Learning and Representing Priors in Visual Cortex

by

Benjamin Lyo

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Center for Neural Science

New York University

August, 2026

Professor Cristina Savin

Professor Eero P. Simoncelli

© Benjamin Lyo

All rights reserved, 2026

Dedication

To my family.

Acknowledgements

Just before I started my PhD, I remember thinking: I will be past 30 by the time I graduate. Who will I be by then? Well, past Ben — you'll have to live the six years to find out. But I can tell you a little about some very important people you'll meet along the way.

The first of these are your advisors, who you'll choose before you know anything about them.

It is an act of trust to choose an advisor. You do not know them, at least not personally, nor what they will be like to work with. They will preside over you during some very difficult moments. With this in mind, I feel extremely lucky and grateful that of all the people I could have worked with, I chose Cristina.

Cristina's qualities as a scientist and as a mentor go hand in hand. I have only developed into the scientist I am now through constant osmosis of her curiosity, rigor, and passion. But her mentorship extends beyond her scientific sensibilities; there is more to science than research. Even when times were tough, when progress was not forthcoming, I have always felt that Cristina has had my best interests at heart. She has shown that she cares about me as a person in so many big and small ways. Good science comes from a place of feeling secure and respected, and she has created a space where these feelings are abundant and never in doubt.

Eero has, more than anyone I have ever met, a way of thinking about science that I find both foundational and aesthetically pleasing. When I meet with him, the worries that accompany the human practice of research fall away, such is his commitment to making sure things are truly understood and properly placed. Because of this he gave me the great gift of taste: what makes

for a good question? and what makes for a good answer? These are questions I expect to be asking for the rest of my life. He also makes me joyful about finding gaps in my own arguments, and brings out a childlike wonder that I think is fundamental to doing good science.

Of course, none of this would have happened were it not for Stephanie, who for some reason took a chance on me after I told her that I wanted to study consciousness. Not only did she introduce me to the field of theoretical neuroscience, but it is only through her kindness, joy, and mentorship that I decided to do a PhD at all. It is the easiest thing in the world to tell everyone I meet that she is perhaps the nicest person I have ever met. She is exactly the sort of mentor I one day hope to become.

I have been supported by a wonderful committee – David, Alex, and now Mike – who have guided me and given me excellent academic and professional advice every step of the way. They have challenged me and pushed my research in directions I would never have taken on my own.

By some miracle of luck I have spent these six years surrounded by people who have made them some of the best years of my life. Cristina's lab is full of extremely smart and kind people. I want to thank Camille for helping me make it through the early years of my PhD, as we emerged from the depths of Covid, and also for handing me her old apartment in Park Slope, which I cherish. I want to thank Erdem for his constant wit, Harsha for our always stimulating conversations, Willy for his empathy and mirth, and Julie for being my lab sibling as we shared academic parents for all six years. I would also like to thank the rest of the group, past and present – David, Colin, Danilo, Matt, Nancy, and Miles – who have made the lab such a pleasant and engaging place for contemplating the brain. In Eero's lab, I would like to mention Teddy, Pierre, Florentin, Zahra, and Guy, who have been a source of great conversations and even better ideas.

I have also met a good number of friends in Meyer Hall, and I think it is one of the great strengths of the department that it attracts so many people worth knowing. Jeroen in particular has been a great friend who brings out the playful and argumentative sides of me in equal measure. This past year I have been especially buoyed by my board game group – Jeroen, Jay, Estelle,

Alisha, Selvin, Willy, and Ella – who have made me look forward to every Friday. I thank Elisa and Alan for commiserating with me in the theory suite; we will probably be forever deficient in vitamin D from our continued lack of sunlight.

Outside of the department, I have also leaned heavily on a number of friends. I do not know how I would've made it through without the incredible warmth and care that Jendayi, Eleanor, David, and August have given me. They are the people I turn to when I need perspective or advice, and I hope for our continued friendship in the years to come. I am lucky too to have made new friends outside of the PhD program, which I was told would be nigh impossible. Jacko, Maddy and Linda have been delightful both as flatmates and as friends. I am grateful for Carol, who helped me through some difficult periods and encouraged me to approach myself with kindness. And finally, I would like to extend my gratitude to all the enumerable people that I have not mentioned above but with whom I have shared so many good memories.

I am also incredibly grateful for the two academic experiences that I've had outside of NYU. The summer school I attended in Norway was, I daresay, the single best experience of my PhD. I was touched by the generosity and candor of all the people I met there, especially Chiara – who hosted me in Milan as well – Teresa, and Galen. I also want to thank Adrienne Fairhall for organizing the French-American NeuroAI exchange program last year; it opened my eyes to the possibilities of research all around the world.

I would also like to thank all the people at NYU and at Flatiron who have made this PhD as easy as it could be. Asa, Jessy, and Meghan have been nothing but the best at helping me jump through all the administrative hoops at NYU. I also want to thank Jessica and Alicja at Flatiron for being so helpful and patient with me.

My mum, dad and sister Sarah have been unwavering in their love since long before my PhD. It is remarkable how much their curiosity and humanity have shaped me, drawing me toward work that is intellectually stimulating and rooted in empathy. I have always looked up to them, and always will.

Abstract

Normative accounts of perception rest on two frameworks: Bayesian inference, which prescribes how a prior ought to be combined with sensory evidence, and efficient coding, which explains why a nervous system should represent the statistics of its environment at all. Though successful in explaining behavioral data, as computational-level theories, neither specifies how the brain comes to learn or represent the prior in neural circuits. The thesis aims to provide a modern computational account of how the brain comes to learn and represent visual priors over natural images, which are high-dimensional, nonlinearly structured, and organized across many scales. I begin by proposing a neural circuit model that represents priors as stochastic attractor dynamics, which are implemented through the use of a recursive denoising operation. By examining denoising through the lens of both variational inference and information theory, I derive a proposal based on InfoMax for how the brain learns general statistical priors in an unsupervised manner. Finally, I show how InfoMax can be extended beyond a single latent layer to learn a full hierarchy. Together these lines of work provide a computational account of how a prior over natural signals might be represented, used, and acquired by circuits in the visual cortex.

Contents

Dedication iii

Acknowledgments iv

Abstract vii

List of Figures xiv

List of Tables xvi

List of Appendices xvii

1 Introduction 1

1.1 Perception as unconscious inference . . . 3

1.2 Bayesian Inference . . . 6

1.3 Efficient coding . . . 8

1.4 Neural Circuit Models of Sensory Priors . . . 10

1.5 Learning priors . . . 12

1.5.1 InfoMax as a solution to epistemic uncertainty . . . 14

1.5.2 Rate-Distortion . . . 17

1.5.3 Generative modeling . . . 21

1.6 Thesis Organization . . . 26

2 A circuit model for representing complex priors and performing flexible inference 27

2.1 Overview . . . 27

2.2 Introduction . . . 28

2.3 Encoding complex priors in recurrent neural circuits . . . 32

2.3.1 Representing priors with diffusion models . . . 32

2.3.2 A diffusion-based circuit for sampling-based probabilistic computation . . . 33

2.4 Flexible inference . . . 36

2.4.1 Using the same prior-encoding circuit for inference . . . 36

2.4.2 Inferring latents from sensory evidence . . . 39

2.4.3 Inference with multiple sources of information . . . 40

2.5 Neural signatures . . . 41

2.6 Discussion . . . 43

3 Learning disentangled representations via score-based variational autoencoders 46

3.1 Overview . . . 46

3.2 Introduction . . . 47

3.3 Methods . . . 50

3.3.1 Implementing the generative model as conditional diffusion . . . 53

3.3.2 Computing the guidance score . . . 54

3.3.3 Inference and guidance . . . 55

3.3.4 Geometric intuition . . . 56

3.4 Results . . . 57

3.4.1 CelebA dataset . . . 58

3.4.2 Feature extraction for pretrained diffusion models . . . 63

3.5 Discussion . . . 65

4 A Rate-Distortion Account of InfoMax **67**

4.1 Overview . . . 67

4.2 Implementing InfoMax using I-MMSE . . . 68

4.2.1 Maximizing Mutual Information using the I-MMSE Relation . . . 70

4.2.2 Implementation of I-MMSE InfoMax . . . 72

4.3 Formalizing the Constraint via Rate-Distortion . . . 73

4.3.1 Casting InfoMax as Wyner-Ziv Distortion . . . 73

4.3.2 Optimal Wyner-Ziv Rate . . . 76

4.3.3 Implementation of the Rate . . . 78

4.3.4 Relationship between Rate and Distortion . . . 78

4.4 Interpreting the objective . . . 80

4.4.1 Geometric Interpretation . . . 80

4.4.2 Biological implications . . . 82

4.4.3 Relation to ELBO . . . 83

4.5 Results . . . 85

4.5.1 Numerical experiments . . . 85

4.5.2 Disentanglement . . . 86

4.6 Discussion . . . 88

5 Hierarchical representations of prior knowledge **90**

5.1 Overview . . . 90

5.1.1 Hierarchy in the world . . . 91

5.1.2 Statistical signatures of Hierarchy in Natural Images . . . 92

5.1.3 Difficulties of Inference . . . 93

5.1.4 Hierarchical Structure in the Brain . . . 94

5.1.5 Hierarchical Models of Visual Processing . . . 95

5.2 Hierarchical InfoMax . . . 97
5.2.1 Objective . . . 97
5.2.2 Implementation . . . 99
5.3 Discussion . . . 102

6 Discussion **104**
6.1 Overview . . . 104
6.2 Inference of complex posteriors in neural circuits . . . 105
6.2.1 What the circuit model does not settle . . . 107
6.3 InfoMax Learning and Representations . . . 108
6.3.1 What this buys . . . 110
6.3.2 Open questions . . . 111
6.4 Experimental predictions . . . 114
6.5 Concluding remarks . . . 116

A Appendix: A circuit model for representing complex priors and performing flexible inference **118**
A.1 Posterior sampling dynamics . . . 118
A.2 Effects of dendritic tree morphology . . . 120
A.3 Testing neural sampling quality by marginal statistics . . . 124
A.4 Posterior autocorrelation . . . 125
A.5 Stochastic neural network . . . 125
A.6 More biologically realistic oscillations . . . 127
A.7 MNIST dataset . . . 128

B Appendix: Learning disentangled representations via score-based variational autoencoders **130**

B.1 Additional results . . . 130
B.1.1 Relation of MMSE solution to conditional score . . . 130
B.1.2 Guiding transition operators . . . 131
B.1.3 Training and conditional generation algorithms . . . 132
B.1.4 Conditional generation as constrained optimization . . . 132
B.1.5 Architecture and training details . . . 136
B.1.6 Reduction in variability from conditioning . . . 138
B.1.7 Metrics . . . 139
B.1.8 Derivation for globally preserved semantic axes . . . 141
B.1.9 Comparisons to other models . . . 146
B.2 Proofs . . . 148
B.2.1 Training on both noisy and clean images smoothens the latent space . . . 148
B.2.2 The SAMI objective encourages disentanglement across scales . . . 158
B.2.2.1 Setup . . . 159
B.2.2.2 Why a diagonal encoder picks the feature axes . . . 160
B.2.2.3 From one noise level to a hierarchy . . . 162

C Appendix: A rate-distortion account of InfoMax **165**
C.1 Derivation of the Barber-Agakov bound . . . 165
C.2 Marginal distortion . . . 166
C.3 Relation of marginal mutual information to Fisher information . . . 167
C.4 Rate as a dot product . . . 171
C.5 Hierarchical disentanglement from noise-level structure . . . 172
C.6 Rate-optimal likelihood bias is orthogonal to the prior drift . . . 183

D Appendix: Hierarchical representations of prior knowledge **184**
D.1 Marginal rate . . . 184

Bibliography **186**

List of Figures

1.1 The "light-from-above" optical illusion. 5
1.2 Coding strategy of a single neuron to maximize its information capacity. 9
1.3 Neural resources are allocated to match the distribution of natural stimuli. 10
1.4 Two hypotheses for how neurons represent priors 11
1.5 Communication channel. 18
1.6 Rate-distortion phase diagram. 19
1.7 Comparison between the variational InfoMax and variational generative modeling perspectives 21

2.1 Diffusion models and their neural circuit implementation. 31
2.2 Neural circuit performing posterior inference using diffusion dynamics. 37
2.3 Flow fields and samples from the prior, likelihood and posterior. 42
2.4 Neural signatures. 43

3.1 Graphical model and schematic of the SAMI network and geometric interpretation of the conditional sampling procedure. 50
3.2 Results of training SAMI on the synthetic disk dataset. 57
3.3 SAMI trained on CelebA. 59
3.4 Extraction of latent representation from pre-trained diffusion model. 63

4.1 Block diagram of our model as a source-coding problem with side information. The encoder $p(z|x)$ sees only the clean source X, while the decoder $p(\hat{x}|x_\gamma, z)$ additionally receives the side information $x_\gamma \sim p(x_\gamma|x)$. 74
4.2 A geometric perspective of the denoiser-based decoder. 80
4.3 InfoMax model trained on the disk dataset 86

A.1 Effect of network architecture on performance over training 122
A.2 Effects of sparse recurrent connections targeting different locations on the dendritic tree 122
A.3 Marginal statistics along the dimension orthogonal to the swiss roll cross-section 125
A.4 Autocorrelation comparison of the posterior samples 126
A.5 Reverse diffusion process using the SNN architecture 127
A.6 Effect of biologically realistic oscillations on sample quality 128
A.7 MNIST results 129

B.1 Reduction in variance from conditioning on the posterior on a CelebA image . . . 139
B.2 Difference in the magnitudes of prior and posterior scores 142
B.3 Comparing the norms of the denoiser score and the guidance score 146

List of Tables

6.1 The obstacles to learning priors as set out in Chapter 1, and where each is addressed. 109

A.1 Morphology of dendritic networks. 121

B.1 Architectural and training details for the SAMI model. 136

B.2 Comparison against other generative models on disentanglement and image quality metrics. 140

B.3 Comparison to other diffusion-based representation learning models in the literature 147

List of Appendices

A A circuit model for representing complex priors and performing flexible inference . . 118

B Learning disentangled representations via score-based variational autoencoders . . . 130

C A rate-distortion account of InfoMax 165

D Hierarchical representations of prior knowledge 184

1 Introduction

> We see with memory. My memory is different from yours, so if we are both standing in the same place we're not quite seeing the same thing. Whether you have been in a place before will affect you, and how well you know it. There's no objective vision ever – ever.
>
> *David Hockney*

What we experience is not the world itself but an interpretation built from our senses. Imagine wading through a lush tropical rainforest. A breeze through the canopy casts dancing shadows on the dark undergrowth, the morning dew glistens beneath the translucent leaves of a giant mango tree, and twigs snap softly beneath our footsteps. At first glance, we appear to perceive this scene with little conscious effort, but even vision, arguably our strongest sense, delivers signals that are riddled with ambiguity and under-determine the state of the world. A flash of orange in our hypothetical rainforest could indicate the presence of a ripe fruit or a predator lurking in the depths. To resolve this uncertainty and construct a coherent percept of the world, our brain draws on our past experience of the world to provide additional information beyond what our senses can provide in the moment. Armed with these prior expectations, the brain weaves together a

narrative about what it believes the external world may be like. Since all of this happens outside of our conscious awareness, we unwittingly experience a subjective world that has been distorted through the lens of our past.

In the fields of neuroscience and cognitive science, this story of perception as acts of unconscious inference has been well established using the framework of Bayesian inference (Helmholtz, 1852). In this narrative, the experience of perception comes about from combining the information from incoming sensory signals with our prior experience-based expectations of the world. This normative framework has been used to explain a wide range of behaviors, such as cue combination (Ernst and Banks, 2002; Landy et al., 1995; Jacobs, 1999), evidence accumulation (Gold and Shadlen, 2007), and choice and estimation under priors (Weiss et al., 2002; Stocker and Simoncelli, 2006; Girshick et al., 2011; Wei and Stocker, 2015). Though this framework is supported empirically at the behavioral level, we encounter difficulties once we peer inside the brain. How is prior information represented by neurons within the brain? Does it reside in the dynamics of neural activity, or is it implicitly stored in the connectivity and synaptic weights of a neural population? And even if we knew *how* the brain represents priors, yet another problem is knowing *what* about the world is worth representing as priors. How does the brain accommodate and account for the inherent complexity of the world? What constraints must the brain obey, and how do these affect the nature of the representation? Finally, our beliefs are a reflection of our lived experience, but which aspects of our sensory input are worth storing?

Why should we be interested in the study of neural priors? Perhaps the most pertinent reason is the potential clinical applications of this knowledge. For example, impairments in inference are symptoms characteristic of both autism spectrum disorder and schizophrenia, and recent empirical results suggest that these may stem from the sub-optimal integration of sensory evidence with priors. Despite this, little is known about the neural underpinnings of inference in neuro-typical brains, and thus a better theoretical understanding of how neural circuits represent priors and perform inference is necessary for downstream clinical interventions. To understand

dysfunction, we must first understand function.

Additionally, understanding how the brain instantiates priors is becoming increasingly important in light of recent developments in the fields of artificial intelligence and machine learning. While current machine learning models display impressive capabilities, they do not yet possess functions that are common in biological organisms, such as planning, abstract reasoning, and the ability to generalize from few examples. Given these shortcomings, there has been a recent but significant push in machine learning to construct so-called *internal world models* that are interpretable, possess useful inductive biases, and carry embodied information about the environment. Biological brains meanwhile constitute an existence proof that such models can and do exist. Basic scientific research is thus complementary to engineering: elucidating the mathematical principles underpinning priors in the brain may provide insight into how best to improve existing forms of artificial intelligence.

Lastly, the study of priors takes on new importance as neuroscience turns to more complex questions. Historically, neuroscientists have contained the complexity of the brain by studying animals performing simple tasks with few known free variables. Large-scale recording techniques now make it possible to study neural computation in naturalistic settings, forcing us to contend with the complexity of the environment itself. Yet existing theories of how the brain represents the statistics of the environment as priors deal predominantly with scalar or low-dimensional stimuli; they do not scale to the high-dimensional latent variables that underlie the natural world. A theory of neural priors that scales to this complexity is thus necessary for understanding the ecological brain.

1.1 Perception as unconscious inference

The world as we perceive it is a subjective construction, assembled by the mind beneath the threshold of conscious awareness. Perception, in this view, is an act of "unconscious inference"

about the state of the world – a principle usually credited to the German physicist and physiologist Hermann von Helmholtz (Helmholtz, 1852), though it appears as early as ca. 1010 in the work of Alhazen (Ibn al-Haytham).

Since these inferences lie beneath conscious awareness, their effects can be difficult to notice. This is especially true of vision – arguably our sharpest and most trusted sense – which seems to hand us an objective portrayal of reality. For this reason it is often more instructive to study the *failures* of inference, where the mechanisms underlying computation are most apparent. In Volume III of his *Physiological Optics*, Helmholtz recounts a childhood memory:

> I can recall when I was a boy going past the garrison chapel in Potsdam, where some people were standing in the belfry. I mistook them for dolls and asked my mother to reach up and get them for me, which I thought she could do. The circumstances were impressed on my memory, because it was by this mistake that I learned to understand the law of foreshortening in perspective.

The "law of foreshortening" the young Helmholtz learns is not merely that a distant object projects a smaller retinal image than a near one of the same size; it is that this projection cannot be inverted without ambiguity. A small image is equally consistent with a small object nearby and a large object far away, and the image by itself cannot resolve this ambiguity. To perceive size at all, the visual system must supply distance as an additional piece of information. Helmholtz's error is instructive because the inference is incorrect: lacking any cue to the belfry's distance, he takes the small images at face value and sees dolls instead of people. Our percepts, veridical or not, are thus inferences about the state of the world; the brain continually creates hypotheses that best explain our ambiguous sensory inputs.

Optical illusions do more than illustrate the failures of inference; they also reveal our prior beliefs about the world. Consider, for example, the "light-from-above" illusion in Fig. 1.1 (Ramachandran, 1988; Adams, 2007). Although the stimulus is a flat image of shaded circles, we see each circle as a three-dimensional surface: a bump when it is dark at the bottom, a dimple when

it is dark at the top. A single assumption accounts for both percepts – that the circles are all lit from above. This indicates that even prior to observing the image, we hold certain expectations that alter what we perceive. In this case, the prior belief is that light predominantly comes from overhead, but it has been shown that human observers possess many other priors, such as surface convexity (Langer and Bülthoff, 2001; Mamassian and Landy, 1998; Goldreich and Peterson, 2012), isotropy and homogeneity of texture (Knill, 1998), slow and smooth motion (Weiss et al., 2002), symmetry (Knill, 2007), smoothness or "good continuation" in space and time (Schwartz et al., 2007; Geisler and Perry, 2009), that cardinal orientations are more frequent than other orientations (Girshick et al., 2011), and that other people's gazes are directed toward them (Mareschal et al., 2013).

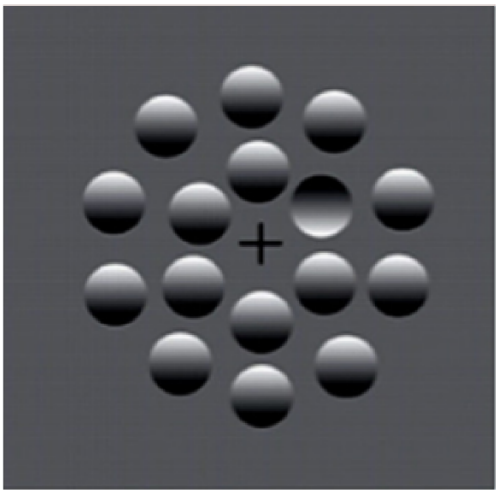


Figure 1.1: The "light from above" optical illusion.

Priors account for the systematic structure of our percepts, but their very existence raises several fundamental questions. First, what is the optimal way to use these priors during perception? Second, why does the brain possess priors at all? And finally, how do we come to acquire them? I address these in turn by introducing three classical theories in neuroscience: how priors are optimally used during perception (Bayesian inference), why the brain represents the statistical structure of its environment (efficient coding), and how those representations are acquired

through experience (unsupervised learning). Along the way I will also introduce rate-distortion theory and the evidence lower bound.

1.2 Bayesian Inference

As we have seen above, perception can be cast as a problem of inference in which we induce the meaning of ambiguous sensory cues by leveraging our prior knowledge of the environment. Though the abundance of behavioral evidence strongly suggests that the brain must be performing this inferential computation, this raises the following question: how should the brain combine our priors with sensory cues?

Since the 1990s, this problem of perception has received a formal treatment using the framework of Bayesian inference, a general and optimal procedure for inductively inferring latent properties of the world (Dayan et al., 1995; Knill and Richards, 1996; Kersten and Yuille, 2003; Geman and Geman, 1984). Suppose we observe some sensory input x and seek to infer the identity of the unobserved or *latent* variable z that caused our observations. According to this framework, inference corresponds to finding a posterior probability distribution over the latent z after observing sensory input x. This depends on two factors: how our sensory observations depend on the latent stimuli – the likelihood $p(x \mid z)$ – and our prior beliefs about the latent – the prior $p(z)$. To get the desired posterior probability distribution $p(z \mid x)$, we combine these factors according to Bayes' rule:

$$\underbrace{p(z \mid x)}_{\texttt{posterior}} \propto \underbrace{p(x \mid z)}_{\texttt{likelihood}} \cdot \underbrace{p(z)}_{\texttt{prior}} .$$

This description is probabilistic: the prior $p(z)$ characterizes the strength of the observer's prior belief over the latent variable; the likelihood function $p(x \mid z)$, read as a function of z, indicates how well each candidate latent accounts for the observed sensory input x; and the posterior distribution $p(z \mid x)$ captures the certainty of our beliefs about the inferred latent stimulus given

our observations. The multiplicative relationship predicts that when sensory information is unreliable, prior knowledge should exert greater influence, while the converse should be true when we are confident about our observations. Bayesian inference thus provides a *normative* account of perception, prescribing how an ideal observer ought to update its beliefs about the latent according to the reliability of sensory evidence.

The quantitative nature of the Bayesian framework allows for phenomenological modeling of inference in human cognitive and perceptual domains, such as cue combination (Ernst and Banks, 2002; Landy et al., 1995; Jacobs, 1999), evidence accumulation (Gold and Shadlen, 2007), and choice and estimation under priors (Stocker and Simoncelli, 2006; Girshick et al., 2011; Wei and Stocker, 2015). Moreover, the framework allows for precise quantification of the effect priors have on human perception. For example, priors can modulate perceptual performance (Sekuler and Ball, 1977; Posner, 1980; Downing, 1988; Bar, 2004), or alter the subjective appearance of visual stimuli, especially when the inputs are ambiguous or can be interpreted in multiple ways (Adams et al., 2004; Haijiang et al., 2006; Sterzer et al., 2008).

At this point, it is worth taking a step back and asking why we believe the brain is performing *optimal* inference. It is plausible that the brain could be inferring in a non-optimal manner, for example by employing heuristics that may be less expensive to compute. Moreover, the existence of a broad normative principle seems unlikely given the immense variability in biological form and function that we encounter in nature. How can we justify the assertion of optimality in inference?

The argument of optimality holds water, even in the face of this heterogeneity, when we consider the role of evolution. Evolution exerts selection pressure on biological organisms through their actions: agents with lower fitness are less likely to propagate their genes. The success of an action, however, is predicated on the organism's ability to properly assess the state of the world through accurate perception. Organisms are thus incentivized to make optimal use of all available information collected by their senses, or risk suboptimal actions due to inaccurate inference.

A species that routinely discards useful information or performs suboptimal inference will be out-competed by other species that make better use of their sensory inputs. As such, biological organisms that depend on their senses are driven to perform optimal inference irrespective of their ecological niche. In general, normative arguments thus allow us to establish theoretical principles in neuroscience despite the heterogeneity of the neural anatomy.

1.3 Efficient coding

The Bayesian framework tells us how knowledge of statistical regularities in the world can help us perform optimal inference at a computational level. However, it does not provide an explanation as to *why* these priors exist in the brain.

More than 70 years ago, Attneave (1954) and Barlow (1961) proposed that sensory systems leverage statistical regularities in our natural environment to efficiently encode sensory information – a theory called the *efficient coding hypothesis*. Attneave intuited that the goal of a perceptual system is akin to that of a scientist, that is, to "describe nature economically". This statement of coding efficiency, as re-expressed by Barlow, says that neurons seek to strip away the redundant information in their inputs to best make use of their limited resources.

By drawing on information theory, a mathematical theory of how signals are communicated (Shannon, 1948), Barlow showed that redundancy reduction is the solution to the objective of information maximization under resource constraints. Redundancy measures the gap between the information a code actually carries and the maximum it could carry given its capacity. However, since a sensory system has only so many neurons, each limited in its range of responses, that maximum is finite. Given a fixed metabolic and representational budget, minimizing redundancy thus pushes a neural code to carry as much information as possible about its sensory environment. Note that this solution/equivalence holds under assumptions of low input noise and variability in neural responses.

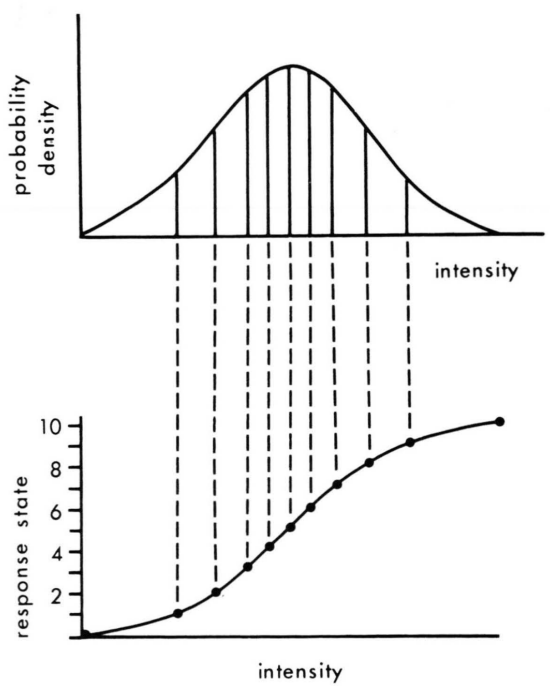


Figure 1.2: Coding strategy of a single neuron to maximize its information capacity. For a given probability density function of stimulus intensities (top), the neuron's response function (bottom) ensures that each response state corresponds to equal areas under the probability density function, such that each state is used with equal frequency. Figure from Laughlin (1981).

One immediate implication of the information maximization principle is that perceptual systems should allocate more of their resources to encode signals that occur most frequently in the environment (Fig. 1.2). This is best illustrated by examining a single neuron responding to a scalar input (e.g. light intensity or constrast) whose frequency of occurrence is given by a continuous probability distribution. To maximize information, the neuron should dedicate more sensitive regions of its response range to encode inputs that are more frequently encountered in exchange for lower fidelity in its encoding of less common inputs. The optimal response function of this neuron should thus be proportional to the cumulative probability distribution of the stimulus. This prediction is borne out in the study by Laughlin (1981), who showed that the contrast-response functions of a large monopolar cell in the fly closely resembled the cumulative probability function for contrast levels in natural scenes.

The same logic applies when we consider a population of neurons, where the resource in

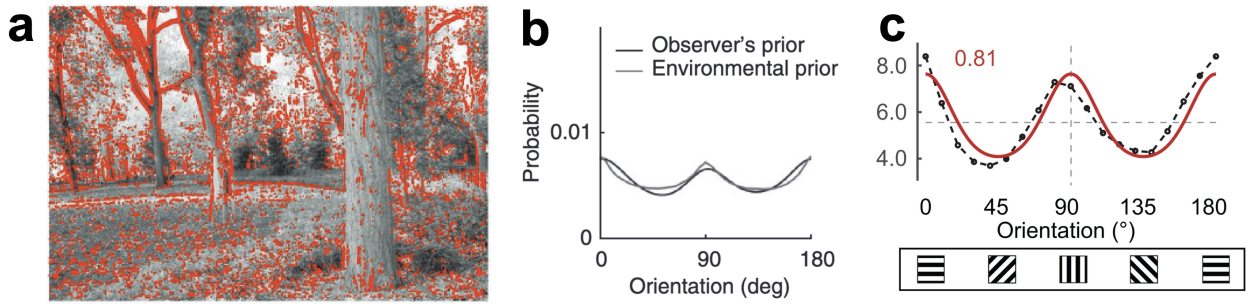


Figure 1.3: Neural resources are allocated to match the distribution of natural stimuli (**a**) Example natural scene with strongly oriented edges marked in red. (**b**) orientation distribution and human observer's prior recovered from discrimination thresholds. (**c**) percentage of V4 neurons preferring orientation within the 10° bin. Figure adapted from Fang et al. (2022) and Girshick et al. (2011).

question is the number of neurons rather than the range of a single one. Orientations in natural images are distributed non-uniformly, with a higher density of edges along the cardinal directions than oblique ones (Fig. 1.3a, b). The distribution of orientation-selective neurons in macaque visual cortex closely matches this environmental distribution, peaking in the cardinal directions (Hubel and Wiesel, 1968; Fang et al., 2022); correspondingly, human orientation perception exhibits biases and discriminability consistent with a cardinal prior (Girshick et al., 2011). Evidence of orientation anisotropy has also been found in the visual cortices of cats (Sengpiel et al., 1999; Wang et al., 2003), ferrets (Chapman and Bonhoeffer, 1998), and owl monkeys (Xu et al., 2006). Whether we consider the case of a single neuron or a population of neurons, we see that the allocation of resources reflects the statistics of the scalar variable they aim to encode.

