Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Artificial Intelligence (37^{th}cycle)

Efficient Neural Coding Under Resource Constraints: A Rate-Distortion Theory Perspective

By

Leo D'Amato

******

Supervisor(s):
Dr. Giovanni Pezzulo, Supervisor
Dr. Aldo Gangemi, Co-Supervisor

Doctoral Examination Committee:
Prof. Antonella Maselli, Referee, University of Messina
Prof. Tom Verguts, Referee, Ghent University
Prof. Simone Cutini, University of Padua
Prof. Stefano Di Carlo, Polytechnic University of Turin
Dr. Francesco Donnarumma, National Research Council of Italy

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Leo D'Amato
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Abstract

Cognitive systems, whether biological or artificial, operate under inherent resource limitations that necessitate efficient coding strategies. This thesis explores the fundamental principles governing efficient coding through the framework of Rate-Distortion Theory (RDT), investigating how limitations on information representation shape the geometry of latent spaces in generative models and contribute to the emergence of cognitive abilities, specifically number sense. The central argument is that resource constraints, formalized by RDT, impose a trade-off between the fidelity of information representation and the capacity of the encoding system, leading to systematic distortions in internal models of the world.

To explore the geometry of efficient codes, a series of experiments were conducted using β-Variational Autoencoders (β-VAEs). By varying model capacity, introducing biases in the training data, and imposing specific task objectives, the research identifies and characterizes key types of geometric distortions, including prototypization (collapsing similar stimuli into prototypical representations), specialization (over-representing frequent stimuli), and orthogonalization (forming task-specific representations). These distortions, which emerge as principled adaptations to RDT constraints, highlight the trade-offs between accuracy and resource allocation in learned representations.

The second line of inquiry investigates the emergence of number sense in β-VAEs trained solely to reconstruct visual scenes. This work demonstrates that a capacity for numerical perception can arise spontaneously under rate-distortion principles, without explicit supervision for counting or numerical tasks. Behavioral and neural analyses reveal key signatures of human number sense, including Weber's law scaling and distinct mechanisms for processing small versus large numerosities. Furthermore, severely limiting model capacity induces deficits in numerical repre-

sentations that parallel those observed in developmental dyscalculia, suggesting a direct link between resource constraints and cognitive impairments.

This thesis provides compelling computational evidence for the unifying role of RDT in understanding the emergence of intelligent behavior. It demonstrates how efficient coding principles manifest in complex, non-linear systems and offers a framework to understand how statistical properties of the environment shape neural representations. By integrating information theory, cognitive science, and artificial intelligence, this work advances our understanding of how resource limitations sculpt cognitive representations and highlights the potential of RDT to explain the emergence of complex cognitive abilities in both biological and artificial systems.

Contents

Chapter 1

Introduction

Operating efficiently under limited resources is a fundamental constraint shared by all intelligent systems, from biological organisms to artificial agents. Energy consumption, neural tissue availability, and processing time itself are finite commodities. To navigate the complexity of the world effectively, these systems must therefore employ strategies for efficient information processing. This imperative for efficiency is not merely a matter of optimization, it is a defining characteristic of intelligent systems, shaping their architecture and their modes of operation. Efficient coding, the ability to represent relevant information using minimal resources, becomes not just advantageous, but essential for survival and for achieving complex cognitive functions. This inherent need for efficiency influences how we perceive, remember, decide, and act within our environments.

However, this drive for efficiency comes at a cost. The very act of compressing information, of distilling the essential from the superfluous, inevitably introduces a degree of distortion. Representations within resource-limited systems cannot be perfect mirror images of the external world. Instead, they are necessarily simplified, abstracted, and shaped by the constraints of the encoding process. This distortion is not simply random noise or error, it is a structured consequence of efficient coding. Understanding the nature of this distortion, how it is systematically introduced and shaped, is crucial to understanding how cognitive systems function. These imperfections in representation are not arbitrary failings, but rather potentially adaptive features, reflecting the system's optimized balance between fidelity and resource expenditure. The challenge then becomes not to eliminate distortion entirely - an im-

possible feat given resource limitations - but to characterize its forms and understand its functional role in efficient cognition.

Rate-Distortion Theory (RDT) offers a robust theoretical framework for addressing this challenge. Originally developed within the field of communication theory to understand the limits of transmitting signals through noisy channels, RDT provides a principled approach to analyzing the trade-off between the accuracy of information transmission and the resources required to achieve it. The core idea is that for any given limitation on resources, there exists an optimal way to encode information that minimizes the loss of relevant details, or 'distortion'. We can draw an analogy between communication systems and cognitive systems: just as a communication channel has a limited capacity to transmit information, cognitive systems have limited resources for processing and storing information. RDT provides a mathematical language to describe and analyze this trade-off, defining 'rate' as the measure of resources used for encoding and 'distortion' as the measure of information loss. Importantly, RDT is a normative theory, meaning it specifies what the best possible encoding strategy is, given certain constraints. This normative aspect makes it particularly valuable for understanding biological cognition, suggesting that brains, through evolutionary and developmental pressures, may have converged on solutions that approximate these theoretically optimal strategies. Therefore, RDT provides not just a descriptive tool, but a predictive framework for understanding the principles governing efficient neural coding and the emergence of structured representations.

To explore these principles in the context of complex, real-world data, this thesis turns to generative models, specifically Beta-Variational Autoencoders (β-VAEs). Generative models, trained using unsupervised learning techniques, learn to capture the underlying statistical structure of data and generate new samples from that structure. This process of learning to generate data inherently involves learning compressed, latent representations. β-VAEs are particularly well-suited for investigating rate-distortion trade-offs in neural systems for several reasons. First, their architecture and loss function are explicitly designed to approximate the principles of RDT. The loss function of a β-VAE directly balances reconstruction accuracy (minimizing distortion) against the complexity of the latent representation (constraining rate). Second, β-VAEs allow for direct manipulation of the encoding capacity through a key parameter, allowing us to systematically study the effects of varying resource limitations. Third, β-VAEs are capable of learning from and generating complex, high-dimensional data like images, providing a powerful tool

to investigate efficient coding in domains relevant to biological perception and cognition. By using β-VAEs as a computational model, we can directly examine how rate-distortion principles shape the emergent properties of latent representations when learning from rich sensory inputs.

This thesis pursues two complementary lines of inquiry to explore the implications of Rate-Distortion Theory for cognitive representations. The first line of investigation focuses on uncovering the general principles that govern how efficient coding shapes the geometry of latent spaces in generative models. Using a novel dataset of spatially structured images, this research investigates how varying model capacity, introducing biases in the training data, and imposing specific task objectives affect the organization and structure of the latent representations learned by β-VAEs. This line of inquiry seeks to identify and characterize fundamental types of geometric distortions that emerge under different RDT constraints. These distortions, including the tendency towards prototypical representations, the specialization of resources for frequent or task-relevant stimuli, and the orthogonalization of representations under task demands, reveal how efficient coding principles manifest in the internal geometry of learned representations. By studying these general principles, we aim to build a foundational understanding of how resource limitations shape the way information is encoded and represented in complex systems.

The second line of inquiry delves into a specific and fundamental cognitive capacity: number sense. Numerosity perception, the intuitive ability to estimate and discriminate quantities without explicit counting, plays a crucial role in adaptive behaviors across species. This line of research investigates whether number sense can emerge spontaneously in unsupervised β-VAEs trained solely to reconstruct visual scenes containing varying numbers of objects. Furthermore, it explores how the encoding capacity of the model influences the accuracy, precision, and psychophysical characteristics of this emergent number sense. Specifically, we examine whether the model's performance aligns with known psychophysical laws, such as Weber's law, which describes the scaling of discrimination thresholds with magnitude. Crucially, this investigation also examines whether imposing severe capacity limitations on the model leads to deficits in numerosity perception that parallel those observed in developmental dyscalculia, a learning disability characterized by difficulties with number processing. By focusing on number sense as a concrete example, this line of inquiry aims to demonstrate how RDT can provide a principled account for the emergence of specific cognitive abilities and their susceptibility to resource constraints.

The contributions of this thesis are multifaceted, advancing our understanding of efficient coding in both general and specific cognitive domains. Firstly, it provides compelling computational evidence for the unifying role of Rate-Distortion Theory in shaping the geometry of latent representations in generative models, demonstrating how fundamental principles of information theory manifest in complex, non-linear systems. Secondly, it identifies and characterizes key types of distortions – prototypization, specialization, and orthogonalization – that emerge as principled adaptations to varying constraints on capacity, data distribution, and task demands. These findings offer a taxonomy of distortions and a framework for understanding their functional significance in efficient coding. Thirdly, this thesis provides a novel computational account for the unsupervised emergence of number sense, demonstrating that a capacity for numerosity perception can arise spontaneously in generative models trained under rate-distortion principles, without explicit supervision for counting or numerosity tasks. Fourthly, it reveals how resource limitations can induce deficits in numerosity perception, suggesting a link between resource constraints and cognitive impairments. Finally, by bridging information theory, computational neuroscience, and artificial intelligence, this thesis offers a valuable framework for future research aimed at understanding the fundamental principles that govern the emergence of intelligent behavior in both biological and artificial systems.

The structure of this thesis is as follows: Chapter 2 presents the computational framework and tools. Specifically it introduces the principles of rate-distortion theory, the beta variational autoencoders and the main applications of rate-distortion theory in cognitive science. Chapter 3 presents the results of the first line of inquiry. Chapter 4 presents the results of the second line of inquiry. Finally, Chapter 5 concludes the thesis.

Chapter 2

Rate-Distortion Theory

This chapter lays the methodological foundation for the investigations presented in this thesis. It begins by providing a comprehensive introduction to Rate-Distortion Theory (RDT), tracing its historical origins and elucidating its core mathematical principles. Following this theoretical exposition, the chapter transitions to a discussion of Beta-Variational Autoencoders (β-VAEs), detailing their architecture and loss function, and critically highlighting their role as a powerful computational tool for approximating the principles of RDT in the context of complex, high-dimensional data. The chapter then explicitly establishes the theoretical link between the β-VAE loss function and the fundamental tenets of Rate-Distortion Theory, justifying the use of β-VAEs as the primary computational instrument for exploring the research questions posed in this thesis. Finally, the chapter broadens the scope to consider the applications of RDT within cognitive science, highlighting how this framework provides a normative perspective on efficient coding in the brain.

2.1 A normative framework for efficient coding

Rate-Distortion Theory (RDT) stands as a cornerstone of information theory, providing a rigorous mathematical framework for understanding the fundamental limits of data compression and the inherent trade-off between data rate and distortion. Its genesis can be traced back to the seminal work [1] of Claude Shannon in the late 1940s and 1950s, during the burgeoning era of information theory and digital communication. Shannon's groundbreaking contributions laid the groundwork for

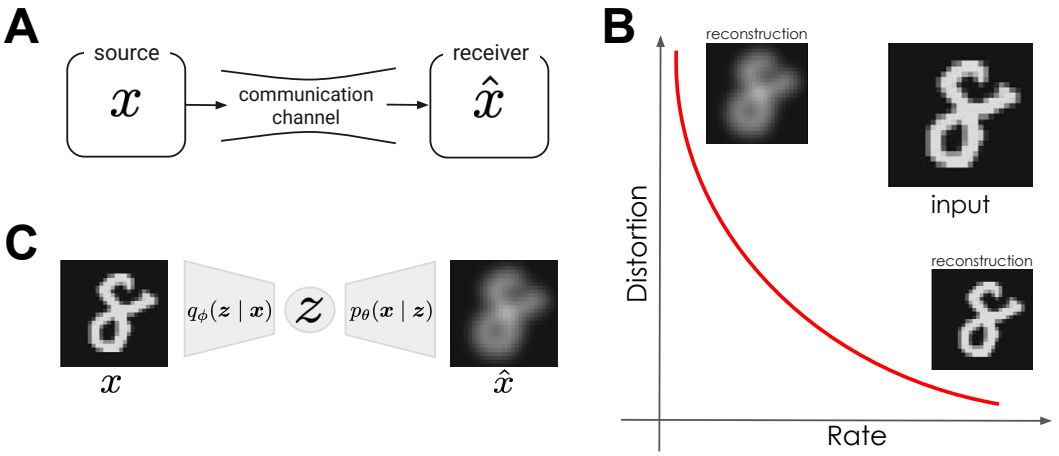

Figure 2.1: Rate-distortion Theory. (A) The problem of transmitting a message from a source to a receiver through a communication channel with limited capacity, as originally formulated by Shannon. (B) Rate-distortion trade-off. (C) Architecture of a Variational Autoencoder.

understanding the theoretical limits of reliably transmitting information across noisy channels. Extending this foundation, Rate-Distortion Theory, formalized in Shannon's 1959 paper "Coding theorems for a discrete source with a fidelity criterion" [2] directly addressed the problem of lossy data compression, moving beyond the realm of perfect, lossless transmission to consider scenarios where some level of information loss is acceptable, or even necessary, to achieve efficient encoding.

Historically, RDT emerged from the practical challenges of communication engineering, where bandwidth limitations and noise inevitably constrain the amount of information that can be transmitted from a source to a receiver over a communication channel [2] (Figure 2.1A). However, its conceptual reach extends far beyond this initial context. The core principles of RDT, concerning the efficient allocation of resources under constraints and the management of information fidelity, are profoundly relevant to a wide range of fields, including signal processing [3], image and video compression [4], and, as we argue in this thesis, cognitive science and neuroscience [5]. Indeed, the brain, operating under severe metabolic and physical constraints, can be viewed as a prime example of a resource-limited information processing system, suggesting that principles analogous to those formalized in RDT may govern the efficiency and structure of neural representations.

At its heart, Rate-Distortion Theory seeks to answer a fundamental question: given a source of information and a constraint on the available resources for encoding

this information (the 'rate'), what is the minimum achievable distortion in the reconstructed information? To formalize this question mathematically, RDT considers a source of information, represented by a random variable X with a probability distribution $p(x)$. The goal is to encode this information into a compressed representation, denoted by Z, and then decode this representation to obtain a reconstruction $\hat{X}$. The encoding process is formally described by a conditional probability distribution $q(z|x)$, representing the encoder, and the decoding process by $p(\hat{x}|z)$, representing the decoder.

The key concepts within RDT are:

- **Rate (R):** The rate quantifies the average amount of information, typically measured in bits or nats, required to represent the source X in the compressed space Z. In the context of RDT, the rate is typically defined as the mutual information between the source X and its encoded representation Z:

$$R(q(z|x)) = I(X;Z) \tag{2.1}$$

  In other terms, the rate R essentially measures the reduction in uncertainty about X gained by observing Z. It also can be interpreted as the average number of bits or nats needed to describe Z.

- **Distortion (D):** The distortion measures the average loss of fidelity in the reconstruction $\hat{X}$ compared to the original source X. This loss is quantified by a distortion function $d(x,\hat{x})$, which measures the discrepancy between an original source value x and its reconstruction $\hat{x}$. The average distortion is then given by:

$$D(q(z|x), p(\hat{x}|z)) = \mathbb{E}_{p(x)q(z|x)p(\hat{x}|z)}\left[d(x,\hat{x})\right] \tag{2.2}$$

  The choice of distortion function $d(x,\hat{x})$ as defined in Equation (2.2) is crucial and depends on the specific application and the nature of the data. Commonly used distortion measures include the squared Euclidean distance for continuous data and the Hamming distance for discrete data. For image data, pixel-wise squared error or more perceptually relevant metrics can be employed.

- **Distortion-Rate Function D(R):** The central object of study in RDT can also be framed in terms of the distortion-rate function $D(R)$. This function represents the minimum achievable distortion for a given rate budget, where

the rate is constrained to be less than or equal to a certain value C. Formally, it is defined as:

$$D(R) = \min_{q(z|x),p(\hat{x}|z)} D(q(z|x), p(\hat{x}|z)) \quad \text{subject to} \quad R(q(z|x)) \leq C \tag{2.3}$$

The distortion-rate function $D(R)$, as shown in Equation (2.3), thus characterizes the fundamental limit of compression from a different perspective. It defines a lower bound on the distortion for any encoder-decoder pair that operates within a given rate constraint C. In other words, for a system with a limited encoding capacity C, RDT seeks to find the encoding and decoding strategies that minimize the unavoidable distortion, leading to the distortion-rate function $D(R)$. The constrained optimization problem in Equation (2.3) can be equivalently formulated using the Lagrangian:

$$\mathscr{L} = D(q(z|x), p(\hat{x}|z)) + \beta \Big(R(q(z|x)) - C \Big) \tag{2.4}$$

where β is the Lagrange multiplier that sets the trade-off between minimizing the distortion $D(q(z|x), p(\hat{x}|z))$ and satisfying the rate constraint $I(X;Z) \leq C$. Minimizing this Lagrangian represents the core objective of rate-distortion theory when framed in terms of the distortion-rate function.

The rate-distortion function typically exhibits a monotonically decreasing relationship: as the allowed resources R increases, the minimum achievable distortion $D(R)$ decreases (Figure 2.1B). This embodies the fundamental trade-off: to achieve higher fidelity (lower distortion), more resources (higher rate) are needed. Conversely, to reduce resource usage (lower rate), some fidelity must be sacrificed (higher distortion). The shape of the rate-distortion function is determined by the statistical properties of the source X and the chosen distortion measure $d(x, \hat{x})$.

In practice, finding the exact rate-distortion function and the optimal encoder-decoder pair is often analytically intractable, especially for complex sources like images. However, RDT provides invaluable theoretical insights and serves as a normative benchmark against which practical compression algorithms can be evaluated. Furthermore, the conceptual framework of RDT, emphasizing the trade-off between rate and distortion, provides a powerful lens through which to understand efficient coding in diverse domains, including the study of neural representations and cognitive systems.

2.2 Approximating the rate-distortion trade-off with deep learning

While Rate-Distortion Theory provides a valuable normative framework, its direct application to complex, high-dimensional data, such as images, poses significant computational challenges. Beta-Variational Autoencoders (β-VAEs) [6] have emerged as a powerful class of generative models that offer a practical approach to approximating the principles of RDT within the domain of deep learning.

β-VAEs are a variant of Variational Autoencoders (VAEs) [7], which are themselves probabilistic generative models based on variational inference. VAEs learn to map high-dimensional input data X into a lower-dimensional latent representations Z and then to generate reconstructions $\hat{X}$ from this latent representation. The architecture of a VAE (Figure 2.1C) typically consists of two main neural network components:

- **Encoder** $q_\phi(z|x)$**:** The encoder network, parameterized by ϕ, learns to map an input data point x to a distribution in the latent space Z. In practice, this is often implemented by parameterizing a Gaussian distribution in Z with a mean and covariance that are functions of the input x, i.e., $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$. The encoder effectively compresses the input x into a probabilistic latent representation z.
- **Decoder** $p_\theta(x|z)$**:** The decoder network, parameterized by θ, performs the inverse mapping, taking a latent vector z and generating a reconstruction $\hat{x}$ in the input space. Similar to the encoder, the decoder can also be probabilistic, often modeled as a Gaussian or Bernoulli distribution depending on the data type, i.e., $p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$ for continuous data or $p_\theta(x|z) = \text{Bernoulli}(x; \sigma(\nu_\theta(z)))$ for binary data, where σ is the sigmoid function and $\nu_\theta(z)$ parameterizes the Bernoulli distribution.

The standard VAE is trained by maximizing the Evidence Lower Bound (ELBO) [7] on the log-likelihood of the data. This is equivalent to minimize the negative ELBO and it can be expressed as the following loss function:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi; x, z) = \mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)] + D_{KL}(q_\phi(z|x)||p(z)) \tag{2.5}$$

where $p(z)$ is a prior distribution over the latent space, typically chosen to be a standard Gaussian $\mathcal{N}(0,I)$. The first term in the ELBO, $\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]$, as seen in Equation (2.5), is the reconstruction term, which encourages the decoder to generate reconstructions $\hat{x}$ that are similar to the original inputs x. This term can be seen as minimizing the distortion in the reconstruction. The second term, $D_{KL}(q_\phi(z|x)||p(z))$, also part of Equation (2.5), is the Kullback–Leibler (KL) divergence between the encoder's posterior distribution $q_\phi(z|x)$ and the prior $p(z)$. This term acts as a regularizer, encouraging the learned latent distribution to be close to the prior distribution. In the context of RDT, this KL divergence term can be interpreted as a constraint on the rate, limiting the complexity of the latent code.

The Beta-VAE (β-VAE) extends the standard VAE by introducing a hyperparameter β that scales the KL divergence term:

$$\mathscr{L}_{\beta\text{-VAE}}(\theta,\phi;x,z,\beta) = \underbrace{\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]}_{\approx D} + \beta \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\approx R} \tag{2.6}$$

The hyperparameter β, in Equation (2.6), provides a crucial control over the trade-off between reconstruction accuracy and latent space regularization. By increasing β to values larger than 1, we increase the weight of the KL divergence term in the loss function. This forces the encoder to learn latent representations that are closer to the prior distribution, effectively reducing the rate or complexity of the latent code, but potentially at the cost of increased reconstruction distortion. Conversely, setting $\beta = 1$ recovers the standard VAE loss, while $\beta < 1$ would prioritize reconstruction over latent space regularization, although this is less commonly explored in the context of disentanglement or efficient coding.

The connection between the β-VAE loss function and Rate-Distortion Theory becomes apparent when we examine the terms in Equation (2.6) in light of the RDT framework. The reconstruction term, $\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(x|z)]$, which encourages accurate reconstruction, directly corresponds to minimizing the distortion D in RDT, as defined in Equation (2.2). Specifically, if we consider the negative log-likelihood of the data under the decoder model, $-\log p_\theta(x|z)$, as a measure of distortion between the original input x and its reconstruction from latent code z, then minimizing the expectation of this quantity over the encoder distribution $q_\phi(z|x)$ effectively

minimizes the average reconstruction distortion. This correspondence becomes even more evident when we consider the common case where the decoder $p_\theta(x|z)$ is modeled as an isotropic Gaussian distribution, i.e., $p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \sigma^2 I)$. In this Gaussian case, the negative log-likelihood term, assuming a fixed variance σ^2, simplifies to:

$$-\log p_\theta(x|z) = \frac{1}{2\sigma^2}||x - \mu_\theta(z)||^2 + \text{constant}.$$

Ignoring the constant term and scaling factor $\frac{1}{2\sigma^2}$, minimizing $-\log p_\theta(x|z)$ becomes equivalent to minimizing the squared Euclidean distance $||x - \mu_\theta(z)||^2$ between the input x and the mean of the decoder's output $\mu_\theta(z)$, which serves as the reconstruction $\hat{x} = \mu_\theta(z)$. This squared Euclidean distance, or Mean Squared Error (MSE), is a widely used distortion measure d in Rate-Distortion Theory, particularly for continuous data.

The KL divergence term, $D_{KL}(q_\phi(z|x)||p(z))$, which regularizes the latent space, can be seen as approximating the rate R in RDT, as defined in Equation (2.1). In the context of variational inference, this term arises from the need to approximate the intractable true posterior distribution $p(z|x)$. VAEs employ an encoder $q_\phi(z|x)$ as a variational approximation to this true posterior. By minimizing the KL divergence between $q_\phi(z|x)$ and a prior distribution $p(z)$, we are essentially constraining the complexity of the approximate posterior and, consequently, the amount of information that the latent representation z can encode about the input x. If the prior $p(z)$ is chosen to be simple, such as a standard Gaussian, then pushing the approximate posterior $q_\phi(z|x)$ towards $p(z)$ limits the capacity of the latent space to deviate significantly from this simple prior, thereby implicitly limiting the rate.

Thus, minimizing the β-VAE loss function can be interpreted as approximately solving the rate-distortion optimization problem. By varying β, we directly manipulate the emphasis on rate versus distortion, exploring the spectrum of solutions along the rate-distortion curve. Increasing β prioritizes rate reduction (latent space regularization) at the potential expense of increased distortion (reconstruction error), while decreasing β allows for lower distortion but potentially at the cost of a more complex and less efficient latent representation.

It is crucial to emphasize that the β-VAE loss function and the VAE framework as a whole provide just an approximation to the true Rate-Distortion Theory optimization. The approximation arises from several factors: First, VAEs rely on variational inference to approximate the intractable true posterior $p(z|x)$. Second,

both the encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$ are parameterized by neural networks with limited capacity and specific architectural biases. These parametric choices can constrain the space of possible encoder-decoder pairs and may not perfectly achieve the optimal solutions predicted by RDT. Third, using the negative log-likelihood as a distortion measure and the KL divergence as a rate proxy are themselves approximations. In particular, the KL divergence term in Equation 2.6 serves as a tractable upper bound on the mutual information, i.e., the rate term in rate-distortion theory (Equation 2.1). In fact, by considering the joint distribution over data and latent codes as $q_\phi(x,z) = p(x)q_\phi(z|x)$ and also considering the mutual information $I_q(X;Z)$ between X and Z under $q_\phi(x,z)$, it is possible to write the KL term of the β-VAE loss as:

$$
\begin{aligned}
&\mathbb{E}_{p(x)} D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) = \\
&\quad = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p(z)} \right] \\
&\quad = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{q(z)} + \log \frac{q(z)}{p(z)} \right] \\
&\quad = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{q(z)} \right] + \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q(z)}{p(z)} \right] \\
&\quad = \mathbb{E}_{p(x)} D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, q(z)) + \mathbb{E}_{q(z)} \left[\log \frac{q(z)}{p(z)} \right] \\
&\quad = I_q(X;Z) + D_{\mathrm{KL}}(q(z) \,\|\, p(z)) \\
&\quad \geq I_q(X;Z)
\end{aligned}
$$

where $q(z) = \int p(x)q_\phi(z|x)\,dx$ is the marginal over the latent variables. Therefore, by minimizing the KL divergence term, the β-VAE is implicitly minimizing an upper bound on the rate. While these approximations are common and effective choices, they may not perfectly capture the ideal distortion and rate measures in all scenarios.

Despite these approximations, the β-VAE framework provides a remarkably powerful and practical tool for investigating the core principles of Rate-Distortion Theory in the context of complex data and neural network models. By training β-VAEs and systematically varying the hyperparameter β, we can empirically explore the trade-off between rate and distortion, observe how different levels of capacity constraint shape the emergent properties of latent representations, and gain valuable

insights into the principles of efficient coding that may be relevant to biological cognitive systems. Therefore, the β-VAE serves as the central computational instrument in this thesis, allowing us to probe the research questions concerning the geometry of efficient codes and the emergence of cognitive functions under resource constraints.

2.3 Applications in Cognitive Science

The conceptual framework of Rate-Distortion Theory, initially conceived for communication systems, has found increasing relevance and application within the field of cognitive science [5].

The brain, as a biological information processing system, operates under stringent resource constraints, including limitations in metabolic energy, neural tissue, and processing time [8–10]. This inherent resource scarcity necessitates efficient coding strategies, leading to the hypothesis that the brain may optimize its neural representations to maximize information content while minimizing resource expenditure [9]. Rate-Distortion Theory provides a normative lens through which to examine this hypothesis, offering a framework for understanding how the brain achieves efficient and adaptive representations of the world.

The core principle of efficient coding, deeply intertwined with Rate-Distortion Theory, suggests that the brain aims to create internal representations that are both informative and economical. This principle is supported by a wealth of evidence across various cognitive domains.

In perception, for example, efficient coding predicts that neural systems should prioritize the encoding of novel or unexpected stimuli, while efficiently representing predictable or redundant information [11, 12].

Attneave [11], in his work on informational aspects of visual perception, argued that sensory processing is designed to reduce redundancy. He specifically highlighted that we tend to focus our perceptual resources on contours and points of high information content, effectively ignoring areas of uniform color or texture, a form of efficient encoding.

Barlow [12] built upon this, proposing the "redundancy reduction hypothesis." He suggested that neurons in the visual system are tuned to represent statistical regularities in the natural environment, effectively "explaining away" predictable

patterns and leaving only the unexpected or novel aspects to be explicitly encoded. This allows the brain to represent sensory information with fewer active neurons. This is reflected in phenomena such as sensory adaptation and predictive coding, where neural responses are attenuated to predictable sensory inputs, freeing up resources for processing novel or behaviorally relevant information [13].

Młynarski and Hermundstad [13] investigated how neurons adapt to dynamic stimuli. They found that adaptation is not simply a reduction in sensitivity, rather, it's an optimization process where neurons become more sensitive to changes in the stimulus relative to the recent history. This adaptation has been observed in the fly visual system [14], allowing the animal to efficiently track moving objects and respond to changes in the visual environment, even in the presence of noisy or fluctuating inputs.

Similarly, in working memory, the limited capacity of this cognitive system necessitates efficient encoding and compression of information to maintain relevant data online for short-term use [15–17].

Jakob and Gershman [15] directly applied rate-distortion theory to model working memory limitations. Their model shows that phenomena like set size effects, stimulus prioritization, serial dependence, and systematic biases in working memory tasks, can be explained as optimal solutions to the problem of encoding information under a limited rate (bit-rate) constraint. The proposed model is a population coding model with spiking neurons and they tested it on five datasets with human subjects and one dataset with monkey subjects performing delayed response tasks.

