org/10.1038/ncomms15027>.
- [61] Charles Spence, Vanessa Harrar, and Betina Piqueras-Fiszman. “Assessing the impact of the tableware and other contextual variables on multisensory flavour perception”. In: *Flavour* 1.1 (May 2012). ISSN: 2044-7248. DOI: 10.1186/2044-7248-1-7. URL: <http://dx.doi.org/10.1186/2044-7248-1-7>.
- [62] Moshe Bar and Elissa Aminoff. “Cortical Analysis of Visual Context”. In: *Neuron* 38.2 (Apr. 2003), pp. 347–358. ISSN: 0896-6273. DOI: 10.1016/s0896-6273(03)00167-3. URL: [http://dx.doi.org/10.1016/s0896-6273\(03\)00167-3](http://dx.doi.org/10.1016/s0896-6273(03)00167-3).
- [63] Melissa Le-Hoa Võ. “The meaning and structure of scenes”. In: *Vision Research* 181 (Apr. 2021), pp. 10–20. ISSN: 0042-6989. DOI: 10.1016/j.visres.2020.11.003. URL: <http://dx.doi.org/10.1016/j.visres.2020.11.003>.
- [64] Hillel Adesnik et al. “A neural circuit for spatial summation in visual cortex”. In: *Nature* 490.7419 (Oct. 2012), pp. 226–231. ISSN: 1476-4687. DOI: 10.1038/nature11526. URL: <http://dx.doi.org/10.1038/nature11526>.
- [65] Mark E. Auckland, Kyle R. Cave, and Nick Donnelly. “Nontarget objects can influence perceptual processes during object recognition”. In: *Psychonomic Bulletin amp; Review* 14.2 (Apr. 2007), pp. 332–337. ISSN: 1531-5320. DOI: 10.3758/bf03194073. URL: <http://dx.doi.org/10.3758/BF03194073>.

- [66] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. “Scene perception: Detecting and judging objects undergoing relational violations”. In: *Cognitive Psychology* 14.2 (Apr. 1982), pp. 143–177. ISSN: 0010-0285. DOI: 10.1016/0010-0285(82)90007-x. URL: [http://dx.doi.org/10.1016/0010-0285\(82\)90007-X](http://dx.doi.org/10.1016/0010-0285(82)90007-X).
- [67] Andrew Hollingworth. “Does consistent scene context facilitate object perception?”. In: *Journal of Experimental Psychology: General* 127.4 (1998), pp. 398–415. ISSN: 0096-3445. DOI: 10.1037/0096-3445.127.4.398. URL: <http://dx.doi.org/10.1037/0096-3445.127.4.398>.
- [68] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. “Putting visual object recognition in context”. In: *CoRR* abs/1911.07349 (2019). arXiv: 1911.07349. URL: <http://arxiv.org/abs/1911.07349>.
- [69] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *European Conference on Computer Vision*. 2014. URL: <https://api.semanticscholar.org/CorpusID:14113767>.
- [70] Javier Portilla. In: *International Journal of Computer Vision* 40.1 (2000), pp. 49–70. ISSN: 0920-5691. DOI: 10.1023/a:1026553619983. URL: <http://dx.doi.org/10.1023/A:1026553619983>.
- [71] Kohitij Kar et al. “Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior”. In: *Nature Neuroscience* 22.6 (Apr. 2019), pp. 974–983. ISSN: 1546-1726. DOI: 10.1038/s41593-019-0392-5. URL: <http://dx.doi.org/10.1038/s41593-019-0392-5>.

- [72] Hamidreza Ramezanpour, Christopher Giverin, and Kohitij Kar. *Low-cost, portable, easy-to-use kiosks to facilitate home-cage testing of non-human primates during vision-based behavioral tasks*. 2024. DOI: 10.48550/ARXIV.2401.03727. URL: <https://arxiv.org/abs/2401.03727>.
- [73] Denis Fize, Maxime Cauchoix, and Michèle Fabre-Thorpe. “Humans and monkeys share visual representations”. In: *Proceedings of the National Academy of Sciences* 108.18 (Apr. 2011), pp. 7635–7640. ISSN: 1091-6490. DOI: 10.1073/pnas.1016213108. URL: <http://dx.doi.org/10.1073/pnas.1016213108>.
- [74] E. Wachsmuth, M. W. Oram, and D. I. Perrett. “Recognition of Objects and Their Component Parts: Responses of Single Units in the Temporal Cortex of the Macaque”. In: *Cerebral Cortex* 4.5 (1994), pp. 509–522. ISSN: 1460-2199. DOI: 10.1093/cercor/4.5.509. URL: <http://dx.doi.org/10.1093/cercor/4.5.509>.
- [75] Mark A. Pinsk et al. “Representations of faces and body parts in macaque temporal cortex: A functional MRI study”. In: *Proceedings of the National Academy of Sciences* 102.19 (Apr. 2005), pp. 6996–7001. ISSN: 1091-6490. DOI: 10.1073/pnas.0502605102. URL: <http://dx.doi.org/10.1073/pnas.0502605102>.
- [76] Nikolaos C. Aggelopoulos, Leonardo Franco, and Edmund T. Rolls. “Object Perception in Natural Scenes: Encoding by Inferior Temporal Cortex Simultaneously Recorded Neurons”. In: *Journal of Neurophysiology* 93.3 (Mar. 2005), pp. 1342–1357. ISSN: 1522-1598. DOI: 10.1152/jn.00553.2004. URL: <http://dx.doi.org/10.1152/jn.00553.2004>.

- [77] Károly Köteles et al. “Coding of images of materials by macaque inferior temporal cortical neurons”. In: *European Journal of Neuroscience* 27.2 (Jan. 2008), pp. 466–482. ISSN: 1460-9568. DOI: 10.1111/j.1460-9568.2007.06008.x. URL: <http://dx.doi.org/10.1111/j.1460-9568.2007.06008.x>.
- [78] Jessica Loke et al. “Human Visual Cortex and Deep Convolutional Neural Network Care Deeply about Object Background”. In: *Journal of Cognitive Neuroscience* 36.3 (2024), pp. 551–566. ISSN: 1530-8898. DOI: 10.1162/jocn_a_02098. URL: http://dx.doi.org/10.1162/jocn_a_02098.
- [79] Bruno Rossion and Jessica Taubert. “What can we learn about human individual face recognition from experimental studies in monkeys?” In: *Vision Research* 157 (Apr. 2019), pp. 142–158. ISSN: 0042-6989. DOI: 10.1016/j.visres.2018.03.012. URL: <http://dx.doi.org/10.1016/j.visres.2018.03.012>.