1.4 Neural Circuit Models of Sensory Priors

A weakness of both the efficient coding hypothesis and Bayesian inference is that they, as computational accounts of perception, do not specify the neurobiological mechanism by which sensory systems encode their inputs and represent the prior. To that end, another way to evaluate the efficient coding hypothesis beyond examining the neural responses under natural stimulation is to create models of early sensory processing. In this "inverse" approach, one assumes a statisti-

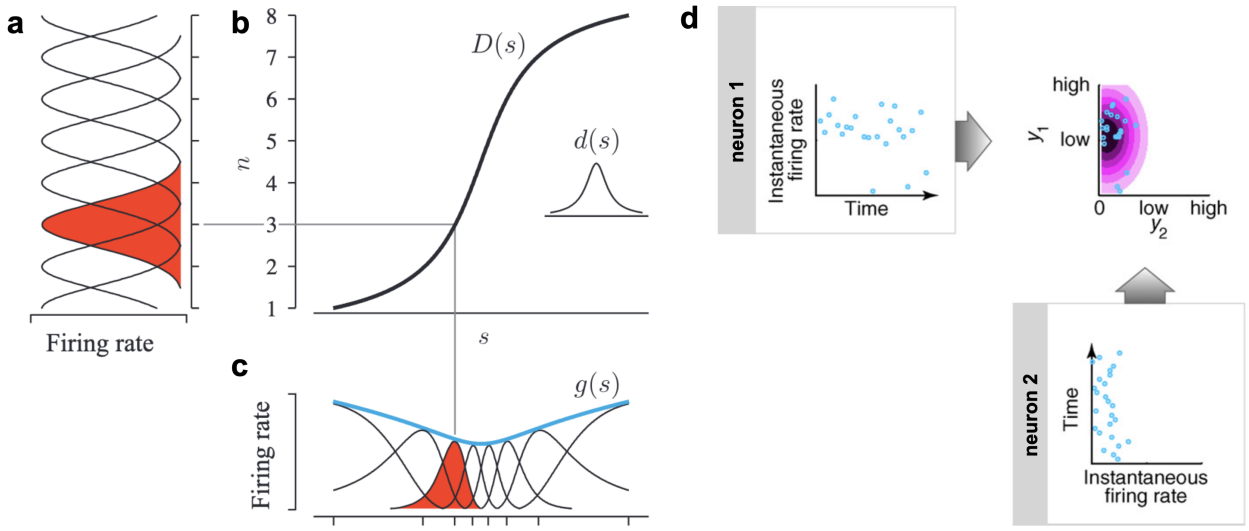


Figure 1.4: Two hypotheses for how neurons represent priors. **Left**: a heterogeneous neural population implicitly represents the sensory prior in the density and gain of its tuning curves. a) a homogeneous neural population whose tuning curves tile the space. b) inset: the tuning curve density $d(s)$. The cumulative integral of this density $D(s)$ alters the positions and widths of the tuning curves in the neural population. c) the warped population, with tuning curve peaks scaled by the gain $g(s)$. **Right**: d) the sampling hypothesis. The neuron's firing rate result in samples from a marginal distribution corresponding to one latent variable; a collection of neurons can thus represent an arbitrary multivariate distribution via samples.

cal description of the environmental signals (i.e. the external prior) and a transformation derived according to a statistical optimization criterion. Showing that the model's response properties well match those of a set of sensory neurons serves as a test of the proposed transformation under a given environmental prior.

Under this framework, priors could be represented in the brain in a number of ways, depending on our assumed choice of transformation and resource under constraint. For example, priors may be implicitly represented in the selectivity of the neurons and the inhomogeneity of their preferred features (Fig. 1.4a-c) (Ganguli and Simoncelli, 2010; Ganguli and Simoncelli, 2014; Fischer and Peña, 2011). Alternatively, neurons representing the expected features of the environment may be present in larger numbers (Girshick et al., 2011), have sharper tuning (Schoups et al., 2001), or be more strongly connected to higher processing stages than non-expected features (Raiguel et al., 2006).

Another intriguing proposal is that the prior may be represented in the spontaneous activity of neurons in sensory cortex. The *sampling hypothesis* states that neural responses to external stimuli represent samples from the posterior distribution (Fig. 1.4d) (Fiser et al., 2010; Berkes et al., 2011). In the absence of external input, the likelihood term in Bayes' rule disappears, so the posterior collapses to the prior. Spontaneous activity – the neural responses when there is no external input – thus corresponds to samples from the prior. One advantage of this proposal is that it can easily accommodate the representation of high-dimensional latent variables since the dimensionality of the variable scales linearly with the number of neurons.

Berkes et al. (2011) provided empirical evidence for the sampling hypothesis by showing that distribution of spontaneous activity in the primary visual cortex in awake ferrets closely resembles that of average evoked activity. Moreover, they showed that the spontaneous activity of ferrets were closely matched as juveniles but diverged in similarity over the course of their lifetimes, in line with the expectation that priors evolve from lived experience.

1.5 Learning priors

Efficient coding and Bayesian inference both presuppose a statistical description of the world, but neither explains how a nervous system comes to possess one. A prior is not handed to the organism, but must be extracted from the very sensory stream it is meant to describe. Let us now suppose that as neuroscientists we wish to build a model of how learning occurs in the brain in response to natural stimuli. We are immediately confronted by several difficulties:

1. The priors we have discussed so far are defined over scalar variables such as orientation or contrast, whereas natural stimuli can only be described by a set of high-dimensional latent variables (e.g. in the case of texture (Bergen and Adelson, 1988; Heeger and Bergen, 1995; Portilla and Simoncelli, 2000)). A prior over such signals cannot be accumulated as a histogram over observed values because of the curse of dimensionality: the number of

independent observations required to get a good estimate scales exponentially with the dimensionality. To overcome this issue we must commit ourselves to a restricted set of transformations that contain useful inductive biases consistent with the set of natural stimuli we wish to learn.

2. The statistics of the natural environment evolve over multiple timescales, which recruit different mechanisms of learning in the brain. Some regularities appear to be inherited by the brain over evolutionary time, some are instantiated during development, and others gathered through experience across a lifetime (Seriès and Seitz, 2013). A model of prior acquisition must therefore specify not only how learning proceeds but what expectations the sensory system begins with and what data it is exposed to.

3. Efficient coding as suggested by Barlow (1961) assumes that inputs have low noise and that the neural responses are related to the input signal in a deterministic manner. However, real sensory neurons can exhibit large variability in response to identical stimuli. Past studies have considered the general case of noisy channels and found that different computational strategies emerge than those predicted by Barlow (Atick and Redlich, 1990; Atick and Redlich, 1992). A general model of learning should thus take both internal and external noise into account.

4. The mechanistic implementation of learning must be specified. Both efficient coding and Bayesian inference are computational frameworks; the manner in which neurons learn through sensory experience depend on how the prior is represented by neurons.

5. Thus far, efficient coding has been invoked as a theory of early sensory processing. The problems discussed above (e.g. the optimal response function to distributions of stimulus intensity or edge orientation) can be in principle performed by a single-stage neural network. But what can we say about the representations at later stages of sensory processing? Can we still treat efficient coding as a guiding principle for learning priors?

6. Neither the Bayesian nor the efficient coding framework specifies which latent variables should be learned. Bayesian inference asserts that priors are useful because they support inference, but presupposes, rather than derives, the latent variables to be inferred. Efficient coding holds that the code should retain as much information as its resources allow, but says nothing about how this information will be used, and hence which variables are worth knowing. Neither, then, tells us which variables are relevant enough to represent. This can be addressed by specifying a downstream task, formalized through Bayesian decision and estimation theory as a cost or reward function over some desired behavior. However, this simply shifts the problem from choosing useful latents to choosing an appropriate task, and it is unclear whether any such task is a "typical" one the organism would encounter in nature. More fundamentally, a representation tailored to any single task is an implausible model of cognition, which must serve many tasks at once.

7. An organism that aims to learn a prior over latent causes $p(z)$ never observes the latents z directly. It only has access to the latents through inference via the sensory transformation, but inference itself depends on the prior. Learning is thus a search for a fixed point of this circular dependence. A possible failure is that learning converges to a fixed point that is not useful. An example is where the latent carries no information about the stimulus, resulting in a self-consistent yet uninformative latent code.

1.5.1 InfoMax as a solution to epistemic uncertainty

One of the central challenges in neuroscience and machine learning is to understand what constitutes a "useful representation" (Bengio et al., 2014). As an illustrative example, consider an agent whose sole objective is to perform classification of natural images. By construction, the agent knows the task through its objective function, which scores how accurately it classifies its input. What, then, are the necessary conditions for a useful representation? If we assume

the inputs are generated by a set of latent features, what the agent must recover is merely those features that distinguish one class from another. A minimally useful representation in supervised learning is therefore one that discards everything in its input irrelevant to the task at hand.

This intuition can been formalized using the Information Bottleneck (IB) method (Tishby et al., 2000). Given some input variable X generated by a "relevance" variable Y, the IB method aims to find the maximally compressed mapping of the input variable, Z, that preserves as much as possible the information on the task variable. This relation is given a Markov chain $Y \to X \to Z$. Finding an optimal representation Z is formulated as a maximization of the Lagrangian

$$\mathcal{J}_{\text{IB}} := I(X;Z) - \beta\, I(Z;Y).$$

where β is a positive scalar Lagrange multiplier that trades off between the complexity of the representation $I(X;Z)$ and the amount of relevant information preserved $I(Z;Y)$. This framework has been used to study representations of deep neural networks (Tishby and Zaslavsky, 2015) and optimal prediction in biological systems (Sachdeva et al., 2020).

Such a setting of supervised learning is hardly applicable to biological agents. Humans and other animals inhabit an ever-changing, highly complex world, subject to a fitness landscape that shifts unpredictably. Functionally, this means they must perform a variety of tasks whose identities are not known in advance. This epistemic uncertainty over the task distribution is a crucial departure from the supervised setting, where the agent is evaluated on a single, prespecified task. And yet, despite this uncertainty, animals display a remarkable ability to adapt quickly to many novel tasks; few-shot learning is often taken as a hallmark of biological intelligence. What kinds of representations make such flexibility possible?

Under epistemic uncertainty, an agent has no way to know which features of its input a future task will require. And because information discarded at one stage cannot be recovered by any subsequent computation, discarding a feature risks permanently losing information that some

later task depends on. Formally, suppose we have a representation Z that is computed from the input X, and a downstream task T that an animal performs using its representation. This describes a Markov chain $X \to Z \to T$. In this case, because we can only lose but never gain new information through each stage of processing, the information that the representation carries about the task is strictly less than or equal to the information about the task that is present in the input. This is formalized by the data processing inequality $I(Z;T) \leqslant I(X;T)$. A representation Z that preserves *all* information in X (i.e. a sufficient statistic) attains $I(T;X) = I(T;Z)$ for any task T. Meanwhile, a representation that loses information about X attains this condition for some T but fails for others.

The rational strategy is therefore to preserve as much information about the inputs as possible. This is given by the InfoMax principle of "maximum information preservation" (Linsker, 1988); the optimal representation is one that maximizes the mutual information between the representation Z and the input X, subject to some arbitrary resource constraints $C(Z)$. This can be found through the maximization of the Lagrangian

$$\mathcal{J}_{\text{InfoMax}} := I(X;Z) + C(Z).$$

where the mutual information is defined as $I(X;Z) = H(X) - H(X \mid Z)$: the entropy of the sensory inputs $H(X)$ measures the total amount of information available to encode about the stimuli variable X in the natural environment, and the conditional entropy $H(X \mid Z)$ denotes the uncertainty about the stimulus that remains after observing the sensory response Z. Here the entropy $H(X) = -\int p(x) \log p(x)\, ds$ is fixed because we assume the input distribution $p(x)$ is stationary. A maximally efficient sensory system has a code Z that captures as much information about X as possible. This is the generalization of Barlow's efficient coding principle, where now neural variability to identical stimuli is not assumed to be negligible.

There is behavioral evidence that sensory systems extract structure in this task-independent

way. Turk-Browne et al. (2010) exposed human observers to sequences of novel shapes in which certain configurations reliably co-occur but are not task-relevant. They found that humans transfer knowledge of statistical regularities across space and time, appearing under conditions absent during encoding. The system thus behaves as though it encodes the input statistics for the possibility of future use rather than a mapping tuned to a particular task.

1.5.2 Rate-Distortion

Implementing InfoMax as an objective is difficult because it requires computing the mutual information $I(X;Z)$, which admits a closed form only in restricted cases such as jointly Gaussian or discrete variables. Though estimators exist (Bialek et al., 1991; Paninski, 2003), they are approximate, data-hungry, and systematically biased in the under-sampled regime.

A more tractable alternative is to require that the representation support reconstruction of its input. Since our input X can be recovered from the representation Z only to the extent that it preserves information about X, reconstruction error serves as a proxy for information loss. This objective alone is trivially satisfied by copying the input, such that $Z = X$. But as established above, the brain has limited capacity, so any useful objective must also constrain how much information the representation carries. Rate-distortion theory formalizes this trade-off by expressing both reconstruction error and coding cost in the common currency of information.

Rate-distortion theory is a branch of information theory that articulates the minimum amount of resources, expressed in terms of information, required to reconstruct a signal to some desired level of fidelity once it has been passed through a noisy communication channel. A communication channel comprises a "source" signal X, together with an encoder $p(z|x)$ and a decoder $p(\hat{x}|z)$ that reconstruct the signal (Fig. 1.5).

The distortion is the cost of representing a sample x with its approximate reconstruction $\hat{x}$.

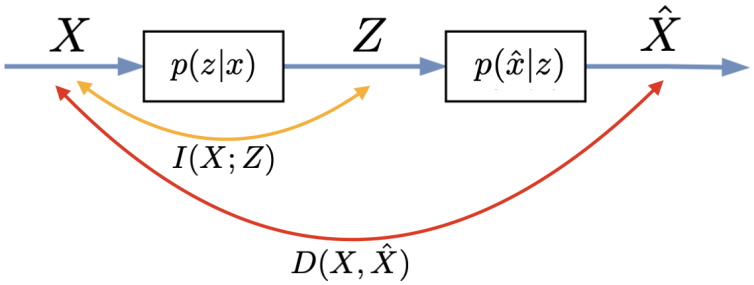


Figure 1.5: Diagram of a communication channel between source X and reconstruction $\hat{X}$ through a stochastic encoder $p(z|x)$ and decoder $p(\hat{x}|z)$. The rate is in yellow and the distortion in red.

We assess this via a distance metric $d(\cdot,\cdot)$, averaged over the joint distribution $p(x,\hat{x})$,

$$D(X,\hat{X}) := \mathbb{E}_{p(x,\hat{x})}\left[d(x,\hat{x})\right].$$

A typical choice is the squared-error distortion, $d(x,\hat{x}) = \|x-\hat{x}\|^2$. The rate R is the average number of bits per sample conveyed by the encoding through the channel, evaluated as the mutual information between the source and the encoded signal. A channel with a low rate caps the information that can be transferred and results in greater distortion; conversely, attaining distortion below a given level requires sending information at a higher rate. If the reconstructed signal perfectly matches the input, we call this *lossless* compression, while any residual distortion is the signature of *lossy* compression.

The relationship between rate and distortion is expressed by a pair of mutually inverse constrained problems, one that minimizes rate subject to a budget on distortion ($\mathcal{L}_{\text{R–D}}$), and its inverse that minimizes distortion subject to a budget on the allowed rate ($\mathcal{L}_{\text{D–R}}$):

$$\mathcal{L}_{\text{R-D}} = \inf_{p(z|x)} I(X;Z) \quad \text{s.t.} \quad D(X,\hat{X}) \leqslant D^*,$$

$$\mathcal{L}_{\text{D-R}} = \inf_{p(z|x)} D(X,\hat{X}) \quad \text{s.t.} \quad I(X;Z) \leqslant R^*.$$

Both can be combined into a single Lagrangian that trades off the two quantities through a mul-

tiplier β,

$$\mathcal{L} = D(X, \hat{X}) + \beta I(X; Z),$$

whose solutions, as β is swept over $(0, \infty)$, trace the boundary of the realizable region on the rate-distortion plane (Fig. 1.6).

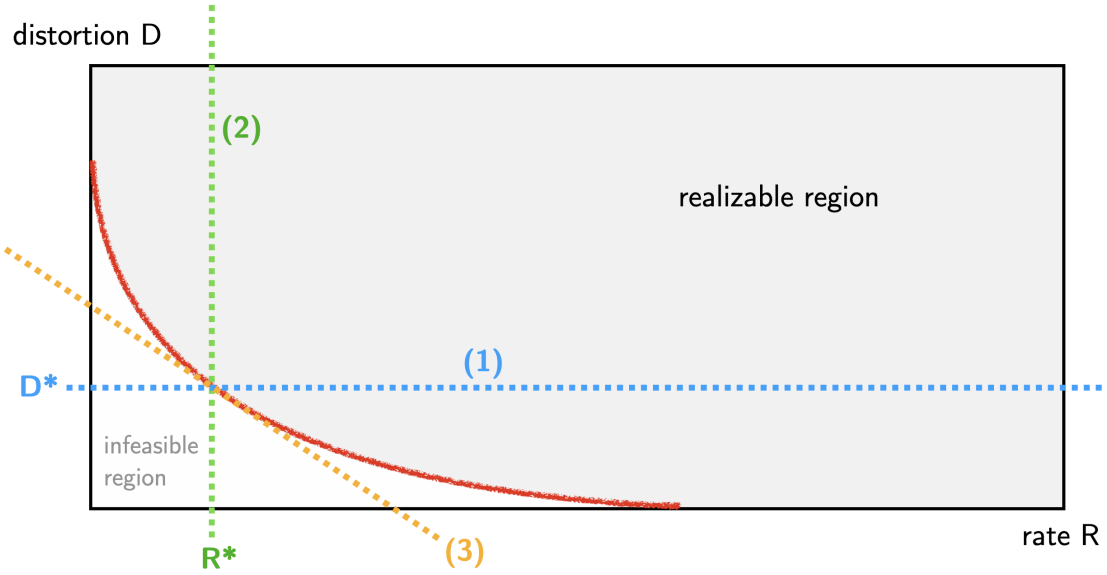


Figure 1.6: A qualitative information plane. The red line indicates the realizable limit of the rate-distortion objective and the shaded area denotes the realizable region. (1) The blue dotted line shows the minimum distortion for the rate-constrained rate-distortion objective. (2) The green dotted line shows the minimum rate for the distortion-constrained objective. (3) The intersection of the yellow dotted line with the red line shows the optimal rate and distortion for a given value of β.

How does this rate-distortion formalism relate to the InfoMax objective? Let us suppose we have a stochastic stimulus encoding from input X to latent Z given by $p(z|x)$. Recall the mutual information term from InfoMax,

$$\begin{aligned} I(X; Z) &= H(X) - H(X \mid Z) \\ &= H(X) + \mathbb{E}_{p(x,z)} \left[\log p(x \mid z) \right]. \end{aligned}$$

Since the entropy $H(X)$ is constant for a fixed stimulus distribution $p(x)$, InfoMax is a matter of minimizing the conditional entropy $H(X \mid Z)$. However, this quantity relies on the mapping

$p(x|z)$, which we have not defined. An ideal mapping that perfectly minimizes the conditional entropy would perfectly invert the stimulus encoding $p(z|x)$ from X to Z. For a stochastic encoding $p(z|x)$, the optimal inversion is given by Bayes' rule: $p(x|z) = {}^{p(z|x)p(x)}/_{p(z)}$. If $p(x|z)$ is conditional Gaussian with fixed covariance, the expected log likelihood reduces to a squared error up to an additive constant, so we can rewrite the mutual information in terms of MSE distortion,

$$I(X;Z) = H(X) - D\big(X, \hat{X}(Z)\big).$$

Thus minimizing the distortion is equivalent to maximizing the mutual information. This identity, however, faces two issues. Satisfying the conditional Gaussian assumption requires overly restrictive assumptions: we must also assume Gaussianity for the stimulus distribution $p(x)$, the marginal over the representation $p(z)$ and the encoding $p(z|x)$, even though empirical measurements show that the statistics of natural images $p(x)$ are highly non-Gaussian (Field, 1987). Moreover, the Bayesian inversion is in general intractable because we do not have direct access to the input distribution $p(x)$.

Variational Approaches to InfoMax Both of these issues stem from our inability to evaluate the true posterior $p(x|z)$. One standard approach is to replace it with a tractable surrogate $q(x|z)$ drawn from a variational family Q, then optimize over Q to find the closest available approximation. The cost of making this substitution on the mutual information is given by the Barber-Agakov bound,

$$I(X;Z) \geq H(X) + \mathbb{E}_{p(x,z)}\left[\log q(x \mid z)\right] = \tilde{I}(X;Z), \tag{1.1}$$

where $\tilde{I}(X;Z)$ is the variational approximation of the mutual information. We see that $\tilde{I}(X;Z) = I(X;Z)$ when $q(x|z) = p(x|z)$; the gap is otherwise the KL divergence $\mathbb{E}_{p(z)}\big[D_{\mathrm{KL}}\big(p(x|z)\|q(x|z)\big)\big]$. As we saw above, restricting Q to conditional Gaussians with fixed covariance reduces the (now

approximate) conditional entropy to MSE distortion. Since $H(X)$ is fixed, minimizing this reconstruction error maximizes a lower bound on the mutual information.

1.5.3 Generative modeling

Models that learn representations by reconstruction have a long history. Much like a communication channel in rate-distortion, they encode input signals into a compressed representation and then reconstruct the signal from it. The roles of the two networks, however, are reversed relative to the InfoMax account above. There, the encoder was the object of interest and the decoder served only to certify that information had been preserved (Fig. 1.7a). Here the decoder is central: the aim is to learn an *internal generative model* $p(x|z)$, a description of the process by which the observed data x were produced (Fig. 1.7b). Inference is then the inversion of this generative process. As in Helmholtz's account of perception as unconscious inference, sensory evidence is evaluated against hypotheses about the latent causes that could have produced it, and perception corresponds to the causes deemed most probable. The Bayesian treatment generalizes this from a single best explanation to a full distribution over possible causes, given by the posterior $p(z|x) = {}^{p(x|z)p(z)}/_{p(x)}$. Since the posterior is generally intractable, these models perform infer-

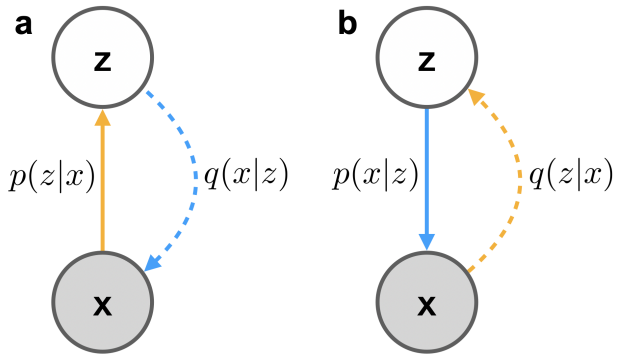


Figure 1.7: Comparison between the variational InfoMax and variational generative modeling perspectives. Both approaches learn representations via reconstruction, but a) in variational InfoMax the encoding is the object of interest, while b) in generative modeling the generation is the object of interest. Inference in yellow, generation in blue. Variational approximation is in dotted lines.

ence either by sampling (Ackley et al., 1985; Neal, 1992) or by variational approximation (Dayan et al., 1995; Kingma and Welling, 2013; Rezende et al., 2014). Variational approaches introduce a surrogate $q(z|x)$ from a simple family Q, chosen to approximate the true posterior as closely as the family allows. Rather than solving a separate optimization problem for each input, the surrogate is *amortized*: a single network, the "recognition model", is trained to map any x directly to the parameters of $q(z|x)$, so that inference reduces to a forward pass (Gershman and Goodman, 2014). The cost of inference is paid once, during learning, and shared across all subsequent inputs.

The aim of generative models is to capture the unknown generative process that gives rise to a distribution over sensory inputs $p_{\mathcal{D}}(x)$. That process is never directly observed; instead, the model must learn from a finite sample, such as a dataset of natural images, which we assume are drawn independently and identically (iid) from the underlying distribution $\{x_i\}_i^N \sim p_{\mathcal{D}}(x)$. Given latent variables z and parameters θ, we fit the model by maximizing the expected marginal log likelihood of the data[1]:

$$\mathcal{L} = \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\log \int p_\theta(x,z)\, dz\right] = \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\log p_\theta(x)\right]$$

which differs only by a constant from $-D_{\mathrm{KL}}\left(p_{\mathcal{D}}(x) \| p_\theta(x)\right)$. The objective is therefore maximized when the model reproduces the data distribution.

When the mean of the generative model is a nonlinear function, the integral admits no closed form. The obvious solution, which is to draw samples from the prior $p(z)$ and average the resulting likelihoods, is unbiased but impractical, since the posterior typically occupies a tiny region of the prior's support (Chatterjee and Diaconis, 2018). Instead, we use the approximate posterior $q_\phi(z|x)$, parameterized by a separate neural network with weights ϕ, and rewrite the marginal

[1]The logarithm is strictly increasing, so it leaves the location of the optimum unchanged; we work with it for numerical stability and for analytic convenience, particularly when the densities involved belong to the exponential family.

likelihood as an importance-weighted expectation under it:

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)} \left[\frac{p_\theta(x,z)}{q_\phi(z|x)} \right].$$

The inner expectation gives an unbiased Monte Carlo estimate, but because the logarithm sits outside it, the estimate is biased below. There are two options: we may retain this objective and reduce the bias by taking multiple posterior samples (Burda et al., 2016), or draw a single sample and move the logarithm inside of the expectation. Since the logarithm is concave, Jensen's inequality tells us the exchange can only lower the value. The resulting quantity is a lower bound on the marginal log likelihood, known as the evidence lower bound (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)} \right] := \mathcal{L}_{\text{ELBO}},$$

Factoring the joint as $p_\theta(x,z) = p_\theta(x|z)\, p(z)$ decomposes the ELBO into two terms, a reconstruction term and a regularizer that penalizes the dissimilarity between the approximate posterior and a fixed prior over the latents, commonly an isotropic Gaussian $p(z) = \mathcal{N}(0, I)$:

$$\log p_\theta(x) \geq \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)} \left[p_\theta(x|z) \right] - D_{\text{KL}} \left(q_\phi(z|x) \| p(z) \right).$$

where the slack in the bound is given by how well the approximate posterior matches the true posterior, $D_{\text{KL}} \left(q_\phi(z|x) \| p_\theta(z|x) \right)$.

A closer inspection reveals that the ELBO reconstruction term is exactly equivalent to the variational approximation of the conditional entropy used to place a lower bound on the mutual information (Eq. 1.1). Thus, under assumptions of conditional Gaussianity on the decoder/generative model, we can now form equivalences between three concepts: conditional entropy, reconstruction, and distortion. In Chapter 4, I will show how these equivalences can be maintained without assumptions of Gaussianity.

How should we interpret the KL term in the ELBO objective? We can show that it upper bounds the *rate* (Hoffman and Johnson, 2016; Cover and Thomas, 2006):

$$D_{\mathrm{KL}}\big(q_\phi(z|x)\|p(z)\big) = I_q(X;Z) + D_{\mathrm{KL}}\big(q(z)\|p(z)\big) \geq I_q(X;Z),$$

where $I_q(X;Z)$ denotes the mutual information between X and Z as computed by the recognition model rather than the generative model. Since $D_{\mathrm{KL}}\big(q(z)\|p(z)\big) \geq 0$, the ELBO KL term is equivalent to the rate $I_q(X;Z)$ when the empirical marginal distribution over the latents, computed as the aggregate posterior $q(z) = \int q(z|x)p(x)\,dx$, matches the assumed prior such that $q(z) = p(z)$. Thus optimizing the ELBO implicitly minimizes the rate, subject to an additional constraint that the aggregate posterior match the assumed latent prior. In Chapter 4 I will show how to optimize the rate directly without the need for additional constraints on the form of the aggregate posterior.

Diffusion models Diffusion models are another, more recent class of generative models (Sohl-Dickstein et al., 2015a; Song and Ermon, 2019; Ho et al., 2020; Kadkhodaie and Simoncelli, 2020a). Relative to VAEs and generative adversarial networks (GANs) (Goodfellow et al., 2014), they are more expressive yet easier and more stable to train, which has made them the dominant approach to generative modeling of images and other complex signals. Their capacity to produce samples that are both probable under the data distribution and perceptually convincing implies that they implicitly carry an expressive prior over natural signals, which makes them a useful object of study for understanding and modeling the priors the brain must itself acquire.

Diffusion models can be explained in a variety of ways, but perhaps the simplest is through denoising. Suppose we have i.i.d. samples from a data distribution $x \sim p_{\mathcal{D}}(x)$, such as a dataset of natural images. We corrupt these samples with isotropic Gaussian noise of varying degrees such that $x_\gamma = \sqrt{\gamma}x + \epsilon$, where x_γ is the noisy image, $\epsilon \sim \mathcal{N}(0, I)$ is a sample of white noise and γ is the signal-to-noise ratio of the noisy image, drawn from some distribution on $(0, \infty)$. A single

denoiser parametrized by a neural network $\hat{x}_\theta$ can be trained to denoise images of *any* SNR. The network takes the noised image and the SNR value as inputs, and is trained with an MSE loss:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,x_\gamma,\gamma}\left[\left\|x - \hat{x}(x_\gamma, \gamma)\right\|^2\right].$$

Once trained, the network can be applied recursively, each pass uncovering only a marginal amount of signal. Chaining these partial denoising operations together yields the *reverse* process, which refines the signal as γ grows until we are left with a "clean" image $\hat{x}_{\gamma=\infty}$. This construction is deceptively simple, and much of the theoretical interest in diffusion models lies in how many separate theoretical traditions it makes contact with (Turner et al., 2024). Three perspectives bear directly on my work.

First, the reverse process is a stochastic dynamical system. Tweedie's or Miyasawa's formula (Miyasawa et al., 1961; Robbins, 1956) relates the optimal denoiser with the score $\nabla_{x_\gamma} \log p_\gamma(x_\gamma)$ of the noisy distribution, so each reverse step stochastically climbs towards regions of higher density under p_γ, the data distribution smoothed by noise at the current SNR. Because these densities sharpen as γ grows, a trajectory that begins in a nearly featureless landscape progressively commits to a basin, and generation amounts to a stochastic selection among attractors.

Second, the forward corruption can be read as a fixed inference model. In this reading, the noisy images $\{x_\gamma\}$ at successive SNRs are latent variables, the reverse process is the corresponding generative model, and the denoising objective, appropriately weighted, is the ELBO for a VAE with an infinite hierarchy of latents (Sohl-Dickstein et al., 2015a; Kingma et al., 2021). What distinguishes it from the VAEs above is that the encoder is not learned but fixed (by construction). The hierarchy is imposed by the noise schedule rather than discovered.

Third, the denoiser is an MMSE estimator, and its loss at a given SNR is the minimum mean-squared error at that SNR. The I-MMSE relation (Guo et al., 2005) equates the derivative of the mutual information between clean and noisy signal with respect to γ to one half of the MMSE, so

integrating the model's own training loss across SNRs recovers information-theoretic properties of the source, including estimates of the log probability of individual samples and the entropy of their distribution (Guth et al., 2025). The quantity the model is trained to minimize thus doubles as an estimator of what the model has learned.

These three perspectives organize the chapters that follow. Chapter 2 takes the attractor framework as a model of perception, making the case that dynamics of this kind underpin inference in the brain. Chapter 3 develops the variational reading of diffusion by combining the diffusion and VAE objectives into a common form. Chapter 4 draws on the information-theoretic connection to operationalize the InfoMax principle.