Brady, Konkle, and Alvarez [16] demonstrated that visual working memory exploits statistical regularities in the environment to compress information. When displays contain redundant structure (e.g., frequently co-occurring color pairs), memory performance improves, not because of increased capacity per se, but because structured stimuli allow for more efficient, compressed encoding. Their results support the view that working memory operates as a resource-limited system that adaptively reallocates representational resources by leveraging regularities to reduce redundancy—thereby increasing the effective number of items stored without violating capacity constraints. More specifically, they presented two experiments contrasting memory capacity for displays where colors were presented in random pairs versus in recurring patterns. In the first experiment, the colors that formed a pattern were presented as part of the same object. In the second experiment, the

colors that formed a pattern were presented on two different but spatially adjacent objects. They proposed a quantitative model of how learning the statistics of the input would allow observers to form more efficient representations of the displays and used a compression algorithm to demonstrate that observers' performance approaches what would be optimal if their memory had a fixed capacity in bits. In addition, they illustrated that a discrete model of chunking also fits experimental data. The degree of compression possible from the display was highly correlated with behavior, suggesting that people optimally take advantage of statistical regularities to remember more information in working memory.

Mathy and Feldman [17] redefined memory "chunks" through the lens of data compression (Kolmogorov complexity), demonstrating that working memory capacity depends on information compressibility rather than fixed item limits. It reconciles the classic "7 ± 2" (Miller) and "4 ± 1" (Cowan) capacity estimates by showing that uncompressed data (e.g., random digits) approaches Miller's limit, while highly compressible patterns (e.g., ordered sequences) align with Cowan's stricter constraint. Memory performance reflects efficient encoding - simpler patterns require fewer chunks, allowing apparent capacity to vary. In their experiments, subjects had to recall digit sequences, with some sequences showing more regular patterns that could be more compressed than others. Their computational analysis calculates the complexity of a string as the sum of the space required to encode each "run" (a monotonic series of digits with a constant increment), thus quantifying the degree to which a sequence can be compressed.

Sims, Jacobs, and Knill [18] developed a theoretical framework to quantify the capacity limitations of human visual working memory using rate–distortion theory. This framework defines visual memory capacity in a task-independent manner and predicts human performance under varying conditions. The authors validated these predictions through empirical studies, demonstrating that a model allowing variability in the number of stored memory representations — without assuming a fixed item limit — adequately accounts for observed data. In the experiments, participants had to recall the angular orientation of arrows or the length of line segments after a brief presentation and a retention interval. The set size (number of items in the display) and variance of the features were manipulated. Their model framed visual working memory as an information transmission channel, where encoding and decoding processes were optimized to minimize distortion subject to a capacity limit, and

where sensory input was corrupted by noise. The model assumed a Bayesian decoder and allowed variability in the number of encoded items.

Sims [19] employed rate–distortion theory to elucidate the trade-off between information retention and error minimization in visual working memory. By applying this theoretical framework, the study derives the optimal memory system tailored to specific task demands, revealing that the inherent loss function deviates from traditional assumptions. This approach underscores the task-dependent nature of memory errors, offering a rational basis for evaluating existing models of visual working memory. In the experiments, participants were asked to remember visual features like color or orientation from a display of triangles, or the length of line segments, and later report their memory in a delayed recall task. The computational model used inverse decision theory to estimate the implicit loss function from empirical data, assuming that observed behavior is optimally efficient for some underlying loss function.

Nagy, Török, and Orbán [20] introduced a semantic compression framework to elucidate systematic distortions in human episodic memory, positing that such distortions arise from the brain's reliance on an internal generative model adapted to environmental statistics. Utilizing deep generative models (β-VAEs), the authors analyze datasets—including chess games, natural text, and hand-drawn sketches—to demonstrate that memory compression via this generative model accounts for phenomena such as expertise-related distortions, gist-based recall errors, contextual influences, and changes in memory over time. This approach suggests that the brain optimally manages memory resources by compressing experiences in a manner that inherently leads to predictable distortions.

In spatial cognition, for instance, the tendency for cognitive maps – mental representations of spatial environments – to exhibit shape compactness and biases towards prototypical forms [21, 22] can be interpreted as a consequence of efficient coding under capacity constraints.

Costa and Bonetti [21] specifically examined systematic distortions in large-scale geographical cognitive maps using a sketching methodology. Key findings reveal consistent geometrical transformations in mental representations of geographic regions. For example, participants rotated Italy's shape toward vertical, overestimated its width, and underestimated its height, leading to a more compact form. Sardinia was similarly distorted and mislocated northward. Across European capitals, dis-

tances were compressed and cities were aligned to horizontal and vertical axes. These findings indicate a strong tendency toward simplification, regularization, and centralization in mental representations of geographic space. This suggests a form of compression where detailed geometric information is lost, but the overall "gist" of the shape is retained.

Tversky [22] demonstrated that cognitive maps are systematically distorted with respect to physical maps. These distortions are not random degradations but arise from cognitive organizing principles, including hierarchical organization, perspective effects, and reliance on landmarks or reference points. Specifically, spatial memory is shaped by categorization (e.g., grouping locations by state or function), leading to biases in direction, distance, and alignment judgments. Additionally, people tend to mentally rotate or align spatial figures toward canonical orientations and frames of reference, resulting in further distortions. These phenomena suggest that cognitive maps integrate multiple sources of information in a non-coherent, constructionist manner, undermining the view of memory as a veridical but compressed/regularized image of the world. This allows to reduce cognitive load, to simplify spatial representations and to facilitate efficient navigation and spatial reasoning.

Muhle-Karbe et al. [23] investigated how goal-directed behavior distorts the neural representation of allocentric space. They found that spatial maps in the human hippocampus and orbitofrontal cortex were "compressed" so that locations cued as goals were coded together in neural state space, and these distortions predicted successful learning. This effect was captured by a computational model in which current and prospective locations are jointly encoded in a place code. In their fMRI experiment, participants navigated an avatar through a grid-world environment consisting of four interlinked rooms. A contextual cue indicated the conditional dependence of one goal location on another, and participants had to visit two successive goal locations. The computational model encoded the avatar's location using simulated Gaussian place fields and was equipped with parameters for orthogonalization, separation, and compression of spatial representations based on context. This model generated representational dissimilarity matrices (RDMs) which were then compared to BOLD signals to infer how spatial codes were distorted.

In decision-making, systematic biases in the representation of value and probability [24] can also be viewed through the lens of RDT, suggesting that these distortions

may reflect optimal strategies for making decisions under limited computational resources, balancing accuracy with cognitive cost [25–27].

Juechems, Balaguer, Spitzer, and Summerfield [24] investigated why humans distort the value and probability of prospects when making economic choices, leading to seemingly irrational decisions. The authors propose that these distortions are optimal adaptations that maximize reward and minimize uncertainty, given the constraint of finite computational precision (irreducible noise in neural computation). Through simulations and experiments, they show that canonical nonlinear forms of utility functions arise from the interplay between reward maximization, uncertainty minimization, and intrinsic decision noise, and that humans adapt optimally to manipulations of outcome certainty.

Bhui, Lai, and Gershman [25] reviewed recent work that aims to reconcile seemingly contradictory views of human decision-making as both rational and irrational. It focuses on a "resource-rational" analysis, which models human decisions as optimal under cognitive resource constraints, connecting psychology and neuroscience to ideas from information theory. The paper highlights the implications of an information-theoretic formalization of cognitive resources for understanding reference dependence, stochastic choice, and perseveration. These phenomena, traditionally viewed as irrational biases, are suggested to arise from a rational solution to resource-limited decision-making.

Gershman, Horvitz, and Tenenbaum [26] discussed the convergence of AI, cognitive science, and neuroscience on a shared view of intelligence as computational rationality. It outlines advancements in addressing challenges of perception and action under uncertainty, emphasizing the development of representations, inference procedures, and mechanisms for reflecting on tradeoffs in effort, precision, and timeliness of computations. Computational rationality is identifying decisions with the highest expected utility, while considering the costs of computation in complex real-world problems, aligning with the principles of rate-distortion theory.

The convergence of evidence from diverse areas of cognitive science – spanning perception, working memory, spatial cognition, and decision-making – strongly supports the principle of efficient coding as a fundamental organizing principle of neural representation. The cited studies, employing a range of experimental paradigms and theoretical frameworks, demonstrate that the brain consistently seeks to represent information in a manner that balances informativeness with resource

expenditure. The observed phenomena, from sensory adaptation and ensemble coding in working memory to distortions in cognitive maps and biases in decision-making, are not merely quirks or limitations of the cognitive system. Instead, they can be interpreted as structured and potentially adaptive consequences of operating under inherent resource constraints. Rate-Distortion Theory provides a powerful normative framework for understanding *why* these phenomena occur, offering a principled account of how limited resources shape the nature of neural representations and the types of distortions that inevitably arise. The findings reviewed here highlight the importance of considering computational limitations and the trade-off between accuracy and efficiency when investigating cognitive processes, suggesting that many seemingly irrational behaviors or cognitive biases may, in fact, reflect optimal solutions within the constraints of a resource-limited brain.

Furthermore, the principle of efficient coding connects to a prominent family of theories that frame the brain as an active inference engine. This perspective posits that the brain builds and maintains an internal generative model of the world to predict and explain the hidden causes of its sensory inputs [28]. Frameworks such as the Bayesian Brain hypothesis [29], Predictive Coding [30], and the Free Energy Principle [31, 32] suggest that the brain continuously generates predictions about incoming sensory data and primarily processes the "prediction error" — the mismatch between expectation and reality. This provides a powerful, unifying account for perception, action, and learning. However, much of this literature has traditionally focused on modeling the inferential process itself, assuming the structure of the generative model is largely known [33, 34]. The central question is often how an agent uses its pre-existing model to perceive and act, rather than how the model itself is acquired from raw sensory experience. Conversely, studies that do tackle the problem of learning the generative model often operate with simplified, low-dimensional inputs or abstract and small state spaces [35–37], which may not fully capture the statistical complexity of the real-world sensory signals, like vision, that the brain must learn from . This thesis, instead, focuses on the learning of the generative model from high-dimensional, unstructured inputs and uses of β-Variational Autoencoders to do so under the principles of efficient coding.

Moreover, the framework of β-VAEs is particularly relevant to this thesis because of its demonstrated property to learn disentangled representations [38]. A disentangled representation is one in which the distinct, underlying generative factors of the observed data are captured by separate dimensions in the model's latent space.

For example, in a visual scene, these factors might correspond to an object's position, shape, or color. By constraining the information capacity of the latent bottleneck, β-VAEs encourage the model to discover and isolate these fundamental sources of variation, as this is a highly efficient way to encode the data. This process of disentanglement is not only a consequence of data statistics but can also be driven by task objectives [39]. When a system needs to solve a specific task, it is often advantageous to develop representations that clearly separate task-relevant information from irrelevant details. This creates a more abstract and robust internal generative model that is specialized for the required computations. Several studies have shown that humans also learn such abstract, disentangled representations, as this strategy is crucial for generalizing knowledge from familiar to novel situations [40, 41, 23]. This property of learning disentangled representations, shaped by the dual pressures of efficient data compression and task utility, provides not only another justification for adopting β-VAEs as the computational models but also justifies the specific design choices for models' architectures used in the experiments.

The following chapters will build upon this foundation, exploring specific instantiations of RDT principles in generative models, to provide a more detailed and mechanistic understanding of *how* efficient coding shapes the geometry of latent representations and influences cognitive functions.

Chapter 3

The geometry of efficient codes

This chapter is adapted from the paper: "D'Amato L., Lancia G.L. and Pezzulo G., *The geometry of efficient codes: how rate-distortion trade-offs distort the latent representations of generative models*" published by PLOS Computational Biology.

3.1 Introduction

As described in Section 2.3, rate-distortion theory elucidates *why* factors such as encoding capacity, data likelihood, and utility for goal attainment can distort latent representations in the brain by formalizing the trade-off between information rate (the average number of bits per stimulus used for encoding) and distortion (the cost of reconstruction errors).

However, while rate-distortion theory provides a normative principle to understand why latent representations are distorted, we still lack a systematic understanding of *how* they are distorted under different experimental conditions.

To address this challenge, in this chapter we systematically examine the distortions in the geometry of latent representations (embeddings) of generative models, specifically Beta Variational Autoencoders (β-VAEs), under varying constraints of model capacity, data distributions, task objectives or their combinations. We focus on two sets of experiments that involve memorizing sets of images representing simple spatial maps composed of two "corridors", placed at the upper and lower parts of the image, respectively (Figure 3.1). The first experiment allows us to test

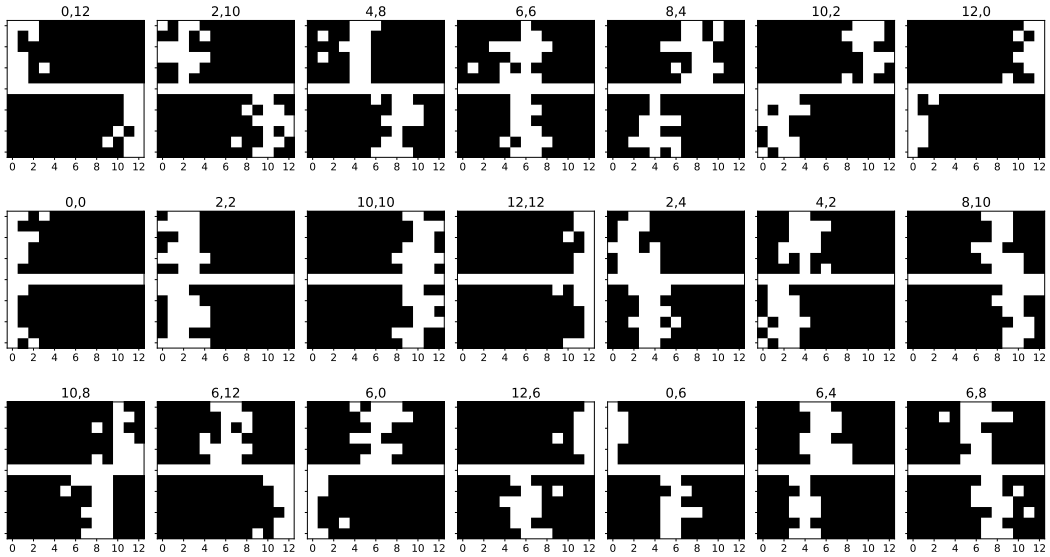

Figure 3.1: The *Corridors* dataset. The figure shows 21 example images in the "corridors" dataset used for this study. Each figure comprises two white corridors, placed at the upper and at the lower parts of the image, with a white horizontal line that is common to all the images. Each corridor is a noisy vertical line of white pixels, whose true center is in one of 13 x positions, coded from 0 (left) to 12 (right). For example, in the first image, the upper corridor is in the position $x_{UC} = 0$ and the lower corridor is in the position $x_{LC} = 12$. The corridor positions are reported on top of each image.

the distortions of latent representations that arise under different capacities and data distributions, such as familiar or unfamiliar images. In this experiment, models are trained to remember images from balanced or unbalanced datasets, where the to-be-remembered image appears more or less frequently. The second experiment enables us to test the distortions of latent representations emerging under different capacities and tasks/goals. Here, models are trained to perform various image classification tasks, such as determining whether the upper corridor is to the left or right of the bottom corridor, or whether the corridors are aligned or not. For instance, for the first image in Figure 3.1, the correct answers would be "yes" in the first case and "no" in the second case.

As motivated in Section 2.2, we employ a β-VAE [6] as computational model since it has been demonstrated to approximate rate-distortion theory [42, 38] when dealing with complex stimuli such as images [43, 20]. Furthermore, in the second experiment set, in order to study the effect of task objectives in shaping the latent space of the model, we modify the β-VAE architecture by supplementing it by

a series of classifiers, each trained in a supervised manner to address a specific classification task. In all the experiments, the dimension of the latent space of the models is fixed to 5 neurons. For more details on the specific architectures used in the two experiment sets, refer to Section 3.4.1.

The dataset, the code and the models' checkpoints of our experiments are available at: https://github.com/damat-le/geom_eff_codes.

3.2 Results

3.2.1 Problem specification: the two experiments

Here, we perform two experiments. *Experiment 1* addresses the distortions induced by statistically biased training sets by comparing the representations learned by the same neural network, first trained on an unbiased dataset, and then trained on a biased one. *Experiment 2* addresses the distortions induced by the classification task assigned to the network. To do this, we compared the representation learned by the network when the only objective was to reconstruct input images against the representation learned when the objective was augmented with a classification task. In both experimental sets, we trained multiple networks by varying their encoding capacity to study how this impacts on the magnitude of distortions (see below).

3.2.2 The Corridors dataset

To study the emergence of disentangled or distorted latent representations in generative models, we designed a novel dataset – the Corridors dataset – in which two generative factors vary orthogonally.

The Corridors dataset consists of 200.000 black and white 13x13 images. Each image is divided into two sections, upper and lower, by an horizontal white line of width 1 pixel. Each section includes a vertical noisy "corridor", i.e., a white vertical core segment with white random pixels around it (Figure 3.1). The white random pixels are independent samples of a Bernoulli distribution $f(g(x;\mu,\sigma))$, where $g(x;\mu,\sigma)$ is a Gaussian probability density function with standard deviation $\sigma = 1.2$ and mean μ equal to the position of the vertical core segment. The position

of the vertical core segment, also referred to as the position of the corridor, is an integer number ranging from 0 to 12, since each corridor can slide horizontally from pixel 0 to pixel 12 across images. Thus, each image is associated with two labels, (x_{UC}, x_{LC}), where x_{UC} denotes the position of the upper corridor and x_{LC} denotes the position of the lower corridor. The positions of the upper and lower corridors represent the two orthogonal factors of variation of the dataset. The Corridors dataset is available at: https://github.com/damat-le/geom_eff_codes.

3.2.3 Experiment 1: distortions induced by varying model capacity and data distributions

In Experiment 1, we examine the distortions of the β-VAE representations, as an effect of model capacity and statistically biased training sets, in which some types of images are more frequent than others.

For this, we train 3 models: baseline, E1M1, and E1M2. The baseline β-VAE model is trained on the Corridors dataset. The E1M1 model is a β-VAE trained on an unbalanced dataset, in which the images with the lower corridor on the left appear 10 times more often than images with the lower corridor on the right. Let x_{LC} be the pixel at which the lower corridor is centered, then "lower corridor on the left" means that $x_{LC} \leq 6$, where $x_{LC} \in \{0, 1, \ldots, 12\}$. The E1M2 model is a β-VAE trained on an unbalanced dataset in which the images having the two corridors aligned ($x_{LC} = x_{UC}$) appear 10 times more often than images with non-aligned corridors.

We repeat the training of all the models (baseline, E1M1, E1M2) for 5 different values of encoding capacity $C_{\max} \in \{0.3, 1, 3, 6, 10\}$ nats.

Performance of the models

Figure 3.2A shows the reconstruction loss of the three β-VAE models trained on the balanced dataset (baseline), the unbalanced dataset with a bias for lower corridor to the left (E1M1), and the unbalanced dataset with a bias for corridors aligned (E1M2), at the 5 different capacities of $C_{\max} \in \{0.3, 1, 3, 6, 10\}$ nats. As expected, for all models, reconstruction loss decreases as the model's capacity increases. The decrease of reconstruction loss follows the same trend across all models.

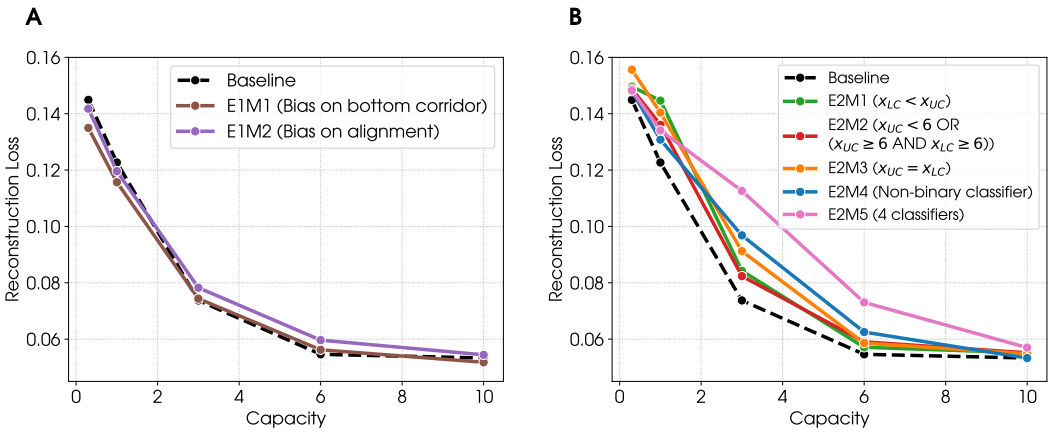

Figure 3.2: Reconstruction loss, all the models. For ease of reading, a brief description of each model along with its label is reported in the legend. (A) Experiment 1. The figure illustrates that increasing capacity reduces reconstruction loss. The trend is similar for the baseline β-VAE model that is trained with a balanced dataset and for the two models (E1M1 - E1M2) that are trained with unbalanced datasets. (B) Experiment 2. The figure illustrates that at any given capacity, the baseline model has a smaller reconstruction loss compared to the hybrid models that are additionally trained to solve classification tasks (E2M1 - E2M5). Furthermore, at any given capacity, reconstruction loss changes across the different tasks and is worst for the E2M5 model, which addresses four classification tasks simultaneously. See the main text for explanation.

Geometry of the latent representations of the models

We next compare the representations learned by the three models, when trained at high ($C_{\max} = 10$ nats) and low capacity ($C_{\max} = 0.3$ nats).

Baseline model, at high and low capacity Figures 3.3A and F show 2D projections of the latent spaces learned by the baseline β-VAE model trained on the balanced dataset, at high and low capacity, respectively. In these and the subsequent figures, each dot corresponds to the position of a specific image within the low-dimensional latent space of the generative model. For ease of interpretation, some images are labeled with corridors positions, such as "6,6" in the center. The colors of the dots correspond to the classes of images that have a high frequency (orange) or low frequency (green) in the E1M1 and E1M2 models. These colors are irrelevant for the baseline model, but are retained for ease of comparison with the E1M1 and E1M2 models, see later.

Figure 3.3A-B shows that, at high capacity, the representation is perfectly *disentangled*. It is possible to notice that the geometry of the latent space is a square, with the two axes corresponding to the two generative factors of variation of the dataset: namely, the positions of the two corridors. Moving along one axis while keeping the other fixed affects the position of only one of the two corridors. For example, moving from the top-left to the top-right angles of the square corresponds to moving from the image (0,0) to the image (12,0). In other words, the first factor (the position of the top corridor, from 0 to 12) changes, whereas the second factor (the position of the bottom corridor, from 0 to 12) remains the same. The opposite is true if one moves from the top-left to the bottom-left angles of the square. The disentanglement of the internal code is confirmed by the activation patterns of the latent channels, shown in Figure 3.3B. The figure shows the average (standardized) activity of each of the 5 neurons of the latent space as a heatmap. Each heatmap reports the neural activity for every possible combination of the corridors' positions (x_{UC}, x_{LC}) of the input images (see Section 3.4.3). In the figure, we observe that only two channels are active, each of which encoding the position of a single corridor (the color changes only along one of the two axes). Moreover, the color transition is smooth, indicating that the model is able to encode every single value of the position of each corridor.

Figure 3.3F-G shows that when baseline model is trained at low capacity, the representation is still disentangled, as evident from the fact that it uses two latent

channels for two orthogonal dimensions of the task (Figure 3.3G). Furthermore, the latent space maintains a square-like shape (Figure 3.3F), even if the scales of the two axes is much smaller than the in Figure 3.3A and some points are flipped between the two latent spaces (e.g., 0,12 and 10,10). Importantly, we observe a *prototypization* effect: the model collapses the representations of images that are similar to one another. This is evident from the fact that the data points are clustered around the vertices of the square that correspond to four main type of images (or prototypes): both corridors on the left (2,2), both corridors on the right (10,10), upper corridor on the left and lower corridor to the right (2,10), upper corridor on the right and lower corridor on the left (10,2). The prototypization effect is also apparent when comparing the activation patterns of the latent channels in the baseline model at high (Figure 3.3B) and low (Figure 3.3G) capacity. At low capacity, the color transitions are no longer smooth as for the model at high capacity; rather, each channel presents a more binarized activation as if it is able to encode only two possible positions of the corresponding corridor. Note that the observations made in this paragraph about panels A, B, F and G hold true for the Figures 3.3, 3.4, 3.6, 3.7, 3.8, 3.9 and 3.10. In fact, they all represent the same baseline model, except for the fact that that data points have been assigned different colors to facilitate comparison with other models.

Comparison of baseline and E1M1 models Figures 3.3C and 3.3H show 2D projections of the latent spaces learned by the β-VAE trained on the unbalanced dataset with a bias for the bottom corridor to the left (E1M1), at high and low capacity, respectively. At high capacity (Figure 3.3C), the data imbalance slightly alters the structure of the latent space and it is still possible to recognize a square-like shape. Despite this similarity, the activation patterns of the latent channels for the unbalanced dataset (Figure 3.3D) differ substantially from the baseline model, with four active channels. The first and the fourth channels encode the position of a single corridor in a binarized fashion (this explains the square-like shape and the two clusters). Rather, the second and third channels enrich the latent representation with fine-grained details about corridors positions, which allow the network to reconstruct more faithfully the input stimuli. Since the number of active channels in the E1M1 model is larger than two, it can be helpful to also inspect the 3D projection of its latent space, reported in Figure A2 (left panel). The figure reveals that, as an effect of the data imbalance, the latent space has been slightly bent to form a cylinder-like

shape. This reconfiguration in a higher-dimensional space leads to a general dilation of the distances between points, an effect that may not be apparent from the 2D projection. The distortion matrix shown in Figure Figure 3.3E provides formal confirmation of this effect.

As a consequence, the data points in the E1M1's latent space are further apart with respect to the baseline model' s latent space, despite in the 2D projection seems to be the contrary, as confirmed by Figure 3.3E. This matrix quantifies the extent to which the latent space of the E1M1 model has been compressed (values < 1) or dilated (values > 1) relative to the latent space of the baseline model. Each element (i, j) represents the average compression/dilation of the pairwise distances between data points of class i and data points of class j. The predominantly larger-than-1 values across the matrix confirm that the E1M1 model allocates more space to its representations compared to the baseline.

At low capacity (Figure 3.3H), the number of active channels returns back to two (one per corridor), as shown in Figure 3.3I. In particular, the first channel encodes upper corridor positions while the the second channel only encodes bottom corridor positions ≤ 6, (i.e. the most frequent images), ignoring images with bottom corridor position >6 (i.e., the least frequent images); note that the color corresponding to $x_{LC} > 6$ is equal to the color of inactive channels. This still results in a square-like shape of the latent space, but in this case the square only comprises the most frequent images. The less frequent images are ignored in the latent representation and when the network is presented with a image of type (x_{UC}, x_{LC}) with $x_{LC} > 6$, it generates an image of type $(x_{UC}, 3)$. This explains why the less frequent images are positioned at the center of the square. In conclusion, at low capacity, the network encodes the most frequent images (with relatively high fidelity), but not the less frequent images – plausibly, because ignoring rare images still ensures a small loss. We call this a *specialization* effect, in the sense that the model is biased in the allocation of resources. The specialization effect is similar to the prototypization, but is specific for frequent stimuli rather than forming averages. The magnitude of such specialization effect with respect to the baseline model at low capacity is quantified by the distortion matrix in Figure 3.3J. The matrix is computed as described above and the larger-than-1 value for the element $(1,1)$ confirms that most frequent stimuli have been dilated, while a smaller-than-1 value for the elements $(0,1)$ and $(1,0)$ confirms that the less frequent stimuli are ignored, in the sense that they are mapped into the same region of the latent space where high frequency stimuli live.

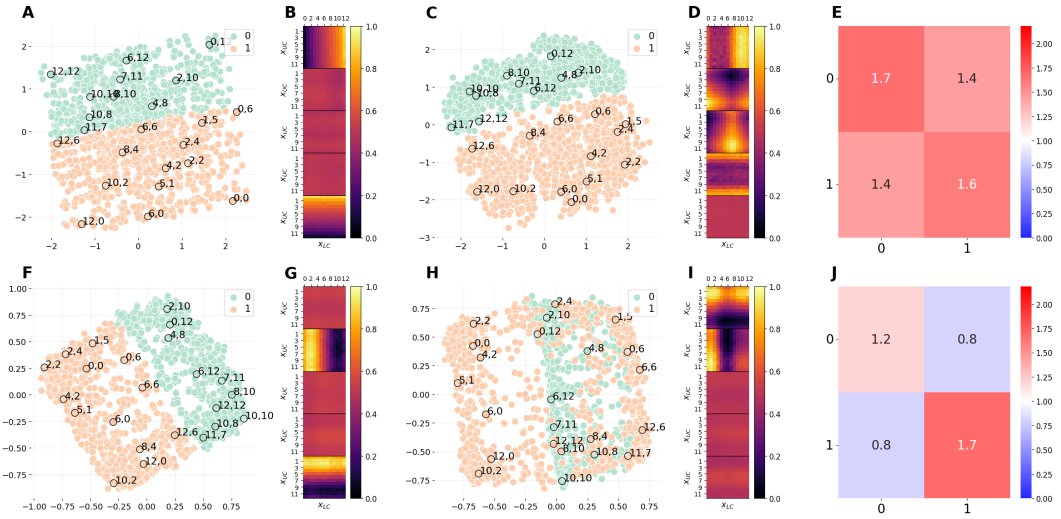

Figure 3.3: Comparison of the latent representations of the baseline model and the model E1M1. The E1M1 model is trained on an unbalanced dataset, in which images with $x_{LC} \leq 6$ (orange) are 10 times more frequent than images with $x_{LC} > 6$ (green). **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E1M1 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E1M1 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E1M1 model compared to the baseline model at high and low capacity, respectively.