1.6 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 proposes a neural circuit model for representing and inferring complex, high-dimensional priors. We show how dendritic computation integrates bottom-up sensory and top-down contextual signals with prior knowledge, and make several testable neural predictions. Chapters 3, 4, and 5 turn to the *learning* of priors, developing new computational theories of unsupervised representation learning in the visual cortex. The three chapters trace the evolution of a single idea: that representations learned without supervision can benefit from better statistical knowledge of the underlying stimulus manifold. Chapter 3 develops this idea from a generative modeling perspective, making use of variational inference and the ELBO, while Chapters 4 and 5 generalize it into a purely information-theoretic account of unsupervised learning. In Chapter 5, I extend the InfoMax principle to a hierarchy of latent variables. Finally, Chapter 6 closes the thesis with a general discussion.

2 A circuit model for representing complex priors and performing flexible inference

2.1 Overview

Despite many successful examples in which probabilistic inference can account for perception, we have little understanding of how the brain represents and uses structured priors that capture the complexity of natural input statistics. Here we construct a recurrent circuit model that can implicitly represent priors over latent variables, and combine them with sensory and contextual sources of information to encode task-specific posteriors. Inspired by the recent success of diffusion models as means of learning and using priors over images, our model uses dendritic nonlinearities optimized for denoising, and stochastic somatic integration with the degree of noise modulated by an oscillating global signal. Combining these elements into a recurrent network yields a stochastic dynamical system that samples from the prior at a rate prescribed by the period of the global oscillator. Additional inputs reflecting sensory or top-down contextual information alter these dynamics to generate samples from the corresponding posterior, with different input gating patterns selecting different inference tasks. We demonstrate that this architecture can sample from low dimensional nonlinear manifolds and multimodal posteriors. Overall, the

model provides a new framework for circuit-level representation of probabilistic information, in a format that facilitates flexible inference.

A version of this work was presented at Computational and Systems Neuroscience (2023), and the main findings are published in the Proceedings of the Twelfth International Conference on Learning Representations (Lyo and Savin, 2023).

2.2 Introduction

The organization of past experience in the form of internal models is crucial for making sense of noisy and incomplete sensory information. The framework of Bayesian inference successfully formalizes this knowledge in terms of probabilistic priors, and perception and action as probabilistic inference. However, despite many successes in explaining behavioral (Mamassian et al., 2002; Li et al., 2003; Gerardin et al., 2010; Girshick et al., 2011; Wei and Stocker, 2012; Ma, 2012; Hahn and Wei, 2022; Angeletos Chrysaitis and Seriès, 2023) and neural (Zemel and Dayan, 1998; Ma et al., 2006; Fiser et al., 2010; Savin and Denève, 2014; Savin et al., 2014; Orbán et al., 2016; Echeveste et al., 2020; Aitchison et al., 2021) observations through the lens of Bayesian inference, the circuit implementation of such computations remains poorly understood. In particular, we know relatively little about how priors are represented and the mechanics of how they affect perception.

Brain priors come in many flavors. For instance, priors over faces are structurally complex and invariant across tasks within subjects, despite substantial across subject variability (Houlsby et al., 2013). Such structural priors reflect natural world statistics; they allow for efficient learning and powerful generalization across a range of tasks (Gerardin et al., 2010; Houlsby et al., 2013). In contrast, contextual priors are relatively short-lived and adapt to reflect the needs of the task at hand (Seriès and Seitz, 2013; Darlington et al., 2018). Due to these differences, it is possible for different types of priors to have different neural representations (Rullán Buxó and Savin, 2021)

and be accessed through separate neural mechanisms during inference.

Neural representation of priors are generally assumed to be implicit, with the exact form reflecting different schools of thought on neural correlates of uncertainty (Ma et al., 2006; Fiser et al., 2010). One set of models take the view that priors are represented in the stimulus encoding population, with more neural resources dedicated to higher probability stimuli (Fischer and Peña, 2011; Girshick et al., 2011; Ganguli and Simoncelli, 2014; Bredenberg et al., 2020). The (log) prior could also be encoded in a separate subpopulation, e.g. in the form of a probabilistic population code and then linearly combined with the (log) sensory evidence (Ma et al., 2006). These theoretical schemes focus largely on the probabilistic representation of a single, one-dimensional stimulus feature, and cannot be readily generalized to the high-dimensional structured priors needed for representing natural stimuli (Houlsby et al., 2013; Durrant et al., 2023). Such complex distributions can be represented in stochastic recurrent circuits whose activity over time encodes samples from the posterior distributions (Hoyer and Hyvärinen, 2002; Fiser et al., 2010), or—in the limit of vanishing sensory evidence—samples from the prior (Berkes et al., 2011). In this neural sampling framework, prior information is implicitly encoded in the recurrent collaterals between the neurons that encode the inferred latents. However, the worked out examples of neural sampling dynamics have so far been restricted to priors and posteriors with relatively simple structure, e.g. Gaussian scale mixtures to account for statistics of gratings in early vision (Orbán et al., 2016), or multimodal distributions to model bistable percepts (Reichert et al., 2011; Savin et al., 2014). Finally, some priors could be encoded downstream of sensory representations (Ma, 2012), affecting perception through feedback mechanisms akin to attention (Whiteley and Sahani, 2012), although the circuit implementation of such effects remains underspecified. Overall, we are still missing circuit level models of probabilistic computation that capture the statistical complexity of natural priors and their flexible use across tasks.

The problem of representing complex high-dimensional distributions has close correspondents in machine learning. For example, digital images lie in a high dimensional pixel space,

yet not all points in this space correspond to veridical images. Within this huge ambient dimensionality, naturalistic images trace low dimensional nonlinear manifolds, whose local coordinates represent continuous deformations or intensity variations. Among recent machine learning attempts at modeling such structure, diffusion models have emerged as particularly expressive and stable generative neural network models (Sohl-Dickstein et al., 2015b; Kadkhodaie and Simoncelli, 2020b; Ho et al., 2020), outperforming previous generative solutions such as VAEs, normalizing flows and GANs on the task of image synthesis (Dhariwal and Nichol, 2021). Their training involves local denoising operations, which can translate into relatively simple local learning objectives (Raphan and Simoncelli, 2011). Furthermore, they can be combined flexibly with different likelihoods to support inference (Sohl-Dickstein et al., 2015b; Kadkhodaie and Simoncelli, 2021; Song and Ermon, 2019; Nichol and Dhariwal, 2021). Taken together, these features make diffusion models a valuable source of inspiration for building new classes of neural circuits for probabilistic computation.

Here we adapt several ideas developed in the context of diffusion models to construct a recurrent circuit model that can implicitly represent sensory priors and combine them with other sources of information to encode task-specific posteriors. Our solution relies on dendritic nonlinearities, optimized for denoising, and stochastic somatic activity modulated by a global oscillation that determines the effective rate of sampling. Additional inputs into the circuit which provide sensory or top-down contextual information shape the dynamics to generate samples from the corresponding posteriors. In simulations, we demonstrate several scenarios of prior and posterior encoding, including nonlinear manifolds embedded in a higher dimensional ambient space as priors and several likelihoods corresponding to bottom-up and top-down evidence. We also identify potential neural signatures of such probabilistic inference that might be testable experimentally.

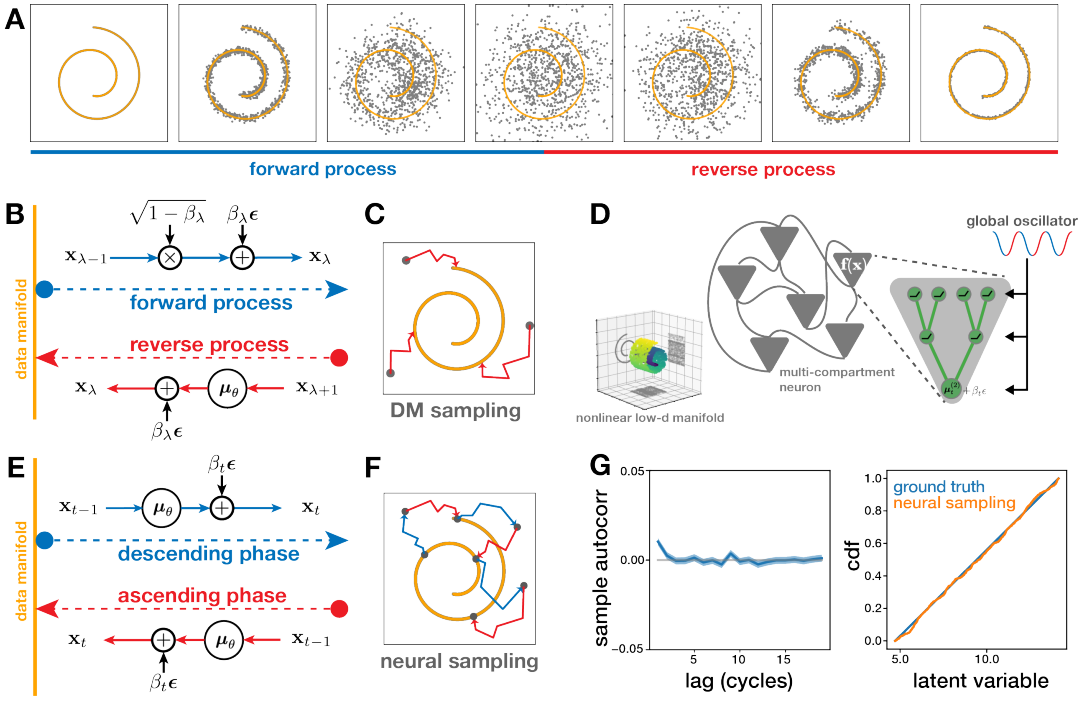


Figure 2.1: Diffusion models and their circuit implementation. **(A)** Simulated diffusion example for a two dimensional data distribution (yellow). Data samples from the manifold (gray dots) are corrupted by increasing levels of noise (blue), while the reverse process pushes the samples back onto the data manifold by recursively removing some of the noise (red). **(B)** Graphic summary of the two core DM operators: forward process adds noise with a predefined variance; dataset-specific reverse process removes the noise to bring the output back to the image manifold. **(C)** DM sampling starts from random points in the state space (gray dots), drawn independently from a standard Gaussian. The reverse process pushes these points back onto the manifold, resulting into independent samples from the data distribution. **(D)** Cartoon of a neural circuit encoding a nonlinear low dimensional manifold embedded in a larger neural feature space ("swiss-roll"). The network includes all-to-all recurrent connections between multi-compartment neurons (gray triangles), each consisting of a dendritic tree with optimized weights and ReLu nonlinearities, modulated by a global oscillatory signal. **(E)** Graphic summary of sampling in the neural circuit: starting from a valid point on the manifold, increasing amounts of noise due to the neural dynamics push the network off manifold, followed by the regular reverse process whose attractor forces brings the dynamics back on manifold, resulting in a new (potentially correlated) sample from the prior. **(F)** Neural sampling alternates between a descending phase which pushes points off the manifold through an increase in noise (blue trajectories), and an ascending phase, where attractor dynamics draw samples towards the manifold (red trajectories). **(G)** Numerical validation of neural sampling for the swiss-roll example. Left: autocorrelation function of the neural circuit generated samples. Right: marginal density recovery along the uniform dimension.

2.3 Encoding complex priors in recurrent neural circuits

2.3.1 Representing priors with diffusion models

Diffusion models (DMs) are generative neural networks that develop an implicit representation of data statistics by learning to denoise noisy versions of the data across multiple noise levels (Sohl-Dickstein et al., 2015b; Kadkhodaie and Simoncelli, 2020b; Ho et al., 2020). While there is more than one way to describe DMs, the essence of the process involves a predetermined *forward process*, which iteratively adds noise with known statistics to the input, and a learned *reverse process* which aims to "undo" the effects of the added noise.

Formally, the forward process transforms N-dimensional samples from a data distribution $\{\mathbf{x}_0\} \sim \mathrm{p}(\mathbf{x}_0)$, e.g. natural images, by progressively adding small amounts of Gaussian noise over a number of steps Λ following a pre-determined variance schedule $\{\sigma_\lambda\} \in (0, 1)$ that increases in magnitude with index $\lambda = 1, \cdots, \Lambda$. Across these steps, noise gradually accumulates to ultimately overwhelm the signal, and the image distribution morphs into a standard N-dimensional multivariate Gaussian (Fig. 2.1A, B; blue). Each step of this process corresponds to the application of a stochastic forward transition operator that maps noisy versions of the data sample $\mathbf{x}_{\lambda-1}$ into noisier versions $\mathbf{x}_\lambda$,

$$p(\mathbf{x}_\lambda \mid \mathbf{x}_{\lambda-1}) = \mathcal{N}(\mathbf{x}_\lambda; \sqrt{1-\sigma_\lambda}\mathbf{x}_{\lambda-1}, \sigma_\lambda I), \tag{2.1}$$

Analogously, the reverse process involves a series of applications of a Gaussian reverse transition operator that partially denoises noisy versions of the image $\mathbf{x}_{\lambda+1}$ into less noisy versions $\mathbf{x}_\lambda$ (Fig. 2.1A, B; red),

$$p_\theta(\mathbf{x}_\lambda \mid \mathbf{x}_{\lambda+1}) = \mathcal{N}(\mathbf{x}_\lambda; \boldsymbol{\mu}_\theta(\mathbf{x}_{\lambda+1}, \lambda), \sigma_\lambda I). \tag{2.2}$$

The mean of the reverse transition operator $\boldsymbol{\mu}_\theta(\mathbf{x}_{\lambda+1}, \lambda)$ is parameterized by a feedforward neural network that takes as input the noisy image from the previous iteration $\mathbf{x}_{\lambda+1}$ and the current

noise index λ, with a matching, but reversed, variance schedule. The network learns this iterative mapping by optimizing a denoising objective that minimizes the reconstruction error, or equivalently, the L2 distance between the estimated and true noise present in the image:

$$\mathcal{L} = \mathbb{E}_{\lambda\sim[1,\Lambda],\,\mathbf{x}_0,\,\boldsymbol{\epsilon}_\lambda}\left[\|\tilde{\boldsymbol{\mu}}_\lambda(\mathbf{x}_\lambda,\boldsymbol{\epsilon}_\lambda) - \boldsymbol{\mu}_\theta(\mathbf{x}_\lambda,\lambda)\|_2^2\right] \tag{2.3}$$

$$= \mathbb{E}_{\lambda\sim[1,\Lambda],\,\mathbf{x}_0,\,\boldsymbol{\epsilon}_\lambda}\left[\|\boldsymbol{\epsilon}_\lambda - \boldsymbol{\epsilon}_\theta(\mathbf{x}_\lambda,\lambda)\|_2^2\right], \tag{2.4}$$

where $\boldsymbol{\epsilon}_\lambda \sim \mathcal{N}(0, I)$, $\tilde{\boldsymbol{\mu}}_\lambda$ is the mean of the forward transition operator at step λ when conditioned on a clean image $\mathbf{x}_0$, and $\boldsymbol{\epsilon}_\theta$ is the network's estimate of the noise present in the noisy image. This objective is equivalent to learning the flow field associated with the score of the reverse transition operator at every noise level $\nabla_{\mathbf{x}_\lambda} \log p(\mathbf{x}_\lambda \mid \mathbf{x}_{\lambda+1})$ (Hyvärinen, 2005; Vincent, 2011). At the end of training, the model can generate novel sample images from the same distribution, by starting with i.i.d. multivariate Gaussian noise and iteratively applying the reverse transition operator following the same variance schedule. Each run of this procedure generates a new independent image sample.

2.3.2 A diffusion-based circuit for sampling-based probabilistic computation

Several properties of diffusion models seem appealing from a biological perspective. In particular, the iterative nature of the operations make them naturally map into recurrent circuit dynamics. In this view, the data distribution follows a nonlinear manifold attractor, embedded in the N-dimensional space of sensory features encoded by the network (not necessarily image pixels). The forward process injects noise that pushes data samples away from the manifold, generating signals for learning, while the reverse process implements denoising. Here, we chose to focus specifically on the circuit dynamics and the mechanism for sample generation. We leave the learning aspect of the problem for future work.

A recurrent circuit can implement something akin to the DM reverse process, by assuming that the the nonlinear reverse operator $p_\theta(\mathbf{x}_\lambda \mid \mathbf{x}_{\lambda+1})$ is computed in the dendritic trees of a set of N cortical pyramidal neurons. At each time step, nonlinear dendritic computations (Poirazi et al., 2003; Polsky et al., 2004; Jadi et al., 2014; Ujfalussy et al., 2018) yield somatic currents that reflect the mean of the Gaussian transition operator $\boldsymbol{\mu}_\theta$, which is subsequently corrupted by somatic noise. As an analogue for the variance schedule, we introduce a global time-varying signal $\beta_t \in (0, 1)$ which modulates the nonlinear operation of the dendrites and the magnitude of somatic noise. Mechanistically, this type of modulation could be implemented by local inhibitory subunits (Murayama et al., 2009; Roux and Buzsáki, 2015). We adopt the convention of associating increasing β_t with a reduction in noise, although this choice is somewhat arbitrary. Scalar β_t is a sinusoid which fluctuates over time between 0 and 1, in anti-phase to the DM noise schedule σ_λ, to account for the fact that downstream circuits read out information at the peak of the local oscillation (Savin et al., 2014):

$$\beta_t = \frac{1}{2}\left(1 + \cos\left(\frac{2\pi t}{T}\right)\right) = 1 - \frac{\sigma_\lambda - \sigma_1}{\sigma_\Lambda - \sigma_1}, \tag{2.5}$$

where T is the number of time steps in a cycle. Putting everything together, recurrent circuit dynamics take the form : $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \beta_t) + (1 - \beta_t)\boldsymbol{\epsilon}_t$, where $f(\cdot, \cdot)$ reflects the nonlinear function computed by the dendrites, itself modulated by the global oscillation, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I)$ are i.i.d. samples from an isotropic Gaussian. Following Poirazi et al. (2003), dendritic computations are modeled using a tree-structured artificial neural network (ANN), whose parameters we train via gradient descent using the standard DM denoising objective (Fig. 2.1D). Unlike previous ANN-based models of dendritic processing, we use ReLU nonlinearities to facilitate training, with the shape of the dendritic tree parameterized by its depth and branching factors at each level.[1] This biologically-motivated solution represents an architectural departure from standard DMs, which

[1]Unless otherwise specified, simulations use a depth of 7 and branching factor of 3, except in the most proximal section, which has a branching factor 4.

assume all-to-all connectivity. The ability to represent priors is relatively robust to variations in tree geometry, although learning speed can vary with architecture. In particular, varying tree depth and width while keeping the total number of parameters constant identifies shallow architectures as the fastest learners in our example (see Suppl. A.2). This implies that one can determine an optimal architecture for encoding any given prior, with potential implications for cortical dendritic morphology.

Another key distinction between traditional DMs and biology is the initial condition for the reverse process during sampling. In DMs, initial conditions are drawn independently from a standard Gaussian, which would yield unrealistic state jumps in the circuit dynamics. In contrast, in the recurrent circuit the same neural units need to support the dynamics at all times. As a result, we use the same *neural transition operator*, $p_\theta(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_\theta, (1-\beta_t)I)$, but under an increasing noise schedule to reinitialize the reverse process and generate another sample from the prior (Fig. 2.1E). If we were to use the DM forward process for this operation, the end effect would be statistically equivalent to the original (since noise dominates the signal, by construction, the initial conditions are forgotten). The proposed neural dynamics do not necessarily come with the same mathematical guarantees, but they do significantly simplify the circuit's operations. In this scheme, sampling from the prior entails running the same stochastic nonlinear recurrent dynamics, at a pace set by an oscillatory β_t signal, with one new sample from the prior generated once per cycle (Fig. 2.1F). [2] The precise structure of these oscillations can be relaxed to account for frequency and amplitude fluctuations seen biologically (Suppl. Sec. A.6).

We numerically tested the quality of the samples generated by our neural circuit in a toy example of a two-dimensional nonlinear manifold (shaped as a "swiss-roll", see Fig. 2.1D inset) with linear dimensionality 3, embedded in an ambient feature space with dimensionality $N = 10$. This simple example both captures the essence of priors of interest (i.e., nonlinear low-dimensional manifold embedded in a larger feature space) and is simple enough to exploit its structure when

[2] In all simulations we initialize the dynamics randomly and discard the first cycle as burn-in period.

assessing the quality of the generated samples, something which is generally difficult for complex distributions (Theis et al., 2016). First, we tested the effects of the neural approximation on the autocorrelation function of the generated samples and found that this remains steadily around zero even with the shortest lags (Fig. 2.1G, left), proving that, despite the approximation, the generated samples are essentially independent. Second, to test the variability of the generated samples we numerically estimated the density along the curved axis of the manifold, which—by construction—has a uniform distribution. We found that the empirical density constructed using the neural samples closely matched the ground truth (Fig. 2.1G). This is also true for the marginal along the orthogonal dimension (Suppl. A.3). While the quality of samples is harder to estimate, we also find good quality representations of a high dimensional prior trained on the MNIST dataset (Deng, 2012) (see Suppl. A.7). Overall, these results indicate that the constraints imposed by biology may have a minimal effect on the quality of neural sampling as compared to traditional DMs.

2.4 Flexible inference

2.4.1 Using the same prior-encoding circuit for inference

Machine learning has proposed several algorithmic solutions for combining DM-based prior representations with different likelihood functions, with practical applications such as in-painting, super-resolution, compressed sensing, or class-specific conditioning (Sohl-Dickstein et al., 2015b; Dhariwal and Nichol, 2021; Kadkhodaie and Simoncelli, 2021; Song and Ermon, 2019; Ho and Salimans, 2021). In order to maximize the flexibility of our circuit implementation, we take advantage of a modular approach that reuses the dynamics that sample from the prior to do flexible inference. In particular, posterior samples are constructed by altering the flow-fields of the DM-prior, in the form of state-dependent additive changes to the DM reverse transition operator (Sohl-Dickstein

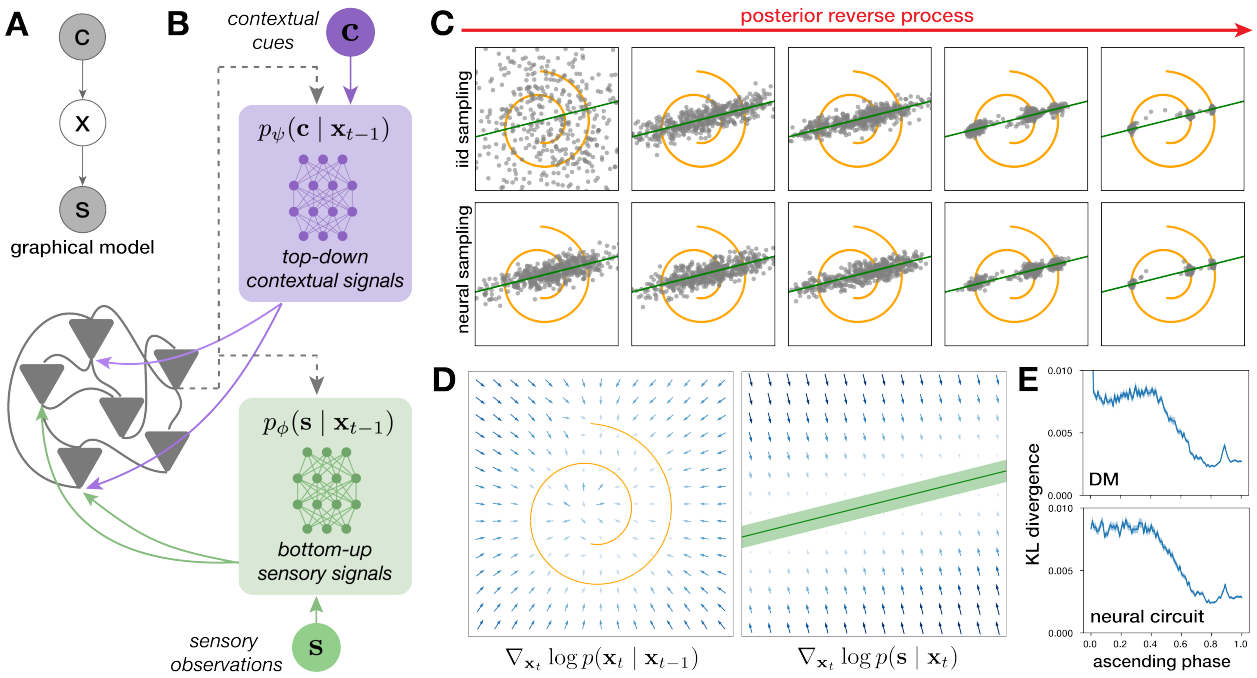


Figure 2.2: Posterior inference. **(A)** Graphical model illustrating dependencies between latent variables $\mathbf{x}$ and sensory inputs $\mathbf{s}$. **(B)** Extended neural circuit for flexible inference: prior dynamics are reshaped by external signals due to either bottom-up sensory observations (green) and/or top-down contextual cues (purple). These inputs are themselves modulated by the state of the circuit, through inter-area interactions. **(C)** Snapshots of posterior samples during the reverse process for the diffusion model (top) and the neural circuit (bottom) under the influence of bottom-up sensory signals. **(D)** Flow fields corresponding to a prior (left) and a linear gaussian sensory likelihood (right; shaded green area 1sd). **(E)** Quantification of the cost of approximating the likelihood when sampling using DM (top) or neural circuit (bottom) dynamics, measured by the KL divergence between the true posterior (evaluated at $\boldsymbol{\mu}_\theta$) and the approximate posterior (evaluated at $\mathbf{x}_{t-1}$).

et al., 2015b; Dhariwal and Nichol, 2021).

In general, our approach entails defining a *posterior neural transition operator* which reflects how additional evidence changes the beliefs about the represented latent variables (Fig.2.2A). Formally, posterior sampling dynamics take the form (see Suppl. A.1 for derivation):

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}) := \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1}) + \gamma(1-\beta_t)\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y}), (1-\beta_t)I), \tag{2.6}$$

where the effects of the prior are reflected in the mean $\boldsymbol{\mu}_\theta$, as before, but shifted by the (log)likelihood gradient $\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y}) = \nabla_{\mathbf{x}_{t-1}} \log p_\phi(\mathbf{y} \mid \mathbf{x}_{t-1})$, which, unlike Dhariwal and Nichol (2021), is evaluated with respect to a noisy data sample at the previous time step $\mathbf{x}_{t-1}$. Scalar γ is a hyperparameter that weighs the relative contribution of the log prior and log likelihood. This hyperparameter is set to 1 in Sohl-Dickstein et al. (2015b) and varied in Dhariwal and Nichol (2021) to optimize a trade-off between sample fidelity and diversity (here, $\gamma = 1$). The relative influence of the additional evidence on the circuit dynamics varies across the cycle, scaled by $(1-\beta_t)$. Mechanistically, this solution means that the local recurrent circuit representing the prior receives an external input drive, itself gated by the global oscillation, which shifts circuits dynamics to sampling from the corresponding posterior. The nature of the input signal is specific to each of the sources of evidence, and implemented in a separate feedforward sub-circuit, which interacts bi-directionally with the primary recurrent network. At each timestep, the sub-circuit receives the activity of the recurrent network from the previous timestep and returns an input drive that is added to the somatic current of each neuron in the recurrent network. Functionally, this sub-circuit implements the computation of $\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y})$ (Fig. 2.2B). The functional form of this circuit is problem specific, determined either in closed form (see sensory inference example below), or itself learned (e.g., in the context cue example). Different sources of evidence can operate in parallel by providing their own additive contribution to the circuit drive.

2.4.2 Inferring latents from sensory evidence

As a first concrete example of inference, we show a simplified version of a sensory perception task in which the input is only partially observed and the corresponding features need to be inferred given this incomplete observation (as would be the case in occlusion, for instance). In particular, we take the sensory observation to induce a linear constraint on the latent variable ($N = 2$ for easy visualization), $\mathbf{M}_s \operatorname{tr}(\mathbf{x} - \mathbf{x}_c) = 0$, parameterized by a unit norm linear operator $\mathbf{M}_s$ and offset $\mathbf{x}_c$, with Gaussian-distributed uncertainty around it (Fig.2.2B, right). Formally, the corresponding likelihood score is orthogonal to the constraint manifold and linearly rescaled by distance to it, $\mathbf{g}_t(\mathbf{x}_{t-1}, s) = \frac{1}{\sigma_s^2}\mathbf{M}_s\mathbf{M}_s \operatorname{tr}(\mathbf{x}_{t-1} - \mathbf{x}_c)$; parameter σ_s defines the degree of sensory uncertainty. As was the case for the prior, the DM implementation of this inference problem generates samples starting from the isotropic N-dimensional Gaussian, with the posterior reverse dynamics bringing the samples towards the posterior, here concentrated at the intersection of the two manifolds corresponding to the prior and likelihood, respectively (Fig. 2.2C, top). The neural posterior dynamics find a similar solution (Fig. 2.2C, bottom), but with initial conditions reflecting the likelihood. This is due to the sequential sampling process which has the same dynamics operating in both forward and reverse phases of the oscillation, and to the fact that the likelihood is gated oppositely to the prior (due to the $1 - \beta_t$ scaling, see Eq. 2.6).

When it comes to inference, the sampling dynamics of the neural circuit approximate DM posterior dynamics in two important ways. First, as for the prior, they introduce sequential effects due to the continuous use of the same operator across all phases of the cycle. Second, the computation of the likelihood is itself approximate, as $\mathbf{g}_t$ is evaluated at point $\mathbf{x}_{t-1}$ instead of the $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$ required in the original posterior flow field derivation (Sohl-Dickstein et al., 2015b; Dhariwal and Nichol, 2021). This switch was needed to account for temporal causality and the fact that the sensory area providing this input does not have access to local somatic currents $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$. In principle, both effects should diminish with a finer discretization of time, but we

also numerically quantified the effects of each approximation individually (Fig. 2.2E, top) and together (Fig. 2.2E, bottom). In particular, we collected samples generated according to either the DM reverse operator (using $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$) or the neural posterior sampler (using $\mathbf{x}_{t-1}$). We quantified the similarity of the resulting sample distributions as a function of β_t, using a naive KL divergence estimator (10000 samples each, with space discretized into 225 bins). We found that discrepancies between the distributions are largest at the start of the reverse process. In this regime, the noise and flow fields are large leading to substantial differences between $\mathbf{x}_{t-1}$ and $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$ (Fig. 2.2E, top). In principle, this might lead to catastrophic accumulation of errors and large sampling biases; however, the attractor dynamics prevent this from happening. The effects of the approximation reduce over the cycle, with the KL divergence approaching zero. We see largely the same effects when the neural likelihood approximation is used in conjunction with the neural prior approximation (same operator throughout; Fig. 2.2E, bottom). Overall, these numerical results suggest that, at least in the simple toy examples considered here, the effects of neural approximations on inference are minimal.

2.4.3 Inference with multiple sources of information

The main benefit of separating the neural substrates implementing the prior from the computation of the likelihood is that it makes inference flexible. Gating signals from different sources of evidence allows the model to reuse the same prior information across multiple tasks. Beyond the sensory example above, another important source of evidence that shapes perception/inference comes from contextual priors, reflecting expectations about which of the latent features are likely to occur (e.g. attention). To illustrate this idea, we introduce a discrete context random variable, c, whose value leads to different expectations about the latent variable $\mathbf{x}$, formally making the prior a mixture of 1D manifolds, with data uniformly distributed along each (Fig. 2.3A, top).[3] This additional complexity does not affect the neural dynamics sampling from this prior (Fig. 2.3A,

[3]The ambient space is 2D, for simplicity.

bottom), but allows us to demonstrate richer inference scenarios.

Sampling from the context conditioned posterior $\mathrm{p}(\mathbf{x}|c)$ proceeds very similarly to the sensory inference example above, but the drive to the circuit is given by a separate top-down feedforward sub-circuit which implements a different $\mathbf{g}_t(\mathbf{x}_{t-1}, c) = \nabla_{\mathbf{x}} \log p_\phi(c \mid \mathbf{x}_{t-1})|_{\mathbf{x}=\mathbf{x}_{t-1}}$. [4] The corresponding flow field and posterior samples are shown in (Fig. 2.3B), where we see the dynamics being strongly attracted to the region of the latent space corresponding to the cued context (the 'S' manifold). Finally, top-down contextual signals can be combined with the bottom-up sensory evidence (Fig. 2.3C) to jointly constrain the posterior distribution from which we draw samples (Fig. 2.3D). When multiple sources of evidence, $y^{(i)}$, are available, their net effect on the likelihood combines additively, $\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y}) = \sum_i \gamma_i g_t^{(i)}(\mathbf{x}_{t-1}, y^{(i)})$. This modular parametrization is at the root of the circuit's inferential flexibility, with individual components gated in or out depending on availability of information and current task demands (Womelsdorf et al., 2014; Kuchibhotla et al., 2017).

2.5 Neural signatures

One unique feature of the inferential process proposed here is how the phase of the circuit oscillation modulates the relative contribution of different inputs to the circuit. At one extreme ($\beta_t = 0$), the effect of the local attractor dynamics is weak and the corresponding prior unstructured, with neural activity dominated by the likelihood component. At the other extreme ($\beta_t = 1$), the effect of the likelihood on the dynamics is weak, and neural activity corresponds to samples from the posterior (Fig. 2.4A; see also Eq. 2.6). This dialing up and down of different sources of evidence opens the door for experimental validation of the model. Concretely testing such dynamics involves assessing patterns of population activity during perception, at different phases of the dominant circuit oscillation, as samples from the intermediate posterior. Samples from

[4] Note here that the direction of the conditioning is switched from the natural causal one, which means that this map can be learned from labeled examples by optimization (see Suppl. A.1 for derivation).

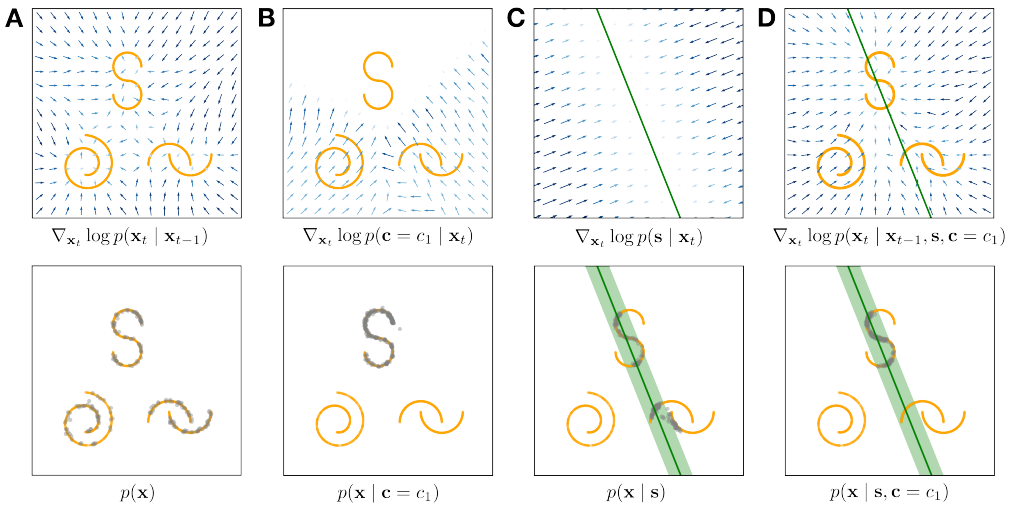


Figure 2.3: (A) Flow fields (top) and samples (bottom) from an example prior mixture distributions, with individual components uniform distributions along low dimensional nonlinear manifolds.**(B)** Likelihood flow field and samples from the corresponding context-conditioned posterior. **(C)** Likelihood flow field and samples from the corresponding sensory conditioned posterior. **(D)** Posterior flow field and samples from posterior conditioned on both sensory evidence and context.

the prior $p(\mathbf{x})$ can be obtained from spontaneous activity, in the absence of a sensory stimulus and outside a specific task context, while a reasonable measure of the effects of the likelihood is given by the activity at the lowest phase of the posterior dynamics, $p(\mathbf{x}_t|\mathbf{y}, \beta_t = 0)$. Given these statistics, the change in the contribution of the likelihood and the prior on this posterior can be estimated as the KL divergence between the activity of the circuit during inference, and these two reference distributions. The precise assessment of differences between potentially complex and structured distributions based on empirical measurements may seem impractical, but there is precedent in the literature for estimating KL divergences in population activity statistics using coarse discretizations of the space (e.g. temporal binning and thresholding of multiunit activity) (Berkes et al., 2011). For simplicity, here we draw 10^5 from each distribution, discretize the latent space into 400 bins and compute a naive estimate just based on the resulting histograms, although more sample efficient estimators are certainly available (Nemenman et al., 2001). Despite the coarseness of the measure, the intuition about the effects of β_t holds: population activity statistics become increasingly similar to the prior, and dissimilar from the likelihood, over the as-

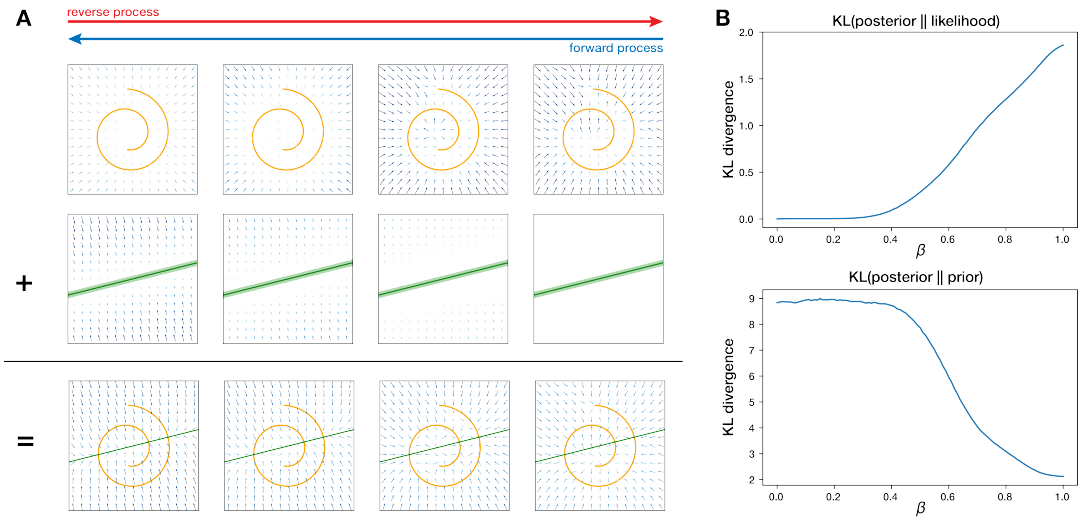


Figure 2.4: Neural signatures. **(A)** Relative contribution of the prior and likelihood scores during the neural forward and reverse processes. **(B)** Fluctuating similarity of neural activity to the likelihood (top) and the prior (bottom) over the ascending phase of the oscillation β_t. Distance between sampled distributions measured by a naive histogram-based KL divergence estimator.

cending phase of the oscillation (Fig. 2.4B). With a more precise understanding of the mechanics of gating different type of signals into the circuit, one could further envision causal manipulations that precisely bias the outcome in one or the other direction.

2.6 Discussion

Natural sensory inputs are often structured into complex nonlinear manifolds in high dimensions (Carlsson et al., 2008; Fefferman et al., 2016; Gorban and Tyukin, 2018; Brown et al., 2023). How does the brain represent such prior knowledge and uses it across tasks? Conceivably, one could solve this problem by learning a mapping between stimuli and responses *de novo* for each task. However, a more efficient representation would consolidate and leverage universally useful information across tasks. Here, we have proposed a circuit implementation of across-task inference in which a common prior is encoded in the form of a recurrent neural circuit with dendritic nonlinearities optimized for denoising. The stochastic dynamics of this circuit provide

samples from the prior at the peak of an ongoing local oscillation, while additional inputs (carrying sensory or contextual information) switch the dynamics to sampling from the corresponding posterior. This solution offers the first circuit model for reusing priors across tasks, as seen behaviorally (Houlsby et al., 2013), and has measurable neural signatures based on neural population recordings in animals.

Theories of probabilistic brain computation based on neural sampling suffer from two main practical challenges. The first is time, as estimating posterior expectations for decision making requires integrating samples over time (Rullán Buxó and Savin, 2021). This makes the speed of sampling a key constraint on plausible neural sampling dynamics (Hennequin et al., 2014; Savin and Denève, 2014; Aitchison and Lengyel, 2016; Masset et al., 2022). Second, although in principle sampling makes most sense for distributions with complex structure that cannot be well-represented parametrically (Fiser et al., 2010), in practice circuit models of sampling tend to be restricted to relatively simple distributions (Chen et al., 2023). The dendritic nonlinearities and DM-inspired oscillatory sampling schedule in our solution naturally overcome both limitations: by construction, the circuit is designed to sample from complex distributions, e.g. involving low dimensional nonlinear manifolds embedded in high dimensional spaces, and mixture distributions with low probability gaps between the components, which pose fundamental challenges for classic Markov Chain Monte Carlo approaches (Aitchison and Lengyel, 2016). Also by construction, DMs are designed to provide independent samples from the target distribution at the end of each reverse run, something that our neural circuit approximation seems to preserve to a large degree.

The specific role of oscillations in sampling resembles a previous model of sampling in the hippocampus (Savin et al., 2014) inspired by tempered transitions (Neal, 1996). In this view, β functions as an inverse temperature annealing the target distribution so as to reduce the gap (and thus increase the likelihood of transitions) between modes; this may inform the theoretical analysis of the sampler's properties. Additionally, theta oscillations naturally set the tempo in

the hippocampus, which may instruct the experimental validation of our proposal. The concrete approach for testing our experimental prediction was strongly influenced by (Berkes et al., 2011), who originally used KL divergences to measure the calibration of (sampling based) probabilistic models of early vision over development. While segregating the responses by oscillatory phase brings additional data and estimation challenges, the Berkes result, together with the evidence of hippocampal theta phase sampling from (Savin et al., 2014), provides reassurance that experimental validation of the model might be possible in practice.

While our proposed solution relies on nonlinearities in the dendrites as an implementation of the reverse operator, it is not clear that this is the only way to map DM operations to biology. In principle, the computations at each timestep could be carried out by a small feedforward subnetwork, recursively interacting between them. This version has the advantage of fewer restrictions on the architecture, but requires additional biological accommodations. In particular, for consistency noise would need to be distributed equally in all the neurons of the network (as opposed to just the last layer in DMs and our solution). Computationally, this solution remains viable (Suppl. A.5), although the architecture is arguably more contrived. Models of learning might better distinguish between these architectures in the future.

Here, we have purposefully abstracted away the nature of the features, so that the modeled circuit could be any processing stage in the cortical hierarchy, receiving evidence from one or several sensory domains bottom-up, and contextual top-down information from higher areas. This modular architecture opens the way for hierarchical probabilistic representations, by stacking such recurrent circuits with appropriate conditioning between them (Qiang et al., 2023). Such an architecture would also help limit the computational complexity of the neural nonlinearities involved in any stage: pixel-level DMs models of images are huge, with parameters on the order of billions (Ramesh et al., 2021; Saharia et al., 2022), but models of abstract image features are likely more tractable. Future work on hierarchical sampling neural circuits might lead not only to better accounts of brain representations but also to more compact machine learning models.

3 Learning disentangled representations via score-based variational autoencoders

3.1 Overview

We present the Score-based Autoencoder for Multiscale Inference (SAMI), a method for unsupervised representation learning that combines the theoretical frameworks of diffusion models and VAEs. By unifying their respective evidence lower bounds, SAMI formulates a principled objective that learns representations through score-based guidance of the underlying diffusion process. The resulting representations automatically capture meaningful structure in the data: it recovers ground truth generative factors in our synthetic dataset, learns factorized, semantic latent dimensions from complex natural images, and encodes video sequences into latent trajectories that are straighter than those of alternative encoders, despite training exclusively on static images. Furthermore, SAMI can extract useful representations from pre-trained diffusion models with minimal additional training. Finally, the explicitly probabilistic formulation provides new ways to identify semantically meaningful axes in the absence of supervised labels, and its mathematical exactness allows us to make formal statements about the nature of the learned representation. Overall, these results indicate that implicit structural information in diffusion models can be

made explicit and interpretable through synergistic combination with a variational autoencoder.

A version of this work has been published as a pre-print (Lyo et al., 2025).

3.2 Introduction

To evaluate the behavioral relevance of their observations intelligent agents must possess a notion of semantics. This requires decomposing complex sensory observations into abstract and semantically meaningful factors of variation, a capability known in machine learning as disentangling (Bengio et al., 2014). Disentangled representations promise several advantages, including improved interpretability, enhanced generalization, and stronger transfer capabilities across related tasks (Higgins et al., 2017). However, given that for both biological and artificial agents, the amount of unlabeled data vastly exceeds that of labeled data, an important consideration is how to learn disentangled representations without relying on explicit supervision.

Latent variational models provide a principled framework grounded in probabilistic generative models. The variational auto-encoder (VAE) has emerged as a particularly promising instantiation of this idea (Kingma and Welling, 2013; Rezende et al., 2014), which allows for amortized inference of latent structure and data generation through optimization of the evidence lower bound (ELBO). Successful variants of this approach have been used for compression (Ballé et al., 2017; Ballé et al., 2021), prediction (Tishby et al., 2000; Alemi et al., 2019; Sachdeva et al., 2020), and as models of neural activity in visual cortex (Csikor et al., 2022; Vafaii et al., 2023). Crucially, VAEs can be encouraged to exhibit disentangled representations via appropriate regularization, as with β-VAEs (Higgins et al., 2017; Alemi et al., 2019).

Despite these successes, there is a fundamental tension between reconstruction fidelity and disentanglement quality (Kumar et al., 2018; Sikka et al., 2019) in such models. Encouraging disentangled representations typically requires stronger regularization that limits the expressiveness of latent variables; this trade-off between compression and reconstruction precision is well

known as the rate-distortion principle. This limitation has motivated various VAE improvements, including enhanced posterior approximations through non-diagonal covariances or complex posterior distributions (Manduchi et al., 2023; Klushyn et al., 2019; Mathieu et al., 2019; Cheng et al., 2020; Rezende and Mohamed, 2015), additional latent constraints (Chen et al., 2019; Zhao et al., 2018), and more flexible latent priors (Klushyn et al., 2019; Wehenkel and Louppe, 2021). Notably, these efforts have largely focused on the inference mechanism rather than the generative model itself, despite the role the diagonal posterior covariance plays in encouraging disentanglement (Rolínek et al., 2019; Dai et al., 2018) and the importance of the generative model in determining the nature of learned representations (Cremer et al., 2018). Moreover, mathematical analyses of VAE optimality typically assume that the data lies on a manifold with fixed intrinsic dimensionality Dai and Wipf (2018), but such an assumption is problematic for natural images, which instead appear to lie in a multi-scale manifold whose intrinsic dimensionality is highly dependent on the signal to noise ratio (SNR) of the image (Guth et al., 2025; Sclocchi et al., 2025). This suggests that to learn good disentangled representations of complex, multi-scale data, we need generative models with the appropriate inductive biases (Locatello et al., 2019).

A possible avenue for building more expressive, multi-scale generative models lies in diffusion models (Sohl-Dickstein et al., 2015a; Ho et al., 2020). These generative models produce high-quality samples in many domains by learning to approximate the data distribution through a denoising objective, and seemingly captures features at multiple spatial scales (Sclocchi et al., 2025). However, their lack of an explicit latent representation poses challenges for tasks requiring embeddings, such as disentangled representation learning. Additionally, guiding the diffusion sampling process to generate samples with specific attributes or classes remains an open research question (Fuest et al., 2024). While techniques like classifier-guided diffusion (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2021) lay the groundwork for sampling from conditional densities in the presence of explicit class labels, achieving robust and flexible conditioning without compromising sample quality or diversity is not fully resolved

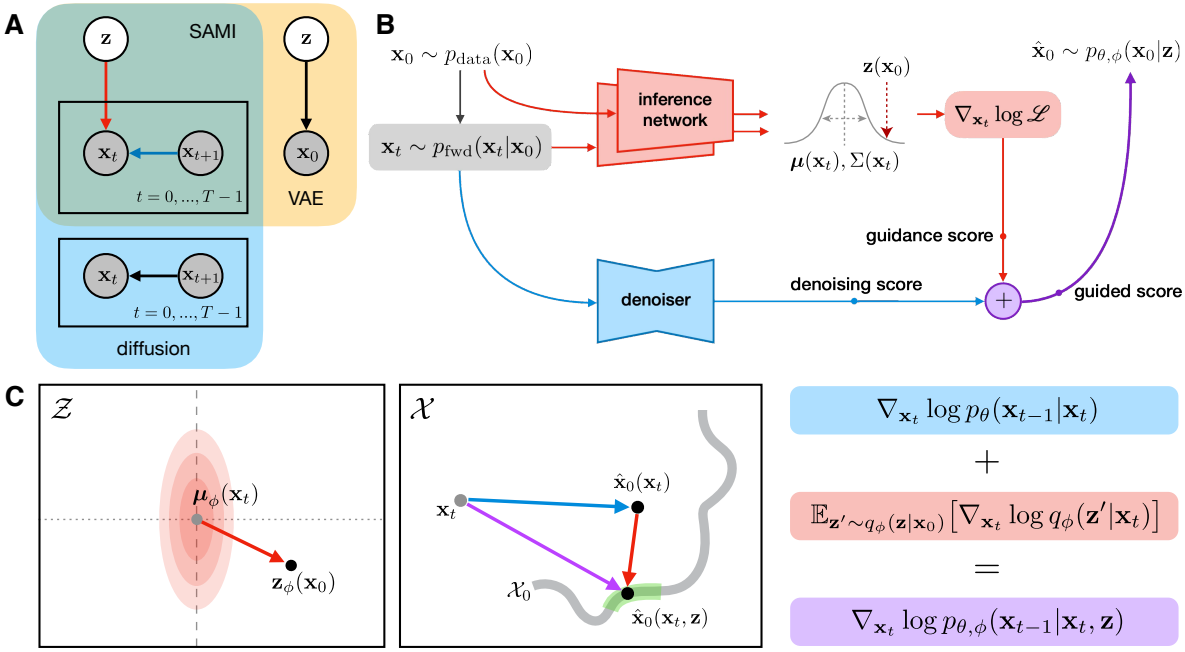
(Chidambaram et al., 2024; Sadat et al., 2024; Kaiser et al., 2024; Ifriqi et al., 2025). Critically, few studies explore the learning of latents that enable diffusion models to effectively navigate the full data manifold *without auxiliary information*.

Here, we introduce a Score-based Autoencoder for Multiscale Inference (SAMI) that combines the strengths of variational autoencoders and diffusion models to achieve unsupervised learning of structured, interpretable latent representations while maintaining high-quality sample generation. Our key innovation lies in employing conditional diffusion as the generative component of a VAE, coupled with a precise mathematical formalization of the objective via the ELBO that enables reuse of the inference network for conditioning.

Leveraging the mathematical exactness of our model, we prove that the diffusion prior on the data space allows us to keep a simple latent posterior approximation that imposes strong factorization constraints on the latent code such that multi-scale generative features are disentangled. We also prove that learning latent representations of images that are coherent across noise levels helps to unify and smooth the latent representation.

We find empirical support for these results. When trained on image data, our framework yields semantically meaningful latent axes and produces straighter trajectories when encoding natural video sequences, all without compromising generative performance. Moreover, a distinctive feature of our approach is that the posterior structure enables identification of semantically meaningful axes without reliance on supervised labels. From the diffusion perspective, incorporating an inference network with an explicit representational space facilitates unsupervised discovery of factorized and interpretable latents that can be used to control sample generation. Notably, this same methodology can be applied to extract semantic latent representations from pretrained diffusion models.

Figure 3.1: **A)** Graphical model of SAMI, contrasted with those of standard VAEs and unconditioned diffusion models. **B)** Schematic of conditional sampling procedure. **C)** Schematic illustration of how movement in the latent space results in guidance of the denoiser score towards regions of the clean image manifold that are semantically similar to the original image. For visualization, both latent and image spaces are depicted as two dimensional.



3.3 Methods

The task of extracting latent representations from data can be formalized as a graphical model in which latent random variables, $\mathbf{z} \in Z$, give rise to observations $\mathbf{x} \in X$ (Fig. 3.1A, yellow box). Representational learning then corresponds to maximizing the marginal likelihood of the observed data under the model, $\prod_k \int p_\theta(\mathbf{x}_k, \mathbf{z}) d\mathbf{z}$, while inference requires calculating the posterior distribution $p(\mathbf{z}|\mathbf{x})$ via Bayes' rule. Unfortunately, directly computing the log likelihood and the posterior distribution in closed form is usually infeasible for expressive models.

Variational autoencoders (VAEs) address this difficulty by approximating the posterior with a simpler parametric form, whose parameters are data-specific and obtained via optimization (Kingma and Welling, 2013; Rezende et al., 2014). The parameters of both the amortized posterior and the generative model are learned via minimization of the *evidence lower bound* (ELBO), a

tractable lower bound on the model marginal likelihood:

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \leq \log p_\theta(\mathbf{x}), \tag{3.1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior (usually computed by a neural network with weights ϕ), the density $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ is the generative model, with prior $p(\mathbf{z})$ given by a simple parametric distribution (e.g., a standard Gaussian), and likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ computed by a separate neural network with weights θ. In practice, the limitations imposed on the likelihood by this parameterization often lead to poor quality of generated samples (Burda et al., 2016; Sønderby et al., 2016; Rezende and Viola, 2018).

In contrast, diffusion models are highly expressive generative models that can estimate and sample from complex densities. Diffusion models such as denoising diffusion probabilistic models (DDPMs) form an implicit prior over data by learning to denoise noisy versions of clean data (Sohl-Dickstein et al., 2015a; Kadkhodaie and Simoncelli, 2020a; Ho et al., 2020). Noisy data are generated by a fixed *forward* operator that is assumed to be additive Gaussian, $\mathbf{x}_t \sim p_{\text{fwd}}(\mathbf{x}_t|\mathbf{x}_0)$, with $t \in [0, 1, ..., T]$ specifying the noise variance. This denoising objective can be written as the mean squared error between the model's estimate of the noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ and the true noise present in the noisy image $\boldsymbol{\epsilon}$, which is typically assumed to be isotropic Gaussian. Minimizing this error has also been shown to be equivalent to maximizing the ELBO (up to a noise-dependent scalar weighting λ_t) for a joint data distribution given by $p(\mathbf{x}_{0:T})$ and a joint "posterior" $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ that captures the distribution over all noising paths (Kingma and Gao, 2023):

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I),\, t \sim [0,T]} \left[\lambda_t \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \right] = -\mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right]. \tag{3.2}$$

It is well known that the minimum mean squared error (MMSE) estimate is given by the posterior mean, and this in turn can be related to the score function via Miyasawa's/Tweedie's formula (Robbins, 1956; Miyasawa et al., 1961). Under a Gaussian variance-preserving noise distribution

given by $p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$, where $\{\bar{\alpha}_t\}$ defines the variance schedule, the MMSE estimate of the noise is given as

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t) = \underset{\boldsymbol{\epsilon}_\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 = \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}[\boldsymbol{\epsilon}] = -\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t). \tag{3.3}$$

Once trained, generation of samples relies on an iterative partial denoising process that samples $\mathbf{x}_t \sim p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ as $t = T-1, ..., 0$ (Fig. 3.1A, blue box). Examples from the training set can be thought of as samples from a complex probability distribution that resembles a low-dimensional, nonlinear manifold lying within the high-dimensional data space (Ho et al., 2020; Kadkhodaie and Simoncelli, 2020a), and the iterative denoising process effectively learns to project noisy samples back onto this manifold. To generate samples with desired characteristics, the denoiser can be guided towards specific regions of the manifold through classifier-based or embedding-based guidance signals (Dhariwal and Nichol, 2021). However, unlike variational autoencoders or GANs (Goodfellow et al., 2014), diffusion models do not specify or infer an explicit latent representation.

By using conditional diffusion models as part of the VAE generative process, we can overcome the limitations of both models. The key idea is to learn the latent variables that will best guide the diffusion process towards the observations. Concretely, we learn a latent representation by augmenting the Markov chain $\mathbf{x}_{0:T}$ of the diffusion model with a latent variable $\mathbf{z}$ that guides the denoising process (Fig. 3.1A, green box). This results in a joint distribution $p(\mathbf{x}_{0:T}, \mathbf{z})$ and approximate posterior $q(\mathbf{x}_{1:T}, \mathbf{z}|\mathbf{x}_0)$ that can be factorized as

$$\begin{aligned} p(\mathbf{x}_{0:T}, \mathbf{z}) &= p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})p(\mathbf{z}) \\ q(\mathbf{x}_{1:T}, \mathbf{z}|\mathbf{x}_0) &= q_{\mathbf{x}}(\mathbf{x}_{1:T}|\mathbf{x}_0)q_\phi(\mathbf{z}|\mathbf{x}_0), \end{aligned}$$

where the latter holds because the noisy samples in $\mathbf{x}_{1:T}$ are independent of $\mathbf{z}$ when conditioned

on $\mathbf{x}_0$. Here, $q_\phi(\mathbf{z}|\mathbf{x}_0)$ is the inference network, used to infer latent $\mathbf{z}$ from observations $\mathbf{x}_0$, while $q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0)$ is the (known) "posterior" that defines the conditional forward process of the diffusion model. As is common for VAEs, we assume that the model prior is an isotropic Gaussian, $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Incorporating these into the ELBO yields a joint objective function of the form:

$$\begin{aligned}
\mathcal{L}_{\text{SAMI}} &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\mathbb{E}_{q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})}\left[\log \frac{p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})p(\mathbf{z})}{q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0)q_\phi(\mathbf{z}|\mathbf{x}_0)}\right] \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\mathbb{E}_{q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{z})}\left[\log \frac{p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})}{q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\left[\log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x}_0)}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0)}\left[D_{\text{KL}}\left(q_\mathbf{x}(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})\right)\right] + D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p(\mathbf{z})\right), \qquad (3.4)
\end{aligned}$$

where the second term is equivalent to the standard regularization term in the VAE objective that encourages the expected approximate posterior distribution to be close to the latent prior. Inputs $\mathbf{x}_0$ to the inference network return the corresponding mean and covariance $\boldsymbol{\mu}_\phi(\mathbf{x}_0)$ and $\Sigma_\phi(\mathbf{x}_0)$ of the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}_0)$, which are then used to analytically calculate the KL divergence term.

3.3.1 Implementing the generative model as conditional diffusion

The first term in SAMI's objective (Eq. 3.4) encourages the latent conditioned generative model $p_{\theta,\phi}(\mathbf{x}_{0:T}|\mathbf{z})$ to be close to the distribution over forward noising paths $q_\mathbf{x}$. Following a similar derivation to Ho et al. (2020), this term can be written as the expected mean squared error between the true and estimated noise across noise levels t:

$$\mathcal{L}_{\text{SAMI}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0),\boldsymbol{\epsilon},t}\left[\lambda_t\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, t, \mathbf{z})\|^2\right] + D_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}_0)\|p(\mathbf{z})\right), \qquad (3.5)$$

where $\boldsymbol{\epsilon}$ is the true noise, $\boldsymbol{\epsilon}_{\theta,\phi}$ is the function estimating it, and λ_t is a scalar hyperparameter that depends on t. For the first term in the objective, we can use Miyasawa's formula to relate the conditional MMSE estimate to the guidance score (full derivation in Appendix B.1.1). Using the

same variance-preserving noise distribution as in Eq. (3.3), we see that

$$\hat{\boldsymbol{\epsilon}}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) = \underset{\boldsymbol{\epsilon}_{\theta,\phi}}{\operatorname{argmin}} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, t, \mathbf{z})\|^2 = -\sqrt{1-\bar{\alpha}_t}\, \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{z}). \tag{3.6}$$

Critically, Bayes' rule can be used to express the guidance score as the sum of two terms, an unconditional denoiser score and a guidance score:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{z}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{z}|\mathbf{x}_t). \tag{3.7}$$

Plugging this back into Eq. (3.6), we see that the MMSE solution can be expressed as

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) &= -\sqrt{1-\bar{\alpha}_t}\big(\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{z}|\mathbf{x}_t)\big) \\ &= \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t) - \gamma_t\, \mathbf{g}_t(\mathbf{x}_t, \mathbf{z}), \end{aligned} \tag{3.8}$$

where $\hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t) = -\sqrt{1-\bar{\alpha}_t}\, \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ is the MMSE solution associated with an unconditional diffusion model (Eq. 3.3), $\mathbf{g}_t(\mathbf{x}_t, \mathbf{z}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{z}|\mathbf{x}_t)$ is the guidance score, and $\gamma_t = \sqrt{1-\bar{\alpha}_t}$ is a noise level-dependent scalar term that weights the two terms appropriately.

3.3.2 Computing the guidance score

From the perspective of the unconditional diffusion model, the VAE posterior has a dual interpretation as a likelihood function $\mathscr{L}_{\phi}(\mathbf{x}_t; \mathbf{z})$, which is a function of the noisy image $\mathbf{x}_t$ over the latent domain $\mathcal{Z}$. The guidance score is the score of the likelihood evaluated at a particular latent value $Z = \mathbf{z}$, which in our case we take to be the latent corresponding to the clean image, since our goal is to reconstruct the original observation. To compute this, we take samples of the clean image latent $\mathbf{z}(\mathbf{x}_0) \sim q_{\phi}(\mathbf{z}|\mathbf{x}_0)$ in accordance with the expectation in Eq. (3.5) and compute its log likelihood, $\log \mathscr{L}_{\phi}(X_t = \mathbf{x}_t; Z = \mathbf{z}(\mathbf{x}_0))$. Taking the derivative of this scalar quantity with respect

to the noisy image $\mathbf{x}_t$ (via auto-diff) then gives the approximate guidance score

$$\mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z}) = \mathbb{E}_{\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x}_0)} \left[\nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}'|\mathbf{x}_t) \right] \tag{3.9}$$

The reuse of the inference network to compute the log likelihood removes the need for a separate neural network to map the representation $\mathbf{z}$ into a guiding signal, and ties the representations of clean and noisy versions of an image together. Plugging this back into Eq.(3.8), and this in turn into Eq. (3.5), we see that the SAMI objective can be written as the influence of two separate networks in our model, the unconditional denoiser parameterized by θ, and the inference network parameterized by ϕ:

$$\mathcal{L}_{\beta\text{-SAMI}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_0), \boldsymbol{\epsilon}_t, t} \left[\lambda_t \| \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \gamma_t \, \mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z}) \|^2 \right] + \beta \, D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}_0) \| p(\mathbf{z})\big), \tag{3.10}$$

where β is a hyperparameter that balances the effective contribution of the KL regularization on the latent space. This is akin to the Lagrange multiplier that controls the reconstruction and regularization terms in β-VAEs (Higgins et al., 2017). Note that if we rewrite this derivation in terms of $\mathbf{x}_0$ estimates instead of $\boldsymbol{\epsilon}$ estimates, we can show, using the relation $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, that $\hat{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{z}) = \hat{\mathbf{x}}_0(\mathbf{x}_t) + {}^{1-\bar{\alpha}_t}/_{\sqrt{\bar{\alpha}_t}} \mathbf{g}_t(\mathbf{x}_t, \mathbf{z})$, which reflects the additive relation between the denoising score and guidance score in Fig. 3.1C, $\mathcal{X}$ space. The expression that details the effect of the guidance score on the DDPM reverse transition operator is given in Appendix B.1.2.