Comparison of baseline and E1M2 models Figures 3.4A and 3.4F show 2D projections of the latent spaces learned by the β-VAE trained on the balanced dataset (baseline), at high and low capacity, respectively. These are identical to Figures 3.3A and 3.3F discussed above, except that the colors of the dots indicate the images that have high frequency (aligned corridors in orange) or low frequency (non-aligned corridors in green) in the E1M2 model.

Figures 3.4C and 3.4H show 2D projections of the latent spaces learned by the β-VAE trained on the unbalanced dataset with a bias for aligned corridors (E1M2), at high and low capacity, respectively. When the E1M2 model is trained with high capacity, both the latent space (Figure 3.4C) and the latent channels (Figure 3.4D) differ significantly from the baseline model (Figure 3.4A). Notably, this reconfiguration requires using more than two latent channels (Figure 3.4D), which implies that the latent space is higher dimensional compared to the baseline model. By inspecting the 3D projection in Figure A3 (left panel), we observe that, as for the E1M1 model, the latent space has been bent to form a cylinder-like shape. The same observations on the dilation of the data points made for the E1M1 model holds true for the E1M2 model, as confirmed by the distortion matrix in Figure 3.4E.

Similar considerations hold for the E1M2 trained with low capacity (Figure 3.4H). However, in this case the *specialization* effect – and the significant compression required to encode the points belonging to class 1 (aligned corridors) with sufficient fidelity – moves them farther apart in latent space compared to the baseline model (Figure 3.4J). As a consequence, the overall latent space appears significantly reconfigured and folded (Figure 3.4H). The folding can be appreciated in greater detail by considering Supplementary Figure A3, which shows the latent representation in 3 dimensions.

Summary of the results of Experiment 1

Experiment 1 explored the geometry of latent space that emerge in β-VAE models, by varying model capacity (i.e., high or low capacity) and data distributions (i.e., balanced vs. unbalanced datasets).

To summarize, our results show that that a baseline β-VAE trained with a balanced data distribution correctly recovers the two generative factors of variation of the dataset (i.e., the positions of the upper and lower corridors, which vary indepen-

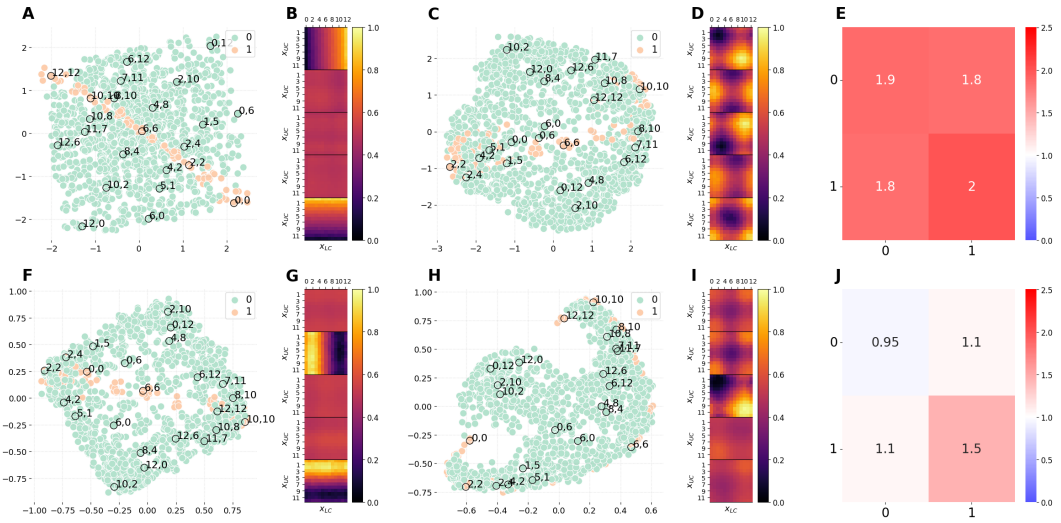

Figure 3.4: Comparison of the latent representations of the baseline model and the hybrid model E1M2. The E1M2 model is trained on an unbalanced dataset in which images with $x_L = x_U$ (orange) are 10 times more frequent than images with $x_L \neq x_U$ (green). **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E1M2 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E1M2 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E1M2 model compared to the baseline model at high and low capacity, respectively.

dently). This is evident from the fact that the geometry of the latent states is a square and the two factors of variation are encoded separately, in two latent channels. When the baseline model is trained at low capacity, we found that the model performance decreases (as expected) and the disentanglement coexists with a prototypization effect: namely, the model collapses the representations of images that are similar to one another.

Rather, the models E1M1 and E1M2 trained on unbalanced datasets develop a latent representation that is distorted (folding) compared to the baseline model and this distortion is more evident at low capacity. In this case, the models E1M1 and E1M2 show *specialization*: they devote more resources to encode the most common stimuli and this changes the position of their corresponding latent compared to the baseline model. Interestingly, not only training on unbalanced datasets disrupts the disentanglement but it also renders the latent representations higher-dimensional. This is plausibly because it is easier to distinguish common from rare stimuli in more dimensions.

3.2.4 Experiment 2: distortions induced by varying model capacity and tasks

In Experiment 2, we examine the distortions of the β-VAE embeddings, as an effect of varying model capacity and including classification tasks.

For this, we compare the same β-VAE baseline model considered above with five models: E2M1, E2M2, E2M3, E2M4, and E2M5. The first four models (E2M1 - E2M4) are hybrid models composed of a β-VAE and one classifier. These hybrid models are trained in a supervised manner on the same dataset. However, for each model the images are relabeled, depending on the classification task that the model has to solve. The classification tasks assigned to the four hybrid models are explained below:

1. The E2M1 model is assigned a binary classification task. We labeled with 1 the images in which $x_{UC} \leq x_{LC}$ and with 0 all the other images. In other words, the correct class is 1 if the upper corridor is to the left of or aligned with the right corridor, and 0 otherwise.

2. The E2M2 model is assigned a binary classification task. We labeled with 1 the images in which $x_{UC} < 6$ (regardless of the value of x_{LC}) OR $(x_{UC} \geq 6$ AND $x_{LC} \geq 6)$ and with 0 all the other images.

3. The E2M3 model is assigned a binary classification task. We labeled with 1 the images in which the corridors are aligned ($x_{LC} = x_{UC}$) and with 0 all the other images.

4. The E2M4 model is assigned a classification task with 25 classes. The 25 classes are defined as follows: we divided both upper and lower part of the image in 5 bins respectively. The upper corridor is assigned a label from 0 to 4 according to which bin it occupies. The bin size is not uniform and the bins for corridor positions are as follows: [0,1] (bin 0), [2,4] (bin 1), [5,7] (bin 2), [8,10] (bin 3), [11, 12] (bin 4). The same holds for the lower corridor. Each image is then assigned one of the 25 possible combinations of labels according to the positions of the two corridors. For example, an image with $(x_{UC}, x_{LC}) = (6, 11)$ has the upper corridor falling into bin 2 and the lower corridor into bin 4 and the it is assigned the label $2 * 5 + 4 = 14$.

Finally, E2M5 is a hybrid model composed of a β-VAE and 4 classifiers, trained to solve the 4 above tasks simultaneously.

For all the five hybrid models (E2M1 - E2M5), the β-VAE and classifier(s) are trained jointly end-to-end in a supervised manner on the same dataset, whose images are labeled depending on the task. All the classifiers are linear, except the one used in the E2M4 model, which is non-linear, since a linear classifier is not sufficient to solve the task. More details about the architecture of the hybrid models are reported in Section 3.4.1.

We repeat the training of all the models (E2M1, E2SM2, E2M3, E2M4, E2M5) for 5 different values of encoding capacity $C_{\max} \in \{0.3, 1, 3, 6, 10\}$ nats.

Performance of the models

Figure 3.2B shows the reconstruction loss of the baseline β-VAE model and the hybrid β-VAE - classifier models, trained in each of the four classification tasks (E2M1 - E2M4) and in all the four classification tasks simultaneously (E2M5). As expected, in all cases, reconstruction loss decreases as the model's capacity increases.

Furthermore, as expected, the reconstruction loss is better for the baseline model compared to the hybrid models, since the latter have to optimize an additional (classification) objective. The decrease of reconstruction loss follows the same trend in all models, but the loss remains slightly higher for the model that solves the four classification tasks simultaneously (E2M5).

We next considered to what extent the baseline and hybrid models generalize to novel tasks, beyond those for which they were trained. For this, we assessed the performance of a frozen baseline β-VAE model, when linear classifiers are trained afterwards (Figure 3.5A). We found that using this sequential training regime permits solving all but one of the tasks, for which the embedding learned through unsupervised learning is not sufficient for downstream classification tasks, using a linear classifier. Furthermore, we found the sequential training regime in which the β-VAE is trained before the classifiers (Figure 3.5A) to be less efficient than the joint training regime of the hybrid models (E2M1 - E2M5), in which the β-VAE and the classifiers are trained jointly (Figure 3.5B,C), across all the model / capacity combinations.

We next asked whether the classification tasks have the same of different capacity demands and what this implies for the latent representations developed by the models. We found that while almost all models reach a high F1 score, they require different capacities to do so. For example, in Figure 3.5B, the models E2M1 and E2M3 reach a F1 score of 0.8 with a low capacity (0.3), whereas the model E2M4 reaches the same F1 score of 0.8 only with a high capacity (10). Since capacity influences reconstruction loss, two models having the same F1 score can have different reconstruction losses. For example, models E2M1 and E2M3 with a capacity of 0.3 (that is sufficient for a F1 score of 0.8) have a reconstruction loss of 0.15, whereas model E2M4 with a capacity of 10 has a reconstruction loss of 0.06 (Figure 3.2B). Interestingly, this results indicate that greater discrimination difficulty during learning induces better and more distinct input representations, which is in keeping with empirical findings [44].

Finally, we considered to what extent the hybrid models trained on one task generalize to other tasks when the representations of the β-VAE are frozen and novel classifiers are trained (Figure 3.5D). Comparing Figure 3.5A with Figures 3.5D permits appreciating that in some instances, training the β-VAE before the classifiers (Figure 3.5A) leads to better performance compared to training jointly the β-VAE

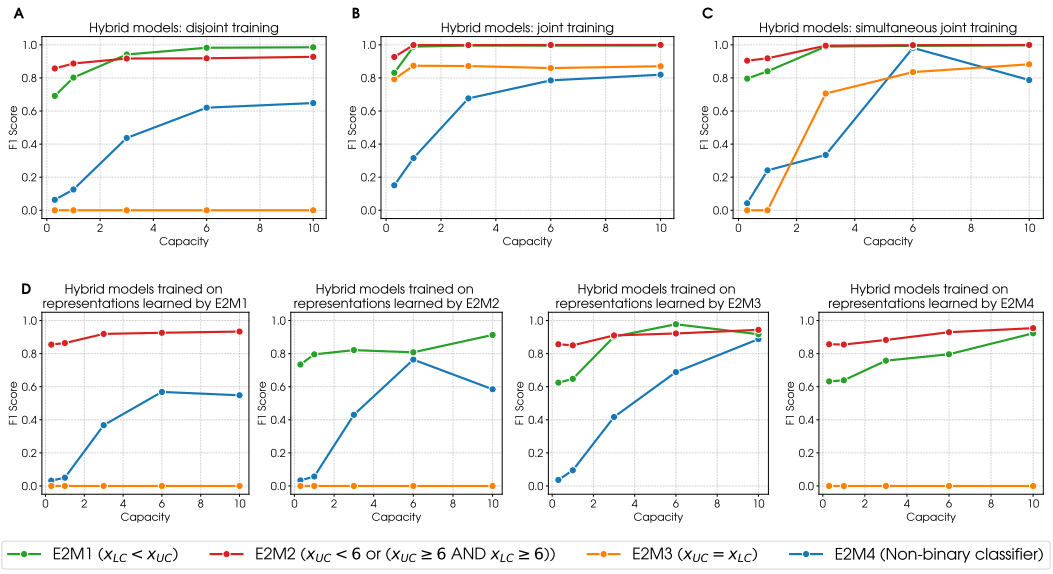

Figure 3.5: Performance analysis. For ease of reading, a brief description of each model along with its label is reported in the legend. **(A)** We first trained the baseline β-VAE model. Its frozen latent embeddings are then used to train separately 4 linear classifiers, one per task. This permits solving all tasks, except one. **(B)** We trained a β-VAE and classifier jointly and we repeated it for each task. In this case the classifier can modify the latent representation of the β-VAE. In fact, with respect to case A, the new representation is optimised for the given task and as result the network is able to solve the task1 and learn faster and better in the other tasks. **(C)** We trained only one β-VAE jointly with 4 classifiers (one per task). Such network is able to solve all the task simultaneously. **(D)** In this plot we asses the overspecialization of the networks trained in B. Consider the network trained on task 0. We use its latent representation (specialised for task0) to train 3 classifiers on the other tasks. What we can see is that task1 can never be solved using the representation learnt from an other task while this is not true for task 0, 2 and 3 (nevertheless, performance is worse than case B).

and one classifier and later train another classifier (Figures 3.5D). However, this advantage is not systematic. Finally, comparing Figure 3.5B and Figure 3.5C permits appreciating that in many instances, jointly training multiple classifiers ((Figure 3.5C)) achieves a performance that approximates models with one single classifier (Figure 3.5B), especially at high capacity.

Geometry of the representations of the models

We next considered what representations emerge in the baseline β-VAE model that only uses unsupervised learning and to what extent the joint training of classification tasks (E2M1 - E2M5) distorts these representations.

Baseline β-VAE model. The representation learned by the baseline β-VAE model at high and low capacity is depicted in Figures 3.6A and F, respectively. This is the same as Figure3.3A and F, except that the dots have different colors. The colors of the dots indicates the class to which the image belongs in the classification task under analysis: green denotes class 0, while orange signifies class 1. While the baseline β-VAE model has no classification task, we color the dots anyway (here, with the classes used by the E2M1 model) for ease of comparison with the other models, see later.

Comparison between baseline and E2M1 models. Figure 3.6 compares the latent representations learned by the baseline model at high capacity (Figure 3.6A) and low capacity (Figure 3.6F) with those learned by the hybrid model E2M1 at high capacity (Figures 3.6C) and low capacity (Figures 3.6H), respectively. These images help understand how the classification task distorts the latent representation of the model.

At high capacity (Figure 3.6C), the representation retains the square-like shape of the baseline network (Figures 3.6A), but the datapoints of the two different categories are more segregated. This is evident by considering Figure 3.6E, which measures an average dilation for images of distinct classes (in red). Interestingly, as elucidated by Figure 3.6D, the activation pattern of the latent channels of the E2M1 model completely changed and it is not aligned anymore with the two factors of variation in the dataset (i.e., there is no disentanglement) but rather it is a 45°-rotation of the baseline's latent representation: the axes of the new latent space correspond to the diagonals of the square in the baseline's latent space. In this way, the E2M1 model is able to solve the task by using only one latent channel and it uses the other latent factor to improve the fidelity of stimulus reconstruction.

When the E2M1 is trained at low capacity, the latent space reduces to a 1-dimensional space and only the axis allowing to solve the task is kept by the network

(see Figure 3.6I). As a consequence, the square-like shape is completely lost an all the data points collapse into two clusters, corresponding to the two categories that the model learned to discriminate (Figure 3.6H). In other words, the classification objective at low capacity induces an orthogonalization of representations, which is quantified in Figure 3.6J: data points belonging to the same class are pushed close to each (in blue) other while data points belonging to distinct classes are more separated (in red). This result is reminiscent of the empirical finding that learning visuomotor associations orthogonalizes population codes in the cortex [45].

In sum, at high capacity, the classification task induces a weak form of orthogonalization of the latent space, which is manifest in a better separation of the points of the two classes compared to the baseline model. Rather, at low capacity, the classification task induces a strong orthogonalization of the latent state, forming two clusters organized around prototypes of the two classes.

Comparison between baseline and E2M2 models. Figure 3.7 compares the latent representations learned by the baseline model at high capacity (Figure 3.7A) and low capacity (Figure 3.7F) with those learned by the hybrid model E2M2 at high capacity (Figures 3.7C) and low capacity (Figures 3.7H), respectively.

Due to the similarity with the task being solved by the E2M1 model (Figure 3.6), the same observations hold true, confirming previous results. However, in this task, the numerosity of stimuli belonging to the two categories is different and this is reflected in the different sizes of their two corresponding subspaces (green and orange areas) in the latent representation.

Comparison between baseline and E2M3 models. Figure 3.8 compares the latent representations learned by the baseline model at high capacity (Figure 3.8A) and low capacity (Figure 3.8F) with those learned by the hybrid model E2M3 at high capacity (Figures 3.8C) and low capacity (Figures 3.8H), respectively. As evident from Figure 3.8D, at high capacity, the E2M3 network presents four active latent channels: one of them (the third from the top in Panel D) only activates when the input image has the two corridors aligned, while the remaining three active channels encode the corridor positions, similar to the baseline model. Thus, interestingly, the E2M3 model does not only encode task positions but also task-relevant information into a separate, almost-binary latent channel (the third one). The usefulness of this

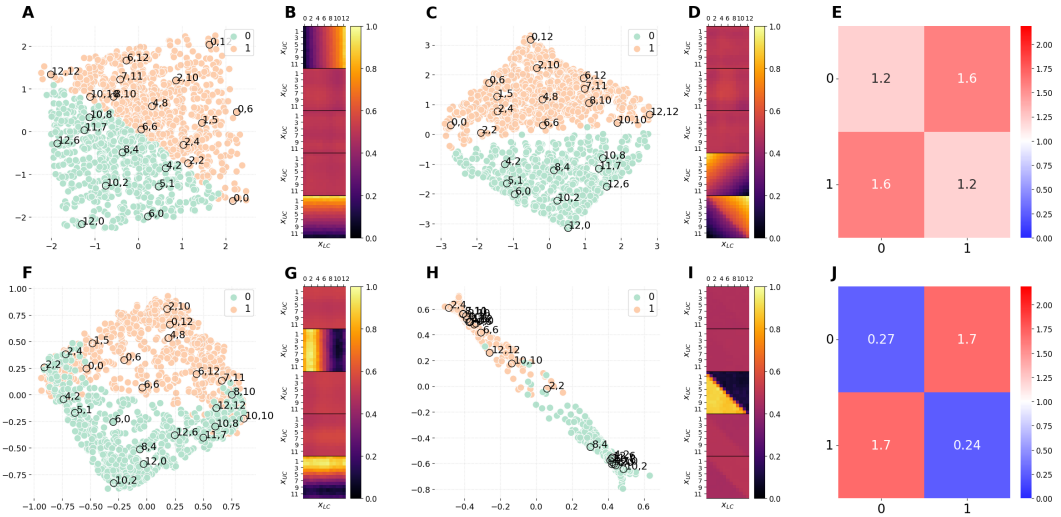

Figure 3.6: Comparison of the latent representations of the baseline model and the hybrid model E2M1. The E2M1 model is assigned a binary classification task, in which images with $x_U \leq x_L$ are labeled as 1, 0 otherwise. **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E2M1 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E2M1 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E2M1 model compared to the baseline model at high and low capacity, respectively.

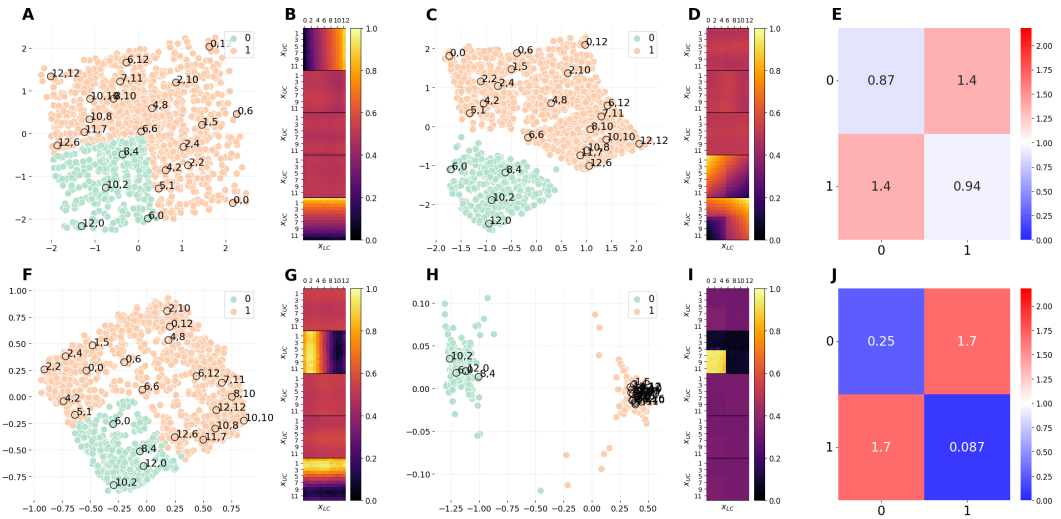

Figure 3.7: Comparison of the latent representations of the baseline model and the hybrid model E2M2. The E2M2 model is assigned a binary classification task, in which images with $x_U < 6$ (regardless of the value of x_L) OR $(x_U \geq 6$ AND $x_L \geq 6)$ are labeled as 1, 0 otherwise. **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E2M2 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E2M2 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E2M2 model compared to the baseline model at high and low capacity, respectively.

latent space representation becomes apparent when considering the nature of the task being solved: while the classification tasks of E2M1 and E2M2 models are linear, recognising corridors alignment represents a non-linear classification task. Since the classifier of the E2M3 model is linear, the classifier alone cannot solve the task on the basis of a representation like the one of the baseline model (see also Figures 3.5A and D) and thus the network adapted the representation to explicitly encode alignment information enabling the linear classifier to correctly solve the task. This also explains the separation in the latent space between class 1 and class 0 images observable in Figure 3.8C and quantified in Figure 3.8E. This result shows that an explicit latent representation specialized for the task can emerge in complex classification tasks and coexist with latent representations that capture the generative factors and allow an accurate reconstruction of stimuli.

Interestingly, at low capacity, only the latent channel encoding task-relevant information is active, resulting in a latent space with two clusters (Figures 3.8H and I). This latent space induces a strong separation of task-relevant inputs compared to the baseline model trained at low capacity, as it can be notice from a plot of their differences (Figures 3.8J).

Comparison between baseline and E2M4 models. Figure 3.9 compares the latent representations learned by the baseline model at high capacity (Figure 3.9A) and low capacity (Figure 3.9F) with those learned by the hybrid model E2M4 at high capacity (Figures 3.9C) and low capacity (Figures 3.9H), respectively.

The channel activations illustrated in Figure 3.9D show that at high capacity, the E2M4 model exploits all the 5 latent channels to encode stimuli. Some channels show horizontal or vertical bands, indicating that they encode information about a single corridor. The presence of these channels indicates that the latent state is still to some extent disentangled; however, the fact that the bands are sharply separated indicates that the representation is distorted compared to the baseline model – plausibly, to favor the performance of the model on the classification task. Furthermore, and interestingly, the channels coding for single corridors coexist with channels coding for both channels simultaneously, as evident from the fact that they show both horizontal and vertical bands. The coexistence of these different types of channels produce a significant rearrangement of the latent space compared to the

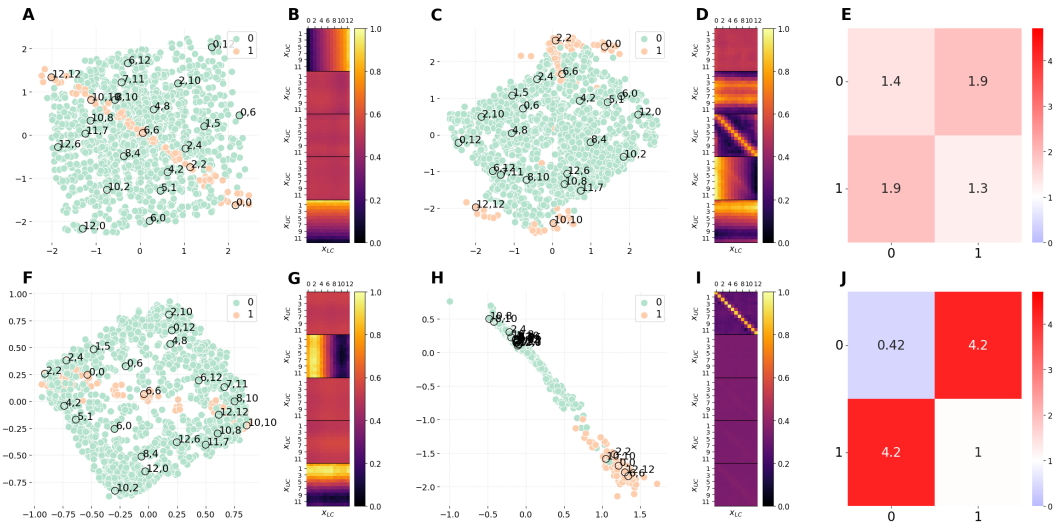

Figure 3.8: Comparison of the latent representations of the baseline model and the hybrid model E2M3. The E2M3 model is assigned a binary classification task, in which images with $x_U = x_L$ are labeled as 1, 0 otherwise. **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E2M3 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E2M3 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E2M3 model compared to the baseline model at high and low capacity, respectively.

baseline model, with different and separated subspaces allocated to the different stimulus classes (Figures 3.9C-E).

When the E2M4 model is trained at low capacity, it does not have sufficient capacity to encode all the stimulus classes: indeed, only 9 of 25 classes are encoded in the latent representation, while remaining classes are collapsed into a single one (Figure 3.9H-J). This effect can be considered analogous to the *specialization* effect observed with unbalanced datasets, in the sense that the model uses its limited resources to perform only partially its task. Note that the patterns shown by the latent channels of the E2M4 model trained at low capacity are more patchy compared to both the baseline model and the E2M4 model trained at high capacity (Figure 3.9I). This limited task representation explains the poor behavioral performance of the model (Figure 3.2B).

Comparison between baseline and E2M5 models. Figure 3.10 compares the latent representations learned by the baseline model at high capacity (Figure 3.10A) and low capacity (Figure 3.10F) with those learned by the hybrid model E2M5 at high capacity (Figures 3.10C) and low capacity (Figures 3.10H), respectively. Note that while in the figures, the label colors reflect the same task as the E2M4 model, the E2M5 model is trained simultaneously on all the tasks solved by models E2M1 - E2M4.

The latent channels of the E2M5 model trained at high capacity (Figure 3.10D) reveal patterns that are compatible with a combination of those observed in the models E2M1 – E2M4 – which also explains the good performance of the E5M5 model across all the tasks (see Figure 3.5C and Figure 3.2B). The first and last latent channels show a complex pattern that partly resembles tho one observed in the baseline model: specifically, the color is constant along one of the axis when considering the upper or lower triangular part only. Rather, the difference in color between the upper and lower triangular parts resembles the activation pattern of the E2M1 model. The second and third latent channels show patterns resembling the E2M3 and the E2M2 models, respectively. It is also possible to observe five color bands in each channel, compatible with the task solved by the E2M4 model. The resulting latent space is square-like, but the data points are further separated to accommodate all the tasks simultaneously (Figure 3.10C). As an example, consider

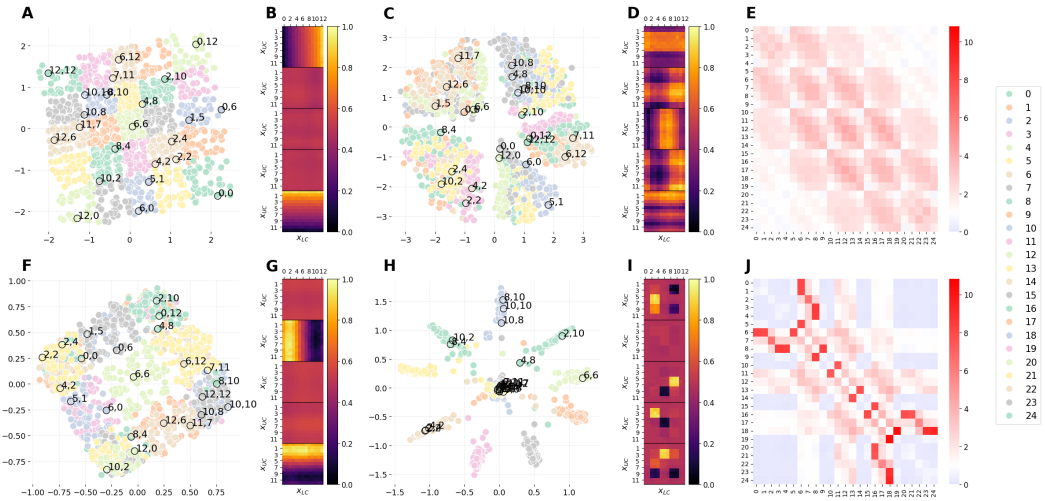

Figure 3.9: Comparison of the latent representations of the baseline model and the hybrid model E2M4. The E2M4 model is assigned a multiclass classification task. The possible positions of each corridor are grouped into 5 bins (as explained in Section 3.2.4) resulting in a partition of input images into 25 possible classes. **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E2M4 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E2M4 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E2M4 model compared to the baseline model at high and low capacity, respectively.