3.3.3 Inference and guidance

Once SAMI is trained using Eq. (3.10), inference follows the general VAE prescription where latent samples are drawn from the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}_0)$ via the reparameterization trick. Generating image samples that resemble the given "guidance" datapoint $\mathbf{x}_0$ involves conditioning the diffusion process on draws from the posterior distribution. As in DDPMs, we start the condi-

tional generation process with a sample from an isotropic gaussian $\mathbf{x}_t \sim \mathcal{N}(0, I)$. At each noise level t, we estimate the noise present in the noisy image via the denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, and compute the guidance score $\mathbf{g}_{t,\phi}$ using Eq. (3.9). Note that while the clean image posterior remains the same at every noise level, the guidance score changes because we evaluate the clean image latent under the noisy image posterior, which changes with the noise level. The resulting guidance score biases the reverse process according to Eq. (3.8), such that as the noise level goes to 0, we arrive at a sample from the conditional distribution $\hat{\mathbf{x}}_0 \sim p_{\theta,\phi}(\mathbf{x}_0|\mathbf{z})$. Algorithms are provided in Appendix B.1.3.

3.3.4 Geometric intuition

In the image space, conditional generation update steps (purple arrow in Fig. 3.1C, $\mathcal{X}$ space) can be thought of as a combination of two forces: starting from a noisy image $\mathbf{x}_t$, attractor dynamics push the state towards the image manifold (blue arrow) while the latent-based guidance selects specific regions of the manifold (red arrow). The addition of these two forces direct the network towards regions of the data manifold that are close to the target $\mathbf{x}_0$ (green shade).

In the latent space (Fig. 3.1C, $\mathcal{Z}$), similarity between the current state $\mathbf{x}_t$ and the conditioning image $\mathbf{x}_0$ (as defined by the likelihood) is the Mahalanobis distance between the mean estimate of the network state in latent space $\boldsymbol{\mu}_\phi(\mathbf{x}_t)$ and the latent corresponding to the conditioning image $\mathbf{z}'(\mathbf{x}_0)$, computed under a metric that is given by the inverse of the covariance Σ_ϕ^{-1} (concentric ellipsoids). Conditional generation can thus be cast as a constrained optimization process, where the network minimizes the latent Mahalanobis distance subject to the constraints posed by the geometry of the latent space and the image manifold (for further details, see Appendix B.1.4).

During the learning process, the network learns to map noisy version of the same image close together in $\mathcal{Z}$ space, while accounting for uncertainty introduced by the noise in the associated posterior covariance, $\Sigma_\phi(\mathbf{x}_t)$. This effectively means learning which directions in latent space correspond to noise-induced variations versus meaningful semantic content.

3.4 Results

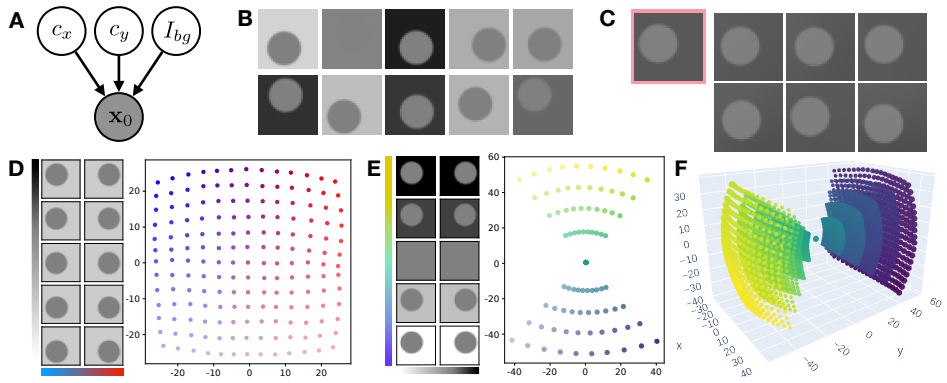


Figure 3.2: Disks dataset. A) Graphical model for the disks dataset, with coordinates of disk center (c_x, c_y) and background intensity I_b as latents. **B)** Random draws from the ground truth generative process. **C)** Samples drawn from the model, conditioned on leftmost image. **D)** Posterior means for a grid of test c_x and c_y ground truth positions, fixed I_b. **E)** Same as D, but c_y fixed during interpolation of the other two factors. **F)** Same as D, for all three factors.

Synthetic disks dataset. We first investigated whether SAMI is able to learn a semantically meaningful latent space by training the network on a synthetic dataset where the ground truth latent factors and generative structure are known, and the number of latent dimensions matches that of the latent factors. The dataset is a simplified version of that used by Kadkhodaie et al. (2024), comprised of images of circles ("disks") of a fixed radius and intensity on a blank background (Fig. 3.2A, B). The images are fully determined by three independent generative factors: the coordinates of the disk center, c_x, c_y, and the intensity of the background, I_{bg}. All three generative factors are uniformly distributed, and the foreground intensity of the disks is held to a constant value of 0.5.

As a first measure of whether SAMI learns a useful representation, we tested the model's ability to faithfully reconstruct images when conditioned on a given guidance image (Fig. 3.2C; see Appendix B.1.5 for architecture details). The conditioned image generation is sensible: it preserves background intensity (which can be accurately estimated from the conditioning image), with small variations in the location of the disc reflecting some posterior uncertainty in that di-

mension. To quantify the latent variable's contribution in "explaining away" the observation, we compared the variability in the test set and in the conditionally generated images by calculating the mean squared distance between the samples and their respective empirical mean. Here the mean squared distance serves as a proxy for the entropy in each distribution. We found a substantial decrease in the variability from 5.6×10^{-1} to 4×10^{-3} when the generation is conditioned on the latent variable, indicating that the representation captures most of the variability in the dataset.

To directly probe the semantic structure of the learned representation, we systematically varied two of the generative factors and extracted the corresponding posterior means provided by the inference network. When varying the x, y coordinates of the disk, we found an approximately Cartesian representation in the latent space (Fig. 3.2D), while varying the background intensity and the x coordinate resulted in an approximately polar representation (Fig. 3.2E). The singular point in the center corresponds to an image where the background intensity exactly matches the foreground intensity, so the x, y coordinates are unspecified. The shrinking of the spatial encoding with contrast is a reflection of increasing uncertainty in position pulling the posterior mean towards the prior. We see this more clearly in the 3d space (Fig. 3.2F), in which we linearly interpolated along all three generative factors. Importantly, the orthogonality of the semantic axes directly reflects the independence of the three generative factors. The constraints imposed by the KL regularization term suffice for the model to learned a factorized, semantically meaningful representation of the data.

3.4.1 CelebA dataset

Next, we applied our method to CelebA, a dataset of diverse, high-quality images of human faces with structured attributes (e.g. facial features, expressions, and accessories) that are qualitatively discernible by human observers, which provides sematic structure that we can post-hoc evaluate in the learned representations. Following Ho et al. (2020), we used a standard UNet ar-

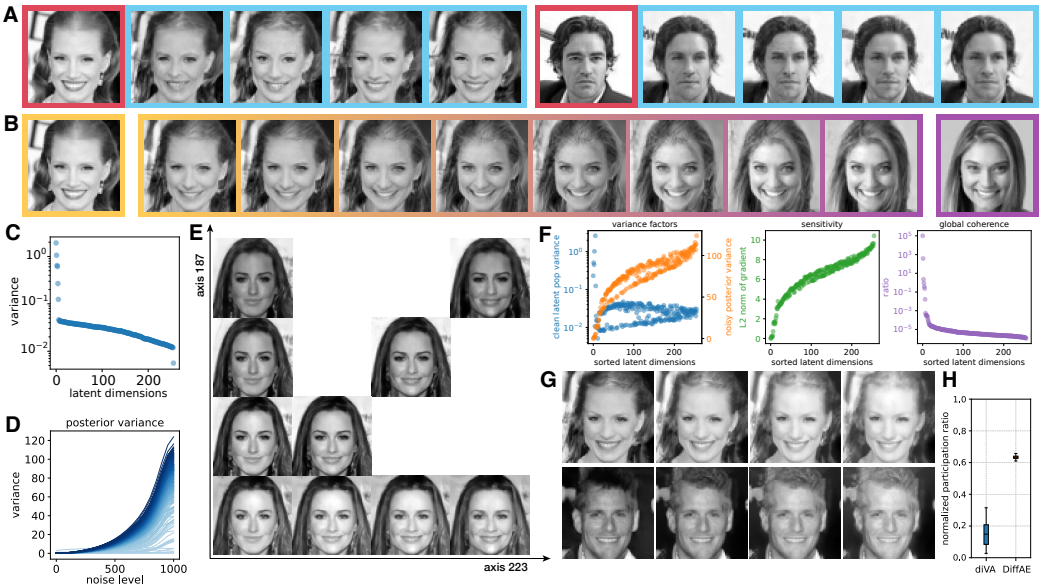


Figure 3.3: CelebA. A) Conditional image generation; conditioning image – red, samples – blue. **B)** Linear traversal between latents of leftmost (yellow) and rightmost (purple) images. **C)** The spectrum of the posterior covariance for one example image from the test set. **D)** The posterior variance as a function of noise level; color intensity marks the ordering of the axes at the largest noise level. **E)** Moving along the linear combination of two latent axes with identified semantics. **F)** Geometric characterization of latent representation. Left: across-test data variability in posterior mean (blue) and posterior variance of a noisy image (orange) for each latent dimension. Middle: norm of the sensitivity of posterior variance to input changes. Right: Global coherence of each axis. The x axes of all three plots are sorted by global coherence. **G)** Example transformation of two different images along a latent axis identified as global. **H)** Weight sparsity of binary classifiers compares how closely the semantics of latent axes align to supervised labels for SAMI vs. DiffAE.

chitecture for the denoiser, but without self-attention layers, so as to focus on the algorithmic effects. The noise level was encoded using a sinusoidal position embedding, provided as an additional input channel (Nichol and Dhariwal, 2021). For the inference network, we used a half-UNet architecture that retains the Down and Mid, but not the Up blocks. The mean and covariance of the posterior are linearly decoded from this network. Unlike Mittal et al. (2023), our inference network does not receive noise level information, which we found to be unnecessary for learning a good representation. As is common in VAEs, we restricted the posterior covariance to be diagonal, $\Sigma_\phi = \text{diag}(\boldsymbol{\sigma}_\phi^2)$. Additional architectural details are provided in Appendix B.1.5.

We trained the joint model from scratch using 60,000 images from the CelebA dataset, and evaluated the learned representation by assessing conditioned reconstruction quality on images

from the test set. If the learned latents are useful, they should encode the pertinent features in the original guiding image that, when used to condition the generation process, reduces the MSE between itself and the generated images. Note that proper conditional generation is not trivial: since guidance scores are added to the denoiser estimates in pixel space, conditioning can easily worsen generation if the latents are non-informative or misaligned with the image prior.

The generated images possess many of the same features as the guiding image, including skin tone, light position, pose, and location and shape of facial features such as the eyes, hairline, and mouth (Fig. 3.3A). As with the disks, we observe variability in conditionally generated image features, stemming from uncertainty in the posterior and the stochastic nature of the diffusion process. Again, we measured the effect of the latent conditioning and found that the variability is reduced from 6.62×10^{-2} to 2.04×10^{-3} (full histogram in Appendix B.1.6). To measure sample fidelity and diversity, we computed the FID based on 10,000 images (Heusel et al., 2017). SAMI's score of 16.25 beats other diffusion-based representation learning models such as DiffAE and InfoDiffusion (Preechakul et al., 2022; Wang et al., 2023) (see Appendix B.1.7 for details).

VAEs exhibit so-called *latent holes*, i.e. discontinuities in the latent space, that emerge from a mismatch between the approximate posterior and latent prior. These holes reduce the capacity of the latent space and negatively affect generative performance and log likelihoods (Li et al., 2021; Xu et al., 2020; Falorsi et al., 2018). While past work account for these holes explicitly through regularization or by enforcing compactness (Zhang et al., 2022a; Glazunov and Zarras, 2023), we prove that training the encoder to satisfy Eq. 3.10 using both clean images and their noisy counterparts implicitly smooths the latent space by minimizing the expected Frobenius norm of the encoder Hessian taken over all noise levels (see Appendix B.2.1 for full proof). To investigate the smoothness of the latent embedding we used latent traversal, linearly interpolating between the latent representations corresponding to two images from the test set (Fig. 3.3B). At regular intervals along this interpolant, we generated sample images conditioned on the corresponding latent. Given two endpoint images (enclosed in yellow and purple boxes in Fig. 3.3B), the in-

termediate images display semantic characteristics that change smoothly between those of the endpoint images, indicating that the learned latent space is smooth and continuous.

Factorized semantics. The results on the disk images suggest SAMI latents can capture independent generative factors in the data. Indeed, by leveraging the additive nature of the score-based guidance, we can show that a disentangled representation of semantic features naturally emerges from this setup. Specifically, under mild assumptions that semantic features are distributed hierarchically with unique characteristic spatial scales, we prove that a diffusion-based generative model optimized to minimize both reconstruction error and the Kullback-Leibler (KL) divergence between a diagonal covariance posterior and an isotropic Gaussian prior yields this disentangled representation (see Appendix **??** for full proof).

Empirically, the posterior variance for any given image exhibits a few dominant modes alongside many smaller ones (Fig. 3.3C), indicating that, within the local neighborhood of the mean latent representation, certain latent axes have significantly higher uncertainty. Our explicitly probabilistic approach point to the structure of the posterior covariance as a natural lens into the interpretability of individual latent axes, and accordingly, adding noise to the image increases variance across these axes (Fig. 3.3D), suggesting that posterior variance reflects the inference network's uncertainty about the corresponding features.

We hypothesize that axes with low posterior variance at low noise levels encode robust semantic features, while the rate of variance increase with noise level indicates the spatial scale of the encoded features. Specifically, large-scale features, still prominent at relatively high noise levels, should correspond to axes with low uncertainty even under significant noise, whereas small-scale features will exhibit overall higher uncertainty.

To test this, we perturbed a clean test image's latent representation along axes with low posterior variance. The resulting images showed consistent changes in specific semantic attributes, suggesting that the model factorizes representations semantically (Fig. 3.3E, smiling horizontally, forward lighting vertically). Interpolating along a linear combination of two axes produced im-

ages reflecting both attributes, suggesting a combinatorial code (Fig. 3.3E, diagonal). Meanwhile, axes that displayed large changes in posterior variance tended to correspond to finer scale features, such as movements of the eyes and mouth. Finally, we measured disentanglement using Total AUROC Difference (TAD) (Yeats et al., 2022), which measures how well a change in one independent data generating factor in the data corresponds to a change in one disentangled latent feature; a disentangled code should have a one-to-one mapping. For each binary attribute ground truth label (such as those that accompany the CelebA dataset), the measure takes the AUROC score of each latent variable in detecting the attribute and takes the difference in the top two values; the final score is the sum over all ground truth attributes. Perfect disentanglement corresponds to a score of 1, while a fully entangled representation has a score of 0. Finally, an attribute is considered "captured" if it has a corresponding latent representation with an AUROC score of at least 0.75.

When trained on CelebA, SAMI attains a TAD score of 0.583 and captures 3 attributes, which significantly outperforms most diffusion-based methods and VAE-based baselines, and has comparable scores to the state-of-the-art EncDiff model (TAD of 0.638 but no reported attributes captured) (Yang et al., 2024). We provide full comparisons in Appendix B.1.7. These findings reveal that SAMI's latent space is structured to disentangle semantic features along specific axes, offering a foundation for interpretable and controllable generative modeling.

How global is the mapping between latent axes and semantic meaning? While perturbation along individual axes is interpretable in the neighborhood of the conditioning image, we found that in general this mapping is not preserved across the full latent space. To identify the axes whose semantic meaning is preserved globally, we derived a formula that provides sufficient conditions for global axes under additive transformations by assessing the $\mathbf{z}$-independence of the linear local approximation of the inference network (see Appendix B.1.8).

When applying this metric (Fig. 3.3F, right) we find the axes with the largest measure to be "junk" dimensions that encode the pixel-level noise, whose semantic mapping is in effect global.

However, most of the identified global axes are interpretable, reflecting e.g. global lightning effects (Fig. 3.3G).

Alignment with explicit labels. A separate way to assess the semantic structure of the representation is by measuring the alignment between latent axes and recognizable semantic labels. To this end, we trained separate logistic regression classifiers on the learned representation to predict each of the 40 binary semantic labels. If the axes are well aligned with the supervised attributes, the weights of the classifier should be sparse, as only a small subset of the axes need to be recruited to form the decision boundary. We used the normalized participation ratio (PR) of each classifier's coefficients as a measure of sparsity; this is 1 if only one latent axis is used for a decision, and it becomes the square root of the dimensionality of the latent space if all axes are used equally. The normalized PR is generally low for SAMI, in particular when compared to state of the art alternative DiffAE (Fig 3.3H), suggesting that our approach is superior in factorizing the latent space into useful semantic axes.

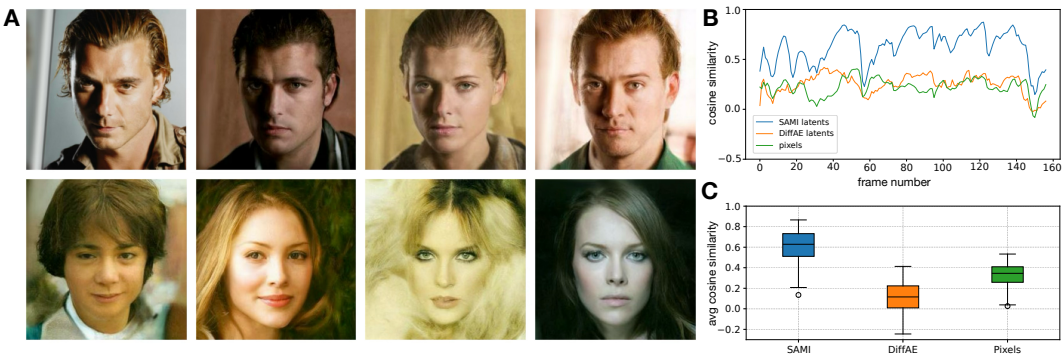


Figure 3.4: Unsupervised extraction of latent representation from a pre-trained diffusion model. **A)** Top: Samples from the trained model, conditioned on leftmost image (green). Bottom: samples drawn from the unconditional diffusion model. **B)** Straightness of latent trajectories over the course of one naturalistic video from CelebV. **C)** Average straightness of latent encoding (over 50 naturalistic videos) for SAMI, compared to trajectories in pixel space and latent trajectories of DiffAE.

3.4.2 Feature extraction for pretrained diffusion models

Since the denoising and guidance terms are parameterized by separate neural networks, one might sensibly wonder if training them jointly is strictly necessary or if the inference network

could be trained separately on top of a pre-trained denoiser. To test this, we used a DDPM from the Huggingface repository pretrained on CelebA-HQ-256 (Ho et al., 2020; Karras et al., 2018) and then trained the inference network to guide the responses of the denoiser towards the corresponding clean image using the same objective as before. We found that samples drawn from the conditioned distribution exhibit similar semantic characteristics as the guiding image, such as lighting direction, facial expression, and hairline, while remaining a naturalistic image with high frequency details (Fig. 3.4). As before, the conditioning reduces variability compared to samples from the unconditional model (Fig. 3.4A, second row), but there is still appreciable diversity among the samples. Overall, separate training of the conditioning network seems to generally preserve the properties seen with joint training with a substantial reduction in training effort.

Encoding of video trajectories. One property that makes for good latent representations is their predictability over time, measured by the ability to represent temporal sequences in a manner that supports linear extrapolation (Niu et al., 2024). Since movement along individual latent axes corresponds to naturalistic transformations in images (Fig. 3.3E), we hypothesized that SAMI may map natural movements in naturalistic videos, such as a mouth opening and closing, to relatively straight trajectories in latent space. To test this, we encoded frames from a naturalistic facial attributes video dataset (Yu et al., 2023) into latent representations using the SAMI model for CelebA-HQ. Following Hénaff et al. (2019), we quantified the "straightness" of temporal sequences in both latent and pixel domains by computing discrete curvature, defined as the cosine similarity between vectors connecting consecutive pairs of frames in the sequence: $s = \mathbf{d}_t \cdot \mathbf{d}_{t+1} / (\|\mathbf{d}_t\| \|\mathbf{d}_{t+1}\|)$, where $\mathbf{d}_t = \mathbf{f}_{t+1} - \mathbf{f}_t$ represents the difference between consecutive frames, or latent posterior means. Despite being trained solely on static images, our model encodes frames of these videos such that their trajectories in the 512-dimensional latent space exhibit near-unity cosine similarity across large portions of a given video (Fig. 3.4B). Compared to pixel-domain trajectories and those from other diffusion-based representation learning models (Preechakul et al., 2022), SAMI produces significantly straighter latent trajectories. This effect holds consistently

across 50 randomly sampled videos from the dataset (Fig. 3.4C). Thus, by leveraging its semantically factorized representation, SAMI naturally straightens dynamic patterns without temporal supervision.

3.5 Discussion

While recent work has demonstrated the promise of unsupervised representation learning with diffusion models (Fuest et al., 2024), past approaches typically rely on (often deterministic) encoders that approximate the log-likelihood (Preechakul et al., 2022; Mittal et al., 2023; Yang et al., 2023a; Kim et al., 2024), leverage additional information beyond the base dataset (Hudson et al., 2024), or use objectives that deviate from the ELBO by adding specialized regularization terms, e.g. to encourage mutual information (Wang et al., 2023), sparsity (Mittal et al., 2023), or to bound entropy (Yang et al., 2023b). See Appendix B.1.9 for detailed comparison to prior methods. SAMI diverges from this paradigm by optimizing an exact ELBO that encodes both clean and noisy images into probabilistic latents.

This dual encoding strategy introduces a novel form of implicit regularization: by requiring the inference network to predict the log-likelihood of clean image features while conditioned on noisy observations, SAMI learns to effectively separate signal from noise and calibrate its posterior variance to capture the irreducible uncertainty inherent in the data. The probabilistic nature of these paired observations constrains the latent space geometry, promoting representations where semantically similar images cluster together. Our empirical analysis reveals that this approach provably and empirically mitigates the "latent holes" problem that commonly afflicts variational autoencoders (Rezende and Viola, 2018; Falorsi et al., 2018; Li et al., 2021; Xu et al., 2020; Zhang et al., 2022a; Glazunov and Zarras, 2023), producing perceptually smoother and more coherent latent spaces.

Pairing a powerful generative model with strong factorization assumptions on the posterior

and prior encourages disentanglement without sacrificing generative performance. The KL regularization term enforces an information bottleneck by encouraging posterior whitening, which promotes maximum entropy representations in the latent space (Burgess et al., 2018). Crucially, the conditional denoising objective encourages the latents to capture meaningful structure in the observed data by leveraging the implicit multi-scale image prior learned by the diffusion model. We prove by leveraging the geometric properties of our model that disentangled representations of hierarchical semantic features naturally arise from the use of this image prior, potentially explaining why diffusion-based representation models offer superior disentanglement scores compared to traditional VAEs.

Moreover, our approach sidesteps the traditional rate-distortion trade-off in VAEs: rather than compromising reconstruction fidelity to achieve better disentanglement, SAMI maintains perceptually high-quality generations through the expressive power of the diffusion prior, even when the latent representation is heavily regularized. This information theoretic perspective suggests interesting connections to the I-MMSE relation (Guo et al., 2005; Kong et al., 2023), which has recently been used to link conditional diffusion scores with estimates of mutual information (Franzese et al., 2024; Kong et al., 2024). Such connections might provide alternative theoretical foundations for understanding representation learning in conditional diffusion models and could potentially allow for more direct optimization of information-theoretic objectives such as InfoMax (Linsker, 1988). We leave the exploration of these connections as an exciting direction for future work. Overall, our results represent an important and mathematically rigorous step towards building learning systems that extract complex semantical structure from data in an unsupervised manner while maintaining competitive generative performance.

4 A Rate-Distortion Account of InfoMax

4.1 Overview

In the previous chapter, I introduced a novel objective for unsupervised learning derived from the ELBO. I showed that augmenting the generative model with auxiliary variables provides several computational benefits: the additional structure allows us to estimate more complex statistical structure, and, when paired with a factorized recognition model, facilitates unsupervised learning of latent factors that are disentangled by spatial scale. However, these computational advantages do not, on their own, clarify what the model tells us about the brain. As an objective, the ELBO makes several key assumptions that make it unclear which of its components should be interpreted as claims about neural computation, and which are merely convenient for computational tractability. This ambiguity subsequently affects the interpretability of the learned representation.

What we would like instead is a model that instantiates the InfoMax objective directly, so that the representations it learns constitute a test of that principle rather than an artifact of our assumptions. In this chapter, I present a new information-theoretic objective based in rate-distortion that directly relates to both InfoMax and efficient coding. By leveraging the I-MMSE relation from information theory, I show that our model can directly optimize this information

theoretic quantity to learn a representation of images without strong approximations or assumptions.

The chapter proceeds as follows. I first show how to learn a representation that maximizes information with its observations using the I-MMSE relation. Next, I formalize an optimal information-theoretic constraint by connecting conditional denoising with rate-distortion. We provide geometric and biological interpretations of our objective, and demonstrate the results of training our model on a synthetic dataset. Finally, I show how the objective relates to the ELBO and prove that, despite making fewer assumptions, the new objective preserves the disentanglement properties of its ELBO-based predecessor.

4.2 Implementing InfoMax using I-MMSE

InfoMax states that the principal goal of sensory processing in neural systems is maximal information transmission across noisy channels, subject to resource constraints. Let us first treat information transmission between the observation variables (i.e. inputs) X and response variables (outputs) Z, then turn to the question of resource constraints within the established framework. The InfoMax objective maximizes the mutual information,

$$I(X;Z) = h(X) - h(X|Z).$$

where $h(X) = -\mathbb{E}_{p(x)}[\log p(x)]$ and $h(X|Z) = -\mathbb{E}_{p(x,z)}[\log p(x|z)]$ are marginal and conditional differential entropies respectively.

Because entropy and mutual information are in general intractable for continuous non-Gaussian distributions, and are difficult to estimate without bias from finite samples (Bialek et al., 1991; Paninski, 2003), past approaches rely on proxies to entropy (Hyvärinen, 1997), computable bounds (Barber and Agakov, 2003; Nguyen et al., 2010; van den Oord et al., 2018; Belghazi et al., 2021;

Poole et al., 2019), nonparametric kNN estimators for continuous variables (Kraskov et al., 2004) or Bayesian estimators for discrete distributions (Archer et al., 2013). However, our aim is not to estimate $I(X;Z)$ for a fixed representation, but to learn the representation Z itself by maximizing the mutual information. This calls for a tractable, differentiable objective that can be optimized with respect to Z, which singles out the class of the computable bounds.

Computable bounds come in two flavors. Critic-based bounds use a learned critic $f(x,z)$ that scores joint against marginal samples, while variational bounds assume a choice of encoder $p(z|x)$ and introduce a variational family $q(x|z)$ to approximate the true generative model $p(x|z)$ via a separate decoder [1]:

$$I(X;Z) \geqslant H(X) + \mathbb{E}_{p(x,z)}[\log q(x|z)] := \tilde{I}(X;Z), \tag{4.1}$$

where the bound is loose by $D_{\mathrm{KL}}\big(p(x|z)\|q(x|z)\big) \geq 0$ with equality iff $q(x|z) = p(x|z)$. The derivation is given in Appendix C.1.

Both classes of mutual information estimators suffer from separate issues. Critic-based methods make no parametric assumptions on $p(x|z)$, but suffer from statistical pathologies: methods with unnormalized bounds (e.g. NWJ (Nguyen et al., 2010), MINE (Belghazi et al., 2021)) have variance that grows exponentially in the true mutual information, while the normalized contrastive bound (InfoNCE (van den Oord et al., 2018)) is bounded above by $\log K$ for K contrastive samples and under-estimates the mutual information above this ceiling (Poole et al., 2019; Song and Ermon, 2020).

Decoder-based bounds are tight only to the extent that their variational family contains the true decoder. That decoder follows from Bayes' rule,

$$p(x \mid z) = \frac{p(z \mid x)\, p(x)}{p(z)},$$

[1] Note that we use $q(\cdot)$ here to denote the variational approximation of the *generative* mapping $Z \to X$, rather than the *inferential* mapping $X \to Z$ as is more customary in the VAE and ELBO literature.

but it is intractable: the marginal $p(z) = \int p(z \mid x)\, p(x)\, dx$ cannot be evaluated when $p(x)$ is known only through samples. A variational decoder $q(x|z)$ removes this intractability, but introduces mis-specification whenever the true decoder lies outside its family. The Gaussian decoder standard in VAEs is a clear instance: natural images are heavy-tailed (Field, 1987), so the true decoder $p(x|z) \propto p(z|x)p(x)$ inherits heavy tails in x that a Gaussian $q(x|z)$ cannot capture. This mismatch is the discrepancy in the bound $D_{\mathrm{KL}}(p(x|z)\|q(x|z)) \geqslant 0$, which closes only when q recovers the true generative model. We thus seek a decoder family expressive enough to eliminate this gap while remaining parametric and tractable to optimize.

4.2.1 Maximizing Mutual Information using the I-MMSE Relation

We address the issues associated with assuming a simple Gaussian variational family by making use of the I-MMSE relation, a seminal result by Guo et al. (2005) that establishes a fundamental connection between the mutual information and minimum mean square error estimators. Let $p(x_\gamma|x)$ be a Gaussian noise channel with $x_\gamma = \sqrt{\gamma}x + \epsilon$ and $\epsilon \sim \mathcal{N}(0, I)$, where γ represents the Signal-to-Noise Ratio (SNR) and $p(x)$ is the data distribution[2]. The I-MMSE relation states that

$$\frac{\partial}{\partial \gamma} I(X; X_\gamma) = \frac{1}{2}\mathrm{mmse}(\gamma) \tag{4.2}$$

where mmse refers to the Minimum Mean Square Error for recovering x from this noisy channel

$$\mathrm{mmse}(\gamma) := \min_{\hat{x}(x_\gamma, \gamma)} \mathbb{E}_{p(x, x_\gamma)} \left[\left\| x - \hat{x}(x_\gamma, \gamma) \right\|^2 \right].$$

Here, $\hat{x}(x_\gamma, \gamma)$ is the denoising function, whose optimum $\hat{x}^*$ is well known to be the posterior mean $\mathbb{E}_{p(x|x_\gamma)}[x]$. Finding this optimum analytically is typically intractable because it requires sampling from the posterior distribution of the noise channel, but this function can be learned

[2] In practice we use the variance-preserving Gaussian channel $x_{\gamma'} = \sqrt{1-\gamma'}x + \sqrt{\gamma'}\epsilon$, where $\gamma' = {}^{1}/_{1+\gamma}$. The results in the rest of this chapter hold up to this global rescaling factor.

via neural network-based denoisers.