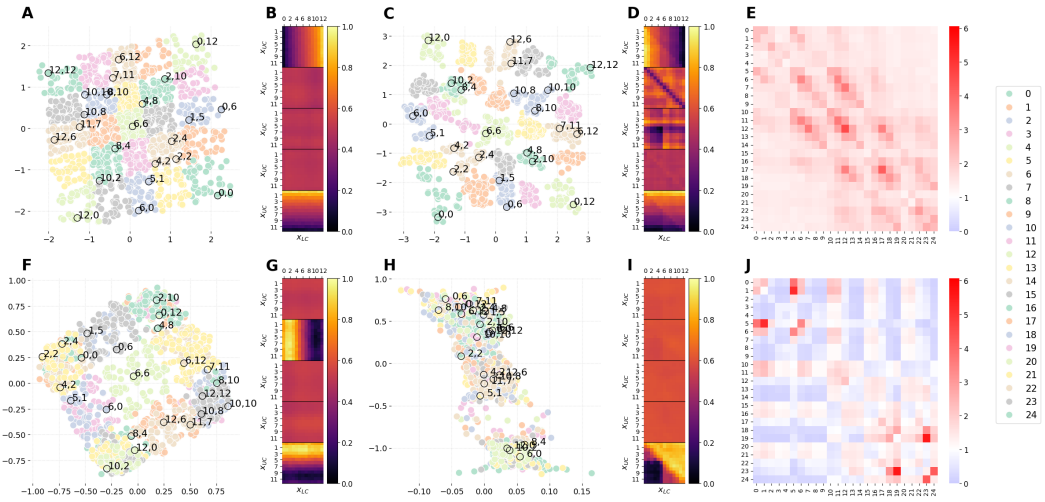

Figure 3.10: Comparison of the latent representations of the baseline model and the hybrid model E2M5. The E2M5 model is required to solve all the previous tasks simultaneously. **(A, F)** 2D projections of the 5D embeddings learned by the baseline model at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(C, H)** 2D projections of the 5D embeddings learned by the hybrid E2M5 model trained at high ($C_{\max} = 10$ nats) and low ($C_{\max} = 0.3$ nats) capacity, respectively. **(B, G)** Activation patterns of the 5 latent channels of the baseline model at high and low capacity, respectively. Each of the five heat-maps is computed as in sec. 3.4.3. **(D, I)** Activation patterns of the 5 latent channels of the E2M5 model at high and low capacity, respectively. Each heat-map is computed as in Section 3.4.3. **(E, J)** Measure of distortions in the latent representations of the hybrid E2M5 model compared to the baseline model at high and low capacity, respectively.

that the violet cluster containing the images with labels (5,1) and (6,0) is divided into two sub-clusters, to accommodate tasks solved by models E2M2 and E2M4.

At low capacity, the E2M5 model learns to solve only 2 tasks (see Figure 3.5C, showing that only the F1 scores of task associated to E2M1 and E2M2 models are high). This effect can be considered analogous to the *specialization* effect observed with unbalanced datasets, in the sense that the model uses its limited resources to perform only partially its task. The latent channels of the E2M5 model resemble those of E2M1 and E2M2 models (Figure 3.10I). This translates into a latent space with three clusters: the first cluster contains images having label 1 on both tasks, the second cluster contains images having label 0 on both tasks, and the third cluster contains images having different labels on the two tasks.

Summary of the results of Experiment 2

Experiment 2 investigates the distortions in β-VAE embeddings resulting from varying model capacity and incorporating classification tasks. Five hybrid models, E2M1 to E2M5, are compared to a baseline β-VAE model, at different capacities. E2M1 to E2M4 are hybrid models combining β-VAE and one classifier, each trained on a specific classification task. E2M5 includes four classifiers trained simultaneously with β-VAE.

Performance analysis reveals that as model capacity increases, reconstruction loss decreases across all models. However, hybrid models consistently exhibit higher reconstruction loss than the baseline model, due to the additional classification objective. Joint training of β-VAE and classifiers enables solving multiple tasks, with different capacities required for optimal performance in each task. Sequential training of β-VAE followed by classifiers yields less efficient performance compared to joint training, except in specific cases. Furthermore, joint training of multiple classifiers approaches the performance of single-classifier models, particularly at high capacity.

Regarding the geometry of representations, the classifiers induce task-specific distortions, as evident in comparisons between baseline and hybrid models. At high capacity, the distortions primarily improve segregation of class representations. Interestingly, explicit latent representations specialized for particular tasks can emerge and coexist in complex classification tasks with latent representations that capture the generative factors and allow an accurate reconstruction of stimuli. At low capacity, the distortions are more severe: the trained models show an orthogonalization of latent channels, with collapsed representations focused solely on task-relevant information – which also implies a severe loss of stimulus reconstruction accuracy of performance. This phenomenon is consistent across models E2M1 to E2M4.

In E2M5, joint training enables the model to simultaneously solve multiple tasks, with latent representations accommodating all tasks. However, at low capacity, the model only learns to solve two tasks effectively, leading to a poor latent space with fewer distinct clusters compared to the number of tasks.

3.3 Discussion

When creating internal models of the world, the brain necessarily compresses information. Formal frameworks like rate-distortion theory provide a normative perspective explaining *why* three different factors – model capacity, data distributions and tasks – distort latent representations. Here, instead, we ask *how* these factors distort latent representations.

To address this question, we employ a Beta Variational Autoencoder (β-VAE) [6], which approximates rate-distortion theory, while allowing for the utilization of real-world stimuli such as images [43, 20]. We use the β-VAE to conduct two computational experiments, in which we vary model capacity and data distributions (Experiment 1) and model capacity and classification task, while also augmenting the β-VAE with additional classification objectives (Experiment 2).

We report six main findings. First, the β-VAE trained at high capacity successfully recovers efficient latent representations that *disentangle* the two dimensions of variation of the stimuli (here, the positions of two corridors). However, lowering capacity distorts latent representations, inducing a *prototypization* effect, which is evident in panels F of Figures 3.3-3.4, 3.6-3.10. These compressed representations can be equally interpreted as *abstract* representations with a categorical bias – in the sense that they abstract away from sensory details [43].

Second, manipulating data distributions (i.e., training the model with unbalanced datasets) distorts latent representations, especially at low capacity. The resulting latent representations lose the disentanglement and show some *specialization*: they faithfully represents the most common stimuli, but not the infrequent ones. Interestingly, this training regime also favors the development of higher dimensional latent representations, plausibly as a way to distinguish rare stimuli from frequent ones. Simultaneously lowering capacity and manipulating data distributions (i.e., using an unbalanced dataset) makes the prototype unbalanced, too.

Third, introducing an additional objective – classification accuracy – penalizes reconstruction accuracy in hybrid models. Furthermore, classification accuracy produces an *orthogonalization* effect – and the increase of distance between latent representations for stimuli of different classes – which is most evident for low capacity, but is also present for high capacity. This result is in line with the empirical finding that task training orthogonalizes visual cortical population codes in rodents

[45]. Interestingly, in the hybrid models, the orthogonalization is not achieved by projecting information in a high dimensional space and then using linear readouts trained for specific computational tasks, as happens in reservoir computing and support vector machines [46]. Rather, it happens in the low dimensional space of VAEs.

Fourth, certain hybrid models can specialize different latent channels for its two main objectives: reconstruction accuracy and classification. For example, in the case of model E2M3, an explicit latent channel specialized for the task coexists with latent channels that capture the generative factors and allow an accurate reconstruction of stimuli (Figure 3.8D).

Fifth, the three kinds of distortions that we described – *prototypization*, *orthogonalization* and *specialization* – can have compound effects. For example, the simultaneous presence of capacity limitations and task produces both prototypization and orthogonalization. This is most evident by looking at the latent representations of hybrid models trained at low capacity (e.g., Figures 3.6H and 3.7H), which show a separation of two orthogonal prototypes. Interestingly, these are not the same prototypes that appear in the baseline model trained at low capacity (Figures 3.6F and 3.7F), but are rather amplifications of the separation between classes already observed when the hybrid models are trained at high capacity (Figure 3.6C and Figure 3.7C).

Sixth, task difficulty significantly affects the specificity of encoding and perceptual learning. Specifically, greater discrimination difficulty during perceptual learning induces more distinct representations of individual images, which is in keeping with empirical findings [44]. This is because easy tasks can be addressed with a greater level of compression.

Taken together, these results illustrate that three main classes of distortions of latent representations – prototypization, specialization, orthogonalization – emerge as signatures of information compression, under constraints on capacity, data distributions and tasks. These distortions can coexist, giving rise to a rich landscape of latent spaces, whose geometry could differ significantly across generative models subject to different constraints.

These results could be potentially useful to interpret neural data. Various empirical studies reported that the geometry of latent representations in neuronal populations is distorted, as in the case of reward-induced compression of spatial representa-

tions of mazes in the hippocampus and orbitofrontal cortex [24] or frequency-induced biases of working memory representations [47]. These and other empirical results can be interpreted in the light of the necessity for the brain to efficiently compress information – keeping with rate-distortion theory [2]. Interestingly, the brain is able to flexibly adapt the type and level of compression to task demands, rather than employing a fixed compression level. For example, the degree of specificity of internal representations along the visual pathway depends on the difficulty of the training conditions, with more specific representations emerging when learning difficult tasks [44]. Furthermore, the prefrontal cortex can rapidly remap the representation of stimuli according to their usefulness [48]. It remains to be established in future studies whether these empirical findings can be characterized formally using rate-distortion theory.

Furthermore, future studies might use the computational framework developed here for the ambitious objective to reverse-engineer the brain's training objective from the geometry of latent representations of neuronal populations. Some researchers have argued that the ventral visual stream pursues the goal of perceptual categorization [49] and others that it pursues efficient data compression or efficient coding [9]. Our analysis shows that these two goals induce different types of distortion; furthermore, it shows that the two goals are not mutually exclusive, but can be pursued in parallel. The relative importance of these two goals for the ventral visual stream (or other brain structures) could be in principle identified, by analyzing the geometry of latent representations.

This work has some limitations that could be addressed in future work. First, while our method assumes that reconstruction is part of the loss function (Equation 3.1), other methods, such as the information bottleneck [50] dispense from reconstruction. It remains to be assessed whether the results obtained here generalize to these methods. Second, for the sake of simplicity, we adopted a simple training scheme for the hybrid models of Experiment 2, in which the β-VAE model and the classifier are trained together (with the exception of the results shown in Figure 3.5A, in which the classifier is trained after the β-VAE was fully trained and frozen). An alternative possibility is adopting a more sequential training regime, in which the classifier is trained after the β-VAE has developed stable latent representations. Preliminary investigations shows that this sequential training regime produces additional changes to the geometry of the representation, but these changes need to be investigated in more detail. Third, future studies might consider more

sophisticated generative architectures (e.g., hierarchical or conditional β-VAEs) , which permit addressing the trade-offs between the capacity and accuracy afforded by different models [51] as well as more sophisticated methods to quantify how well latent representations support classification objectives [52]. Fourth, while we have focused on relatively simple measures of the geometry of latent representations of β-VAEs and their distortions, future studies might consider more advanced analysis methods, which are based on interventions [53]. Finally, we focused on a specific dataset – the Corridors dataset – which we created specifically to study the distortions of generative models. We opted to create a new dataset instead of using standard datasets like MNIST [54] because the lack of clearly defined orthogonal factors of variation would make it difficult to interpret the effects of rate-distortion trade-offs. While other less standard existing datasets like dSprites [55] do feature orthogonal factors, they usually include a large number of independent dimensions of variation, significantly increasing the complexity of the analysis. By contrast, the Corridors dataset was designed to contain only two orthogonal generative factors, allowing us to systematically study how capacity constraints, data distributions, and task requirements distort latent representations. We expect our results to generalize to other datasets with a larger number of disentangled dimensions. Furthermore, we expect our methods to be particularly useful to study neural representations in biological systems and their distortions, which have been demonstrated in previous empirical studies [23, 48, 45]. However, the generalization of our methods to other datasets and their application to address biological systems remain to be addressed in future research.

3.4 Methods

3.4.1 Computational Models

As explained in Section 2.2, we use the β-VAE architecture (Figure 3.11A) as computational model since it represents an effective approximation of RDT for high-dimensional data, such as images [43, 20].

The β-VAE used for both Experiments 1 and 2 comprises an encoder block and a decoder block. The encoder block consists of 3 convolutional layer, each followed by a leaky ReLU activation [56, 57]. Similarly, the decoder block consists of 3

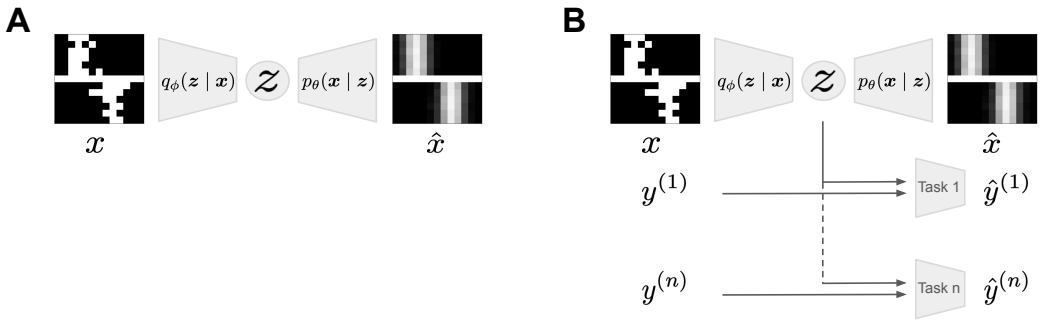

Figure 3.11: Computational models. (A) The standard β-VAE model used in Experiment 1. (B) The general hybrid model used in Experiment 2. It extends the standard β-VAE architecture with one or more classifiers that are jointly trained with the β-VAE.

transposed convolutional layers, also followed by a leaky ReLU activation function, except for the last layer having a sigmoid activation function to output grayscale images. The latent dimension is fixed to 5, since there are only two generative factors of variation in the data. We added three extra dimensions to see whether the models are able to use them efficiently or not. All the models are trained with Adam optimizer and a learning rate of $5 \cdot 10^{-5}$.

Inspired by [38], we trained all the models using the following variant of the loss function in equation 2.6:

$$\mathscr{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z|x)}\Big[-\log p_\theta(x \mid z)\Big] + \beta\Big|D_{KL}\left(q_\phi(z \mid x) \| p(z)\right) - C\Big| \quad (3.1)$$

where C represents the encoding capacity available to the model. In fact, the amount of information that each latent dimension of the VAE can encode about the input depends on the KL divergence between the approximate posterior $q_\phi(z \mid x)$ and the prior $p(z)$. By penalizing with a factor β any deviation of the KL divergence from the required capacity C, the loss function in Equation 3.1 allows to adjust the encoding capacity of the model during training. Following the training recipe introduced in [38], during training, we gradually increased the capacity C from 0 to a target value $C_{\max}$ to favor the disentanglement of latent representation [38]. All the models have been trained 5 times, each time with a different value for the maximum encoding capacity $C_{\max}$. Possible values for $C_{\max}$ are 0.1, 1, 3, 6 and 10 nats.

Note that, with respect to Equation 2.4, in Equation 3.1 we introduced the absolute value in the rate term. This is equivalent to require that the KL divergence

is exactly C and not just less or equal to C. This choice is motivated by a well-known issue in VAE training called "posterior collapse". In fact, a powerful decoder can learn to ignore the latent code entirely, causing the information rate (R) to collapse to zero [58]. By enforcing a target capacity $C > 0$, we force the model to use the latent channel to transmit a specific amount of information, measured in nats. This makes C a directand interpretable measure of the channel's capacity.

However, in Experiment 2, we extend the β-VAE architecture with one or more classifiers (Figure 3.11B) that are jointly trained with the β-VAE by means of the following loss function:

$$\mathscr{L}(\theta, \phi; x, y, z, \beta) = \\ \mathbb{E}_{q_\phi(z|x)}\Big[-\log p_\theta(x \mid z)\Big] + \beta\Big|D_{KL}\left(q_\phi(z \mid x)\|p(z)\right) - C\Big| + \sum_{i=1}^{n} \mathscr{L}_{\text{clf}}^{(i)}\left(z, y^{(i)}\right), \tag{3.2}$$

where the last term represents the sum of the losses $\mathscr{L}_{\text{clf}}^{(i)}$ of the classifiers that are jointly trained with the β-VAE. It is possible to have only one classifier ($n = 1$), as for models E2M1-E2M4, or more than one ($n > 1$), as for model E2M5 which features $n = 4$. The i-th classifier receives as input the latent vector z from the β-VAE and the label $y^{(i)}$ associated to z in the i-th task. This joint training procedure allows us to study the distortion induced by the classification task on the latent representation learnt by the β-VAE models. Depending on the specific task, the classifier can be linear or non-linear. The linear classifier is a feedforward neural network with no activations and a single hidden layer with 1024 units. The non-linear classifier is a feedforward neural network with leaky ReLU activations and a single hidden layer with 1500 units.

3.4.2 Comparing latent representations (or embeddings) learned by the different models

Our goal is to compare the geometry of the latent representations (or embeddings) learned by different models in our experiments. Since all models present a 5-dimensional latent space, we employ Multidimensional Scaling (MDS) [59] as a dimensionality reduction technique to render 2D representations of the latent space structure.

To execute MDS, we input a dissimilarity matrix derived from pairwise Euclidean distances between data points in the high-dimensional space. For each trained model, we generate 2D projections of the latent space by applying MDS to the pairwise Euclidean distance matrix computed from the 5-dimensional vectors associated with each image via the encoder block of the respective network. Then, we compare the 2D embeddings learned by the different models considered (see panels A, C, F and H of figures 3.3, 3.4, 3.6, 3.7, 3.8, 3.9 and 3.10); see also Supplementary Figures A2 - A8 for 3D plots of the embeddings.

Class distortion matrix and Item distortion matrix Given two models, M_1 and M_2, a way to quantitatively measure how much the latent space of model M_2 is distorted with respect to that of model M_1 is to measure, for each pair of images u and v, how far the distance between $M_2(u)$ and $M_2(v)$ deviates from the distance between $M_1(u)$ and $M_1(v)$, where $M_\bullet(u)$ denotes the embedding of image u according to model $M_\bullet$. More specifically, we compute the ratios

$$\rho(u,v) = \frac{d\left(M_2(u), M_2(v)\right)}{d\left(M_1(u), M_1(v)\right)}, \qquad \forall u \neq v. \tag{3.3}$$

for all pairs of images employing the Euclidean distance as $d(\cdot,\cdot)$. When $\rho > 1$, the two images are farther apart in the latent space of M_2 with respect to M_1. When $\rho < 1$, the two images are closer in the latent space of M_2 with respect to M_1. When $\rho = 1$, the two images have the same distance in the latent space of both models. This allows us to measure the magnitude of dilation or compression of the embeddings of a model with respect to those of an other model.

Panel E of Figures 3.3, 3.4, 3.6, 3.7, 3.8, 3.9 and 3.10 reports the distortion matrix arising when having an unbalanced dataset or a task assigned to the network at high capacity. In particular, we consider the baseline model at high capacity as M_1 and a "hybrid" model at high capacity as M_2. In this case, we group images into classes depending on the experiment. In experiment 1, for example, it is possible to divide images according to their frequency of appearance in the dataset (low and high frequency) and thus we have two classes (0 and 1 labels). In experiment 2, it is possible to divide images into classes according to the specific classification task being solved by the network. The number of classes will hence depend on that task. By aggregating images according to the class they belong to, we can measure the average distortion for couple of images within the same class or belonging to distinct

classes. This tells us how much, on average, the relative distance between an images of one class and a image of an other class is compressed ($\rho < 1$) or dilated ($\rho > 1$) when passing from the latent space of model M_1 to that of model M_2. Similarly for panel F, the models involved are the same of panel E but they are trained at low capacity.

3.4.3 Comparing neurons' activity of the different models

Another way to qualitatively assess the distortion induced by a task (or stimuli statistics) onto the embedding space is to visualize the activation patterns of the individual latent channels of the 5D latent space when specific stimuli are passed as input to the model. For each latent channel, we create a $13x13$ matrix whose (i, j) element is the average value of that channel when a image with corridors in position (i, j) is passed as input, where the average is computed on all the images on the test set. For each model we thus obtain 5 matrices, one for each dimension of the latent space, and we plot them as heat maps (panels B, D, G and I of figures 3.3, 3.4, 3.6, 3.7, 3.8, 3.9 and 3.10). This approach facilitates the identification of significant differences between models, such as variations in the number of active channels within the representation or their distinct activation patterns.

Chapter 4

Number Sense in unsupervised generative models

In Section 2.3, we examined existing applications of rate-distortion theory in cognitive science. In this chapter, we specifically focus on number sense, the cognitive ability to estimate the number of items in a visual scene without explicit counting. We begin with a general overview of number sense, exploring its cognitive and neural underpinnings, and its characterization through established psychophysical laws. We then transition to a rate-distortion perspective, exploring how RDT can provide a normative explanation for the emergence of number sense and the observed psychophysical phenomena, particularly in the context of limited cognitive resources. We thus present the results of the second line of inquiry as described in Chapter 1.

4.1 Introduction

Number sense, the intuitive ability to estimate and discriminate quantities [60], is a fundamental cognitive capacity shared by humans and animals [61–63]. This innate ability allows for quick, approximate judgments about the number of objects in a scene, without relying on symbolic counting or formal mathematical knowledge. This "approximate number system" (ANS) [62] supports crucial adaptive behaviors, such as foraging, predator avoidance, and social interaction, where rapid assessments of quantity can be critical for survival [64].

From a neural perspective, neuroimaging and lesion studies have identified specific brain regions associated with number sense. The intraparietal sulcus (IPS) in the parietal lobe is consistently implicated in numerical processing, showing activation during both symbolic and non-symbolic numerical tasks [65]. Furthermore, the prefrontal cortex, known for its role in executive functions and working memory, also contributes to numerical cognition, particularly in tasks requiring numerical comparison and manipulation [66]. The interplay between these brain regions suggests a distributed neural network supporting various aspects of number sense.

At behavioral level, the performance of both humans and animals in numerical tasks is characterized by specific psychophysical laws that describe the relationship between objective quantities and subjective perception. In this context, the Weber's law is a foundational principle in psychophysics, stating that the just-noticeable difference (JND) between two stimuli is proportional to the magnitude of the stimuli [67]. In the specific context of number sense, Weber's law predicts that the ability to discriminate between two numerosities decreases as the numerosities increase. For example, it is easier to discriminate between 2 and 3 than between 22 and 23, even though the absolute difference is the same. Empirical studies have consistently demonstrated that Weber's law holds for number sense, with discrimination accuracy declining as numerosity increases [68, 69].

While the aforementioned psychophysical laws provide a descriptive account of number sense, Rate-Distortion Theory offers a normative framework for understanding *why* these laws emerge and how they reflect efficient coding under resource constraints. As highlighted in Section 2.3, the brain operates under stringent resource constraints, including limitations in metabolic energy, neural tissue, and processing time [8–10]. This necessitates efficient coding strategies that maximize the amount of relevant information represented while minimizing resource expenditure. In the context of number sense, this suggests that the brain may have evolved representations that optimally balance the accuracy of numerosity estimates with the cost of encoding and processing numerosity information. Theoretical studies suggest that numerosity representations in the brain might emerge from the efficient (unsupervised) coding of statistical regularities of sensory inputs, under limited computational resources [70–72]. Such representations indeed emerge in neural networks that imperfectly capture visual scenes, highlighting that an explicit, supervised objective to count objects is not required for learning [73–77]. While these previous works revealed many insights, they have limitations that prevent a full mechanistic under-

standing of numerosity perception. For example, theoretical investigations have linked the psychophysics of numerosity perception to limited information capacity [71], or more specifically to bandwidth-limited scene memory [70], using abstract mathematical (Bayesian) models that cannot take real images as input. Conversely, computational studies based on unsupervised neural networks have not systematically explored the link between numerosity perception and information capacity (in the information-theoretic sense) [78], nor have they compared the solutions that emerge in unsupervised systems with those of supervised models explicitly trained to estimate the number of items in the images.

In sum, the precise computational principles governing how these numerosity representations emerge in resource-limited neural systems, and their relationship to known psychophysical laws, remain poorly understood. One question that remains unaddressed is whether psychophysical signatures of numerosity, such as Weber's law, are an unavoidable effect of information compression and therefore can occur regardless of the amount of resources available at the encoding stage, or if instead they only occur in systems with severe encoding capacity constraints as an effect of resource limitations. Another question is whether unsupervised learning is sufficient to develop numerical codes with the capability to generalize beyond the training set and to generate novel numerical stimuli, hence fully capturing the complexity of numerical perception – or whether these capabilities necessarily require the supervised objective of learning to explicitly estimate visual numerosity. Finally, another question is what is the effect of severely limiting the capacity of unsupervised systems – and whether there are relations between these capacity limitations and deficits in numeosity perception as observed in developmental dyscalculia, a learning disability characterized by difficulties in number sense, memorization of arithmetic facts, accurate or fluent calculation, and accurate mathematical reasoning [79].

To address the above questions, here we investigate the emergence of number sense in unsupervised generative models through the lens of RDT [2]. While applying RDT to visual image learning can be challenging, as demonstrated in Section 2.2, β-Variational Autoencoders (β-VAEs) approximate RDT principles by learning compressed latent representations of high-dimensional data under capacity constraints [43]. In this chapter, we exploit the capability of β-VAEs to incorporate RDT principles and examine how numerosity information extracted from visual images is represented in the latent space of β-VAEs trained in an unsupervised manner to reconstruct the images. This allows us to assess whether the statistical structure of

numerosity can be inferred without explicit supervision, whether the latent representations that emerge during β-VAE training at high capacity align with known psychophysical laws of numerosity perception, and whether low capacity learning impairs performance in ways similar to that observed in dyscalculic children.

To assess the ability of β-VAE models trained at different capacities to learn numerosity as a summary statisticof visual scenes, we conducted a series of behavioral analyses using a dataset commonly employed in numerosity perception research, which includes images containing items that vary in numerosity, size and spatial arrangement. Our analyses include numerosity estimation and discrimination tasks, examinations of latent space structure using dimensionality reduction techniques, and tests of generalization to novel stimuli. Additionally, we evaluate the models' generative capabilities, assessing whether numerical representations can be flexibly used to generate images with a specified numerosity. Our main analysis focuses on the β-VAE model trained at high capacity, which we call β-VAE-high, to assess whether it captures key aspects of human numerosity perception. As a control analysis, we compare the behavior and latent representations of the unsupervised β-VAE-high to those of a supervised network explicitly trained to estimate the number of items in the image. We refer to this supervised model as β-VAE-sup-high. Finally, we progressively reduced the encoding capacity of the β-VAE models, observing a gradual degradation in behavioral performance and, at low capacity, we investigated whether the models exhibited maladaptive numerosity perception abilities comparable to dyscalculia.

To preview our main results, we found that the unsupervised β-VAE trained at high capacity captures numerosity perception in a manner akin to human perception, demonstrating sensitivity to numerosity. In contrast, the supervised model exhibits superhuman precision in estimating numbers, highlighting the impact of task-specific constraints. Additionally, the unsupervised β-VAE develops a robust coding for numerosity that generalizes to novel datasets and permits generating novel images having a desired number of items, by simply sampling from the latent space of the model. Furthermore, in keeping with RDT, we observe that varying encoding capacity decreases behavioral performance in numerosity perception. Finally, the β-VAE trained at low capacity displays impairments in numerosity perception and limited generalization, offering insights into the role of resource limitations in numerosity perception.

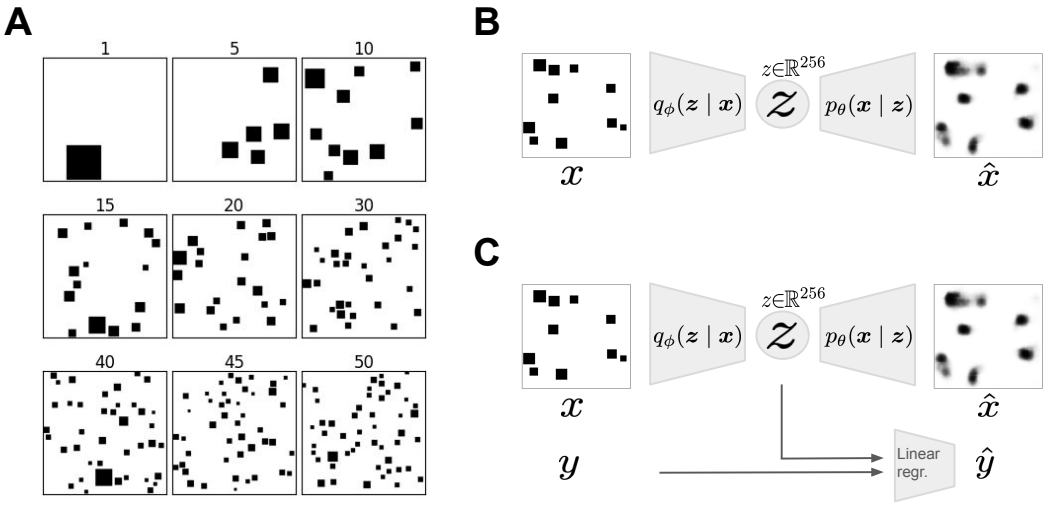

Figure 4.1: Dataset and computational approach. (A) Samples from the dataset used in the study. The training dataset consists of 50.000 256x256 black-and-white images, each containing *N* items (black squares), with *N* ranging from 1 to 50. (B) Schematic illustration of the unsupervised learning model (β-VAE). (C) Schematic illustration of the supervised learning model (β-VAE-sup) used for control analyses. It extends the standard β-VAE architecture with a linear regression layer attached to the latent space, explicitly trained to predict the numerosity of the input images. See the Section 4.4 for explanation.

The code to reproduce the simulations, the models' checkpoints and the datasets used in this study are available at:https://github.com/damat-le/num_sense_rdt.git

4.2 Results

To explore the emergence of number sense in unsupervised generative models, we trained a β-Variational Autoencoder (β-VAE) at high capacity. The training dataset comprised 50,000 black-and-white images, each containing *N* items (black squares), with *N* ranging from 1 to 50. Representative samples are shown in Figure 4.1A. Our primary goal was to assess the capabilities of this unsupervised β-VAEs (Figure 4.1B), trained with the sole objective of reconstructing the input stimuli, in addressing phenomena related to numerosity perception. Furthermore, we aimed to study the effect of varying encoding capacity on numerosity perception.

To this end, we conducted six distinct sets of analyses. The first four sets of analyses focus on the main model considered in this study: a β-Variational Autoencoder trained at high encoding capacity (5000 nats). We refer to this model as

β-VAE-high. First, in Section 4.2.1, we evaluate the performance of linear readouts trained on top of the model's latent representations to solve numerosity estimation and numerosity discrimination tasks. Second, in Section 4.2.2, we explore the structure of the latent space that emerges in the model. Third, in Section 4.2.4, we assess the robustness and abstractness of the learned latent representations by measuring the model's ability to generalize to a novel set of images. Fourth, in Section 4.2.4,we evaluate the model's ability to generate visual numerosities through top-down sampling from the latent state. The fifth set of analyses, illustrated in Section 4.2.5, focuses on the rate-distortion trade-off that emerges when varying the encoding capacity of the β-Variational Autoencoder. For this, we train seven β-VAE models with various capabilities, from high (5000 nats) to low (50 nats). Finally, in the fifth set of analyses in Section 4.2.6, we examine whether low-capacity training impairs numerosity perception abilities in a β-VAE model trained at low capacity (50 nats). We refer to this model as β-VAE-low.

As control simulations, we also trained a supervised counterpart of the β-VAE, at both high and low capacity, by augmenting the loss function with an additional term that requires to explicitly estimate the numerosity of input images (Figure 4.1C). We refer to these supervised models as β-VAE-sup-high and β-VAE-sup-low, respectively. This allows to compare the latent representations of the unsupervised β-VAE-high model with the numerosity-specific representations of the β-VAE-sup-high model, which had access to numerosity ground truth during training. We thus repeated all the analyses above, except for the fourth, on the β-VAE-sup model and report the results in Section B.1 of Supplementary Materials.

Technical details about the training recipes of the models, the dataset and the six sets of analyses are reported in Section 4.4.

4.2.1 Behavioral Analyses

The behavioral analyses were structured to mimic experimental paradigms used in human psychophysics studies [79, 80], allowing direct comparisons between the model's performance of the β-VAE-high model and human number sense. We conducted two tests, one to assess the model ability to represent an estimate of the number of items in a given input image (numerosity estimation) and the other to

assess the model ability to support the discrimination between the numerosity of items in an input image and a reference number (numerosity discrimination).

Numerosity estimation. To quantify the ability of the unsupervised model to encode numerosity relevant information, we trained a linear regression model to predict the number of items in an image from the corresponding latent representation generated by the pre-trained β-VAE-high. Figure 4.2A compares the actual numerosity of the input images (x-axis) with the numerosity predicted by the linear regression trained on the model's latent representations (y-axis). The data points, which represent the linear regression's average predictions for each numerosity class, align closely with the diagonal, indicating strong agreement between predicted and true numerosities. As evident from Table 4.1, the low mean square error (MSE) of regression (MSE $= 4.98$) confirms the model's ability to support accurate numerosity estimation. Interestingly, for small numerosities ($N < 5$), the linear regression exhibits low bias (predictions closely align with the diagonal) and low variance (average standard deviation of 1.3). For large numerosities ($N > 46$), the linear regression exhibits a slightly higher bias (the mean predictions exhibit minimal deviation from the true values) and also the variance grows (the average standard deviation of 2.8) with respect to small numerosities.

Numerosity discrimination. To assess the model's ability to support the discrimination between images with distinct numerosity, we trained a logistic regression to determine whether the numerosity of an input image exceeds a fixed reference numerosity $n_{\text{ref}} = 25$. The logistic regression is trained on the latent representations of the input images, obtained through the pre-trained model. Figure 4.2B presents the probability of the logistic regression correctly discriminating whether the numerosity of an input image exceeds 25. The curve is smooth and sigmoidal, with a steep transition around the reference numerosity, indicating that the model is highly confident in its discrimination ability even for numerosities near 25. The slope of the curve implies fine sensitivity to numerical differences in this range, supporting the model's capability for accurate numerosity discrimination. As reported in Table 4.1, the Weber fraction ($w = 0.12$) derived from this analysis is consistent with empirical findings in human psychophysics [79], reflecting a number acuity comparable to that observed in adult human populations.

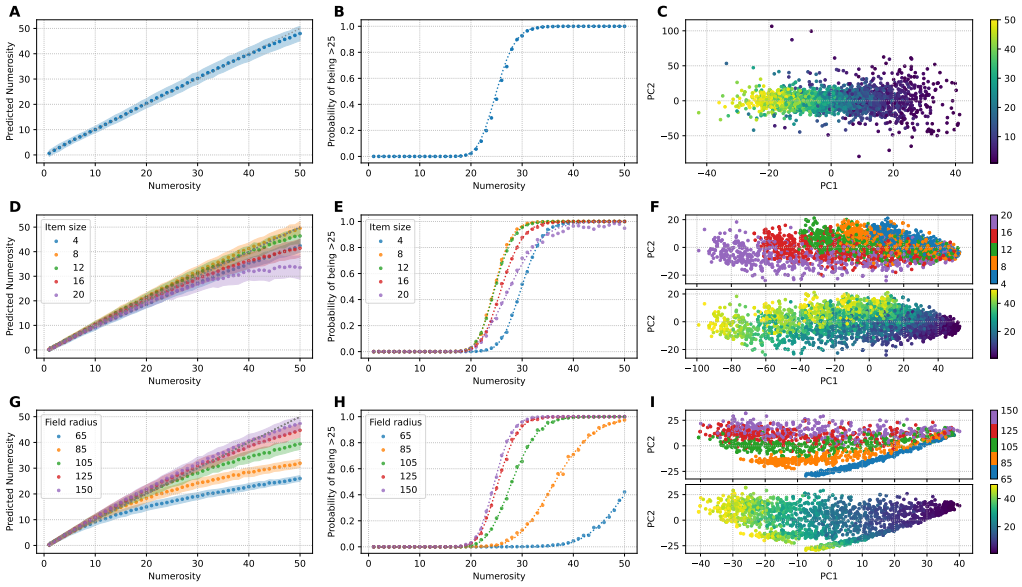

Figure 4.2: Analysis of the β-VAE-high model. (A) Results of the numerosity estimation task on the test dataset. (B) Results of the numerosity discrimination task on the test dataset. (C) Latent space visualization: first two principal components (PCs), colored by numerosity. (D-F) Zero-shot generalization tests on the "item size" dataset: (D) Numerosity estimation, (E) Numerosity discrimination, (F) Latent space (PC1 vs. PC2), colored by item size (top) and numerosity (bottom). (G-I) Zero-shot generalization tests on the "field radius" dataset: (G) Numerosity estimation, (H) Numerosity discrimination, (I) Latent space (PC1 vs. PC2), colored by field radius (top) and numerosity (bottom). See main text for details.

		zeroshot			retrained		
		MSE	w	n_ref	MSE	w	n_ref
model	dataset						
β-VAE-high	test	4.98	0.12	25.00	4.98	0.12	25
	itemsize_4	15.48	0.11	29.93	3.74	0.10	25
	itemsize_8	5.52	0.11	24.61	3.31	0.10	25
	itemsize_12	5.65	0.11	24.76	2.96	0.09	25
	itemsize_16	15.38	0.13	26.07	2.30	0.08	25
	itemsize_20	48.05	0.18	27.68	1.54	0.06	25
	fieldradius_65	148.09	0.15	51.89	1.28	0.07	25
	fieldradius_85	69.33	0.16	36.42	2.76	0.08	25
	fieldradius_105	21.30	0.14	28.03	4.10	0.09	25
	fieldradius_125	7.52	0.12	25.35	4.03	0.09	25
	fieldradius_150	5.81	0.11	24.63	3.89	0.10	25
β-VAE-low	test	40.71	0.38	25.00	40.71	0.38	25
	itemsize_4	564.57	0.44	111.76	62.72	0.50	25
	itemsize_8	114.98	0.25	37.57	24.25	0.29	25
	itemsize_12	34.29	0.24	26.97	15.33	0.21	25
	itemsize_16	63.41	0.23	22.36	10.49	0.18	25
	itemsize_20	143.28	0.22	19.51	7.92	0.15	25
	fieldradius_65	610.73	1.76	4382.17	21.03	0.25	25
	fieldradius_85	444.57	0.67	149.83	19.31	0.23	25
	fieldradius_105	322.23	0.45	73.52	18.53	0.23	25
	fieldradius_125	226.64	0.32	49.73	19.97	0.24	25
	fieldradius_150	168.97	0.27	42.72	22.07	0.27	25
β-VAE-sup-high	test	1.10	0.05	25.00	1.10	0.05	25
	itemsize_4	8.22	0.08	24.64	2.75	0.11	25
	itemsize_8	4.48	0.05	22.69	1.39	0.09	25
	itemsize_12	1.13	0.05	24.66	0.75	0.07	25
	itemsize_16	1.10	0.05	26.26	0.51	0.06	25
	itemsize_20	7.92	0.05	25.54	0.51	0.05	25
	fieldradius_65	93.34	0.10	39.71	0.87	0.06	25
	fieldradius_85	35.29	0.09	29.33	1.55	0.08	25
	fieldradius_105	10.68	0.07	25.44	1.59	0.07	25
	fieldradius_125	4.38	0.06	23.76	1.51	0.08	25
	fieldradius_150	3.81	0.06	23.11	1.47	0.08	25
β-VAE-sup-low	test	21.90	0.25	25.00	21.90	0.25	25
	itemsize_4	192.88	0.24	48.19	30.31	0.31	25
	itemsize_8	29.27	0.22	28.89	16.87	0.23	25
	itemsize_12	21.40	0.20	26.23	10.66	0.18	25
	itemsize_16	22.51	0.18	24.32	6.48	0.14	25
	itemsize_20	46.05	0.18	21.81	4.43	0.12	25
	fieldradius_65	307.95	0.56	99.93	25.15	0.25	25
	fieldradius_85	171.96	0.31	46.93	19.03	0.23	25
	fieldradius_105	100.41	0.26	37.05	16.66	0.22	25
	fieldradius_125	62.21	0.25	32.78	16.15	0.22	25
	fieldradius_150	42.84	0.23	30.58	17.04	0.23	25

Table 4.1: Performance metrics for numerosity estimation and discrimination tasks across different models and datasets. MSE: Mean Squared Error for numerosity estimation task. w: Weber fraction for numerosity discrimination task. n_{ref}: Reference numerosity used for discrimination. Results are shown for both zero-shot evaluation (without retraining readouts) and after retraining the linear readouts on each dataset, with models' weights frozen. Each model is evaluated on three datasets: test dataset, a dataset with the same distribution of the training dataset, item size dataset, a dataset in which item size and numerosity vary orthogonally and field radius dataset, a dataset in which field radius and numerosity vary orthogonally (see Section 4.4.5 for more details).

Control analyses. To establish a performance baseline and further probe the behavioral capabilities of the β-VAE-high model, we conducted two control analyses. First, we evaluated the supervised β-VAE-sup-high model, which was explicitly trained using numerosity labels, on the same numerosity estimation and discrimination tasks. The β-VAE-sup-high model exhibited good performance in both tasks (Figure B.1.1A-B, Table 4.1). However, its performance substantially surpassed human psychophysical capabilities, as evidenced by a "superhuman" Weber fraction as low as $w = 0.05$ in the numerosity discrimination task and a MSE of 1.10 in the numerosity estimation task. This underscores the considerable impact of task-specific supervised training.

Second, to evaluate the reconstruction quality of the unsupervised β-VAE-high model and its impact on numerosity perception, we administered more challenging versions of the numerosity estimation and discrimination tasks. Instead of using the original input images, the β-VAE-high model first reconstructed the images, and then performed the numerosity estimation and discrimination tasks on these reconstructed images in a zero-shot manner (i.e., without retraining any part of the model). The results demonstrate that the model performs these more challenging tasks with accuracy comparable to the original tasks (Figure B.4.1), showcasing its effective reconstruction capabilities. This control analysis guarantees that the β-VAE-high model has correctly optimized both the rate and distortion terms during training (see Equation 3.1).

To summarize, our behavioral analyses reveal that the unsupervised β-VAE-high model supports successful visual numerosity estimation and discrimination, without explicit training. Moreover, it exhibits excellent reconstruction capabilities, supporting accurate numerosity tasks not only with original input images but also with reconstructed images. Finally, and in contrast to the supervised β-VAE-sup-high model, the unsupervised β-VAE-high model demonstrates signatures of human visual numerosity perception, including greater precision in estimating smaller versus larger numerosities and a discrimination ability (measured by the Weber fraction) within the range of human psychophysics.

4.2.2 Neural Analyses, β-VAE-high model

The good performance in numerosity perception tasks suggests that the β-VAE-high model could develop numerosity-related features in its latent space through unsupervised learning. To test this hypothesis, we applied a Principal Component Analysis (PCA) to the 256-dimensional latent space of the model. Figure 4.2C shows the first two principal components (PC1 and PC2). The points are colored according to the numerosity of the corresponding input images and range from blue (for smaller numbers) to yellow (for larger numbers). Notably, the plot reveals distinct clustering of points based on numerosity, with a gradual and smooth transition of colours along the first principal component (see Figure B.5.4 for another illustration of the first PC). This result indicates that the model has successfully organized its latent representations in a way that preserves numerosity information – as a one-dimensional, population-level "number line" – in the first PC. Moreover, the spread of points along PC2 highlights the model's ability to encode variance in numerosity while preserving separability between different numerosities. Interestingly, the variance along PC2 is greater for smaller numbers, possibly reflecting the greater diversity of input images with lower numerosity (i.e., images containing fewer items tend to be more varied compared to those with higher numerosity). However, importantly, this increased variance does not hinder the model's ability to estimate numerosity for images with fewer items, indicating a robust numerical representation.

To obtain more insight about how numerosity information is coded in the β-VAE-high model's latent space, we performed three additional analyses. In these analyses, we asked whether numerosity is coded in a distributed manner across the 256 dimensions of the model's latent space or localized in a small number of dimensions. In the first additional analysis, we sorted the 256 neurons of the latent representation according to the Spearman correlation between their neural activity and the numerosity of the input images, then we trained 256 linear regressions, each with an increasing number of neurons, and computed the respective MSE. Figure B.5.1A shows that while the first 20 neurons give the greatest contribution, the MSE scores degrade gracefully, indicating that also all the other neurons contribute to a distributed numerosity coding. This result can also be appreciated by looking at the Spearman correlation score between neurons and numerosity which is higher for the first 20 neurons and degrades gracefully for the remaining ones.

In the second additional analysis, we used mutual information, to assess whether smaller and larger number of items are coded by the same or different neurons. Our analysis suggests that the latent code that emerges in the β-VAE-high includes a mixture of neurons that code specifically for smaller ($N \leq 4$) versus larger ($N \geq 47$) number of items and neurons that show a mixed selectivity for both (Figure B.5.2A). Furthermore, there are more neurons that code specifically for smaller compared to larger number of items.

In the third additional analysis, we examined the average activity of individual latent neurons in response to images containing varying numbers of items. Specifically, neural activity was first standardized using z-score normalization. Subsequently, we computed the average activity across images, grouped by numerosity. Figure B.5.3A displays the average activity of each neuron within the latent space of the β-VAE-high model for each numerosity class, visualized as a heatmap. The heatmap reveals that many neurons exhibit high activity (red) when presented with images depicting small numerosities, and low activity (blue) in response to large numerosities. Conversely, other neurons exhibit the opposite pattern. These observations corroborate the findings of the mutual information-based analysis. Furthermore, the majority of neurons exhibit a discernible pattern that varies with the numerosity of the input images, which is consistent with prior results suggesting a distributed representation of numerosity.

Finally, we replicated all the analyses for the control β-VAE-sup-high model. Interestingly, the PCA analysis applied to the control β-VAE-sup-high model reveals a very similar structure of the latent space, albeit with slightly less variance along PC2, as evident from the scale of the y-axis in Figure B.1.1C. The supervised β-VAE-sup-high model also develops a more specialized neural code, in which the MSE scores degrade more abruptly (Figure B.5.1B). The higher Spearman correlations (in absolute value) also confirm this specialization. Instead, no significant difference emerges in the neurons' selectivity for small or large numbers (Figure B.5.2B) compared to the β-VAE-high model.

To summarize, this analysis the β-VAE-high model's latent space reveals a structured representation of numerosity, encoded in a distributed manner across latent dimensions, with specialized neurons for small and large numbers.

4.2.3 Robustness of the β-VAE-high with respect to non-numerical magnitudes

A robust numerosity code should generalize to visual stimuli that are out-of-distribution with respect to the training data. Furthermore, it should not be confounded by additional factors, such as the size and spatial distribution of items, which often correlate with numerosity (e.g., larger and smaller items could correlated with low and high numbers, respectively).

To evaluate the generalization capability of the β-VAE-high model and to ensure its numerosity estimation and discrimination abilities stemmed from a robust numerosity code rather than correlated non-numerical visual features, we conducted the same behavioral and neural analyses on two novel datasets: "item size" and "field radius". Crucially, these tests were performed as zero-shot evaluations, without any fine-tuning or re-training of the β-VAE-high model or of the linear readouts.

Item size. In this test, we evaluate the β-VAE-high model using a dataset where all items in a single image share the same size, but such size can vary across images (a condition never encountered during training). Samples from the dataset are shown in Figure B.2.1. This experiment tests whether the model can estimate numerosity independently of item size. The size of an item in this novel dataset is given by the length (measured in pixels) of the edge of the square-shaped items. Possible edge lengths are $4, 8, 12, 16, 20$. Figure 4.2D shows the predicted versus true numerosity for different item sizes. As reported in Table 4.1, the model maintains strong performance, as evidenced by the small MSE values (ranging from 5.52 to 15.48) across all item sizes, except for edge length equal 20 on which case the predicted numerosity consistently underestimates the true numerosity when the number of items in a image is large. Similarly, as evidenced by Figure 4.2E, the model preserves a good number acuity since the Weber fraction remains low ($w \sim 0.11$ to 0.13), except for edge length equal to 20.

Field radius. In this test, we evaluate the β-VAE-high model using a dataset where all items share the same size and they are confined to circular fields of varying radii at the center of the image (a condition never encountered during training). Samples from the dataset are shown in Figure B.2.2. This experiment tests whether the model

can estimate numerosity independently of items' spatial distribution. Possible values of the field radius are $65, 85, 105, 125, 150$. Figure 4.2G show the performance of the model in the numerosity estimation task. As reported in Table 4.1, the model shows improved performance as the radius increases, with MSE values decreasing from 148.09 (for the smallest radius) to 5.81 (for the largest). This trend suggests that spatial crowding in smaller radii introduces challenges for accurate numerosity encoding. Figure 4.2H shows a similar pattern for number acuity. The sigmoidal curves become sharper with increasing radii, reflecting better discrimination of numerosity at larger radii.

Considering the zero-shot nature of these tests, the β-VAE-high model demonstrates a remarkable ability to generalize numerosity estimation to novel item sizes and spatial distributions without any additional training. Despite being trained only on images with randomly sized items and uniform distributions, the model retains strong performance across a wide range of item sizes and field radii. However, we found the generalization ability of the model to break down with stimuli that are very large in size or that are spatially constrained into very small areas. We asked whether this drop in performance depends on a poor latent code or on a poor readout of the latent code using the regressions. To address this question, we repeated the same tests by retraining the linear and logistic regressions on the new datasets (while keeping the weights of the β-VAE-high model frozen). Interestingly, our results show that after retraining the regressions, the model's performance becomes as high as with the training dataset (Figure B.3.1, Table 4.1). This indicates that the model's latent representations capture essential numerosity-related information – and the decline in performance during generalization is due to the read-out (i.e., the regression component), not to the latent representations. A similar trend is observed when testing the control β-VAE-sup-high model, both zero-shot (Figure B.1.1) and after retraining the regressions (Table 4.1).

Finally, to shed further light on the robustness of the β-VAE-high model's numerical code, we repeated the PCA analyses, but considering the two new datasets, item size and field radius. Interestingly, our results show a marked disentanglement of numerosity with respect to both item size (Figure 4.2F) and field radius (Figure 4.2I). This is evident by considering that in the item size dataset, five partially independent subspaces in PC space become apparent (color-coded in Figure 4.2F, top panel), one for each of the five item sizes in the dataset. Importantly, each of these

five subspaces includes data-points that span numerosities from 1 to 50 (color-coded in Figure 4.2F, bottom panel). The same result is apparent when considering the field radius dataset (Figure 4.2I).

Furthermore, the structure of the latent space of the β-VAE-high model is similar to that of the β-VAE-sup-high control model (Figures B.1.1F and I) but, in the latter case, a greater disentanglement of features is evident.

To summarize, the results illustrated here show a remarkable generalization capability of the β-VAE-high model to stimuli that have different statistical properties compared to the training set. This is due to the fact that the unsupervised model forms disentangled representations of numerosity and of other, non-numerical factors of variation, such as the size and spatial distribution of items, despite numerical and non-numerical factors correlate in the training dataset.

4.2.4 Generative capability of the β-VAE-high model

We next assessed the capability of the β-VAE-high model to generate stimuli having a given target numerosity, by selectively activating the model's latent numerosity code. We should note that generating novel stimuli having a given target numerosity is significantly more challenging than simply reconstructing the input image (which is the model's unsupervised objective), and it is considered a challenging tasks even for state-of-the-art generative models based on transformer architectures [81, 82].

To generate novel stimuli, we "prompted" the β-VAE-high with target numerosities, from 1 to 50, by sampling from the model's latent space (see Section 4.4.7 for a detailed description of this sampling procedure). Figure 4.3A shows that the β-VAE-high model tends to generate stimuli with the required numerosity, albeit (as expected) more noisy compared to the stimuli in the training set. Then, we computed the probability of generating a image with M items when the target numerosity is N, with $M,N = 1,\ldots,50$, and we report it in Figure 4.3B. The result shows that generation accuracy is very high for stimuli with low numerosity and decreases for stimuli with high numerosity – with the network systematically generating stimuli having a lower numerosity. These results are coherent with the other analyses of the model's behaviour in numerosity perception tasks.

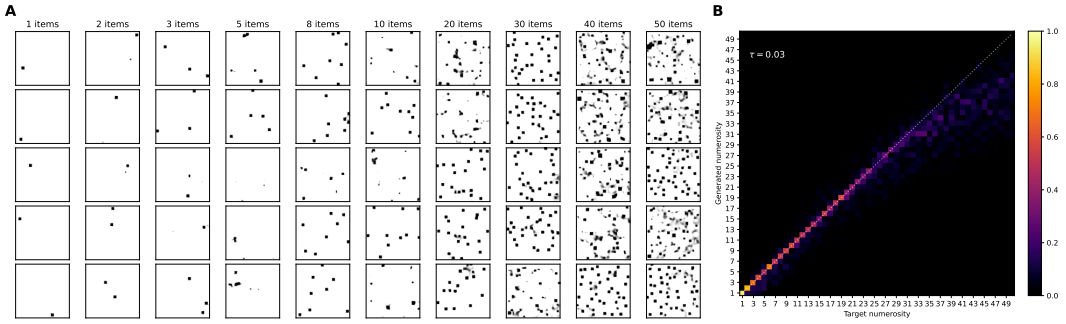

Figure 4.3: Generative capability of the unsupervised β-VAE-high model. (A) Representative generated samples for various target numerosities. (B) Confusion matrix quantifying generation accuracy: For each target numerosity N, 100 images were generated. The matrix shows the probability of generating an image with M items, given the target numerosity N. See Section 4.4.7 for more details.

4.2.5 Rate-distortion trade-off when varying encoding capacity

The β-VAE-high evaluated so far was trained to reconstruct images, with a high encoding capacity (5000 nats). As discussed, this unsupervised objective endows the model with a robust numerosity code. Here, we aim to explore whether a rate-distortion trade-off emerges in numerosity perception, implying that reducing the information that the model can encode about the input reduces its numerosity perception capability.

Specifically, RTD formalizes the fact that low capacity training produces a compression of the neural code and we are interested in assessing the effects of such compression on the model's numerosity perception capability. To explore the impact of encoding capacity on numerosity perception, we exploit the formal relation between the RTD and the encoding capacity of β Variational Autoencoders (Section 4.4.1) and we train β Variational Autoencoders at seven levels of capacity (50, 100, 300, 500, 1000, 2500 and 5000 nats; note that the β-VAE trained at 5000 nats is the one considered in the previous analyses).

The results, summarized in Figure 4.4, reveal a clear rate-distortion trade-off: as the encoding capacity increases, the MSE of the numerosity estimation task decreases, and the Weber fraction of the numerosity discrimination task also decreases. This analysis provides a direct visualization of the trade-off between compression efficiency and the accuracy of numerosity representation, confirming that the perfor-

mance of the model is directly related to the amount of information it can encode about the input.

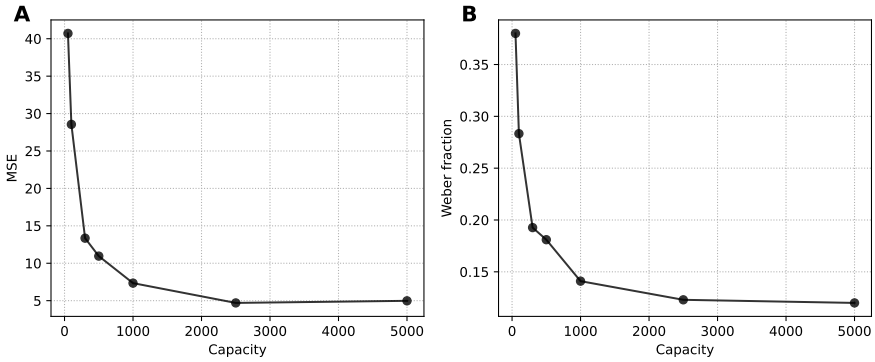

Figure 4.4: Rate-distortion trade-off in numerosity perception when varying the encoding capacity of β Variational Autoencoders . (A) Mean Squared Error (MSE) of the numerosity estimation task as a function of the encoding capacity of the β Variational Autoencoder. (B) Weber fraction for the numerosity discrimination task as a function of the encoding capacity of the β-VAE model. The tested capacity levels were 50, 100, 300, 500, 1000, 2500 and 5000 nats. Note that the model trained at 5000 nats corresponds to the β-VAE-high model illustrated in Figure 4.2. Rather, the model trained at 50 nats corresponds to the β-VAE-low model illustrated in Figure 4.5.

4.2.6 The effects of low capacity training in the β-VAE-low model

The previous analysis showed a rate-distortion trade-off in numerosity perception. Here, we aimed to get more insight on how low capacity training influences the numerosity perception abilities and the neural codes exhibited by the model. Furthermore, we are interested in assessing whether the effects of low capacity training could be comparable to deficits of numerical cognition, like dyscalculia.

To address these questions, we performed detailed behavioral and neural analyses (Figure 4.5) using the β Variational Autoencoder trained at the lowest level of capacity (50 nats) in the previous analysis. We refer to this model as the β-VAE-low model.