Kong et al. (2023) showed that integrating Eq. 4.2 over all values of SNR allows us to numerically estimate the differential entropy of an arbitrarily complex data distribution of ambient dimensionality d from which we have i.i.d. samples (Kong et al., 2023):

$$h(X) = \mathbb{E}_{p(x)}[-\log p(x)] = \frac{d}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left(\frac{d}{1+\gamma} - \text{mmse}(\gamma)\right) d\gamma. \tag{4.3}$$

This identity can easily be generalized to estimate the approximate conditional entropy $h(X|Z)$. Defining $\text{mmse}(\gamma|z) \equiv \mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|x - \hat{x}(x_\gamma, \gamma, z)\right\|^2\right]$, where $\hat{x}(x_\gamma, \gamma, z)$ is the z-conditional denoiser, we get

$$h(X|Z) = \mathbb{E}_{p(x)p(z|x)}[-\log q(x|z)] = \frac{d}{2}\log(2\pi e) - \frac{1}{2}\int_0^\infty \left(\frac{d}{1+\gamma} - \text{mmse}(\gamma|z)\right) d\gamma. \tag{4.4}$$

Combining these two identities, we get a denoiser-based estimate of the mutual information between observation X and latent Z:

$$\begin{aligned} I(X;Z) &= h(X) - h(X|Z) \\ &= \frac{1}{2}\int_0^\infty \mathbb{E}_{p(x_\gamma,x,z)}\left[\|x - \hat{x}(x_\gamma,\gamma)\|^2 - \|x - \hat{x}(x_\gamma,\gamma,z)\|^2\right] d\gamma \end{aligned} \tag{4.5}$$

This identity has been used to measure the influence of a conditioning variable on diffusion models during conditional generation (Kong et al., 2024), the effect of stimuli on sensory coding (Laquitaine et al., 2025), and the role of non-linear transforms (Yu et al., 2025).

Our contribution is to invert this I-MMSE estimator to learn a representation Z. Here we assume an encoding distribution $p(z|x)$ that provides a conditioning signal to the conditional denoiser $\hat{x}(x_\gamma, \gamma, z)$. The quality of the latent is then evaluated by how well it maximizes the I-MMSE estimate of the mutual information. How the conditioning is implemented in practice is the subject of the following section.

4.2.2 Implementation of I-MMSE InfoMax

We assume $p(x)$ is a fixed distribution, so we can treat its entropy as a constant and the optimization problem is reduced to one of minimizing the conditional entropy such that $\mathcal{L} = \max I(X;Z) = \min h(X|Z)$. Eq. 4.4 can be written as

$$h(X|Z) = \frac{1}{2}\int_0^\infty \mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|x - \hat{x}(x_\gamma,\gamma,z)\right\|^2\right] d\gamma + \text{const.} \tag{4.6}$$

where the constant term is independent of the conditional denoiser.

Ideally, we would like to rewrite this expression in terms of the encoding distribution $p(z|x)$, so that we can directly optimize its parameters. Tweedie's formula and Bayes' rule tells us that the *z-conditional* posterior mean is equal to the *unconditional* posterior mean plus the likelihood score, $\hat{x}(x_\gamma,\gamma,z) = \hat{x}(x_\gamma,\gamma) + g_\gamma \nabla_{x_\gamma} \log p(z|x_\gamma)$, where $g_\gamma = 1/\sqrt{\gamma}$ is a γ-dependent scalar that appropriately weighs the two terms according to Bayes' rule. This means that we can parameterize this optimization problem with two neural networks, one with parameters θ that estimates the unconditional posterior mean $\hat{x}_\theta(x_\gamma,\gamma)$, and another with parameters ϕ that computes the encoding distribution $p_\phi(z|x)$. As is typical in VAEs, we set the encoding distribution to be a Gaussian with diagonal covariance, which, as we will later see, is necessary for learning a disentangled representation. To derive the guidance score, we compute the likelihood of the clean input representation $z \sim p_\phi(z|x)$ with respect to the latent distribution of the noisy input $x_\gamma \sim p(x_\gamma|x)$, given by $\mathbb{E}_{p_\phi(z|x)}\left[\log p_\phi(z|x_\gamma)\right]$. Note that the same encoder network is used to output the latent distributions of both the clean inputs $p_\phi(z|x)$ and the noisy inputs $p_\phi(z|x_\gamma)$. The guidance score $\nabla_{x_\gamma} \log p_\phi(z|x_\gamma)$ comes from taking the gradient of the log likelihood with respect with respect to the noisy input x_γ, which we perform via auto-differentiation through the encoder network.

The two networks are then trained jointly to minimize the objective

$$\mathcal{L}(\theta, \phi) = \frac{1}{2} \int_0^\infty \mathbb{E}_{p_{\mathcal{D}}(x),\, p(x_\gamma|x),\, p_\phi(z|x)} \left[\left\| x - \hat{x}_\theta(x_\gamma, \gamma) - g_\gamma \nabla_{x_\gamma} \log p_\phi(z|x_\gamma) \right\|^2 \right] d\gamma. \tag{4.7}$$

As it stands, the objective has two issues. The first is one of identifiability: because only the sum of the denoiser's output and the encoder's guidance term is matched to the clean signal x, nothing forces the denoiser to output the unconditional posterior mean on its own as the interpretation above would suggest. This can be ameliorated by first pre-training the model on the MSE loss used for unconditional diffusion models.

The second issue was alluded to earlier: absent any constraint on the total information carried by the representation, the objective is trivially minimized by a Z that copies all the information in X. Reducing the dimensionality of Z does not help, since a sufficiently expressive encoder can retain all information about X through a nonlinear transformation. What we need instead is a second, information-theoretic constraint that caps the information held by Z. But what is the optimal information theoretic constraint? For our answer, we turn to rate-distortion theory.

4.3 Formalizing the Constraint via Rate-Distortion

4.3.1 Casting InfoMax as Wyner-Ziv Distortion

Rate-distortion theory is a branch of information theory that articulates the minimum amount of resources, expressed in terms of information, required to reconstruct a signal to some desired level of fidelity once it has been passed through a noisy communication channel. A classic communication channel comprises a "source" signal X, together with an encoder $p(z|x)$ and a decoder $p(\hat{x}|z)$ that reconstruct the signal (Fig. 1.5).

The distortion is the cost of representing a sample x with its approximate reconstruction $\hat{x}$.

We assess this via a distance metric $d(\cdot, \cdot)$, averaged over the joint distribution $p(x, \hat{x})$,

$$D(X, \hat{X}) := \mathbb{E}_{p(x,\hat{x})} \left[d(x, \hat{x}) \right] .$$

A typical choice is the squared-error distortion, $d(x, \hat{x}) = \|x - \hat{x}\|^2$. The rate R is the maximum number of bits per sample that can pass through the channel, evaluated as the mutual information between the source and the encoded signal. These two terms can combined into a single Lagrangian that trades off the two quantities through a multiplier β,

$$\mathcal{L} = D(X, \hat{X}) + \beta \, I(X; Z),$$

whose solutions, as β is swept over $(0, \infty)$, trace the boundary of the realizable region on the rate-distortion plane (Fig. 1.6).

Our model departs from the classic single-channel picture in two respects. First, at a given SNR γ the decoder additionally receives the noisy observation X_γ. A source-coding problem in which the decoder alone holds correlated side information is the setting of *Wyner-Ziv* coding (Wyner and Ziv, 1976) (Fig. 4.1). Additionally, our model operates across a range of SNRs $\gamma \in (0, \infty)$, meaning that the decoder must receive side information X_γ that is correlated with X to different degrees.

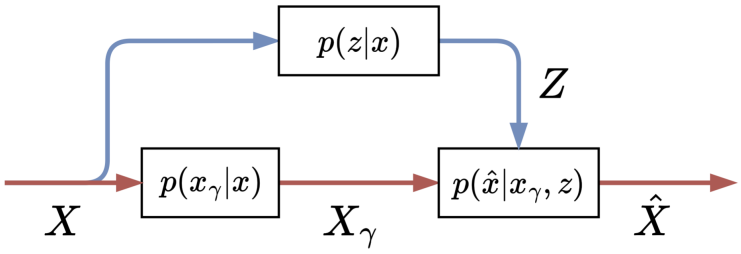


Figure 4.1: Block diagram of our model as a source-coding problem with side information. The encoder $p(z|x)$ sees only the clean source X, while the decoder $p(\hat{x}|x_\gamma, z)$ additionally receives the side information $x_\gamma \sim p(x_\gamma|x)$.

Our model consists of a recognition model ("encoder") $p(z|x)$ that maps the clean source X to a latent representation Z, and a conditional denoiser ("decoder") $\hat{x}(x_\gamma, \gamma, z)$ that reconstructs the source from the latent z together with a noisy observation $x_\gamma \sim p(x_\gamma|x)$ at signal-to-noise ratio γ. We emphasize that the encoder conditions only on the clean signal: since $Z \to X \to X_\gamma$ forms a Markov chain, the representation Z is a property of X, not of any particular noise level. Thus the encoder distribution $p(Z|X = x)$, which we optimize, is distinct from the likelihood $p_\phi(Z = z|X = x_\gamma)$, which scores the latent against a *noisy* observation and supplies the guidance term. The guidance term is an inferential quantity that is computed as part of the decoder rather than the message being communicated by the encoder.

The integrand in the I-MMSE relation for fixed γ – i.e. the conditional MMSE – corresponds to a choice of a squared-error distortion metric. Thus we can define the γ-specific distortion, or the *marginal* distortion, as

$$D_\gamma \;:=\; \frac{1}{2}\,\mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|x - \hat{x}(x_\gamma, \gamma, z)\right\|^2\right] \;=\; \frac{1}{2}\mathbb{E}_{p(x,x_\gamma,z)}\Big[d\big(x, \hat{x}(x_\gamma, \gamma, z)\big)\Big].$$

Since integrating the conditional MMSE over all γ results in the conditional entropy up to a constant (Eq. 4.7), we can show that D_γ can be expressed as the SNR-derivative of the conditional entropy

$$D_\gamma \;=\; \partial_\gamma h(X_\gamma|Z) \;=\; -\,\partial_\gamma h(X \mid X_\gamma, Z) \;\geqslant\; 0.$$

The full derivation is given in Appendix C.2.

4.3.2 Optimal Wyner-Ziv Rate

With the distortion established, the Wyner-Ziv theorem tells us the optimal rate. In general, for codeword Z, side information Y, and distortion D, the optimal rate is given as

$$R^*(D) = \min_{p(z|x)p(\hat{x}|x_\gamma,z):\,\mathbb{E}[d(x,\hat{x})]\leqslant D} \Big[\, I(X;Z) - I(Y;Z) \,\Big] = \min_{p(z|x)p(\hat{x}|x_\gamma,z):\,\mathbb{E}[d(x,\hat{x})]\leqslant D} I(X;Z \mid Y),$$

where the second equality follows from the Markov chain. This is the *residual* information the decoder still needs from Z once it holds Y, smaller than the unconditioned $I(X;Z)$ by $I(Y;Z)$. With $Y = X_\gamma$, the optimal rate is $R^*(D) = \min I(X;Z \mid X_\gamma)$.

This rate is defined at a single operating SNR, whereas our model sweeps the whole range $\gamma \in (0,\infty)$. Rather than evaluate our model at one point, we track how the rate is resolved as the SNR grows. Define the *marginal rate* as the information about Z that the observation resolves per unit SNR,

$$R_\gamma \equiv \partial_\gamma I(X_\gamma;Z) = -\,\partial_\gamma I(X;Z \mid X_\gamma) \geqslant 0,$$

which is non-negative because a higher-SNR observation can only reveal more information about the latent. As γ runs from 0 to ∞, the observation X_γ sweeps from pure noise, which carries no information about Z, to the clean source X, which carries all of $I(X;Z)$. Integrating the marginal rate across all SNR therefore accumulates these increments to the full information content of the representation,

$$\int_0^\infty R_\gamma \, d\gamma = I(X_\infty;Z) - I(X_0;Z) = I(X;Z).$$

The Wyner-Ziv rate at any operating point is the remaining tail upon observation of X_γ, $I(X;Z \mid X_\gamma) = \int_\gamma^\infty R_{\gamma'}\, d\gamma'$, which vanishes as $\gamma \to \infty$ when the side information becomes the source itself.[3]

[3] Under the incremental-channel coupling that underlies the I-MMSE relation (Guo et al., 2005), the SNR family can be realized as an increasing observation process, so that X_γ is a sufficient statistic for the accumulated history $\{X_{\gamma'} : \gamma' \leqslant \gamma\}$. This is what makes the increments R_γ additive across scales rather than double-counting; the decomposition itself needs only the marginal Markov chain $Z - X - X_\gamma$ at each γ.

The same conclusion follows from another simpler argument. Fix a single operating point γ and consider how many bits must be conveyed to the decoder. Through the side information the decoder already shares $I(X_\gamma; Z)$ bits with the codeword which now no longer needs to be sent. Wyner-Ziv coding states that the rate still to be transmitted is $I(X; Z \mid X_\gamma)$. By the chain rule, these two parts sum to the total information the codeword carries about the source,

$$\underbrace{I(X_\gamma; Z)}_{\text{free to the decoder}} + \underbrace{I(X; Z \mid X_\gamma)}_{\text{Wyner-Ziv transmission rate}} = I(X; Z) \qquad \forall\, \gamma.$$

Though we fixed γ at the start, note that each term on the left depends on the operating point while their sum does not. This means that this decomposition holds at every γ and sums to the same total. The total rate $I(X; Z)$ is thus a property of the encoder $p(z \mid x)$ alone, independent of any decoder, side information, and the SNR at which we read it out.

Efficient coding tells us why this sum, rather than the transmitted rate, is the quantity to constrain. The free term $I(X_\gamma; Z)$ is free only in the sense of *transmission*: a decoder already holding a noisy copy of the stimulus no longer needs to receive those bits. They are not free to the *encoder*, which has to commit them to the representation regardless of the operating point. A capacity- and energy-limited neural code thus pays for these bits whether or not a downstream process can reconstruct the clean signal from x_γ. Limited neural resources constrain the cost of *representing* the stimulus, not of *reconstructing* it.

The total rate-constrained InfoMax objective is thus given as

$$\mathcal{L} = D + \beta R = h(X \mid Z) + \beta\, I(X; Z), \tag{4.8}$$

which can be expressed in terms of γ-specific Lagrangians:

$$\mathcal{L} = \int \mathcal{L}_\gamma \, d\gamma, \qquad \mathcal{L}_\gamma = D_\gamma + \beta R_\gamma = \partial_\gamma h(X_\gamma | Z) + \beta\, \partial_\gamma I(X_\gamma; Z).$$

4.3.3 Implementation of the Rate

How should we implement the rate in our model? The instantaneous rate can be related to the norm of the guidance score via de Bruijn's identity, which connects mutual information to Fisher information:

$$R_\gamma = \partial_\gamma I(Z; X_\gamma) = \frac{1}{2\gamma} \mathbb{E}_{p(x,x_\gamma,z)} \left[\left\| \nabla_{x_\gamma} \log p(z \mid x_\gamma) \right\|^2 \right] \geqslant 0. \tag{4.9}$$

The proof is given in Appendix C.3. Penalizing the squared norm of the guidance score therefore restricts the information the representation carries about the source. The same quantity can be recast as a statement about the *direction* of the guidance score,

$$R_\gamma = \mathbb{E}_{p(x,x_\gamma,z)} \left[\nabla_{x_\gamma} \log p(z \mid x_\gamma) \cdot \partial_\gamma x_\gamma \right], \tag{4.10}$$

where the equivalence of the two forms of marginal rate is given in Appendix C.4. This form can be implemented in our model by minimizing the dot product of the guidance score with $\partial_\gamma x_\gamma$, which can be expressed in terms of x and x_γ depending on the form of the noising schedule.

Both forms of the rate try to minimize the amount of guidance required to steer the dynamics toward the observed sample x. We provide a geometric interpretation below.

4.3.4 Relationship between Rate and Distortion

Upon close inspection, one might notice an apparent tension within the objective: originally we sought to *maximize* the mutual information, whereas the rate term now *penalizes* it. However, this tension is resolved by two observations.

First, these two terms correspond to different components of our model. The distortion is computed via the I-MMSE relation using the latent-conditioned denoiser (the decoder), while the rate pertains solely to the recognition model (the encoder). For this reason, minimizing distor-

tion and rate have different effects: minimizing rate maximizes information in Z about X via its *decodability*, which means the encoder is encouraged to represent bits the decoder can turn back into X. Minimizing rate, meanwhile, penalizes the total bit count of the encoder. In theory, the encoder could raise $I(X;Z)$ in ways the decoder cannot use, but because those bits would raise the rate term without lowering distortion, they are not represented. Thus the model selects for information that is relevant while minimizing the total. This ties back to our goals of unsupervised learning: the conditional entropy is essentially a downstream *task* of the representation, while the second term sets a limit on how many resources the representation can have in the first place.

Second, the trade off between these two terms are in terms of the learned representation is controlled by the Lagrange multiplier β. Substituting $h(X \mid Z) = h(X) - I(X;Z)$ gives

$$\mathcal{L} = h(X) + (\beta - 1)\, I(X;Z).$$

Since $h(X)$ is a constant of the fixed data distribution, minimizing $\mathcal{L}$ over the encoder is equivalent to optimizing $(\beta - 1)\, I(X;Z)$. Let us consider three regimes of β:

- $\beta < 1$: the coefficient is negative, so minimizing $\mathcal{L}$ still maximizes $I(X;Z)$. This is constrained InfoMax.
- $\beta > 1$: minimizing $\mathcal{L}$ compresses the representation by driving down $I(X;Z)$.
- $\beta = 1$: the two terms cancel and $\mathcal{L} = h(X)$.

When $\beta < 1$, the network is encouraged to care more about minimizing reconstruction error over creating a compressed code; the converse is true when $\beta > 1$. At $\beta = 1$ we have no learning signal for Z, which results in a non-identifiable representation. This has been observed by Alemi et al. (2018) when training β-VAEs: when $\beta = 1$, the objective is indifferent to how much information the representation carries, and β must be set away from one to express how compressed we wish

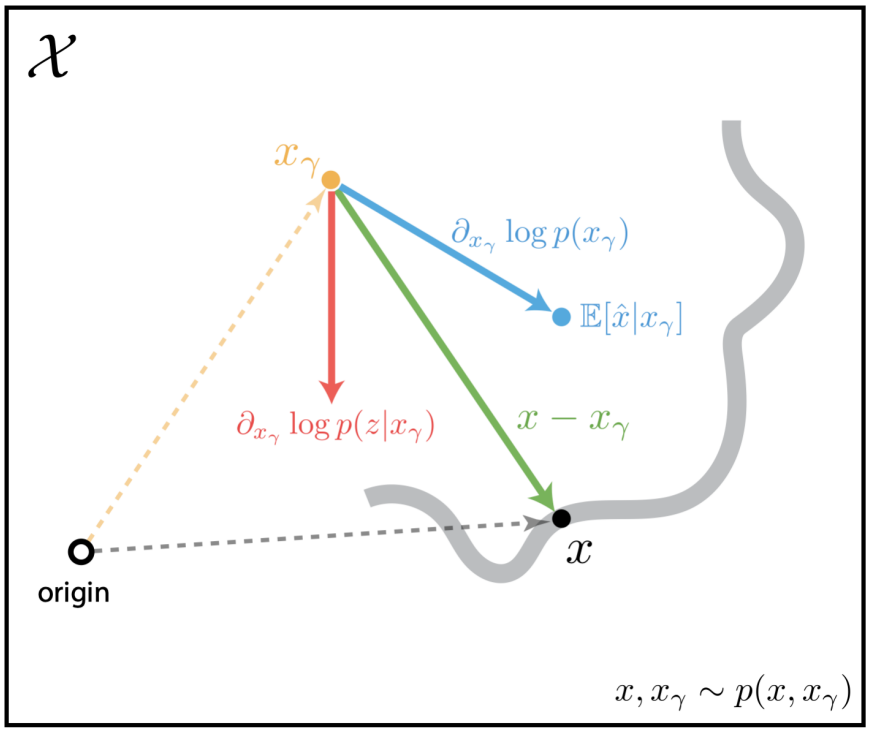


Figure 4.2: A geometric perspective of the denoiser-based decoder. Given our current operating point γ and current noisy state x_γ (given to the decoder as side-information), the goal is to reconstruct the original source x. The unconditional denoiser provides a denoising estimate that pushes the decoder towards the posterior mean (blue vector), while the guidance score corrects the dynamics (red vector). The InfoMax objective tries to force the sum of the two vectors to be as close as possible to the green arrow.

our representation of X to be.

4.4 Interpreting the objective

4.4.1 Geometric Interpretation

In Fig. 4.2, we consider a single step of guided denoising at SNR γ. Noising a clean image $x \sim p_{\mathcal{D}}(x)$ (black dot) results in a noised image x_γ (yellow dot). The cartoon corresponds to a variance-exploding noising process $x_\gamma = x + \gamma\epsilon$ where $\gamma = 1$. The aim of the model is to reconstruct the original image via conditional denoising; the optimal solution is given by $x - x_\gamma$ (green line). The optimal unconditional denoiser gives us the unconditional posterior mean $\mathbb{E}[x|x_\gamma]$

(blue dot) which relates to the unconditional score $\partial_\gamma \log p(x_\gamma)$ (blue line) via Tweedie's formula: $\mathbb{E}[x|x_\gamma] = x_\gamma + \partial_{x_\gamma} \log p(x_\gamma)$. The latent z that comes from the recognition model $p(z|x_\gamma)$ is used to produce a guidance score $\partial_{x_\gamma} \log p(z|x_\gamma)$ (red arrow). The sum of the unconditional score and the guidance score gives us the conditional score $\partial_{x_\gamma} \log p(x_\gamma|z) = \partial_{x_\gamma} \log p(x_\gamma) + \partial_{x_\gamma} \log p(z|x_\gamma)$, which via Tweedie produces the *conditional* posterior mean $\mathbb{E}[x|x_\gamma, z]$. The distortion term encourages this conditional estimate to be as close as possible to the original clean image, which, for a fixed unconditional denoiser, means encouraging the encoder to produce a guidance score that is the vector difference between the green and blue arrows. In other words, the information contained by the optimal z should capture the remaining variability not already captured by the unconditional denoiser.

The rate term places a restriction on how useful this guidance score can be. If we assume the dot-product form of the rate (Eq. 4.10), zero rate corresponds to a guidance score *orthogonal* to the green arrow since $\partial_\gamma x_\gamma = \epsilon = x_\gamma - x$. From this picture, we see that the optimal guidance score has large cosine dissimilarity with the hyperplane orthogonal to $x - x_\gamma$, and thus requires a non-zero rate. Thus by constricting the *direction* of the guidance score to lie close to the hyperplane, we can limit how much useful information is contained in the representation Z.

Another piece of intuition that we can glean from this picture is the effect of the noising schedule. Under the variance exploding noising schedule corresponding to this cartoon, we see that the guidance score makes more efficient use of bits at high noise levels (low SNR regime): at high SNR, the unconditional posterior mean is close to the clean image, and an orthogonal guidance score provides almost no benefit. Meanwhile, at low SNR when the unconditional posterior mean and clean image are far apart, even an orthogonal guidance score may still be useful. However, under a different noising schedule (such as a variance preserving schedule), the distances can be scaled by some function of γ such that rate is equally useful across all noise levels.

Meanwhile, the norm form of the rate (Eq. 4.9) does not penalize the direction of the guidance score, but rather its *magnitude*. This means that even though it can face the optimal direction of

guidance, the rate restricts how useful it can be at given γ by limiting the amount by which it can contribute to the total conditional score.

4.4.2 Biological implications

Though the rate is given as an information theoretic term $I(X;Z)$, the decomposition of this term into likelihood *vectors* has a natural biological interpretation. If we interpret the guidance score as the additive likelihood computed by a separate neural population providing sensory or contextual signals (as suggested by our circuit model in Chapter 2), then the rate restricts the influence of this signal on the prior circuit dynamics in one of two ways. Namely, if we assume the norm form of the rate, the rate restricts the total energy or variance of the neural population encoding the likelihood signal. Alternatively, if we assume the dot product form of the rate, it encourages the likelihood population responses to live in a subspace orthogonal to the prior attractor dynamics.

There is biological evidence to support both forms. The norm form corresponds to the metabolic constraint on total spiking power (response variance) of the likelihood population, the standard variance-based proxy for metabolic cost in efficient coding (Atick and Redlich, 1990; van Hateren, 1992; Doi et al., 2012). Meanwhile, if we assume the dot-product form of the rate, we penalize the alignment between the prior drift s and the likelihood bias g, so that under a resource constraint the likelihood population is encouraged to inject its bias along an axis orthogonal to the prior drift. This encourages a non-redundant, steering signal rather than a gain modulation the prior population could realize on its own. This geometry in which a contextual signal occupies an orthogonal dimension of the prior population response is reported by Mante et al. (2013) in prefrontal cortex, in which task-relevant sensory evidence and contextual selection signals are integrated within a single population while remaining along near-orthogonal directions of state space. The broader principle that cortex routes interacting signals through mutually orthogonal subspaces to prevent interference is further supported by the organization of motor preparatory

activity (Kaufman et al., 2014) and by the communication subspaces through which cortical areas selectively relay activity (Semedo et al., 2019).

These two forms need not be regarded as independent constraints. If the resource budget is imposed not on the likelihood bias alone but on the *total* drive $u = s + g$, then the norm and dot-product forms are simply the two terms of a single quadratic,

$$\|u\|^2 = \underbrace{\|s\|^2 + \|g\|^2}_{\text{norm form}} + \underbrace{2\langle s, g\rangle}_{\text{dot-product form}} .$$

Under a fixed budget on $\|u\|^2$, any component of the likelihood bias that is parallel to the prior drift increases the combined power through the dot-product term, while conveying nothing the prior population could not produce by gain modulation alone. Orthogonalizing the likelihood bias and the prior drift is therefore the allocation that maximizes the non-redundant influence of the likelihood influence per unit of metabolic cost (Proposition C.6). Thus both forms of the rate – the norm constraint on neural activity and the orthogonality of the likelihood signal – emerge from assuming that the combined population adheres to the efficient coding principle.

4.4.3 Relation to ELBO

We can show that our InfoMax objective derived from rate-distortion in Eq. 4.8 is closely related to the ELBO objective from the previous chapter. The InfoMax objective is given as

$$\mathcal{L} = h(X \mid Z) + \beta\, I(X; Z).$$

The distortion term (conditional entropy) is upper bounded by the reconstruction term in the ELBO by the Barber-Agakov bound,

$$h(X|Z) = \mathbb{E}_{p(x,z)}\left[-\log q(x|z)\right] - \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(p(x|z)\|q(x|z)\right)\right],$$

while the rate term (mutual information) is upper bounded by the ELBO's KL divergence term (Hoffman and Johnson, 2016),

$$I(X;Z) = \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(p(z|x)\|m(z)\right)\right] - D_{\mathrm{KL}}\left(p(z)\|m(z)\right),$$

where we have kept our current notation of using $p(z|x)$ to indicate the true encoder and $q(x|z)$ for the variational decoder. The distribution $m(z)$ is the *marginal* that is a variational approximation to $p(z)$ and typically chosen to be an isotropic Gaussian $N(0, I)$ in VAEs.

Putting the two terms together, we get:

$$\begin{aligned} h(X|Z) + \beta\, I(X;Z) &= \mathbb{E}_{p(x,z)}\left[-\log q(x|z)\right] + \beta\, \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(p(z|x)\|m(z)\right)\right] \\ &\qquad - \Big(\mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(p(x|z)\|q(x|z)\right)\right] + \beta\, D_{\mathrm{KL}}\left(p(z)\|m(z)\right)\Big) \\ &= \mathcal{L}_{\beta\text{–VAE}} - \Big(\mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}\left(p(x|z)\|q(x|z)\right)\right] + \beta\, D_{\mathrm{KL}}\left(p(z)\|m(z)\right)\Big) \\ &\leq \mathcal{L}_{\beta\text{–VAE}}. \end{aligned}$$

Since KL divergences are non-negative, we see that the InfoMax objective, which we wish to minimize, is upper bounded by the β-VAE loss.

Here it is worth noting that the InfoMax objective also uses the Barber-Agakov bound in its computation of the conditional entropy, but because of the expressivity of the denoiser-based generative model, it does not suffer the same degree of slack as VAEs that employ conditional Gaussian generative models.

4.5 Results

4.5.1 Numerical experiments

To assess how the InfoMax objective affects the learned representation, we trained our model on a simple synthetic 28×28 "disk" dataset that is similar to the synthetic data in the previous chapter. The disk dataset comprises greyscale images of a circle on a blank background. These images come from a generative process with only two independent variables: the x and y coordinate of the disk.

The recognition network is a convolutional neural network and the denoiser is a UNet. We set the number of latent dimensions to two to equal the number of independent latent variables and used a static value of $\beta = 1 \times 10^{-3}$.

After training the model, we investigated the representation learned by the model by feeding the encoder images with known causal factors. The images comprise disks of constant intensity but varying position on a blank background; the x and y coordinates of the disk are the two causal factors (Fig. 4.3a-b). To test the quality of the learned representation, we drew posterior samples from the recognition model and synthesized samples using the generative model via the latent-conditioned diffusion process. The synthesized images closely resemble the input image, suggesting that the model has learned a good representation of the causal factors (Fig. 4.3b). We investigated the learned representation by generating images from an exhaustive search over the two causal factors (Fig. 4.3c) and examining the corresponding posterior mean. We observe that the latents form a near-Euclidean grid (Fig. 4.3d); changes in the disk's position trace out roughly straight lines in the latent space. The curvature near the edges are likely due to the rate penalty; compression in a Gaussian channel draws the distribution close to an isotropic Gaussian. This suggests that the model has learned a factorized representation of the data, with each latent axis representing a single causal factor.

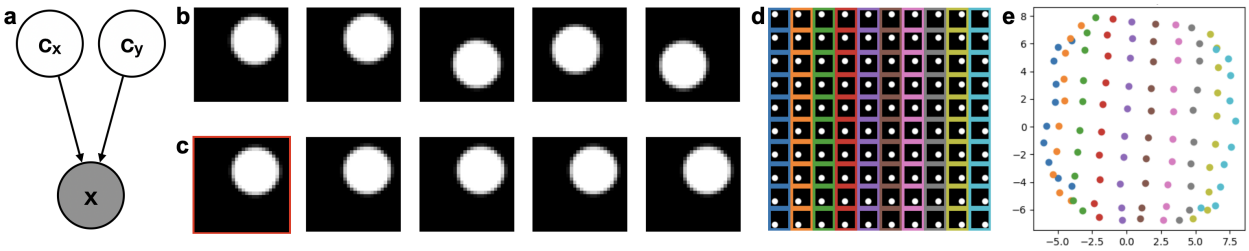


Figure 4.3: Disk dataset. a) graphical model describing the image generation process, b) samples drawn i.i.d. from the disk dataset, c) conditional generation given a "target" image (red outline), d) disk images fed into the encoder, with x coordinate coded by color, and e) latent space of model. Each dot corresponds to the latent of one disk image. The dots share the same colors as subplot d.

4.5.2 Disentanglement

Learning a representation that is *disentangled* is a common goal of unsupervised learning. It rests on the assumption that the world consists of few generative independent causes that can be recovered by latent variable models. If this assumption holds, then disentangled representation offer many potential benefits: they should be more interpretable and compact, generalize to unseen data during inference, answer counterfactual questions, and allow for transfer, few-shot learning, and supervised learning of downstream tasks (Pearl, 2009; Kumar et al., 2018; Chen et al., 2019; Locatello et al., 2019).