Our analyses shows that compared with the β-VAE-high model, the β-VAE-low model shows an impaired performance in numerosity perception tasks, which is evident by the large variance and systematic underestimation of large numbers in the numerosity estimation task (Figure 4.5A), where it achieves a MSE $= 40.71$. The

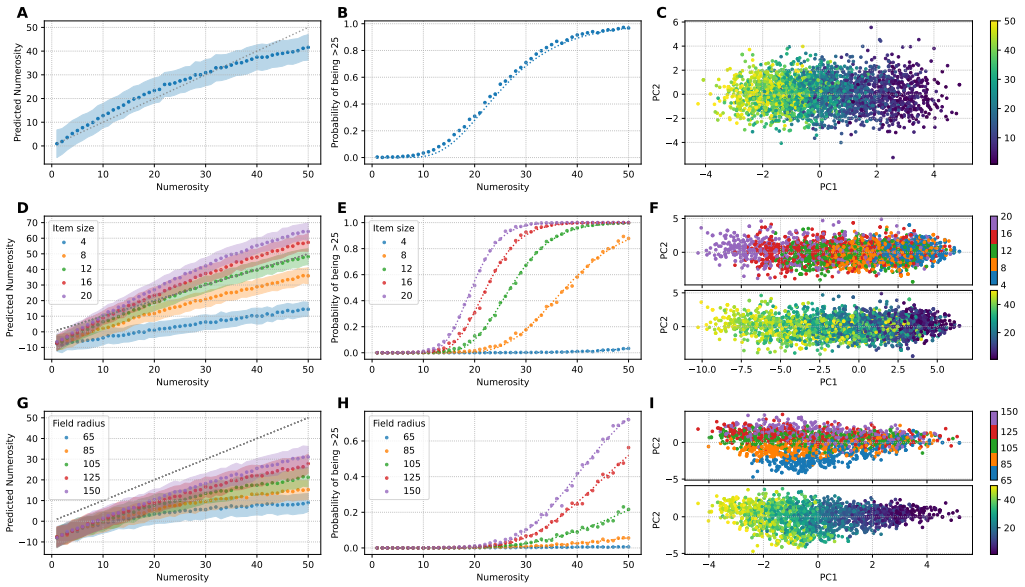

Figure 4.5: Analysis of the β-VAE-low model. Panels A-I are the same as Figure 4.2.

large Weber fraction ($w = 0.38$) also confirm a poor number acuity in numerosity discrimination task (Figure 4.5B). Furthermore, the neural space is significantly more compressed, with the range of variation of PC1 and PC2 being one order of magnitude smaller than in the β-VAE-high model (Figure 4.5C). The capacity constraint also significantly reduces the number of neurons involved in numerosity perception (~ 70) compared to the β-VAE-high model, as evident from Figure B.5.1C. Finally, the β-VAE-low show a reduced ability to generalize to the novel datasets "item size" (Figure 4.5D-E) and "field radius" (Figure 4.5G-H) and to correctly disentangle their latent factors of variation (Figure 4.5F and I). Retraining the readout significantly improves generalization, but the severe capacity constraints do not allow the model to perform as the β-VAE-high model, suggesting that (unlike the β-VAE-high) part of the problem is in the latent representation (Figure B.3.2). These results are also confirmed by the numerical values in Table 4.1.

As control, we trained the supervised counterpart of β-VAE-low model, named β-VAE-sup-low, using the same low-capacity constraints of $C_{\max} = 50$ nats. The results are illustrated in Figure B.1.2 and numerical values are reported in Table 4.1. As before, the β-VAE-sup-low model certainly benefits from supervision during training showing superior performance compared to the β-VAE-low model in both numerosity estimation and discrimination tasks and in generalizing to novel images.

Interestingly, the two models have a similar latent space structure.

In sum, our results indicate that the ability of the β-VAE-high model to support numerosity perception tasks and generalization to novel inputs is severely limited when the model is trained under strict capacity constraints.

Potential relationship with dyscalculia

The Diagnostic and Statistical Manual of Mental Disorders [83] defines dyscalculia as a specific learning disability characterized by difficulties in number sense, memorization of arithmetic facts, accurate or fluent calculation, and accurate mathematical reasoning . Within this broad definition, various classifications of dyscalculia have been proposed [84], including verbal dyscalculia (difficulties in naming numbers and mathematical symbols), lexical dyscalculia (challenges in reading mathematical symbols), operational dyscalculia (difficulties in performing calculations), and ideognostical dyscalculia (difficulties in understanding mathematical concepts and performing mental calculations). This heterogeneity suggests that a singular computational model might not fully capture the diverse manifestations of dyscalculia, and different subtypes could potentially exhibit distinct information processing characteristics.

Several core cognitive deficits have been consistently observed in individuals with dyscalculia [85, 86]. One prominent theory posits a deficit in number sense, specifically an impairment in the Approximate Number System (ANS) [85]. The ANS is responsible for the intuitive understanding of quantity and the ability to compare numerical values without counting [62]. Individuals with dyscalculia often show a less precise ANS, indicated by a reduced sensitivity to numerical differences in comparison tasks [79]. This lower precision in the internal representation of numerical magnitudes could be interpreted as a higher level of distortion for a given amount of encoded information. Additionally, difficulties with subitizing, the rapid identification of small quantities, have been noted in individuals with dyscalculia [85], possibly reflecting a lower rate of information processing for small numerical sets.

Existing computational models of dyscalculia have attempted to capture these various cognitive deficits but they often target specific aspects of numerical cognition.

Rate distortion theory could potentially offer a more overarching framework to understand how limitations in information processing efficiency and accuracy across these different aspects contribute to the overall disorder.

To the best of our knowledge, there is currently a lack of research explicitly applying rate distortion theory to develop computational models of dyscalculia. While several studies discuss rate distortion theory in the context of working memory and neural coding, which are pertinent to cognitive deficits observed in dyscalculia [87], a direct application of this framework to model the specific numerical processing challenges in dyscalculia appears to be a significant gap in the current research. This absence presents a notable opportunity for future research to explore the potential of rate distortion theory in providing a novel perspective on the information processing limitations in this learning disability.

In our results, when training the β-VAE at low capacity, we observed a weber fraction of 0.38 that is compatible with the one observed in [79] for 10 years old dyscalculic children. This observation reinforces the fact that rate distortion theory may be a good perspective for examining learning disabilities like dyscalculia in the future.

4.3 Discussion

In this study, we investigate the emergence of number sense in unsupervised generative models through the lens of rate-distortion theory (RDT), a normative framework to understand information processing under limited resources. We use RDT to examine whether known psychophysical phenomena related to numerosity emerge in unsupervised generative models with bounded resources and how capacity limitations influence the models' performance in numerosity perception tasks and their numerical representations.

For our investigation, we trained a β-Variational Autoencoder (β-VAE) on a dataset commonly used in number sense experiments, in an unsupervised manner. This specific class of models incorporates fundamental principles of RDT and permits addressing visual image learning. As control, we also designed a supervised counterpart of the model (β-VAE-sup) that was explicitly trained to estimate nu-

merosity. The latter allowed us to better understand the impact of task-specific training on the latent representations of the model.

We then conducted a series of behavioral and neural analyses to evaluate the extent to which the model's representations can support linear readouts in numerosity perception tasks. These analyses were structured to mimic experimental paradigms used in human psychophysics studies, allowing for direct comparisons between the model's performance and human numerosity perception. Linear readouts allow to extract numerosity information from the model's latent space while minimizing the potential for confounding factors arising from more complex, non-linear decoding architectures. In fact, complex decoders could potentially compensate for poor latent representations by learning intricate mappings, obscuring the true nature of the encoded information. By using linear readouts, we ensure that any observed performance reflects the inherent structure of the latent space itself.

The results of our behavioral analyses (numerosity estimation and discrimination tasks) indicate that the unsupervised β-VAE-high model successfully supports numerosity perception tasks without being explicitly trained to do so. Interestingly, the unsupervised β-VAE-high model shows key signatures of human number sense, such as a greater sensitivity to smaller compared to larger numbers. For small numbers, the accuracy is excellent, whereas for larger numbers the mean predictions are accurate but their standard deviation increases. The greater sensitivity to smaller compared to larger numbers is a ubiquitous feature of human numerosity perception and it is typically explained as an effect of a greater familiarity with smaller numbers [88]. In contrast, in our model, smaller and larger numbers are equally present in the training dataset. The lower sensitivity to larger numbers emerges from lossy information compression during the encoding phase. This process happens also at high capacity, suggesting that the information compression – and the noise in the neural representation of stimuli – is more impairing for the discrimination or larger compared to smaller numbers [89]. This effect is mitigated but not eliminated in the supervised model β-VAE-sup-high that is explicitly trained to estimate numerosity and hence (differently from the unsupervised β-VAE-high) devotes a large portion of its resources to this task. Another key feature exhibited by the unsupervised β-VAE-high model is a discrimination ability (Weber fraction $w = 0.12$) in the range of human psychophysics, reflecting a number acuity comparable to that observed in adult human populations [79]. In contrast, the supervised β-VAE-sup-high model shows a superhuman performance (Weber fraction $w = 0.05$). Finally, the β-VAE-

high model shows excellent reconstruction capabilities, affording accurate numerical perception tasks not only with input images but also with reconstructed images.

We next assessed the nature of the neural representations that emerged during the training of the unsupervised β-VAE-high model. Our findings reveal that despite the absence of explicit numerosity labels during training, the β-VAE-high model encodes robust numerical representations that are comparable to the neural code developed by a model trained with the supervised objective to estimate numerosity, but in which numerosity information is encoded in a more distributed manner. The fact that robust numerical codes emerge in unsupervised systems is important since it indicates that, first, these codes need not to be innate, and second, that living organisms could learn latent codes implicitly supporting a "number sense" from statistical learning unrelated to numerical perception, without (or before) having the explicit goal to count. These codes are not qualitatively different from those acquired through supervised learning with explicit labels, suggesting that supervised or "goal-driven" [49] learning approaches are not always necessary to develop internal codes supporting efficient solutions to perceptual and cognitive tasks. This perspective is in keeping with mounting evidence about the importance of unsupervised (pre)training in the brain [90].

Furthermore, we assessed the robustness and generality of the learned latent representations by measuring the ability of the unsupervised β-VAE-high model to generalize to novel images with respect to the training data in terms of size and spatial distribution of the items. We conducted all the evaluations in a zero-shot manner, i.e., without re-training or finetuning the model or the linear readouts. Considering the zero-shot nature of tests, our results show a good generalization capability of the β-VAE-high model to encode novel inputs even though we observe a substantial performance degradation on inputs that have very different statistical properties compared to the training set. We thus conducted further experiments to understand if the performance degradation is due to a poor latent representation or to oversimplified readout. To this end, we retrained the readouts on the novel datasets but keeping model's representations frozen as before. We observed a significant improvement of all the metrics to the level of the training set, indicating that probably the model representations are robust enough and that the problem is in the linear readouts. Further indications of a robust latent space are given by the dimensionality reduction, that shows a rather disentangled latent space.

We also tested the generative aspects of the unsupervised β-VAE-high model and its capability to exploit the structure of the latent space to generate images having a given target numerosity. This is a particularly challenging task for existing generative models addressing numerical tasks [73, 72]. Our analysis shows that in the unsupervised β-VAE-high model, the accuracy of this generative process is very high for stimuli with low numerosity and decreases for stimuli with high numerosity – with the network systematically generating stimuli having a lower numerosity. These results are coherent with the other analyses of the model's behaviour in numerical perception tasks.

Finally, we compared the numerosity perception capabilities of β-Variational Autoencoders trained at seven different encoding capacities, from high to low. Our results reveal a clear rate-distortion trade-off, with the model's number acuity decreasing as a function of the amount of information it can encode about the input. A more detailed behavioral and neural analysis was then carried out on a model with severe encoding capacity constraints (β-VAE-low). Our analyses shows that the low capacity training significantly impairs performance in numerosity perception tasks. Furthermore, low capacity training produces neural codes for numerosity that show a much reduced generalization to novel inputs that, however, improves when retraining the linear readouts. The latent space appears more compressed and much less disentangled. The numerosity information is encoded in a less distributed manner among neurons compared to the high capacity β-VAE-high, since only few neurons appear to be correlated with numerosity and contributing to form a representation of numerosity estimates.

Taken together, these results show that interesting numerosity perception capabilities and a robust encoding of numerosity emerge in an unsupervised β-VAE-high model trained at high capacity to reconstruct the input, without labels related to numbers. The β-VAE-high model shows signature of human psychophysical experiments, including an underestimation of high numbers. The fact that these effects emerge also at high capacity indicates that they are intrinsic properties of statistical learning of numerosity stimuli. These effects are mitigated but not eliminated in the supervised β-VAE-sup-high model that, being explicitly trained to estimate numerosity, dedicates a large portion of its resources to build numerosity-specific codes. Instead, training at low capacity (β-VAE-low) significantly impairs numerosity perception.

More broadly, the results presented here are relevant to theories that emphasize the emergence of sophisticated cognitive abilities in generative models through unsupervised learning objectives. Previous studies have shown that inferential and predictive processing mechanisms, closely related to the variational approach of the β-VAE, can support various cognitive abilities, ranging from (active) perception to planning and cognitive control [31, 91–93]. Our study extends this line of inquiry by demonstrating that numerosity perception can be conceptualized within the same framework but future studies are needed to further assess the empirical validity of this claim.

In sum, By integrating insights from numerosity perception and information theory, our study aims to contribute to the understanding of how abstract numerical concepts can emerge in generative models, under resource constraints. More broadly, this work sheds light on the computational principles that may underlie numerosity perception in biological systems and offers a framework for investigating the intersection between cognitive science and artificial intelligence.

4.4 Methods

4.4.1 Models

As explained in Section 2.2, we use the β-VAE architecture (Figure 3.11A) as computational model since it represents an effective approximation of RDT for high-dimensional data, such as images [43, 20].

Unsupervised Model. The core of our analysis relied on a classical β-VAE architecture, which operates in an entirely unsupervised manner. The model learns to compress input images (256x256 pixels) into a 256-dimensional latent space, capturing the essential statistical structure of the data without explicit supervision. The β-VAE consists of two main components: an encoder and a decoder. Our encoder block comprises 5 convolutional layers, each followed by a leaky ReLU activation function [56, 57]. The encoder maps input images to a 256-dimensional latent space, where each dimension represents a compressed feature of the input data. Our decoder block consists of 5 transposed convolutional layers, also followed by leaky ReLU activations, except for the final layer, which uses a sigmoid activation to

output grayscale images. The decoder reconstructs the input images from the latent representations.

In our experiments, inspired by [38], we trained the model using the loss function in Equation 3.1, since it allows to control the encoding capacity C of the network by penalizing with a factor β any deviation of the KL divergence from the required capacity C. However, differently from the training procedure in Section 3.4.1, C is not gradually increased from 0 to a target value $C_{\max}$ during training but it is fixed at $C = 5000$ nats. In addition, we used Adam optimizer with a learning rate of 10^{-3}.

Supervised Model. We also implemented and tested a supervised version of the β-VAE, which we refer to as β-VAE-sup. This model incorporates an additional numerosity estimation module f that is explicitly trained to predict the (ground truth) numerosity y associated to the latent vectors z. The numerosity estimation module is implemented as a linear regression layer attached to the latent space of the β-VAE (Figure 4.1E). The supervised model is thus trained using a modified loss function:

$$\begin{aligned}&\mathscr{L}(\theta,\phi;x,z,y,\beta) = \\ &\quad \mathbb{E}_{q_\phi(z|x)}\left[-\log p_\theta(x \mid z)\right] + \beta\left|D_{KL}\left(q_\phi(z \mid x)\|p(z)\right) - C\right| + \text{MSE}(f(z),y),\end{aligned} \tag{4.1}$$

where $\text{MSE}(f(z),y)$ is the mean squared error of the numerosity estimation module. This joint training procedure allows the supervised model to learn numerosity-informed latent representations, while still maintaining the generative capabilities of the β-VAE. The training procedure and hyperparameters are the same as the unsupervised model.

Low-Capacity Model To investigate the impact of resource limitations on numerosity perception, we also trained a low-capacity version of the β-VAE-high, referred to as β-VAE-low. This model is identical in architecture to the high-capacity β-VAE-high, uses the loss in Equation 3.1, but is constrained to an encoding capacity of $C = 50$ nats. This severe capacity constraint forces the model to compress the input data more aggressively, leading to a loss of fine-grained information in the latent space. We used this model to study the effects of capacity limitations on the emergence of numerical representations and to draw parallels to conditions like dyscalculia, which is characterized by deficits in numerical processing.

Rate-distortion trade-off. To explicitly demonstrate the rate-distortion trade-off predicted by RDT, we conducted experiments varying the encoding capacity, C, of the β-VAE model. Specifically, we trained multiple β-VAEs with different fixed capacity values, ranging from 50 nats to 5000 nats. More specifically, the capacity values used for training are 50, 100, 300, 500, 1000, 2500 and 5000 nats. For each capacity level, we then evaluated the model's performance on both the numerosity estimation task (quantified by the Mean Squared Error, MSE) and the numerosity discrimination task (quantified by the Weber fraction, w) as a proxy for distortion measure. By plotting MSE and w as a function of C, we generated rate-distortion curves (Figure 4.4) that visually illustrate the trade-off between compression efficiency and the accuracy of numerosity representation and perception. These curves provide empirical evidence for the theoretical predictions of RDT within the context of unsupervised numerosity perception.

4.4.2 Dataset

All the models are trained on a dataset commonly used in number sense experiments. The dataset consists of 50.000 black-and-white 256x256 pixel images. Each image contains a random number of black squares, ranging from 1 to 50. The positions of the squares are uniformly distributed across the image. Notably, the dataset generation ensures that the total number of black pixels is not correlated with the numerosity, preventing the model from using pixel density as a proxy for numerosity. This design facilitates the investigation of robust numerosity representation in the latent space.

4.4.3 Behavioral Analysis

To evaluate the capabilities of the generative models to support numerical perception tasks, we conducted a two behavioral tests. The behavioral tests were structured to mimic experimental paradigms commonly used in human psychophysics studies, allowing for a direct comparison between the models' performance and human numerical perception.

Numerosity estimation. To quantify the ability of the unsupervised model to encode estimates of numerosity, we trained a linear regression model on top of the latent representations of the pre-trained β-VAE. This regression model predicts the number of items N in the encoded images z. For each numerosity class N (ranging from 1 to 50), we computed the mean and standard deviation of the predicted estimates. This allows us to evaluate both the accuracy and uncertainty of the predictions for each numerosity class. For the supervised model β-VAE-sup, we did not train a separate linear regression, as we directly used the linear module that was jointly trained with the model (Figure 4.1C).

Numerosity discrimination. To quantify the model's ability to support numerosity discrimination, we estimated the Weber fraction [94], a measure commonly used in numerical perception research. In numerosity discrimination tasks, participants are typically asked to judge whether the numerosity represented in a given image is larger or smaller than a reference numerosity n_{ref}. In our study, we trained a logistic regression to classify latent vectors z based on whether they corresponded to images with numerosity N smaller (class 0) or larger (class 1) than $n_{\text{ref}} = 25$. This task can be modelled using the logarithmic number line framework, where the internal representation of numerosity is assumed to be noisy and follows a Gaussian distribution on a logarithmic scale [88]. More specifically, the numerosity N of a set of objects is represented internally by a Gaussian random variable $R \sim \mathcal{N}(\log(N), w^2)$ where the standard deviation w is known as Weber fraction and quantifies the precision of the internal representation. A smaller Weber fraction indicates better acuity in numerosity discrimination. In a larger-smaller judgment task, the optimal decision strategy under a maximum likelihood criterion is to respond "larger" whenever the internal sample R exceeds a criterion value c. The criterion c should coincide with the internal representation of the reference numerosity n_{ref}. For the experiments in which the training and test data of the logistic regression have the same distribution, we assume $c = \log(n_{\text{ref}})$, i.e., the representation of the reference numerosity is unbiased. Otherwise, when the logistic regression is tested on a dataset statistically different than the training one (like in the zero-shot tests on item size and field radius datasets), we estimate c from the data. This explains why in some rows of Table 4.1 the value for n_{ref} is different than 25 (it represents the exponential of c, when c is treated as a free parameter). Thus, the probability of responding "larger" to a target numerosity N is given by the integral of the Gaussian distribution from c

to infinity:

$$P_{\text{larger}}(N, n_{\text{ref}}) = \int_c^{\infty} \frac{1}{\sqrt{2\pi}w} \exp\left(-\frac{(r-\log(N))^2}{2w^2}\right) dr \tag{4.2}$$

This integral can be simplified using the error function (erf):

$$P_{\text{larger}}(N, n_{\text{ref}}) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{\log(N)-c}{\sqrt{2}w}\right)\right) \tag{4.3}$$

We then fitted the model to behavioral data (logistic regression outputs) by treating w, and in some cases as explained above also c, as a free parameter. This approach allows us to directly compare the unsupervised and supervised models under varying capacity conditions.

4.4.4 Neural Analysis

We conducted several neural analyses to examine the latent representations learned by the models. This involved:

Principal Component Analysis. To analyze the structure of the latent space and assess the disentanglement of numerosity from other visual features, we performed Principal Component Analysis (PCA) on the 256-dimensional latent representations of the input images. To visualize the organization of the latent space, we plotted the projected latent vectors in the 2D subspace spanned by PC1 and PC2, coloring each point according to the numerosity N or other relevant features (e.g., item size or field radius). This analysis allowed us to evaluate the extent to which numerosity and other visual features are disentangled in the latent space, providing insights into the structure of the learned representations.

Spearman Correlation. To quantify the relationship between individual neuron activations in the latent space and the numerosity of the input images, we computed the Spearman rank correlation coefficient. For each neuron z_i ($i = 1, \ldots, 256$), we calculated the correlation between its activation values and the numerosity N of the corresponding input images. The Spearman correlation coefficient ρ_i for neuron z_i is defined as $\rho_i = \text{Spearman}(z_i, N)$, where $\text{Spearman}(z_i, N)$ measures the strength

and direction of the monotonic relationship between the neuron's activations and the numerosity. Neurons with high absolute values of ρ_i are considered particularly relevant for numerosity estimation. We report the Spearman correlation coefficients for all neurons in Figure B.5.1.

Mutual Information. To further investigate the encoding of numerosity in the latent space, we computed the mutual information $I(z_i;N)$ between the activations of individual neurons z_i and the numerosity N of the input images. Mutual information quantifies the reduction in uncertainty about the numerosity N given the knowledge of a neuron's activation z_i. Since the numerosity N is a discrete variable, we compute the mutual information I using the non-parametric method based on entropy estimation from k-nearest neighbors distances as described in [95] and [96]. To examine whether the models learned distinct mechanisms for processing small and large numerosities as observed in human cognition [97], we divided the training set into two subsets, small ($N \leq 4$) and large ($N \geq 47$) numerosities. Then, on each subset, we computed the mutual information $I(z_i;N)$. This allowed us to identify neurons that were selectively tuned to either small or large numerosities, providing insights into the specialization of the latent space for different numerical ranges. We report the results in a 2D plane, where each point is a neuron, with the x-axis denoting the mutual information between its activity and small numerosities, while the y-axis denotes the mutual information between its activity and large numerosities (Figure B.5.2). The range $1-4$ for small numbers aligns with previous studies in the field [97, 63]. Consequently, we choose the range $47-50$ for large numbers to preserve symmetry and balance in the data.

Heatmap of neural activity. To further investigate the relationship between latent neuron activity and numerosity, we conducted an analysis to quantify the average activity of individual latent neurons in response to images containing different numbers of objects. First, we standardized the activity of each neuron across the entire dataset using z-score normalization. This ensured that all neurons were on a comparable scale, mitigating the effects of differing baseline activity levels. Subsequently, the images were grouped based on their object count (numerosity). For each numerosity class, we computed the average activity of each latent neuron across all images belonging to that class. This resulted in a matrix of average neuron activities, with rows representing individual neurons and columns representing

different numerosities. This matrix was then visualized as a heatmap to reveal patterns of neuron activity as a function of numerosity, as depicted in Figure B.5.3A. This analysis allowed us to identify neurons that exhibited preferential activation for specific numerosity ranges, providing insights into how numerosity information is encoded within the latent space of the models.

4.4.5 Robustness with respect to non-numerical magnitudes

To assess the models' ability to generalize beyond the training distribution and to determine if they rely on spurious correlations instead of robust numerosity representation, we evaluated their behavioral and neural performance on novel datasets with varying item sizes and field radii. These datasets present conditions not encountered during training, allowing us to examine the robustness of the learned numerosity representations. Specifically, we performed zero-shot evaluations, i.e., directly performing inference on the new datasets without retraining any model component (linear or logistic regressions).

Item Size Dataset. We created a dataset where all items within a single image have the same size, but this size varies across different images (Figure B.2.1). The size of an item (square) is given by the length of its edge measured in pixels. Possible values of item size are $4, 8, 12, 16, 20$. This contrasts with the training data, where item sizes within each image were randomly varied. This variation challenges the models to estimate numerosity independently of item size.

Field Radius Dataset. We also created a dataset where items are constrained to appear within a specific field radius measured from the center of the image (Figure B.2.2). The field radius is measured in pixels and possible values are $65, 85, 105, 125, 150$. This spatial constraint, not present in the training data (where items are uniformly distributed), probes the models' ability to estimate numerosity independently of the spatial distribution of the items.

For both Item Size and Field Radius datasets, we conducted the numerosity estimation and numerosity discrimination tasks as described in Section 4.4.3. We analyzed the performance metrics (mean, standard deviation, Weber fraction) to

quantify the degree of generalization achieved by each model. Furthermore, we performed Principal Component Analysis (PCA) on the latent representations obtained from these new datasets to examine how the internal organization of the latent space adapts to the novel stimuli and whether numerosity remains disentangled from item size and field radius. The results of the unsupervised β-VAE-high on the new datasets are displayed in Figure 4.2D-I, while the results of the β-VAE-sup-high model are reported in Figure B.1.1D-I. Instead, Figure 4.5D-I shows the results of the β-VAE-low model.

4.4.6 Reconstruction capability

To rigorously assess the models' ability to reconstruct the input data from their latent representations, we performed a reconstruction fidelity analysis. This analysis evaluated the quality of the reconstructed images and served as a complementary assessment of the information preserved within the latent space. Specifically, we passed the images from the test set through the model's encoder to obtain their latent representations. These latent representations were then fed into the model's decoder to generate reconstructed images. The reconstructed images are then encoded again through the model's encoder to obtain a latent representation of the reconstructed images. To quantify the fidelity of the reconstructions, we repeated the numerosity estimation and numerosity discrimination tasks, as detailed in Section 4.4.3, using the latent representations of reconstructed images as input. This allowed us to directly compare the performance of the models on the original images with their performance on the reconstructed images. Any significant degradation in performance on the reconstructed images would indicate a loss of relevant information during the encoding and decoding process, highlighting limitations in the reconstruction fidelity of the model. The performance metrics obtained from these analyses (mean, standard deviation of estimates, Weber fraction) were used to quantitatively evaluate the reconstruction capability of each model. The results are reported in Figure B.4.1.

4.4.7 Generative capability

To evaluate the generative performance of the models, we established a procedure for generating novel images conditioned on a specified target numerosity, without providing any input stimulus. This assessment aimed to determine whether the

models could synthesize realistic and controllable images that accurately reflected the desired numerical quantity. This involved characterizing the learned latent space representation of each numerosity and subsequently sampling from these representations to generate new visual stimuli.

Latent Space Characterization. For each target numerosity class N, we modeled the distribution of latent vectors z corresponding to images with that given numerosity using a Gaussian Mixture Model (GMM) with $K = 28$ components. The GMM provides a compressed probabilistic representation of the latent space for each numerosity, allowing us to sample representative latent vectors. The model is defined as:

$$p(z|N) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z|\mu_k, \Sigma_k),$$

where π_k are the mixture weights, μ_k are the mean vectors, and Σ_k are the covariance matrices for the k-th Gaussian component. The matrices Σ_k are assumed to be diagonal. The parameters of the GMM (mixture weights, means, and covariances) were estimated using the Expectation-Maximization (EM) algorithm, iteratively optimizing the likelihood of the latent representations of the training images with numerosity N.

Conditional Image Generation. To generate a novel image conditioned on a target numerosity N, we first sample a latent vector z from the corresponding GMM, $z \sim p(z|N)$. The sampled latent vector z then serves as input to the decoder network $p_\theta(x|z)$, producing the generated image $\hat{x}$. This process allows us to synthesize images whose latent representations are statistically consistent with the specified numerosity.

Evaluation of Generated Images. To quantitatively evaluate the accuracy of the generated numerosity, we used the following procedure: given a generated image $\hat{x}$, since it is a grayscale image, we first binarize it by applying a threshold $t = 0.016$. Pixels with intensity values less than or equal to t were set to black (item), while pixels with intensity values greater than t were set to white (background). Then we count the number of connected components in the binary image. Each connected component corresponds to an item. In this way, we can count the number of items

in a generated image. For each target numerosity N, we generated 100 images and computed the probability of generating an image with M items when the target numerosity is N. We report the probabilities in a confusion matrix (see Figure 4.3B), providing a quantitative measure of the model's ability to control numerosity during generation. Note that the threshold t and the number of components (K) were determined via a grid search to maximize the average F1-score across all numerosity classes.

Chapter 5

Conclusions

5.1 Summary of Findings

This thesis has explored the fundamental principles governing efficient coding in cognitive systems, with a particular focus on how Rate-Distortion Theory (RDT) shapes the geometry of latent representations in generative models. By bridging information theory, computational neuroscience, and artificial intelligence, this work has provided compelling evidence for the unifying role of RDT in understanding the emergence of structured internal representations under resource constraints.

The core argument of this thesis is that cognitive systems, whether biological or artificial, operate under inherent resource limitations that necessitate efficient coding strategies. These limitations, formalized by RDT, impose a trade-off between the fidelity of information representation and the capacity of the encoding system. As a consequence, internal models of the world are not perfect mirror images, but rather simplified, abstracted, and systematically distorted by the constraints of the encoding process.

Chapter 3, "The Geometry of Efficient Codes," systematically investigated how varying model capacity, introducing biases in the training data, and imposing specific task objectives affect the organization and structure of latent representations learned by β-VAEs. This research identified key types of geometric distortions, including prototypization, specialization, and orthogonalization, that emerge as principled adaptations to different RDT constraints. Specifically, we found that:

- **Prototypization:** Under limited capacity, models tend to collapse similar stimuli into prototypical representations, sacrificing fine-grained details for efficient encoding.
- **Specialization:** Biased training data leads to an over-representation of frequent stimuli at the expense of rare ones, reflecting an adaptive allocation of resources.
- **Orthogonalization:** Task objectives drive the formation of orthogonal representations that facilitate discrimination and classification, enhancing task performance.

Chapter 4, "Number Sense in Unsupervised Generative Models," delved into a specific cognitive capacity: numerosity perception. This research demonstrated that a capacity for number sense can emerge spontaneously in unsupervised β-VAEs trained solely to reconstruct visual scenes containing varying numbers of objects. The models exhibited key signatures of human numerosity perception, including Weber's law scaling and distinct mechanisms for processing small versus large numerosities. Furthermore, we found that imposing severe capacity limitations on the model led to deficits in numerosity perceptionsuggesting a link between resource constraints and cognitive impairments. For this reason we advance the hypothesis that resource limitations can play a role in developmental dyscalculia, but this must be accurately tested in future works. This work further showed that the unsupervised model developed robust coding for numerosity that generalized to novel datasets and permitted generating novel images having the desired number of items.

These findings, taken together, provide a cohesive picture of how efficient coding principles shape cognitive representations. The geometric distortions observed in the latent spaces of β-VAEs are not arbitrary failings, but rather adaptive features that reflect an optimized balance between fidelity and resource expenditure. The emergence of number sense in unsupervised models demonstrates that complex cognitive abilities can arise spontaneously from the efficient coding of sensory inputs, without explicit supervision for specific tasks.

5.2 Theoretical Implications

This thesis has significant implications for Rate-Distortion Theory and the efficient coding framework in cognitive science. By providing compelling computational evidence for the unifying role of RDT in shaping latent representations, this work strengthens the argument that the brain operates under principles analogous to those formalized in information theory. The brain, operating under severe metabolic and physical constraints, can be viewed as a resource-limited information processing system, suggesting that RDT provides a normative lens through which to examine how the brain achieves efficient and adaptive representations of the world.

The identified geometric distortions – prototypization, specialization, and orthogonalization – offer a taxonomy of distortions and a framework for understanding their functional significance in efficient coding. These distortions are not merely imperfections in representation, but rather potentially adaptive features that reflect the system's optimized balance between fidelity and resource expenditure. These concepts offer valuable insights for deep learning models since they can be used to diagnose and address issues such as mode collapse, overfitting, and lack of generalization.

Specifically, the emergence of number sense in unsupervised models demonstrates that complex cognitive abilities can arise spontaneously from the efficient coding of sensory inputs. This finding challenges the traditional view that such abilities require explicit supervision or innate mechanisms, suggesting that the brain may leverage statistical learning and efficient coding principles to develop fundamental cognitive capacities. The dyscalculia analogy further underscores the importance of resource constraints in shaping cognitive abilities, suggesting that deficits in numerosity processing may arise from limitations in encoding capacity.

Importantly, this work contributes to a growing trend in neuroscience that moves beyond the analysis of single neurons towards understanding the collective behavior of neuronal populations. Traditional neuroscience has often focused on the function of individual neurons, but recent advances in recording and analysis techniques have enabled researchers to study the activity of large populations of neurons simultaneously. This shift has led to the realization that cognitive functions are often implemented by the coordinated activity of distributed neuronal ensembles, rather than by the isolated activity of single cells [98–101]. A central aspect of

this population-level approach is the analysis of the geometry of neural activity [102, 41, 23], using methods such as dimensionality reduction and manifold analysis [103, 104]. These techniques reveal the underlying structure and organization of neural activity, allowing researchers to identify the low-dimensional manifolds that capture the essential dynamics of cognitive processes. Results from this line of research have provided insights about the neural coding underlying control of movement [105], working memory [106], value-based decision making [107] and other cognitive functions. The methods developed in this thesis, which characterize the distortions in latent representations, can directly support this type of analysis by revealing the structure and characteristics of neuronal populations under different computational constraints. By providing a framework for understanding how resource limitations shape the geometry of neural activity, this thesis contributes to the emerging field of "geometric neuroscience" and offers a set of tools for analyzing the complex dynamics of neuronal populations.

5.3 Limitations

This thesis has several limitations that should be acknowledged. First, as explained in Section 2.2, the β-VAE is only an approximation of the true Rate-Distortion Theory optimization. The approximation arises from the use of variational inference, the limited capacity and architectural biases of the neural networks, and the choice of negative log-likelihood as a distortion measure and KL divergence as a rate proxy.

Second, the findings are based on specific datasets and tasks, and may not generalize to other domains. The "Corridors" dataset, while carefully designed to isolate and control generative factors, is a simplified representation of real-world sensory inputs. Similarly, the number sense experiments focused on a specific visual numerosity task, and may not capture the full complexity of numerosity perception.

Third, the β-VAE architecture, while effective for exploring rate-distortion trade-offs, is a relatively simple model compared to the intricate architectures of biological neural networks. More complex models, such as hierarchical or conditional β-VAEs, may reveal further insights into the mechanisms of efficient coding.

Fourth, while this work provides valuable insights into the emergence of number sense in unsupervised generative models, more sophisticated analysis methods,

which are based on interventions [53], may be needed to characterize the causal relationships between latent representations and cognitive abilities.

5.4 Concluding Remarks

In conclusion, this thesis has provided compelling evidence for the unifying role of Rate-Distortion Theory in shaping cognitive representations. By demonstrating how efficient coding principles manifest in the geometry of latent spaces in generative models, this work has offered a valuable framework for understanding the emergence of intelligence in both biological and artificial systems. The findings of this thesis contribute to a deeper understanding of the fundamental principles that govern the emergence of intelligent behavior in both biological and artificial systems and offer a valuable framework for future research aimed at reverse-engineering the objective functions that shape the neural codes of the brain.

This research has highlighted the importance of considering resource limitations when studying cognitive systems and has demonstrated that the seemingly imperfect representations that arise under these constraints are not arbitrary failings, but rather potentially adaptive features that reflect an optimized balance between fidelity and resource expenditure. By bridging the gap between information theory, computational neuroscience, and artificial intelligence, this thesis has laid the groundwork for a more principled and unified understanding of the nature of efficient coding.

References

[1] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[2] Claude E Shannon et al. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[3] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[4] Antonio Ortega and Kannan Ramchandran. Rate-distortion methods for image and video compression. *IEEE Signal processing magazine*, 15(6):23–50, 1998.

[5] Joseph J Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[6] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[8] Simon B Laughlin, Rob R de Ruyter van Steveninck, and John C Anderson. The metabolic cost of neural information. *Nature neuroscience*, 1(1):36–41, 1998.

[9] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961.

[10] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[11] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[12] HB Barlow. Redundancy and perception. In *Physics and Mathematics of the Nervous System: Proceedings of a Summer School organized by the International Centre for Theoretical Physics, Trieste, and the Institute for Information Sciences, University of Tübingen, held at Trieste, August 21–31, 1973*, pages 458–468. Springer, 1974.

[13] Wiktor F Młynarski and Ann M Hermundstad. Adaptive coding for dynamic sensory inference. *Elife*, 7:e32055, 2018.

[14] Adrienne L Fairhall, Geoffrey D Lewen, William Bialek, and Robert R de Ruyter van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–792, 2001.

[15] Anthony MV Jakob and Samuel J Gershman. Rate-distortion theory of neural coding and its implications for working memory. *Elife*, 12:e79450, 2023.

[16] Timothy F Brady, Talia Konkle, and George A Alvarez. Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138(4):487, 2009.

[17] Fabien Mathy and Jacob Feldman. What's magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, 122(3):346–362, 2012.

[18] Chris R Sims, Robert A Jacobs, and David C Knill. An ideal observer analysis of visual working memory. *Psychological review*, 119(4):807, 2012.

[19] Chris R. Sims. The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, 15(3):2–2, 03 2015.

[20] David G Nagy, Balázs Török, and Gergő Orbán. Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, 16(10):e1008367, 2020.

[21] Marco Costa and Leonardo Bonetti. Geometrical distortions in geographical cognitive maps. *Journal of Environmental Psychology*, 55:53–69, 2018.

[22] Barbara Tversky. Distortions in cognitive maps. *Geoforum*, 23(2):131–138, 1992.

[23] Paul S Muhle-Karbe, Hannah Sheahan, Giovanni Pezzulo, Hugo J Spiers, Samson Chien, Nicolas W Schuck, and Christopher Summerfield. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, 2023.

[24] Keno Juechems, Jan Balaguer, Bernhard Spitzer, and Christopher Summerfield. Optimal utility and probability functions for agents with finite computational precision. *Proceedings of the National Academy of Sciences*, 118(2):e2002232118, 2021.

[25] Rahul Bhui, Lucy Lai, and Samuel J Gershman. Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41:15–21, 2021.

[26] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

[27] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

[28] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[29] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, 27(12):712–719, 2004.

[30] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[31] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[32] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[33] Samuel J Gershman, David M Blei, and Yael Niv. Context, learning, and extinction. *Psychological review*, 117(1):197, 2010.

[34] Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7):e6421, 2009.

[35] Karl Friston, Rosalyn J Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. World model learning and inference. *Neural Networks*, 144:573–590, 2021.

[36] Toon Van de Maele, Bart Dhoedt, Tim Verbelen, and Giovanni Pezzulo. A hierarchical active inference model of spatial alternation tasks and the hippocampal-prefrontal circuit. *Nature Communications*, 15(1):9892, 2024.

[37] Giovanni Pezzulo, Leo D'Amato, Francesco Mannella, Matteo Priorelli, Toon Van de Maele, Ivilin Peev Stoianov, and Karl Friston. Neural representation in active inference: Using generative models to interact with—and understand—the lived world. *Annals of the New York Academy of Sciences*, 1534(1):45–68, 2024.

[38] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[39] W Jeffrey Johnston and Stefano Fusi. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, 14(1):1040, 2023.

[40] Zeming Fang and Chris R Sims. Humans learn generalizable representations through efficient coding. *Nature Communications*, 16(1):3989, 2025.

[41] Silvia Bernardi, Marcus K Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967, 2020.

[42] Seonho Park, George Adosoglou, and Panos M Pardalos. Interpreting rate-distortion of variational autoencoder and using model uncertainty for anomaly detection. *Annals of Mathematics and Artificial Intelligence*, 90(7):735–752, 2022.

[43] Christopher J Bates and Robert A Jacobs. Efficient data compression in perception and perceptual memory. *Psychological review*, 127(5):891, 2020.

[44] Merav Ahissar and Shaul Hochstein. Task difficulty and the specificity of perceptual learning. *Nature*, 387(6631):401–406, 1997.

[45] Samuel W Failor, Matteo Carandini, and Kenneth D Harris. Visuomotor association orthogonalizes visual cortical population codes. *bioRxiv*, pages 2021–05, 2021.

[46] Wolfgang Maass. Searching for principles of brain computation. *Current Opinion in Behavioral Sciences*, 11:81–92, 2016.

[47] Matthew F Panichello, Brian DePasquale, Jonathan W Pillow, and Timothy J Buschman. Error-correcting dynamics in visual working memory. *Nature communications*, 10(1):3366, 2019.

[48] G Castegnetti, M Zurita, and B De Martino. How usefulness shapes neural representations during goal-directed behavior. *Science advances*, 7(15):eabd5363, 2021.

[49] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[50] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[51] Kei Akuzawa, Kotaro Onishi, Keisuke Takiguchi, Kohki Mametani, and Koichiro Mori. Conditional deep hierarchical variational autoencoder for voice conversion. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 808–813. IEEE, 2021.

[52] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. biorxiv, 2019.

[53] Felix Leeb, Stefan Bauer, Michel Besserve, and Bernhard Schölkopf. Exploring the latent space of autoencoders with interventional assays. *Advances in Neural Information Processing Systems*, 35:21562–21574, 2022.

[54] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[55] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[56] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[57] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[58] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pages 159–168. PMLR, 2018.

[59] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[60] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. OUP USA, 2011.

[61] Andreas Nieder and Stanislas Dehaene. Representation of number in the brain. *Annual review of neuroscience*, 32(1):185–208, 2009.

[62] Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314, 2004.

[63] Giovanni Anobile, Guido Marco Cicchini, and David C Burr. Number as a primary perceptual attribute: A review. *Perception*, 45(1-2):5–31, 2016.

[64] Andreas Nieder. The adaptive value of numerical competence. *Trends in Ecology & Evolution*, 35(7):605–617, 2020.

[65] Andreas Nieder. The neuronal code for number. *Nature Reviews Neuroscience*, 17(6):366–382, 2016.

[66] Stanislas Dehaene, Manuela Piazza, Philippe Pinel, and Laurent Cohen. Three parietal circuits for number processing. In *The handbook of mathematical cognition*, pages 433–453. Psychology Press, 2005.

[67] Stanislas Dehaene, Ghislaine Dehaene-Lambertz, and Laurent Cohen. Abstract representations of numbers in the animal and human brain. *Trends in neurosciences*, 21(8):355–361, 1998.

[68] Colin V Newman. Detection of differences between visual textures with varying number of dots. *Bulletin of the Psychonomic Society*, 4(3):201–202, 1974.

[69] Susannah K Revkin, Manuela Piazza, Véronique Izard, Laurent Cohen, and Stanislas Dehaene. Does subitizing reflect numerical estimation? *Psychological science*, 19(6):607–614, 2008.

[70] Samuel J Cheyette, Shengyi Wu, and Steven T Piantadosi. Limited information-processing capacity in vision explains number psychophysics. *Psychological Review*, 2024.

[71] Samuel J Cheyette and Steven T Piantadosi. A unified account of numerosity perception. *Nature human behaviour*, 4(12):1265–1272, 2020.

[72] Ivilin Stoianov and Marco Zorzi. Emergence of a'visual number sense'in hierarchical generative models. *Nature neuroscience*, 15(2):194–196, 2012.

[73] Marco Zorzi and Alberto Testolin. An emergentist perspective on the origin of number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1740):20170043, 2018.

[74] Alberto Testolin, Serena Dolfi, Mathijs Rochus, and Marco Zorzi. Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports*, 10(1):10045, 2020.

[75] Alberto Testolin, Will Y Zou, and James L McClelland. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental science*, 23(5):e12940, 2020.

[76] Celestino Creatore, Silvester Sabathiel, and Trygve Solstad. Learning exact enumeration and approximate estimation in deep neural network models. *Cognition*, 215:104815, 2021.

[77] Tom Verguts and Wim Fias. Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, 16(9):1493–1504, 2004.

[78] Serena Dolfi, Gisella Decarli, Maristella Lunardon, Michele De Filippo De Grazia, Silvia Gerola, Silvia Lanfranchi, Giuseppe Cossu, Francesco Sella, Alberto Testolin, and Marco Zorzi. Weaker number sense accounts for impaired numerosity perception in dyscalculia: Behavioral and computational evidence. *Developmental Science*, 27(6):e13538, 2024.

[79] Manuela Piazza, Andrea Facoetti, Anna Noemi Trussardi, Ilaria Berteletti, Stefano Conte, Daniela Lucangeli, Stanislas Dehaene, and Marco Zorzi. Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, 116(1):33–41, 2010.

[80] Michèle MM Mazzocco, Lisa Feigenson, and Justin Halberda. Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child development*, 82(4):1224–1237, 2011.

[81] Tommaso Boccato, Alberto Testolin, and Marco Zorzi. Learning numerosity representations with transformers: number generation tasks and out-of-distribution generalization. *Entropy*, 23(7):857, 2021.

[82] Alberto Testolin, Kuinan Hou, and Marco Zorzi. Visual enumeration is challenging for large-scale generative ai. *arXiv preprint arXiv:2402.03328*, 2024.

[83] FIFTH EDITION. Diagnostic and statistical manual of mental disorders. *American psychiatric association, Washington, DC*, pages 205–224, 1980.

[84] Jan Viktorin. Specific learning disabilities. In Wendy Troop-Gordon and Enrique W. Neblett, editors, *Encyclopedia of Adolescence (Second Edition)*, pages 575–585. Academic Press, Oxford, second edition edition, 2024.

[85] Anna J Wilson and Stanislas Dehaene. Number sense and developmental dyscalculia. *Human behavior, learning, and the developing brain: Atypical development*, 2:212–237, 2007.

[86] Brian Butterworth. Foundational numerical capacities and the origins of dyscalculia. *Trends in cognitive sciences*, 14(12):534–541, 2010.

[87] Janet F McLean and Graham J Hitch. Working memory impairments in children with specific arithmetic learning difficulties. *Journal of experimental child psychology*, 74(3):240–260, 1999.

[88] Manuela Piazza, Véronique Izard, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene. Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3):547–555, October 2004.

[89] Jingyang Zhou, Lyndon R Duong, and Eero P Simoncelli. A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences*, 121(25):e2312293121, 2024.

[90] Lin Zhong, Scott Baptista, Rachel Gattoni, Jon Arnold, Daniel Flickinger, Carsen Stringer, and Marius Pachitariu. Unsupervised pretraining in biological neural networks. *Nature*, 2025.

[91] Kevin S Walsh, David P McGovern, Andy Clark, and Redmond G O'Connell. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, 1464(1):242–268, 2020.

[92] Giovanni Pezzulo, Francesco Rigoli, and Karl J Friston. Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, 22(4):294–306, 2018.

[93] Matteo Priorelli and Ivilin Peev Stoianov. Flexible Intentions: An Active Inference Theory. *Frontiers in Computational Neuroscience*, 17:1 – 41, 2023.

[94] Kenneth H. Norwich. On the theory of weber fractions. *Perception & Psychophysics*, 42(3):286–298, May 1987.

[95] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

[96] Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

[97] Daniel C Hyde. Two systems of non-symbolic numerical cognition. *Frontiers in human neuroscience*, 5:150, 2011.

[98] Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current opinion in neurobiology*, 55:103–111, 2019.

[99] Rafael Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8):487–497, 2015.

[100] Howard Eichenbaum. Barlow versus hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neuroscience letters*, 680:88–93, 2018.

[101] David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

[102] SueYeon Chung and Larry F Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

[103] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[104] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

[105] Mark M Churchland, John P Cunningham, Matthew T Kaufman, Justin D Foster, Paul Nuyujukian, Stephen I Ryu, and Krishna V Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

[106] EH Baeg, YB Kim, K Huh, I Mook-Jung, HT Kim, and MW Jung. Dynamics of population code for working memory in the prefrontal cortex. *Neuron*, 40(1):177–188, 2003.

[107] David Raposo, Matthew T Kaufman, and Anne K Churchland. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17(12):1784–1792, 2014.

Appendix A

Supplementary materials of Chapter 3

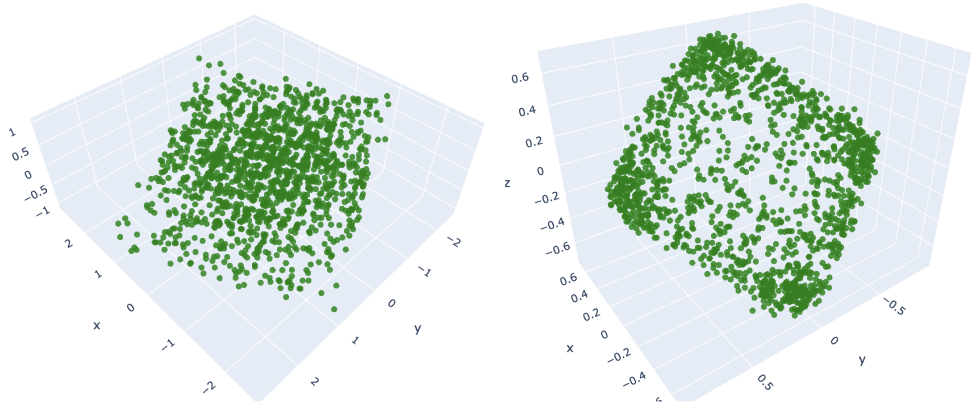

Figure A1: Comparison of the latent representations of the baseline model, trained at high capacity (left) and low capacity (right), in three dimensions.

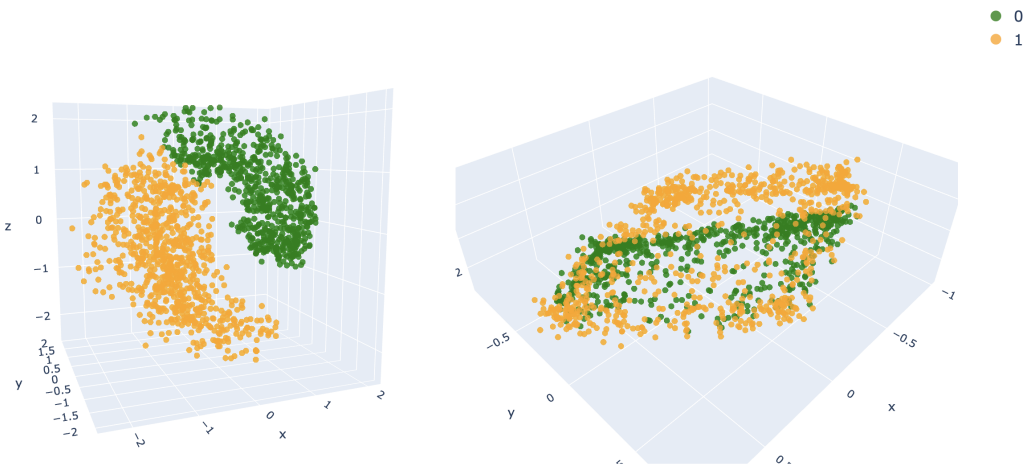

Figure A2: Comparison of the latent representations of the model E1M1, trained at high capacity (left) and low capacity (right), in three dimensions.

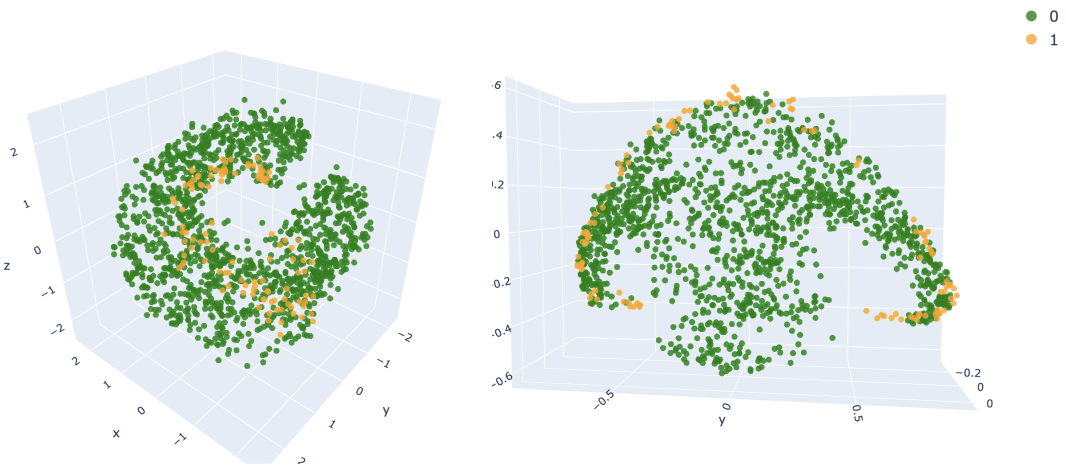

Figure A3: Comparison of the latent representations of the model E1M2, trained at high capacity (left) and low capacity (right), in three dimensions.

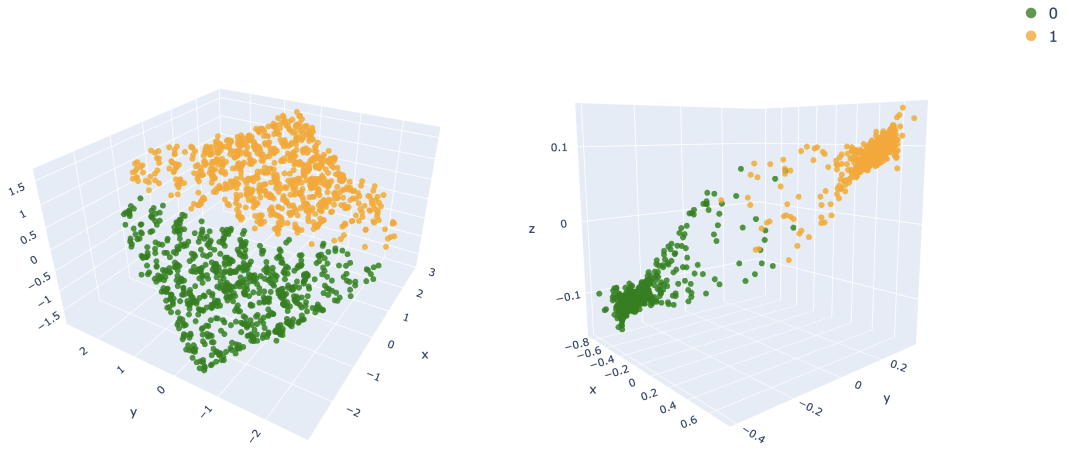

Figure A4: Comparison of the latent representations of the model E2M1, trained at high capacity (left) and low capacity (right), in three dimensions.

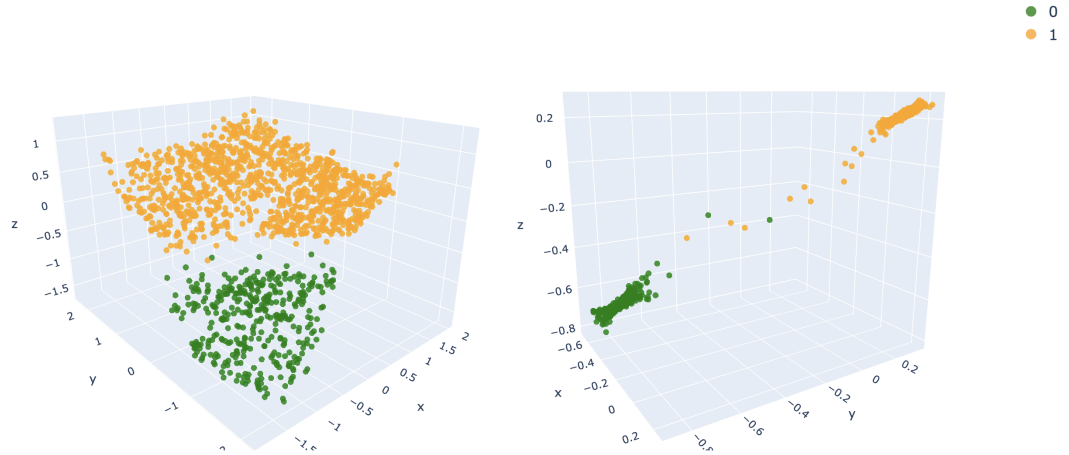

Figure A5: Comparison of the latent representations of the model E2M2, trained at high capacity (left) and low capacity (right), in three dimensions.

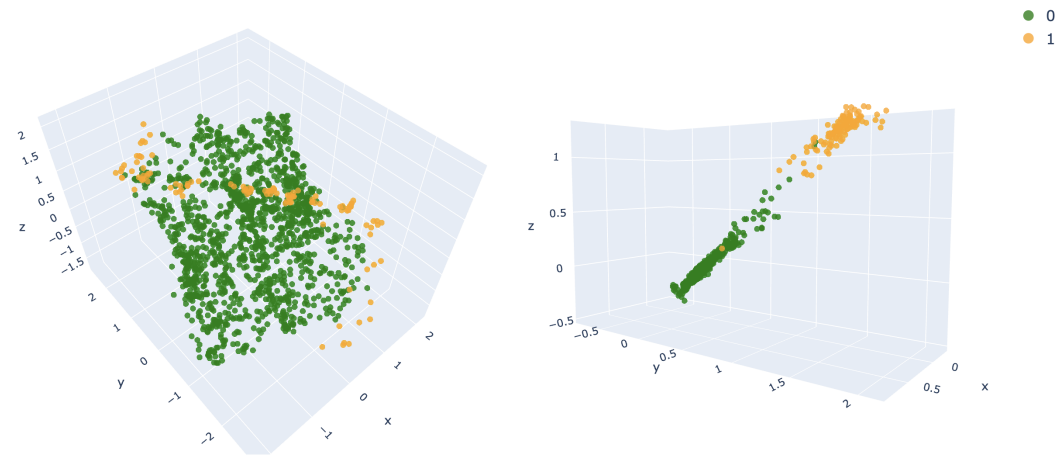

Figure A6: Comparison of the latent representations of the model E2M3, trained at high capacity (left) and low capacity (right), in three dimensions.