While there are no formal definitions of disentanglement, a key intuition is that disentangled representations should be able to resolve *factors of variation* in the data (Bengio et al., 2014), such that a single change in an underlying generative factor results in a corresponding change in a single factor in the learned representation. Many metrics of disentanglement have been proposed based on variations this idea (Higgins et al., 2017; Eastwood and Williams, 2018; Chen et al., 2019; Ridgeway and Mozer, 2018; Kumar et al., 2018).

Most approaches in unsupervised learning are variants of the VAE (Kingma and Welling, 2013) that encourage disentanglement via encouragement of a factorized aggregate posterior $q_\phi(z) = \int q_\phi(z|x)p(x)\,dx$ (Higgins et al., 2017; Burgess et al., 2018; Kumar et al., 2018; Chen et al., 2019; Kim and Mnih, 2019). This is usually accomplished by tweaking the relative strength of the KL

divergence term $D_{\mathrm{KL}}\big(q_\phi(z|x)\|m(z)\big)$ as compared to the reconstruction term.

Attempts at learning disentangled representation have not been promising (Locatello et al., 2019; Locatello et al., 2020). In particular, Locatello et al. (2019) show theoretically that unsupervised learning of disentangled representations is fundamentally impossible without invoking inductive biases that are specific to the particular dataset or learning algorithm, and demonstrate empirically that most models that claim disentanglement are sensitive to randomness and the choice of hyperparameters. Moreover, nonlinear Independent Component Analysis (ICA) theory shows that when the function from latent to observed variables is nonlinear, the general problem is ill-posed and it is impossible to recover the independent latents (Hyvärinen and Pajunen, 1999). However, more recent work in nonlinear ICA has shown, with some not very restrictive conditions, identifiability guarantees for deep latent variable models (Hyvarinen and Morioka, 2016; Hyvarinen and Morioka, 2017), with extensions to VAEs with additional observed variables (Khemakhem et al., 2020).

Unsupervised representation learning based on diffusion models suggest that disentanglement is heavily dependent on the expressivity of the generative model (Liang et al., 2024; Liang et al., 2025; Yang et al., 2023b; Kim et al., 2024). In the previous chapter we presented SAMI, a latent variable model where the generative model is given by a conditional diffusion model. SAMI was trained in an unsupervised manner on a generalization of the ELBO objective and learns representations that are more strongly disentangled than traditional VAEs and β-VAEs and near state-of-the-art compared to other diffusion-based latent variable models despite fewer inductive biases. We provided theoretical justifications for where this improvement in disentanglement performance may come from. Our disentanglement proof relied on three major components: the existence of a factorized posterior, multiscale training data, and a KL divergence penalty that implicitly penalizes total correlation. An open question is whether the InfoMax objective, which eschews the KL regularizer for an exact information theoretic rate, still promotes disentanglement.

We prove that under similar assumptions of multiscale data and a factorized posterior, and despite the lack of an implicit total correlation penalty, our model is incentivized to learn a disentangled representation. The theorem is as follows:

Theorem 4.1 (Disentangled hierarchical representation)**.** *Under a reasonable set of assumptions on the data generation process* (Assumption 2)*, the maximally informative code for noisy image x_γ with multi-scale structure is the assignment $z_i = \hat{y}_{k(i)}$ of one latent per feature, and this assignment is preserved across noise levels, with the latents activating in the order $y_K, \ldots, y_1$ as γ increases. The assignment holds up to a permutation and a per-coordinate reparametrization.*

The full proof is given in Appendix C.5.

4.6 Discussion

In this chapter I proposed an unsupervised learning objective based on the InfoMax principle. The I-MMSE relation allows InfoMax to be written in a form that a denoiser-based generative model implements exactly, up to a variational gap that an expressive decoder renders small. Recasting the objective in the language of rate-distortion theory then yields the optimal rate, and I showed numerically that the representations learned under this objective are both informative and disentangled.

The relation between information and MMSE also admits a geometric interpretation, which I used to sketch the biological implications of the objective. Given separate neural populations encoding the likelihood and the prior, the rate can be read either as a constraint on signal power or as an orthogonality constraint on the likelihood bias relative to the prior drift; the two readings are unified by a single constraint on the total power of both populations.

Finally, I proved that for multiscale data and a factorized recognition model, the learned representation is disentangled up to a permutation. Unlike previous disentanglement results for VAEs and for SAMI (Chapter 3), my proof does not appeal to the KL term of the ELBO.

Several directions follow from this work. The compression literature has extended the rate-distortion tradeoff to include *realism*, a constraint on the divergence between the distribution of reconstructions and the data distribution rather than on any pairwise notion of fidelity. Decoders that approximate the data distribution well — diffusion models chief among them — saturate the realism constraint almost by construction. Since my decoder is of exactly this kind, the theoretical relationship between InfoMax and perceptual realism would be interesting to study.

A second direction is empirical. The disentanglement proof assumes that data is multiscale in a specific sense, and that assumption can be tested directly against natural image datasets: measuring the density of feature onsets as a function of SNR would provide one such test.

A third concerns the choice of relevance variable. Here I have taken the data itself as the target, so that reconstruction stands in for information. The information bottleneck (Tishby et al., 2000) shows that the same rate-distortion machinery applies to *any* relevance variable, but the resulting objective is intractable for all but the simplest choices. The I-MMSE relation offers a route to optimizing the *exact* bottleneck objective, and with it a route to relevance variables that are high-dimensional rather than categorical: prediction on natural video, for instance, is a biologically plausible objective that the bottleneck has never been able to accommodate. The same approach would place supervised learning on a more rigorous information-theoretic footing.

5 Hierarchical representations of prior knowledge

5.1 Overview

The principle of efficient coding has served primarily as a model of early vision, but to date no model has used exact InfoMax as an organizing principle for the visual hierarchy as a whole. Here I extend the single-layer InfoMax objective of the previous chapter to a hierarchy of latent variables. In principle, this unsupervised objective supports learning of modular, layer-wise priors and a decomposition of images into a set of conditionally independent generative features. By pairing a hierarchical feedforward recognition model with recurrent and feedback connections, the resulting architecture assigns a functional role to both intra- and inter-layer connectivity and blends discriminative and generative accounts of the visual hierarchy. The model remains unimplemented, but the objective predicts that it should learn layer-wise representations of progressively more abstract variables, each conditionally independent given its parent latent.

I begin with an overview of hierarchy and its signatures in natural images. I then turn to hierarchical structure in the brain and computational efforts to explain it. Next, I present the hierarchical InfoMax objective and show how it might be implemented in a neural network. I close with a short discussion about the proposed model and directions for future work.

5.1.1 Hierarchy in the world

It is an intuitive fact that the world has hierarchical structure. In the physical world, atoms are organized into molecules, molecules into crystals, crystals into rocks, so on and so forth, all the way up to macroscopic objects such as planets and galaxies. But moving downwards from the atom, we find another hierarchy of the microscopic: nuclei, quarks, and other elementary particles. Similar hierarchies exist in biology, social systems, and language. Many people have speculated as to why hierarchies are so ubiquitous, but among the proposed reasons perhaps the most salient comes from Herbert Simon (Simon, 1962):

> On theoretical grounds we could expect complex systems to be hierarchies in a world in which complexity had to evolve from simplicity... The complex forms can arise from the simple ones by purely random processes. Direction is provided to the scheme [of hierarchy] by the stability of the complex forms, once these come into existence. But this is nothing more than survival of the fittest i.e., of the stable. The effect of the existence of stable intermediate forms exercises a powerful effect on the evolution of complex forms.

In essence, complex objects that have hierarchical structure – in the sense that they are comprised of stable intermediate objects – are more likely to persist than complex objects that do not, and thus "win out" in numbers over time. Hierarchy is thus an expected organizing principle of a world where we observe complexity.

We do not, however, encounter this hierarchy directly. Our access to the physical world is mediated through our senses, and our senses register only the end product of a hierarchical process. For example, images are the two-dimensional patterns of light that surfaces cast onto the retina. But images are formed from a hierarchical generative process: the scenes we observe are themselves composed of objects, objects contain surfaces, and surfaces have local pattern of reflectance, and only then, at the final stage, do these result in light intensities. As such, images

have embedded in them a statistical signature of the generative process, and it is only through these statistics that we get access to the hierarchical structure in the world.

5.1.2 Statistical signatures of Hierarchy in Natural Images

There are several notable examples of statistical signatures of natural images arising from hierarchical structure. We mention three that reveal progressively more about the generative structure of the world. In images we observe a power spectrum of $S(f) \propto |f|^{-\alpha}$, where $\alpha \approx 2$ (Field, 1987). This arises from the scale invariance of potential causes, which implies that there is no privileged spatial frequency and the image appears self-similar across scales. Yet the power spectrum is only a second-order statistic; a Gaussian random field with the same spectrum contains no edges, objects, or recognizable structure. While scale invariance tells us that causes span all scales, they do not indicate that they are hierarchically organized.

A stronger signature appears in higher-order statistics. The responses, or coefficients c, of localized bandpass filters applied to natural images, such as those given by Gabor or wavelet filters, are highly non-Gaussian, with sharply peaked, heavy-tailed marginal distributions well fit by a generalized Gaussian $p(c) \propto \exp(-|c/s|^{\beta})$ with $\beta < 1$ (Field, 1987; Mallat, 1989; Olshausen and Field, 1996). This sparsity reflects the fact that scenes are composed of a relatively small number of localized structures, such as edges, contours, and textures, so that any given filter is silent most of the time and responds strongly only when its feature is present. However, while marginal statistics describe each coefficient in isolation and give us a sense of the primitives out of which images are built, they by themselves do not suggest a hierarchy.

The clearest indication of the hierarchical generative process appears in the dependencies between coefficients. Measured in a representation with joint localization in space, scale, and orientation, such as the steerable pyramid (Simoncelli and Freeman, 1995), neighboring bandpass responses are approximately de-correlated, $\mathbb{E}[c_i c_j] \approx 0$, and yet their magnitudes are strongly coupled, $\mathbb{E}[|c_i|\,|c_j|] > \mathbb{E}[|c_i|]\,\mathbb{E}[|c_j|]$ (Wainwright and Simoncelli, 1999; Schwartz and Simoncelli,

2001). A single underlying cause such as an object boundary, or a change in illumination, modulates the amplitude of a neighborhood of coefficients, keeping the second-order structure intact while inducing higher-order dependencies. This shared cause is a sign of hierarchical structure in the world; it is captured by hierarchical latent-variable models such as the Gaussian scale mixture, in which a shared latent variable v scales a vector of Gaussian coefficients, $\mathbf{c} = \sqrt{v}\,\mathbf{u}$ (Wainwright and Simoncelli, 1999).

5.1.3 Difficulties of Inference

To get access to the hierarchical causes of the world from these statistical observations, we must perform inference of the underlying hierarchical causes. But there are two potential difficulties. If the generative process has dense causal structure, exact inference becomes intractable. A posterior over dense causes does not factorize, and disentangling any one cause requires reasoning jointly about all the others it interacts with, at a cost that grows exponentially with the number of causes. Moreover, if the generative function is non-linear – as we know they are – then the inferred structure may be non-identifiable (Hyvärinen and Pajunen, 1999).

We escape the first difficulty by noticing that most causes in the world are only weakly connected with most other things, and that an effective description of reality only requires an account of a tiny fraction of all possible interactions – a notion called the "empty world hypothesis"[1] (Simon, 1962). There is ample evidence to suggest this is true: Lin et al. (2017) points out that the physical generative hierarchy is a sequence of simple, often Markov, local operations. This sparsity is formalized most completely by the renormalization group, under which coarse-graining suppresses all but a few "relevant" couplings, so that the effective description at each scale depends on only a handful of parameters irrespective of the microscopic details (Wilson, 1975). This explains the success of reductionism in the physical sciences; seemingly complex systems can be well-described by the sparse interaction of simpler, constituent parts. Thus when perform-

[1] This hypothesis should give hope to all who strive to build simple theories of natural phenomena.

ing inference, we can build hypotheses assuming a sparsity of causes and interactions between causes.

We overcome the second difficulty in two ways. The non-identifiability statement is predicated on observing independent samples from the world, but our visual system receives a temporally continuous stream of inputs in which the latent causes are non-stationary. This auxiliary signal allows for identifiability even in non-linear settings (Hyvarinen and Morioka, 2016; Hyvarinen and Morioka, 2017). Moreover, we can again invoke the empty world hypothesis to assert that the generative function of images is not an arbitrary nonlinearity; the process of image formation is local, compositional, and sparse in how the latents interact over time (Klindt et al., 2021). Recent work has shown that leveraging sparsity of this kind allows for identifiability (Lachapelle et al., 2022) even when applied to natural data (Klindt et al., 2021).

5.1.4 Hierarchical Structure in the Brain

Biological brains possess structural features that suggest inferential computations may be performed hierarchically. The visual cortex of primates is organized as an anatomical hierarchy, defined by the characteristic laminar patterns of feedforward and feedback connectivity between areas (Felleman and Van Essen, 1991), and the patterns of neural activity along this cortical hierarchy reflect progressively more complex image structure. In primary visual cortex (V1), neurons have receptive fields well described by Gabor filters that are selective for local oriented edges (Hubel and Wiesel, 1962; Jones and Palmer, 1987). In V2, neurons become sensitive to the higher-order statistical dependencies that characterize naturalistic textures, a selectivity that is largely absent in V1 (Freeman et al., 2013). These correspond to the coefficient correlations across scale, orientation, and position that were identified as a signature of hierarchical structure. Area V4 represents intermediate shape features such as curvature and boundary conformation (Pasupathy and Connor, 2001). Meanwhile, at the top of the ventral stream the inferotemporal cortex (IT) represents complex objects and faces, while remaining invariant to changes in position, scale, and

pose (DiCarlo et al., 2012). Each stage untangles object identity from other nuisance factors that obscure it, such that by the end of the hierarchy, identity becomes approximately linearly separable (DiCarlo et al., 2012). Taken together, ascending the ventral stream can be interpreted as an inversion of the generative process that produced the image.

In general, forming a hierarchical model of the world is an economical way to describe a world that is itself hierarchical. Given such a world, a modular and compositional set of representations allows combinatorially many objects to be described in a compact format by recombining a small set of reusable parts. A hierarchy also reorganizes the dependencies among elements: rather than modeling many dense interactions among elements within a single layer, it captures those relations through a shared higher-level element, rendering them conditionally independent. The interactions are not eliminated but rather expressed as a few vertical dependencies rather than many horizontal ones. This allows each description within each layer to be simpler and sparser. Finally, a hierarchy allows grouping of features by scale, so that an object can be described to any desired level of granularity without invoking lower level detail, allowing the description to be as economical as the task demands.

5.1.5 Hierarchical Models of Visual Processing

If hierarchy is such an economical way to represent a hierarchical, sparse world, it is unsurprising that it has become a central organizing principle in models of vision. These models fall into two broad families, reflecting two conceptions of inference (Peters et al., 2024): either as *discriminative* inference, which map an image bottom-up onto the behaviorally relevant variables, or as *generative* inference, in which sensory evidence is evaluated against a generative model of how the image was formed.

The discriminative conception treats vision as a largely feedforward transformation that progressively discards nuisance variation in an image for the purpose of formatting it for a downstream behavior rather than for reconstruction. This idea has a long lineage, starting with the

simple-to-complex cell scheme of Hubel and Wiesel (1962) and the neocognitron of Fukushima (1980); but across all these models they utilize a motif of alternating feature selectivity and pooling that builds representations of increasing complexity and invariance (Riesenhuber and Poggio, 1999; Serre et al., 2007; Lecun et al., 1998; Krizhevsky et al., 2017). Of these, the models that learn their features from natural images are the best models of cortex. Such models with learned intermediate representations account for a range of ventral-stream selectivities (Serre et al., 2007), and, somewhat surprisingly, deep networks optimized purely for object recognition, with no fitting to neural data, turn out to be the best available predictors of population responses in V4 and IT cortex (Yamins et al., 2014).

The generative conception casts vision instead as the inversion of an internal generative model of images, based on Helmholtz's account of perception as unconscious inference (Helmholtz, 1852). Sensory evidence is evaluated against hypotheses about the latent causes that could have produced it, and perception corresponds to the causes that are deemed most probable. In its hierarchical form, higher cortical areas encode beliefs about more abstract causes and send top-down predictions to lower areas, which return the residual mismatch (Mumford, 1992; Lee and Mumford, 2003). Rao and Ballard's predictive coding model provides further biological support: a hierarchical network trained to minimize prediction error on natural images develops simple-cell-like receptive fields, and the units carrying residual error reproduce extra-classical receptive-field effects that a purely feedforward account struggles to explain (Rao and Ballard, 1999). The corresponding lineage in machine learning, beginning with the Helmholtz machine (Dayan et al., 1995) and Wake-Sleep algorithm (Hinton et al., 1995), has led to hierarchical latent-variable models, such as the Ladder VAE (Sønderby et al., 2016) and other deep VAEs (Vahdat and Kautz, 2021), in which each level of latents generates the level below.

In practice these two conceptions are less distinct than one might initially suspect (Peters et al., 2024), and one way to bridge them is through amortized inference. A variational autoencoder is generative in its philosophy, positing an explicit model of how images arise, yet the recognition

network that performs inference is a feedforward, discriminative map from image to latent estimate, trained to approximate the inversion of the generative model rather than on a behavioral objective. Predictive coding blurs the line from the other direction, combining top-down generative prediction with bottom-up error signals that are themselves a discriminative readout of prediction mismatch. A model can therefore be generative in aim while remaining discriminative in its mechanism.

The model, which I present below, sits in this middle ground. Its aims and mechanisms blend both philosophies: where the VAE elevates the generative decoder and treats the encoder as an amortized approximation to the posterior, an InfoMax objective makes the encoder itself the target, maximizing the mutual information $I(X;Z)$ carried by the recognition model. Meanwhile, my model possesses recurrence and feedback connections from higher to lower layers, suggesting a generative view of inference. This places the objective on the discriminative side of vision, even as the learned representation captures the generative causes of sensory input.

5.2 Hierarchical InfoMax

The goal of this model is to extend the principle of InfoMax to learn a *hierarchy* of representations rather than a single latent layer. Much like the single layer model presented in the previous chapter (Chapter 4), the nature of our solution is a feedforward encoder paired with a denoiser-based generative model that aims to maximize the mutual information between the representation and inputs. As we shall see, it is jointly optimized on a generalization of the InfoMax objective to multiple hierarchically organized latent variables.

5.2.1 Objective

We investigate the simplest form of hierarchy where the generative model possesses two latent variables Y and Z. For images X, we assume a Markov generative structure $Z \to Y \to X$,

as in Ladder VAEs (Sønderby et al., 2016). Our goal is InfoMax, so we wish to maximize the mutual information between the joint distribution over the latents (Y, Z) and the observation X, given by $I(X; \{Y, Z\})$. Since we consider a stationary distribution, we can assume the entropy of X is fixed, and simply minimize the conditional entropy $h(X \mid \{Y, Z\})$. In the parlance of rate-distortion, this is our *distortion* term.

As before, we formalize the resource constraint of efficient coding (the *rate*) as the mutual information $I(X; \{Y, Z\})$. Altogether, we aim to minimize the objective

$$\mathcal{L} = h(X \mid \{Y, Z\}) - \beta\, I(X; \{Y, Z\}).$$

A quick sanity check that this is the right generalization of the single layer InfoMax objective is by defining a substitute variable $U := \{Y, Z\}$; writing the above objective with U instead of $\{Y, Z\}$ recovers the single layer objective.

Under a Markov generative model, the conditional entropy reduces to

$$h(X \mid \{Y, Z\}) = h(X \mid Y).$$

The rate, meanwhile, corresponds to a constraint on the recognition model, which we do not assume to be Markovian. Thus we can decompose the mutual information constraint as

$$I(X; \{Y, Z\}) = I(X; Z) + I(X; Y|Z). \tag{5.1}$$

Thus we minimize the objective

$$\mathcal{L} = h(X \mid Y) + \beta \left(I(X; Z) + I(X; Y \mid Z) \right).$$

5.2.2 Implementation

Model parameterization There is a large family of models that can implement this objective. To constrain this family, we must make several modeling choices. One such choice is to follow the philosophy laid out in Chapter 2, where we represented local priors using attractor dynamics. This suggests parameterizing each prior over X and Y with a separate denoiser. The idea is to learn an attractor in the space of Y such that inference occurs by integrating prior dynamics with bottom-up and top-down sensory signals, which take the form of guidance scores.

For the recognition model, we choose a particular parameterization that is inspired by the hierarchical encoder in the Ladder VAE (Sønderby et al., 2016). It specifies a posterior over the joint latents $p(y, z|x)$ that is factorized in a "top-down" manner,

$$p_\phi(y, z \mid x) = p_\phi(z \mid x)\, p_\phi(y \mid x, z). \tag{5.2}$$

We adopt a variational approach to the recognition model such that all latent variables are encoded using a single feedforward neural network with parameters ϕ. The neural network parametrizes the following maps: (i) deterministic maps $x \to y_d$ and $y_d \to z_d$, (ii) linear decoder heads $y_d \to (\mu_y, \Sigma_y)$ and $z_d \to (\mu_z, \Sigma_z)$, and (iii) a linear encoder head $y \to y_d$. Through parameter sharing, we can express the map $p(z|x)$ using the combination of maps $x \to y_d \to z_d \to (\mu_z, \Sigma_z)$. The likelihood $\mathscr{L}(y; z)$ is given by the maps $y \to y_d \to z_d \to (\mu_z, \Sigma_z)$. Meanwhile, the map $p(y \mid x, z)$ is never expressed other than in through score form $\nabla_{y_\eta} \log p(y_\eta \mid x, z)$; we will see later how this is expressed using the encoder maps.

For the denoisers, we assume that both noising operators follow $x_\gamma = \sqrt{\gamma}x + \epsilon_x$ and $y_\eta = \sqrt{\eta}y + \epsilon_y$ where ϵ_x and ϵ_y are drawn from isotropic Gaussians with the dimensionality of x and y respectively.

Computing the distortion The distortion is the conditional entropy $h(X \mid Y)$, which is given by an integral over MMSE terms according to the conditional I-MMSE relation:

$$\begin{aligned} h(X \mid Y) &= \frac{1}{2} \int_0^\infty \mathbb{E}_{p(x,x_\gamma,y)} \left[\left\| x - \hat{x}(x_\gamma, \gamma, y) \right\|^2 \right] d\gamma + \text{const.} \\ &= \frac{1}{2} \int_0^\infty \mathbb{E}_{p(x,x_\gamma,y)} \left[\left\| x - \hat{x}(x_\gamma, \gamma) - g_\gamma \nabla_{x_\gamma} \log p_\phi(y|x_\gamma) \right\|^2 \right] d\gamma + \text{const.} \end{aligned}$$

where the constant term is independent of the conditional denoiser, and $g_\gamma = 1/\sqrt{\gamma}$ is a γ-dependent scalar that appropriately weighs the two terms according to Bayes' rule. This is essentially the same formulae as Eqs. 4.4 and 4.7 in Chapter 4, but with z replaced by y. We compute the y-posterior score with respect to x_γ via auto-diff. As before, the marginal distortion is defined as the expected MMSE at SNR γ:

$$\partial_\gamma h(X_\gamma \mid Y) := \mathbb{E}_{p(x,x_\gamma,y)} \left[\left\| x - \hat{x}(x_\gamma, \gamma) - g_\gamma \nabla_{x_\gamma} \log p_\phi(y|x_\gamma) \right\|^2 \right]. \tag{5.3}$$

Computing the rate We can now invoke the I-MMSE relation to show that the marginal forms of the rate terms in Eq. 5.1 are

$$\begin{aligned} \partial_\gamma I(X_\gamma; Z) &= \mathbb{E}_{x,x_\gamma,z} \left[r_\gamma^2 \left\| \nabla_{x_\gamma} \log p(x_\gamma|z) \right\|^2 \right] := \mathbb{E}_{x,x_\gamma,z} \left[r_\gamma^2 \left\| \mathrm{g}_{x_\gamma|z} \right\|^2 \right] \\ \partial_\eta I(X; Y_\eta|Z) &= \mathbb{E}_{x,y,y_\eta,z} \left[r_\eta^2 \left\| \nabla_{y_\eta} \log p(x|y_\eta, z) \right\|^2 \right] := \mathbb{E}_{x,y,y_\eta,z} \left[r_\eta^2 \left\| \mathrm{g}_{x|y_\eta,z} \right\|^2 \right]. \end{aligned}$$

The derivation is given in Appendix D.1. We know how to compute the first marginal rate term $\partial_\gamma I(X_\gamma; Z)$ from the previous chapter: given a fixed denoiser $\nabla_{x_\gamma} \log p(x_\gamma)$, we can minimize the squared norm of the guidance score $\nabla_{x_\gamma} \log p(z|x_\gamma)$. The second marginal rate term is trickier.

We can use the following identities

$$\nabla_{y_\eta} \log p(y_\eta|x,z) = \nabla_{y_\eta} \log p(x|y_\eta,z) + \nabla_{y_\eta} \log p(y_\eta|z) = \nabla_{y_\eta} \log p(z|y_\eta,x) + \nabla_{y_\eta} \log p(y_\eta|x) \tag{5.4}$$

that fall out from Bayes' rule. Rearranging, we get

$$\begin{aligned}\nabla_{y_\eta} \log p(x|y_\eta,z) &= \nabla_{y_\eta} \log p(y_\eta|x) + \nabla_{y_\eta} \log p(z|y_\eta,x) - \nabla_{y_\eta} \log p(y_\eta|z) \\ &= \nabla_{y_\eta} \log p(y_\eta|x) + \nabla_{y_\eta} \log p(z|y_\eta,x) - \nabla_{y_\eta} \log p(z|y_\eta) - \nabla_{y_\eta} \log p(y_\eta)\end{aligned}$$

where in the second line we have again used Bayes' rule.

Let us consider each of these four terms in turn. Since the mapping $x \to y$ is a Gaussian and the forward noising operator $p(y_\eta|y)$ is also Gaussian, the distribution $p(y_\eta|x) = \int p(y|x)p(y_\eta|y)\,dy$ is given by the Gaussian $p(y_\eta|x) = \mathcal{N}(\sqrt{\bar{\alpha}_\eta}\mu_{y|x}, \bar{\alpha}_\eta\Sigma_{y|x} + (1-\bar{\alpha})I)$. Thus $\nabla_{y_\eta} \log p(y_\eta|x)$ has a closed-form solution:

$$\nabla_{y_\eta} \log p(y_\eta|x) = -\frac{y_\eta - \sqrt{\bar{\alpha}_\eta}\mu_{y|x}}{\bar{\alpha}_\eta\Sigma_{y|x} + (1-\bar{\alpha}_\eta)}.$$

Next, we know how to compute $\nabla_{y_\eta} \log p(y_\eta)$ using the unconditional y-denoiser, and w can calculate the guidance score $\nabla_{y_\eta} \log p(z|y_\eta)$ by taking the gradient of the log likelihood of $z(y)$ with respect to y_η, using auto-diff of the mapping $y_\eta \to z_d \to (\mu_{z|y}, \Sigma_{z|y}) \to z$ where we sample $z(y_\eta) \sim \mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$.

We compute the last term $\nabla_{y_\eta} \log p(z|y_\eta,x)$ by taking the gradient of the log likelihood of $z(x)$ – the z that comes from x – with respect to y_η, using auto-diff of the same mapping $y_\eta \to z_d \to (\mu_{z|y}, \Sigma_{z|y}) \to z$ where we sample $z(y_\eta) \sim \mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$.

The rate is the L2 norm of the sum of all four scores.

Training The loss is computed in expectation over γ using the marginal form of the loss, which is defined as

$$\partial_\gamma \mathcal{L} = \partial_\gamma h(X_\gamma \mid Y) + \beta \left(\partial_\gamma I(X_\gamma; Z) + \partial_\eta I(X; Y_\eta \mid Z) \right), \qquad \mathcal{L} = \int \partial_\gamma \mathcal{L} \, d\gamma.$$

Conditional generation In order to perform conditional generation (conditioned on the latents of a clean target image x), we follow the Ladder VAE prescription of sampling from the posteriors in turn: first we sample $z \sim p(z|x)$, then sample from $p(y|x,z)$ via the conditional diffusion model. Tweedie's formula tells us how to get the γ-specific y estimate in terms of the score:

$$\hat{y}(y_\eta, \eta, x, z) := \mathbb{E}[y|y_\eta, x, z] = y_\eta + \gamma \, \nabla_{y_\eta} \log p(y_\eta|x,z).$$

Since we know how to compute all four terms on the RHS of Eq. 5.4, we can compute the score $\nabla_{y_\eta} \log p(y_\eta|x,z)$. We get a clean sample of y by performing iterative denoising using the DDPM reverse process (Ho et al., 2020).

5.3 Discussion

We have generalized the single-layer InfoMax objective of the previous chapter to a hierarchy of latent variables. This unsupervised objective can be used to learn modular, layer-wise priors that combine with bottom-up sensory and top-down contextual signals to infer local latent variables.

While we have considered the simplest hierarchy of two latent variables, we can easily generalize the model to three or more latent variables. Because of the Markov assumption between the latent variables, each latent variable only depends on its neighbors and hence the number of mappings scales linearly with the number of latent variables. Moreover, we see that the global rate-distortion objective is a series of self-similar local rate-distortion objectives, with each layer

aiming to reconstruct its own latent variable under rate constraints.

The modularity of the objective is inspired by the modular neural circuit account of prior representation and inference proposed in Chapter 2. The decomposition of the objective into a series of local reconstruction objectives suggests that the role of feedback connectivity in cortex is not only for top-down guidance during inference, but also for learning via reconstruction. This has two implications. First, it avoids the common criticism levied against reconstruction as a learning objective, which is that the brain does not possess the requisite anatomical structure for comparing reconstructed against sensory evidence. In contrast, *local* reconstruction requires only the transmission of signals between cortical layers rather than to the retina. Second, the re-use of the same circuitry for both learning and inference suggests either a bi-phasic alternation of these two functions or a separation of the two over different timescales. I have both explored both possibilities in the context of a VAE in a separate paper (that is not part of this thesis) (Bredenberg et al., 2021) that considers a biologically-plausible generalization of the Wake-Sleep algorithm (Hinton et al., 1995). A promising and interesting direction would be to extend this work to the denoising regime.

Beyond implementing the model and training it on toy data, natural next steps include scaling to naturalistic datasets and characterizing the representations the model learns. Remarkably little work has examined what the layers of a hierarchical VAE come to represent, as opposed to whether a single latent space is disentangled; Csikor et al. (2022) is a notable exception. In theory, the diagonal encoder covariance should encourage the model to learn *conditionally disentangled* latents. Verifying this will require both empirical and theoretical work.

6 Discussion

6.1 Overview

This thesis began with a question that the Bayesian account of perception raises but does not answer. If perception is inference, and inference requires a prior, then the prior must be something that is physically present in the brain: represented by neurons, used within the time available for a percept, and acquired from experience. The cognitive literature has treated the prior as a term in an equation, but once we ask where that term lives inside the brain, the existing theories thin out quickly. No single model provides an explanation for the sort of priors that have been measured behaviorally (Houlsby et al., 2013). A few models capture priors over scalars such as orientation or contrast (Laughlin, 1981; Ganguli and Simoncelli, 2014). Circuit models of sampling tell us how a distribution might be represented in activity, but the distributions they represent are, in practice, Gaussian or close to it (Hennequin et al., 2014; Aitchison and Lengyel, 2016). Neither scales to the objects that actually generate our sensory data: nonlinear manifolds of high dimension, with structure at many scales and mass concentrated far from any single mode (Fefferman et al., 2016; Brown et al., 2023).