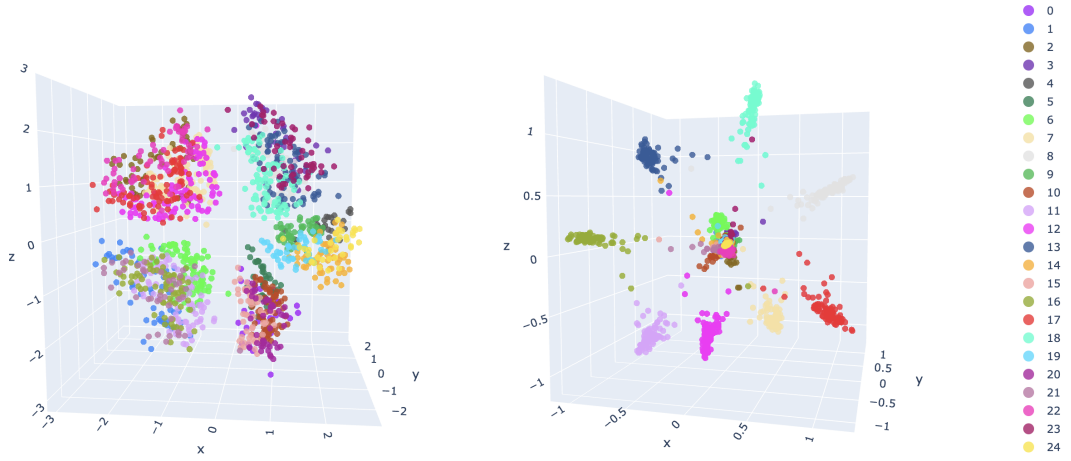

Figure A7: Comparison of the latent representations of the model E2M4, trained at high capacity (left) and low capacity (right), in three dimensions.

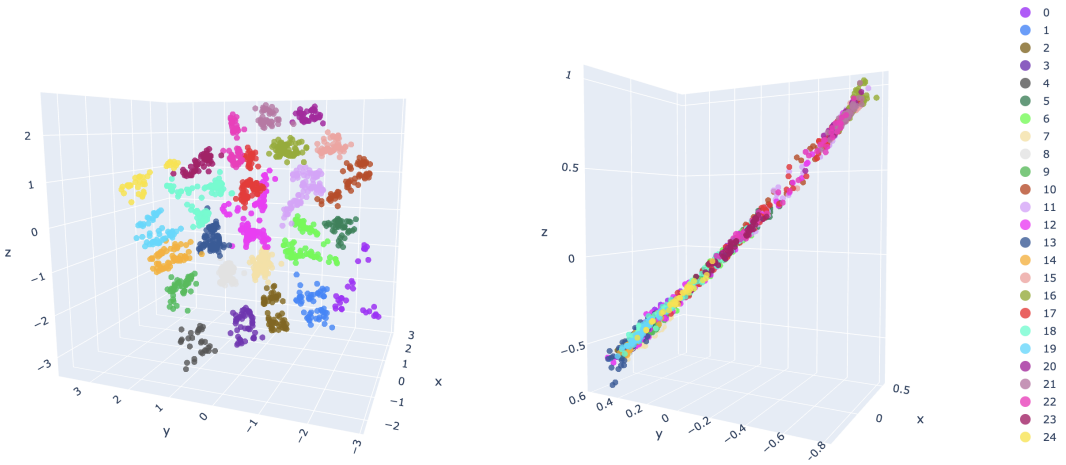

Figure A8: Comparison of the latent representations of the model E2M5, trained at high capacity (left) and low capacity (right), in three dimensions.

Appendix B

Supplementary materials of Chapter 4

B.1 Analysis of the supervised β-VAE-sup models

To provide a comparative baseline and assess the impact of explicit supervision on numerosity representation learning, we analyzed the supervised counterpart of our β-VAE model, denoted as β-VAE-sup. This model was trained with additional information about the ground truth numerosity, enabling us to examine how supervision influences the emergent latent space structure and behavioral performance. The subsequent sections detail the results obtained from this analysis, mirroring the investigations performed on the unsupervised β-VAE model.

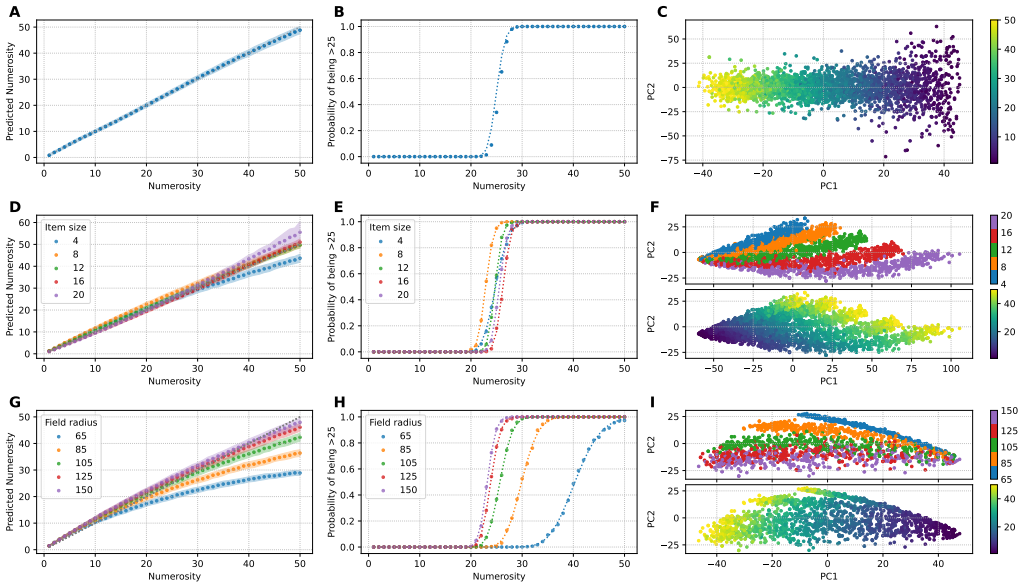

Figure B.1.1: Analysis of the β-VAE-sup-high model with zero-shot generalization tests. All panels are analogous to those described in Figure 4.2. The generalization tests in panels D, E, G and H are conducted without re-training the model or the readouts. This allows comparison of the model's performance in numerosity estimation and discrimination tasks and the organization of its latent space under supervised training conditions.

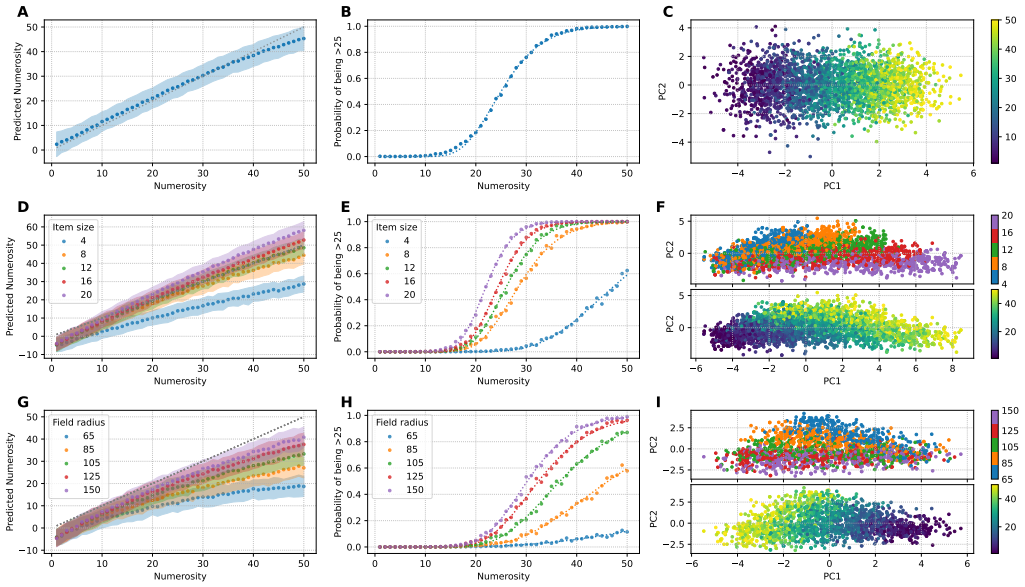

Figure B.1.2: Analysis of the β-VAE-sup-low model with zero-shot generalization tests. All panels are analogous to those described in Figure 4.2. The generalization tests in panels D, E, G and H are conducted without re-training the model or the readouts. This allows comparison of the model's performance in numerosity estimation and discrimination tasks and the organization of its latent space under supervised training conditions and severe resource constraints.

B.2 Datasets used in generalization tests

To evaluate the generalization capabilities of our models, we employed two novel datasets, "item size" and "field radius," which present conditions not encountered during training. These datasets systematically vary the size and spatial distribution of items, allowing us to assess the robustness of the learned numerosity representations to changes in these visual attributes. Figures B.2.1 and B.2.2 illustrate representative samples from each dataset.

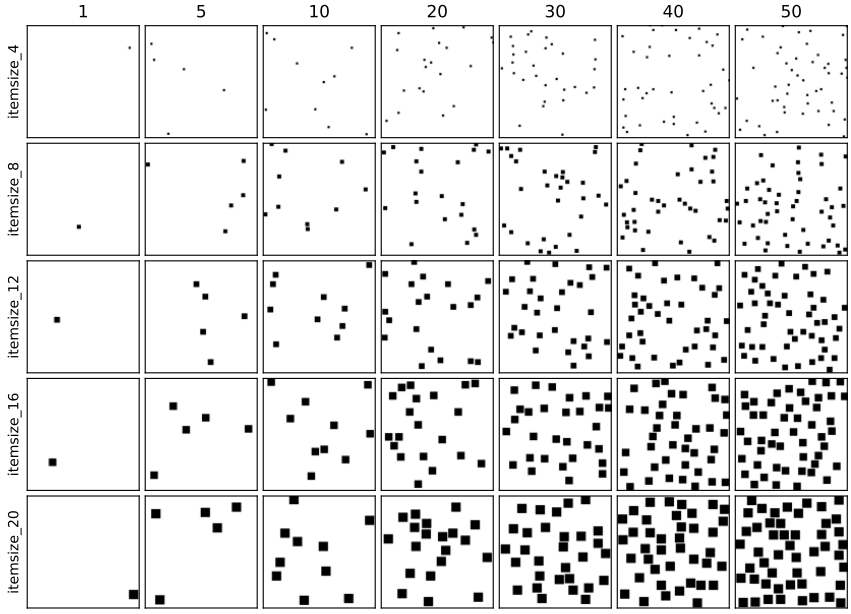

Figure B.2.1: "Item size" dataset used for assessing generalization capabilities. Each subplot displays a representative sample with a specific combination of numerosity and item size. The size refers to the length measured in pixels of the edge of the squared items.

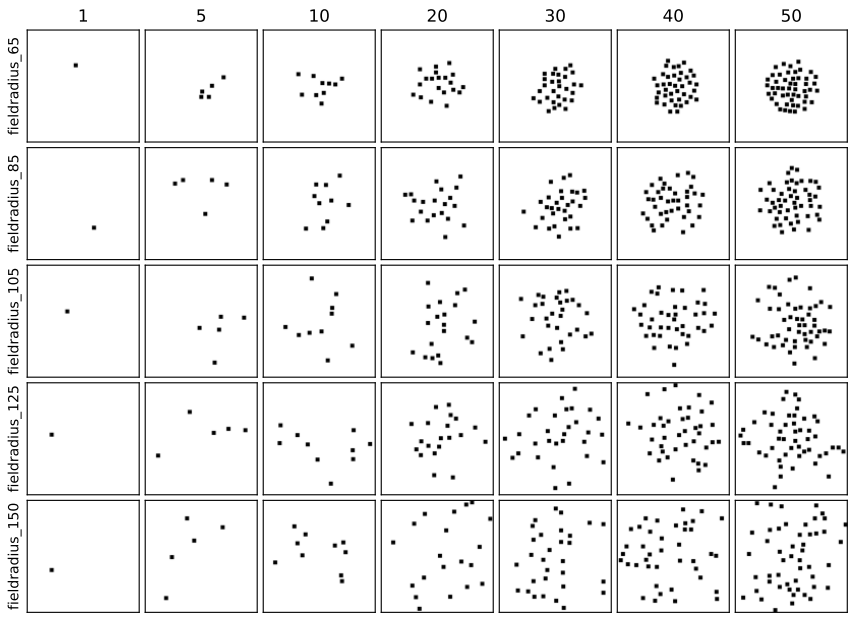

Figure B.2.2: "Field radius" dataset used for assessing generalization capabilities. Each subplot displays a representative sample with a specific combination of numerosity and field radius. The field radius is measured in pixels from the center of the image and constrains the location of the items.

B.3 Generalization tests on the β-VAE-high model, when retraining the linear readouts

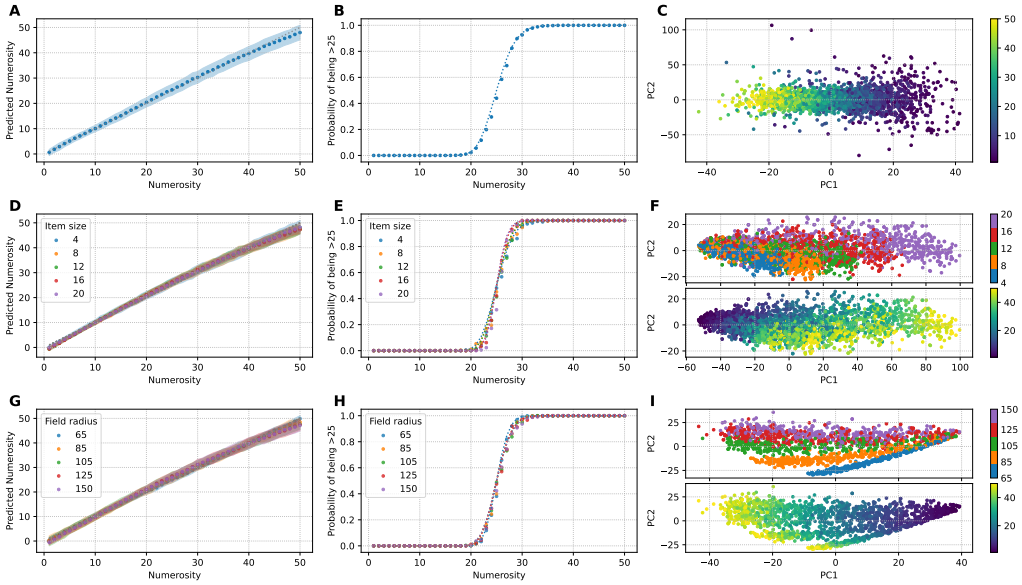

Figure B.3.1: Analysis of the β-VAE-high model with re-trained readouts in the generalization tests. All panels are analogous to those described in Figure 4.2. The generalization tests in panels D, E, G and H are conducted by re-training only the readouts. Retraining the readouts improves the model's generalization performance, indicating that the initial decline in performance was primarily due to limitations in the readout layer rather than fundamental deficiencies in the learned latent representations.

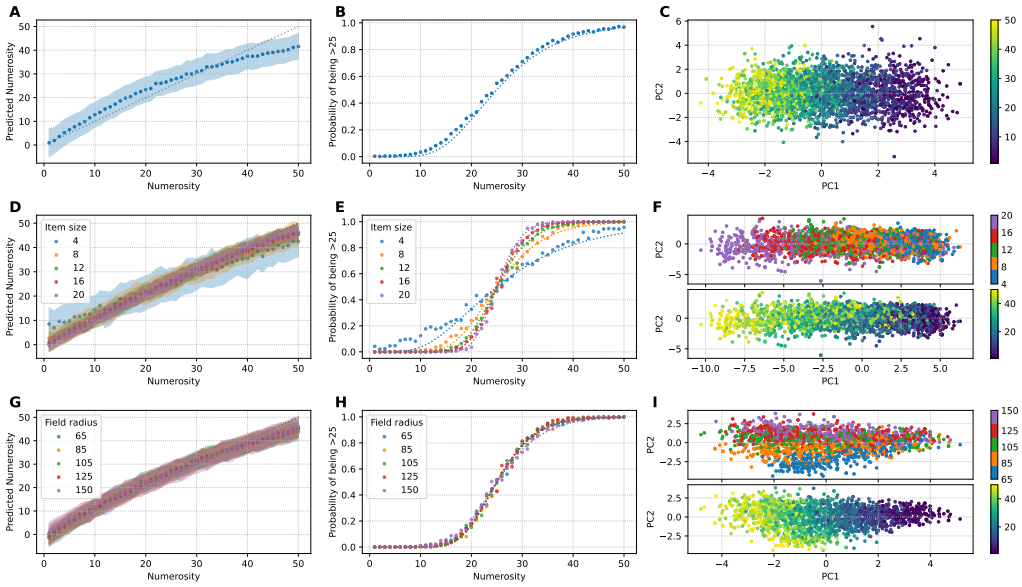

Figure B.3.2: Analysis of the β-VAE-low model with re-trained readouts in the generalization tests. All panels are analogous to those described in Figure 4.2. The generalization tests in panels D, E, G and H are conducted by re-training only the readouts. Retraining the readouts improves generalization performance, but the underlying limitations of the latent representation due to the low encoding capacity still constrain the model's overall abilities.

B.4 Analysis of reconstructions of the β-VAE-high model

Imposing capacity constraints on the β-VAE, while beneficial for studying rate-distortion trade-offs, introduces the risk that the model might prioritize optimizing the rate term of the loss function at the expense of the distortion term (i.e., reconstruction fidelity). To ensure that both terms are adequately optimized during training, we evaluated the reconstruction capability of the β-VAE-high model. Specifically, instead of directly analyzing the latent representations of the input images, we repeated the behavioral analyses (numerosity estimation and discrimination) on the *reconstructions* of the input images. This allows us to verify that the model not only compresses the information effectively but also produces high-fidelity reconstructions.

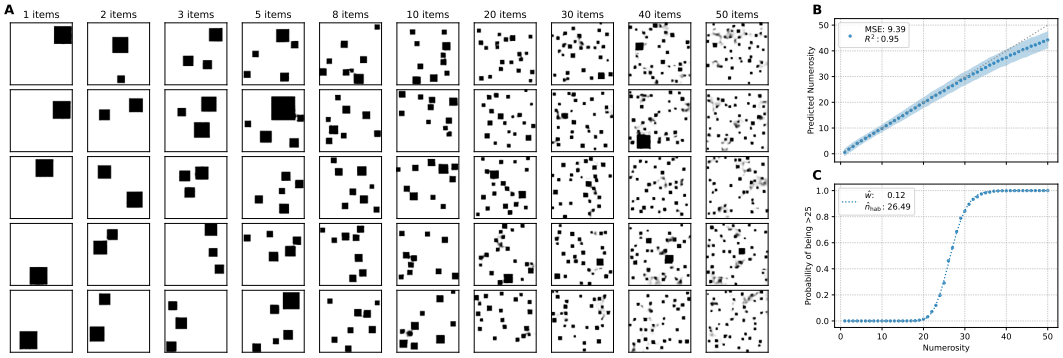

Figure B.4.1: Analysis of reconstructions of the β-VAE model, with zero-shot evaluations. (A) Reconstruction samples (B) Zero-shot numerosity estimation test on latent representations of reconstructed samples (C) Zero-shot numerosity discrimination test on latent representations of reconstructed samples.

B.5 Supplementary neural analyses on the β-VAE model

To gain further insights into the neural coding strategies employed by the models, we performed various supplementary analyses on the latent representations. These analyses, including investigations of the relationship between neuron activity and numerosity, aim to characterize how numerosity information is encoded and organized within the models' latent space. The following figures present the results of these supplementary neural investigations.

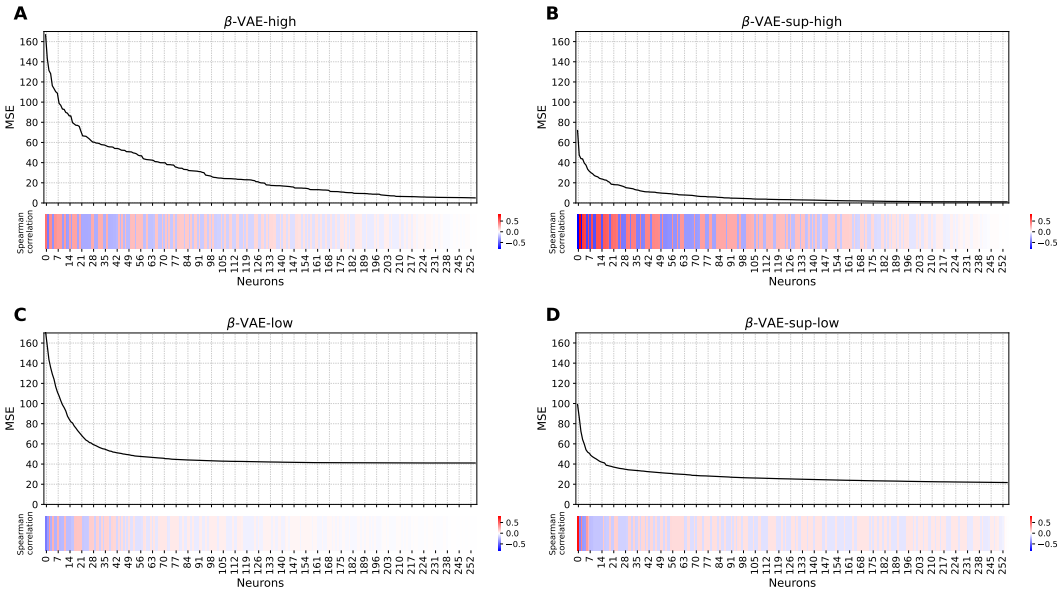

Figure B.5.1: Distributed nature of the neural code. For each model, the plots show the MSE of 256 linear regressions trained to estimate numerosity from neural activity (top), using an increasing number of neurons, sorted by absolute value of correlation scores. The correlation score between neurons and numerosity is computed as spearman correlation (bottom).

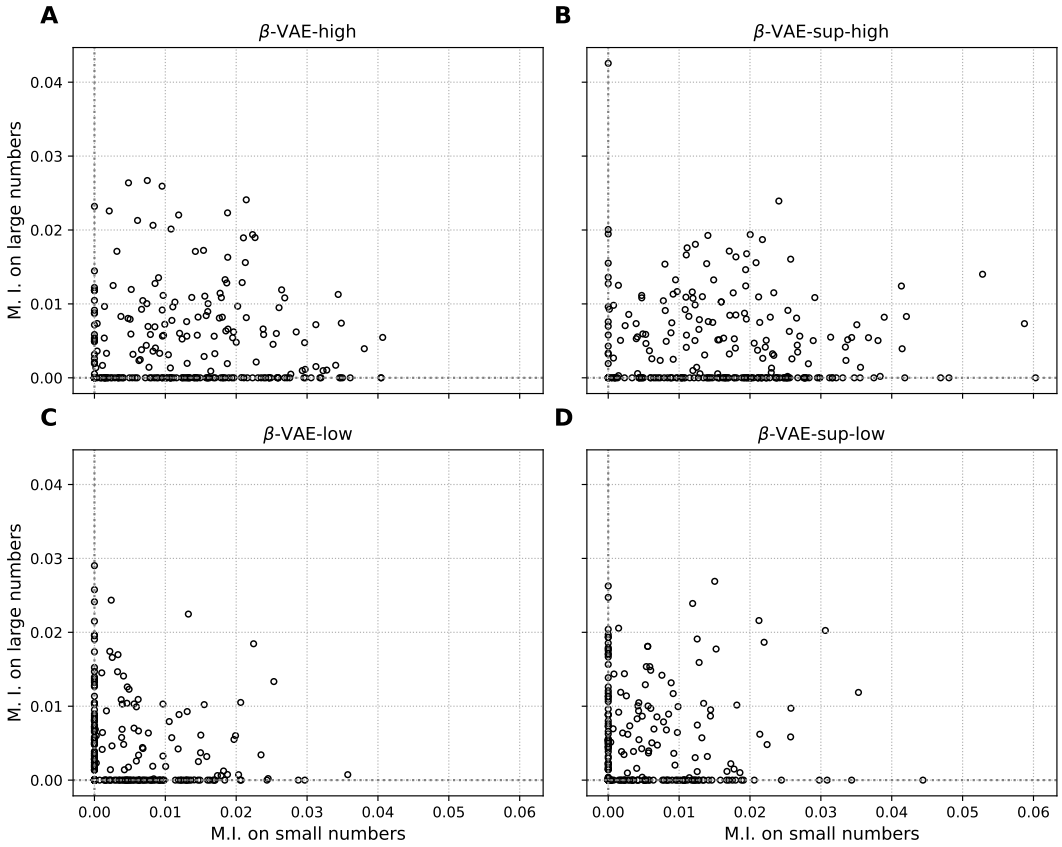

Figure B.5.2: Selectivity of neurons for small versus large numerosities. Each data point represents a neuron in the latent space, with its coordinates indicating the mutual information between its activity and either small ($N \leq 4$) or large ($N \geq 47$) numerosities. This visualization helps identify neurons that are preferentially tuned to specific ranges of numerosity, suggesting potential specialization within the latent space for processing different numerical magnitudes.

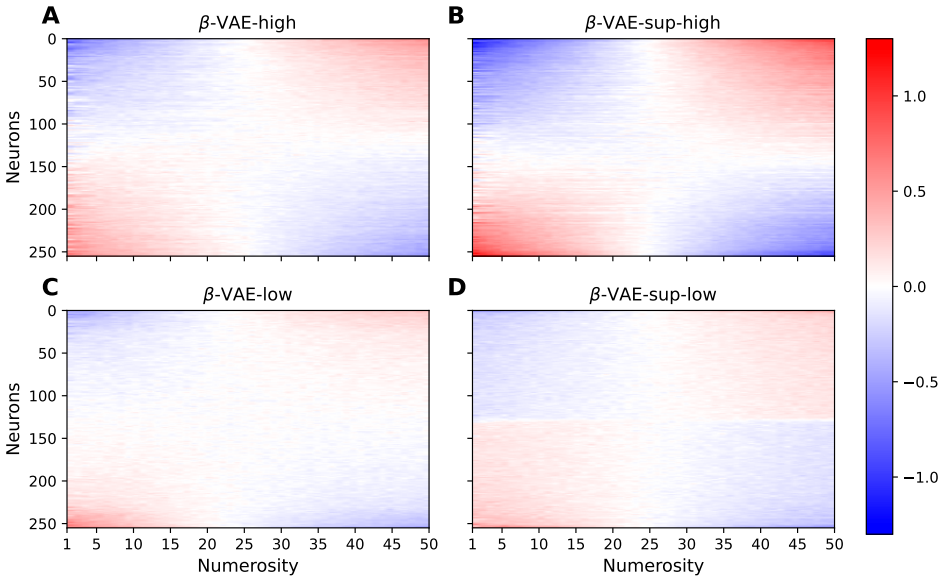

Figure B.5.3: Heatmaps of average neuron activity as a function of numerosity. Each heatmap represents the average activity of each neuron in the latent space for images containing different numerosities. Activity levels have been standardized (z-scored). These heatmaps reveal patterns of neuron activation that correlate with numerosity, illustrating how different neurons respond preferentially to specific numerical quantities. The color scale represents the standardized activity, with red indicating high positive activations and blue indicating negative activations.

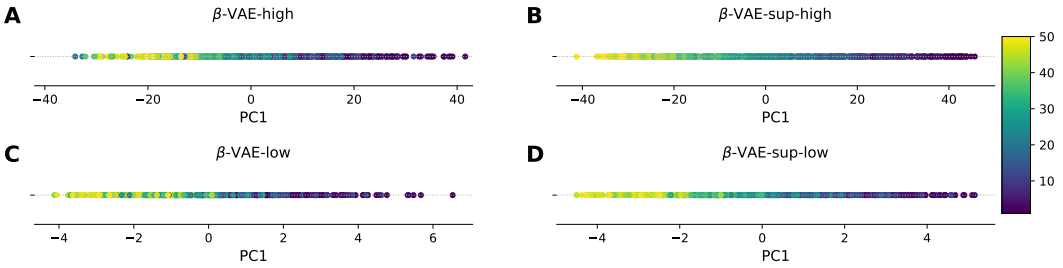

Figure B.5.4: First principal component of the PCA applied to the latent space of each model. Data points are colored according to the numerosity of the input image.