The wager of this thesis is that the gap is not necessarily conceptual but technical, and that it is accounted for by a specific class of models. Diffusion models learn complex priors over the kind of variables the brain must represent, and they do so through a single operation, denoising. Section 1.5.3 laid out three implications of this operation: used iteratively, it corresponds to

stochastic attractor dynamics; used in conjunction with latent variable models, it can learn a multiscale representation; seen through the lens of the I-MMSE relation, it can provide an estimate of mutual information.

Each perspective corresponded to a research chapter. Chapter 2 asked what a circuit that ran such stochastic attractor dynamics would look like. Chapter 3 asked what kinds of representations one could learn when the diffusion and VAE objectives are written in a common form. Chapters 4 and 5 used the connection to information theory to turn InfoMax from a principle into an objective that can be optimized exactly.

Read together, the chapters divide into two topics. Chapter 2 concerns the *use* of a prior that is already in place, while Chapters 3 through 5 concern how such a prior might be *acquired*. The two halves are connected: the prior that Chapter 2 assumes is carried by the weights of a denoiser, and the objectives in Chapters 4 and 5 train such a denoiser.

This thesis addresses many of the difficulties in learning priors raised in Section 1.5, but Table 6.1 is explicit about which difficulties this work addresses and which it does not.

6.2 Inference of complex posteriors in neural circuits

Chapter 2 proposed that a prior over a complex, high-dimensional latent space is stored in the recurrent weights and dendritic nonlinearities of a cortical circuit, and is expressed as samples by the circuit's stochastic dynamics. An ongoing local oscillation supplies the noise schedule; over the course of this oscillation the effective noise anneals, and at its peak the circuit's state corresponds to a prior sample. Inputs, whether bottom-up sensory evidence or top-down context, enter as an additive bias that redirects the same dynamics from the prior to the corresponding posterior. Across tasks, the circuit does not change aside from what is fed into it.

Set against existing accounts of neural priors, this proposal occupies a position that was previously empty. In the tuning-curve accounts, the prior is implicit in a population's preferred fea-

tures (Ganguli and Simoncelli, 2010; Ganguli and Simoncelli, 2014), but each neuron's preferred feature is static and the number of neurons required scales exponentially with the dimension of the feature space. In the sampling accounts, the prior can in principle be high-dimensional, since dimensionality scales with neuron count (Fiser et al., 2010; Berkes et al., 2011). But circuit-level implementations of sampling trade expressiveness for tractability, and the distributions they sample are simple enough that a parametric representation would have served (**chen.etal.2023**). Chapter 2 keeps the explicitness of the sampling accounts and drops the restriction on the distribution.

While I concede that not all details of my model are necessarily the ones implemented by neural circuits, there are several properties that address fundamental difficulties faced by the brain.

Speed. Estimating a posterior expectation requires accumulating samples, so sampling speed is a hard constraint on any sampling account of neural computation (Rullán Buxó and Savin, 2021; Savin and Denève, 2014; Masset et al., 2022). Markov chain methods pay for the complexity of the target in mixing time, and the cost is worst exactly for the cases that motivate sampling in the first place: multi-modal distributions with low probability "holes", and distributions concentrated on nonlinear manifolds. The reverse process of diffusion models does not have this problem by construction. Each run is independent of the last, and the annealing schedule blurs the distribution at low SNR so as to bridge the low probability barriers. The circuit implementation of Chapter 2 preserves this property.

Complexity. The dendritic nonlinearity and the recurrent dynamics give us expressivity. Because the reverse operator is a nonlinear function of the population state, the represented prior can be an arbitrary distribution on a low-dimensional nonlinear manifold embedded in a high-dimensional space. This is the first circuit-level account I am aware of in which the class of representable priors is not the binding constraint on the model.

Reuse. The same weights serve every task, and a task is specified entirely by the bias. This matches the behavioral observation that humans carry priors across tasks rather than relearning a stimulus-response map for each one (Houlsby et al., 2013; Turk-Browne et al., 2010), and it is the property the introduction argued an agent under epistemic uncertainty must have. It also means the model makes a prediction that is unusual for a circuit model: the spontaneous statistics should be *invariant* across tasks that share a sensory domain, with all task dependence confined to the evoked condition.

6.2.1 What the circuit model does not settle

The proposal is an existence argument, not a uniqueness argument. Dendritic nonlinearities are one way to implement the reverse operator, but each timestep could equally be carried out by a small feedforward subnetwork with recurrent interaction between subnetworks (Suppl. A.5). That alternative hypothesis constrains the architecture less but demands that noise be injected uniformly across the network rather than confined to a single circuit, which is not, on the face of it, what the anatomy suggests. I do not think the computational results distinguish these options. A model of learning might, since the two architectures assign credit differently, but one would need to measure the noise correlations across the brain to rule one out over the other.

Three further gaps deserve naming. First, the schedule is exogenous. I have taken the oscillation as given, but I do not suggest where the schedule comes from (though I speculate that this type of modulation could be implemented by local inhibitory subunits), or how it would adapt when the sensory evidence is ambiguous or noisy. The tempered-transitions view (Savin et al., 2014; Neal, 1996) suggests one possibility: if β is an inverse temperature, then the schedule is a free parameter that a circuit could in principle tune, and an adaptive schedule would predict oscillation frequency or amplitude to covary with sensory uncertainty. Second, the readout is unaccounted for. If samples arrive at a particular phase, some downstream mechanism must know to read at that phase, and I have not defended this assumption of a phase-dependent readout.

Third, the conditioning operation is stipulated rather than derived. I have assumed the likelihood enters additively, and in Chapter 4 I provide a normative reason as to why an additive bias orthogonal to the prior drift is rate-optimal, but I have not joined these two arguments into a single claim.

6.3 InfoMax Learning and Representations

The second half of the thesis asks how the brain learns a prior. I argued that an animal that does not know which tasks it will face should preserve as much information about its input as its resources allow, since information discarded during encoding cannot be recovered later. This calls for learning a representation via InfoMax.

Until recently, the mutual information $I(X;Z)$ had no closed form outside restrictive cases. Estimators are data-hungry and biased, and so the principle has only been implemented through proxies or under assumptions in service of analytic tractability. The contribution of Chapters 4 and 5 is to remove that obstacle: the I-MMSE relation (Guo et al., 2005) turns the quantity a denoiser is already trained to minimize into the derivative of the information it has captured, so InfoMax can be optimized exactly rather than approximated.

The three chapters trace this idea through three formulations.

Chapter 3: the generative account. SAMI optimizes an exact ELBO in which both the clean and the noisy image are encoded into probabilistic latents. This dual encoding allows the inference network to predict the log-likelihood of clean features while conditioned on noisy observations. This has several benefits: it induces the representation to separate signal from noise, and to calibrate posterior variance against the irreducible uncertainty in the data. The "latent-hole problem" that afflicts VAEs (Rezende and Viola, 2018; Xu et al., 2020) is also provably and empirically mitigated. Moreover, the learned representation is both sparse and disentangled across features,

Difficulty (§1.5)	Where addressed
(1) Priors over high-dimensional latents	Ch. 2–5: the prior is learned by a denoiser
(2) Statistics evolve over several timescales	Partially: see §6.3.2
(3) Efficient coding assumes low noise and deterministic responses	Ch. 4: the rate is defined for a stochastic encoder
(4) The mechanism of learning must be specified	Partially: Ch. 5 decomposes the objective layer-wise, but biologically plausible local rules remain to be derived
(5) Efficient coding is a theory of early processing only	Ch. 4: InfoMax implemented using a deep convolutional network as the recognition model; Ch. 5: extended InfoMax to a hierarchy of latents
(6) Neither efficient coding nor Bayesian inference says which latents to learn	Ch. 4: the relevance variable is the input itself, and the rate does the selection
(7) Learning the prior requires inference, which requires the prior	Partially: taking the data as its own target avoids the uninformative fixed point, but posterior collapse is still possible

Table 6.1: The obstacles to learning priors as set out in Chapter 1, and where each is addressed.

allowing for automatic straightening of natural videos (Hénaff et al., 2019). The disentanglement result follows from three ingredients: multiscale structure in the data, a factorized encoder, and the total correlation penalty implicit in the ELBO.

Chapter 4: the information-theoretic account. This model generalizes on the previous account by leveraging the relationship between denoising and mutual information. By recasting this problem in rate-distortion terms we derive the optimal rate. The disentanglement result that then follows is stronger than the one in Chapter 3 in a specific way: it does not appeal to the KL term at all, only to the multiscale structure of the data and the factorization of the recognition model. If the result holds, then disentanglement in diffusion-based representation learning is simply a consequence of the data's structure interacting with the structure of the recognition model under information maximization.

Chapter 5: the hierarchical reading. The single-layer objective generalizes to a hierarchy of latents, and does so in a way that decomposes into layer-wise rate and distortion terms. The intermediate layer learns a local prior that combines bottom-up evidence with top-down context to infer its own latents. This ties the two halves of the thesis together: the modularity is a reflection of the circuit proposal in Chapter 2, and the objective is an attempt to say how this circuit proposal would learn its prior.

6.3.1 What this buys

The efficient coding as formulated by Barlow (1961) had several restrictions, which my model addresses. The encoding is stochastic, so the low-noise assumption can be dropped. The variable is high-dimensional and nonlinearly embedded, so the theory contends with the latents of natural images rather than with scalar contrast or orientation. And the code is hierarchical, so efficient coding becomes a principle for the visual system rather than for its first stage.

It is true that many machine learning models learn hierarchical representations of naturalistic images; these are thus far the best fit to neural data (Yamins et al., 2014; Schrimpf et al., 2020). Two differences are worth drawing out.

The first is the absence of a task. Those models are typically trained on a supervised or self-supervised objective, and the resulting representation is sensitive to the choice of task or data augmentations, which induces the representation to discard nuisance variables. This requires the practitioner to make a set of arbitrary decisions that are not well founded in theory. My account approaches the problem of learning a hierarchical set of features without making such decisions.

The second is anatomical. The models that best fit neural data are feedforward, but the cortex is not: feedback and lateral connections outnumber feedforward ones at every stage of the visual hierarchy (Felleman and Van Essen, 1991), and yet their computational purpose is still unclear. My models have both pathways, and offers a functional role for each: top-down input is the

conditioning that turns a layer's local prior into a local posterior, and lateral recurrence is the iterative denoising operator that samples from it. A feedforward network, by contrast, has no prior to condition on and no local distribution to sample.

6.3.2 Open questions

Learning rules. Chapter 5 decomposes the objective into layer-wise rate and distortion terms, which is a precondition for locality but not locality itself. Deriving a canonical local rule that implements those terms, and asking whether it can be expressed in quantities a synapse has access to, is the natural next step and the one that would most strengthen the link to Chapter 2.

Some of this work is already done. Bredenberg et al. (2021), on which I was also an author, derives a local learning rule for VAEs by generalizing the wake-sleep algorithm, and the layer-wise decomposition of Chapter 5 supplies the structure such a rule needs. What it does not cover is the denoiser. A denoising objective is not a wake-sleep objective, and it carries a difficulty of its own: as written it requires clean samples as targets, which may not be biologically plausible.

This is less damaging than it first appears, because the same problem has been solved elsewhere. Stein's unbiased risk estimator gives an unbiased estimate of the mean squared error to a clean target using only the noisy observation and the divergence of the estimator (Stein, 1981), which has been used both to train denoisers without ground truth (Metzler et al., 2020), and, in the empirical Bayes tradition, to derive optimal estimators without access to a prior at all (Raphan and Simoncelli, 2011). Whether a SURE-like quantity can be computed locally, and what its divergence term would correspond to in a dendrite, is an open problem left by this thesis. Dendritic plasticity is the obvious place to look for the substrate, given that the dendrite is where the reverse operator lives in the circuit model.

A related worry concerns reconstruction itself. I have used reconstruction as the distortion term throughout, but read literally as a claim about anatomy it is hard to defend. No pathway returns fully from visual cortex to the retina, so there is nothing in the brain that is positioned

to reconstruct the image. The objective is better understood as applying between cortical stages than to the sensory periphery. A cortical stage reconstructing the activity of the stage below, or of a neighboring area, requires only the feedback connectivity that already exists, and all of the information-theoretic justification can be readily transferred; the data processing inequality holds along any Markov chain, and nothing privileges the retina as its source. In this view, the distortion at each layer is defined against that layer's input rather than against the world.

The relevance variable. I have taken the data as its own target, so that reconstruction stands in for information. The information bottleneck (Tishby et al., 2000) shows that the same machinery accommodates any relevance variable, but the resulting objective is intractable for all but the simplest choices, which is why the bottleneck literature has been confined to categorical or scalar targets. The I-MMSE route offers a way to optimize the exact bottleneck objective for high-dimensional relevance variables. Prediction on natural video is a promising case. Prediction is arguably an even more biologically plausible objective for sensory systems (Palmer et al., 2015), and a video frame is exactly the sort of relevance variable that the information bottleneck has not been able to accommodate.

Realism. The compression literature has extended the rate-distortion tradeoff to include realism, a constraint on the divergence between the distribution of reconstructions and the data distribution rather than on any pairwise fidelity. Diffusion-based decoders, such as mine, saturate the realism constraint almost by construction, so the theoretical relationship between InfoMax and perceptual realism are worth investigating. I suspect it may be informative as to why the perceptual quality of these models does not degrade under compression.

Perceptual effects. We can test the model by trying to produce behavioral effects. Wei and Stocker (2015) showed that perceptual biases for orientation move *away* from the peak of the prior rather than toward it. This effect comes about because the likelihood is sufficiently asymmetric

in stimulus space so as to overwhelm the pull of the prior, so an observer is repelled from the values it encounters most often.

Wei and Stocker assume efficient allocation, which my model encourages through a constraint on rate. The test is whether an encoder trained on a scalar variable induces a likelihood with the right asymmetry, and whether the resulting estimates reproduce the measured biases in size and not only in sign. A negative result would be as informative, since it would implicate the resource constraint: their allocation follows from a budget on Fisher information and mine from a budget on the rate, and the two may differ.

Timescales. The statistics of the environment are inherited over evolutionary time, instantiated during development, and refined across a lifetime (Seriès and Seitz, 2013), whereas the objective has a single learning process running at a single timescale.

The multiscale disentanglement result is a start, since it is a statement about features organizing by scale, and empirically that organization extends to time: when fed frames of natural videos, SAMI recovers features ordered by both spatial and temporal scale due to the innate correlation between the two scales in natural statistics. Whether larger, coarser features are learned first during training, however, is left as future work.

The combination I would like to try is the bottleneck with prediction as the relevance variable, applied to the hierarchical model. Prediction sorts features by how long they stay informative and the hierarchy sorts them by abstraction, so together they should sort features by the timescale over which each is worth learning at all. My expectation is that regularities holding across all natural video, and therefore across the whole of an organism's experience, would settle into the early layers, while features that hold only within a context would be confined to the late ones. That would yield a correspondence between depth and timescale of acquisition, and with it a prediction about where in the visual system to look for structure that is developmentally fixed as opposed to experience-dependent. It would not explain why the layers differ in plasticity, but it

would say which ones ought to.

Conditional disentanglement. The diagonal encoder covariance should, in theory, push the hierarchical model toward *conditionally* disentangled latents, that is, latents that are independent given the layer above rather than marginally independent. Verifying this needs both empirical work on naturalistic datasets and theoretical work on what the layer-wise objective identifies. Remarkably little is known about what the layers of a hierarchical VAE represent, as opposed to whether a single latent space is disentangled.

6.4 Experimental predictions

There are several experimental predictions that come out of my models.

Phase-resolved sampling The strongest prediction concerns the relationship between spontaneous and evoked activity. Berkes et al. (2011) showed that the distribution of spontaneous activity in ferret V1 comes to match the average evoked distribution over development, and used the KL divergence between the two as a measure of calibration. My model predicts that this match is also *phase-dependent* over the course of inference. Spontaneous activity is a sample from the prior only at the peak of the oscillation, at the end of a reverse run; at other phases the state is a partially denoised intermediate and its distribution is a smoothed version of the prior. The prediction is therefore that the Berkes analysis, repeated with spikes segregated by the phase of the local oscillation, should show a systematic modulation of the KL across phase, with a minimum at one particular phase.

Two secondary predictions follow from the same schedule. Across-trial variability should decline systematically within a cycle, since the effective noise is annealed, so Fano factors computed in a phase-resolved manner should be highest early in the cycle and lowest at the readout phase. And samples should be close to independent across cycles but strongly correlated within them,

so the autocorrelation of the population state should show a sharp drop at cycle boundaries.

Dendritic manipulation The expressiveness of the represented prior is carried entirely by the dendritic nonlinearity. This yields a prediction that is unusual in being about higher-order statistics rather than tuning: suppressing dendritic nonlinearity, pharmacologically or genetically, should preferentially degrade the structure of the spontaneous distribution while leaving mean rates and first-order tuning comparatively intact. Concretely, the spontaneous distribution should become more Gaussian and less concentrated on a low-dimensional manifold. A linear circuit samples from a Gaussian, and this is the closest thing the model has to a falsification test.

Synthesizing illusions We can also make behavioral predictions using psychophysics. An illusion is a percept in which the prior has overruled the sensory evidence. Given a model with a trained hierarchical prior, we can search for stimuli on which the posterior departs maximally from the likelihood, which should correspond to images impoverished or ambiguous enough that the sampled posterior is dominated by the structure of the prior rather than by the input. If the model prior resembles the observer's, those stimuli should be misperceived in the direction the model predicts. The light-from-above illusion of Fig. 1.1 is the kind of illusion that such a procedure ought to detect. The interesting case is the generation of illusions that no one has yet cataloged.

Hierarchical priors allow for the generation with respect to a cortical depth. Fixing the state at one level while sampling the levels below yields stimuli whose prior-driven distortion originates at a known layer. If we believe that different layers capture features at different spatial and temporal scales, we should be able to predict the scale at which the illusions fool human observers.

Predictions from the InfoMax account The computational models of learning make less direct contact with neural data, but three predictions are possible.

The geometric reading of the rate in Chapter 4 predicts that the likelihood bias should be, on average, orthogonal to the prior drift. In population terms, the displacement of the evoked state relative to the matched spontaneous state should have little component along the direction the spontaneous dynamics are already moving. This is measurable with the tools used to identify task-relevant subspaces in prefrontal cortex (Mante et al., 2013), and the same analysis would test the reading of the rate as a constraint on the total power of the likelihood population.

The multiscale assumption underlying the disentanglement proof is testable directly against data rather than against brains: measuring the density of feature onsets as a function of SNR in natural image datasets would establish whether the assumption required by the proof holds for the signals the visual system actually encounters.

Finally, conditional disentanglement makes a prediction about noise correlations. If latents within a layer are conditionally independent given the layer above, then a substantial share of the noise correlation within a cortical area should be attributable to shared top-down input, and correlations conditioned on the state of the higher area should be markedly weaker than unconditioned ones. Simultaneous recordings across two areas, with the higher area used as the conditioning variable, would test this.

6.5 Concluding remarks

In the introduction, I observed that neuroscience has sought to contain the complexity of the brain by studying animals doing simple tasks, and that large-scale recording in naturalistic settings has removed that constraint. Simultaneously, the theories of neural priors that I build on were built for this simpler regime. They describe scalars, often assume low noise, and concern early stages of visual processing. Yet none of these restrictions are essential to the underlying principles.

This thesis presents a set of models that chip away at these restrictions. A denoiser gives us

a prior over signals of arbitrary complexity, dynamics that sample from it quickly, and, through the I-MMSE relation, a way to say exactly how much information a representation carries and what it costs. By providing a computational account of how priors are learned and represented in the brain, I provide some answers to the questions posed by Helmholtz and Barlow. As with any theory, whether the brain builds priors in the way I have described is now an empirical matter.

A | Appendix: A circuit model for representing complex priors and performing flexible inference

A.1 Posterior sampling dynamics

We modify the original posterior flow field derivation by Sohl-Dickstein et al., 2015b to allow for a biologically plausible posterior sampling algorithm. We wish to derive the *posterior neural transition operator* $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y})$ by combining the prior neural transition operator $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ with an arbitrary likelihood signal $p_\phi(\mathbf{y} \mid \mathbf{x}_t)$. Recall that our prior neural transition operator is Gaussian:

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_\theta, (1-\beta_t)I)$$
$$\log p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = -\frac{1}{2(1-\beta_t)}\|\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t}\|^2 + C$$

where C is a constant. Under some mild assumptions, we can approximate the log likelihood $\log p_\phi(\mathbf{y} \mid \mathbf{x}_t)$ by Taylor expanding around $\mathbf{x}_t = \boldsymbol{\mu}_{\theta,t}$, where $\boldsymbol{\mu}_{\theta,t} = \boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$:

$$\begin{aligned}\log p_\phi(\mathbf{y} \mid \mathbf{x}_t) &\approx \log p_\phi(\mathbf{y} \mid \mathbf{x}_t)\big|_{\mathbf{x}_t=\boldsymbol{\mu}_{\theta,t}} + (\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t})\nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y} \mid \mathbf{x}_t)\big|_{\mathbf{x}_t=\boldsymbol{\mu}_{\theta,t}} \\ &= (\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t})\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y}) + C_1\end{aligned}$$

where the likelihood flow field $\mathbf{g}_t(\boldsymbol{\mu}_{\theta,t}, \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y} \mid \mathbf{x}_t)\big|_{\mathbf{x}_t=\boldsymbol{\mu}_{\theta,t}}$ and C_1 is a constant. Note that in this derivation, the gradient of the log likelihood is evaluated at $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$, the local somatic current at the current time step. This makes a direct biological interpretation of the model problematic since the sensory area providing this likelihood signal cannot access the somatic current and only has access to information from the previous time step. In order for our model to obey temporal causality and local access to information, we suppose instead that the likelihood flow field receives as input a point $\mathbf{x}_{t-1}$ from the previous time step which is used to evaluate the likelihood flow field $\mathbf{g}_t(\mathbf{x}_{t-1}, \mathbf{y}) = \nabla_{\mathbf{x}} \log p_\phi(\mathbf{y} \mid \mathbf{x}_{t-1})|_{\mathbf{x}=\mathbf{x}_{t-1}}$. This approximation becomes more exact the closer we are to the data manifold: since $\boldsymbol{\mu}_\theta(\mathbf{x}_{t-1})$ can be thought of as the prior network's current guess of a point on the prior manifold, the error from substituting $\mathbf{x}_{t-1}$ for $\boldsymbol{\mu}_{\theta,t}$ goes to zero as $\mathbf{x}_t$ approaches the prior manifold. We demonstrate this numerically in Fig. 2.2E.

The posterior neural transition operator combines the prior transition operator and likelihood signal using Bayes rule:

$$\begin{aligned}\log p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{y}) &= \log p_\theta(\mathbf{x}_t \mid \mathbf{x}_{t-1}) + \log p_\phi(\mathbf{y} \mid \mathbf{x}_t) + C_1 \\ &= -\frac{1}{2(1-\beta_t)}\|\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t}\|^2 + (\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t})\mathbf{g}_t + C_2 \\ &= -\frac{1}{2(1-\beta_t)}\|\mathbf{x}_t - \boldsymbol{\mu}_{\theta,t} - (1-\beta_t)\mathbf{g}_t\|^2 + \frac{\mathbf{g}_t^2(1-\beta_t)}{2} + C_2 \\ &= -\frac{1}{2(1-\beta_t)}\|\mathbf{x}_t - (\boldsymbol{\mu}_{\theta,t} + (1-\beta_t)\mathbf{g}_t)\|^2 + C_3 \\ &= \log \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\theta,t} + (1-\beta_t)\mathbf{g}_t,\ (1-\beta_t)I) + C_4\end{aligned}$$

where C_{1-4} are constants that do not contain $\mathbf{x}_t$. As we can see, the posterior transition operator is again a Gaussian and C_4 corresponds to its normalizing constant. The effect of the external likelihood signal is to shift the mean of the prior transition operator by an amount $(1 - \beta_t)\mathbf{g}_t$.

To sample from the context conditioned posterior $p(\mathbf{x} \mid c)$, we learn a map from $\mathbf{x}_t$ to c by training a feedforward neural network to classify points $\mathbf{x}_t$ via optimization. For the trimodal prior in Fig. 2.3, this is done by assigning one of three class labels to samples from the corresponding mode. We construct the classifier training set by adding i.i.d. multivariate Gaussian noise of varying degrees of variance to clean samples from the prior using the DM forward process. As with the unimodal swiss-roll prior, at large degrees of noise most of the signal is removed and the distribution resembles a white Gaussian. The classifier network is then trained via backpropagation to receive $\mathbf{x}_t$ as input and identify the class of the corresponding clean data sample. Since the last layer of the classifier is a Softmax layer that normalizes the outputs to one, we interpret the values of the classifier output as the probability distribution $p(c \mid \mathbf{x}_t)$. $\mathbf{g}_t(\mathbf{x}_{t-1}, c)$ is evaluated using the torch autograd package.

A.2 Effects of dendritic tree morphology

We studied how the morphology of the dendritic networks affected the speed of learning by conducting a grid search over dendritic architectures that differed in width and depth. To isolate the effect of the architecture choice, we kept the total number of parameters roughly the same across all networks, meaning that as we increased network depth we also decreased network width.

A dendritic architecture can be fully specified by an array of branching factors, which denotes the number of children nodes at each layer. The array gives branch factors in order of distal to proximal layers. The six architectures are defined as follows:

We trained four iterations of each network architecture with different random seeds. The

Depth	Morphology	Num. of parameters
2	[59, 59]	35048
4	[8, 8, 7, 7]	33154
5	[5, 5, 5, 5, 5]	34372
6	[4, 4, 4, 4, 4, 3]	34814
7	[3, 3, 3, 3, 3, 3, 4]	34986
10	[2, 2, 2, 2, 2, 2, 2, 2, 3, 3]	32234

Table A.1: Morphology of dendritic networks.

networks were trained on the 2D swiss roll manifold using the Adam optimizer with a learning rate of 3e-4 over 1.5e6 epochs. We quantified the similarity of the network to the ground truth manifold as a function of training epoch using a naive KL divergence estimator (10000 samples each, with space discretized into 255 bins).

We see that shallower but wider networks learned faster than deeper but thinner networks, and converged to a KL divergence value much lower than the other networks. However, we cannot conclude that shallower networks perform better overall: the KL divergence value for the 6 layer network even towards the end of the 1.5e6 training epochs indicates that the networks have not reached the global minimum of their loss functions, and suggests that with more training epochs the network performance may continue to improve.

One limitation of the architecture is that the resulting recurrent network has dense connectivity, linking each neuron to all distal branches of all neurons, which is biologically unrealistic. We adjusted the tree architecture so that inputs from other neurons can project to different locations on the target neuron's dendritic tree. Architecturally, this was achieved in the form of skip connections from the neurons in the first layer (representing the output of the somatic compartment at the previous timestep) to intermediate subunits in the dendritic arbor of all neurons. These skip connections, unlike those commonly found in deep neural network models (He et al., 2015), have learnable weights. The synaptic inputs from these skip connections are summed by the subunit together with incoming activity from more distal subunits. To encourage sparsity in the recurrent inter-neuron connectivity, we added L1 regularization on the synaptic strengths.

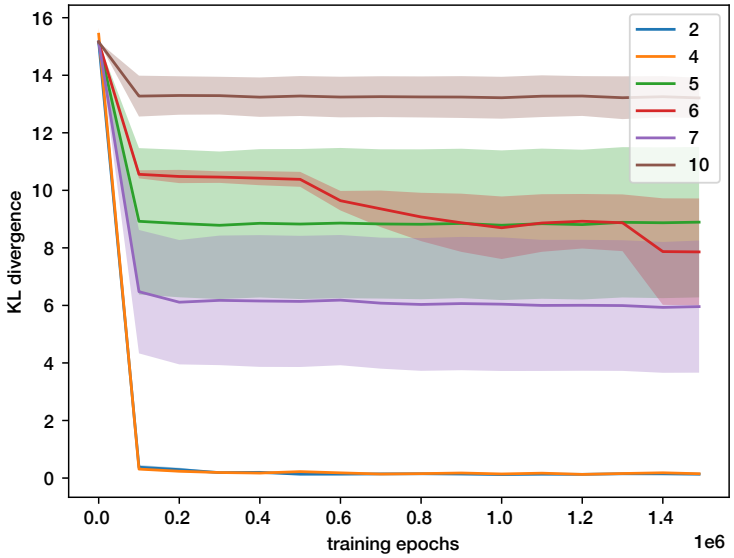


Figure A.1: Network performance over training. Each line corresponds to the performance of a model architecture of a particular depth. We see that the KL divergence of shallower networks (depths 2, 4) decreased the fastest, while deeper network converges slower. Note that the speed of learning does not scale monotonically with network depth; for example, the 7-layer network learns faster than the 5 layer network.

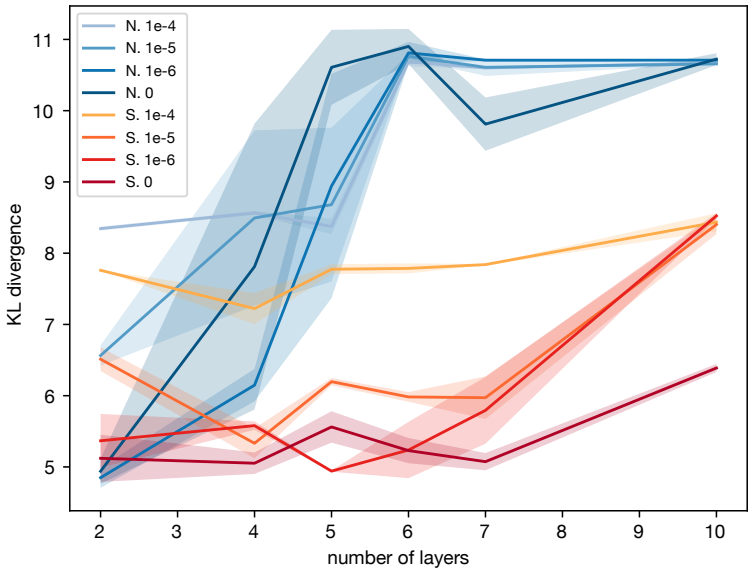


Figure A.2: Effects of sparse recurrent connections targeting different locations on the dendritic tree. Quality of prior samples as a function of the dendritic tree depth. *S* and *N* in the legend indicate architectures with and without skip connections, respectively. Number denote the strength of the L1 regularization. We trained two models for each architecture. Shaded regions indicate variability in the KL divergence resulting from variability in generated samples.

We studied prior representational quality of the biologically-refined version of the model while varying the depth of the dendritic tree for the same task (10 dimensional network representing a 3 dimensional "swiss-roll" latent manifold). All models were trained over 2e6 epochs with different degrees of L1 regularization. We compared model generated samples against the ground truth by calculating the KL divergence between the two distributions formed by projections onto a two dimensional cross section of the manifold. The rest of the KL computation is equivalent to the process described in Section 2.5. Here we chose this evaluation method over calculation of marginal statistics (as in Fig. 2.1G and Fig. A.3) because some models produced samples far away from the ground truth manifold. Models with stronger L1 regularization generally showed lower performance across most layer depths (Fig. A.2). Skip connections have a more nuanced effect. In deeper networks, they result in large performance gains when compared to models without skip connections. However, in shallower networks, skip connections result in comparable or sometimes worse performance than their counterparts with no skip connections.

Why is this the case? Skip connections such as those used by deep networks allow for better performance by preventing vanishing/exploding gradients; by giving the network another way to route information, they induce a smoother loss landscape that is easier to traverse during training (Li et al., 2018). Even though our skip connections have learnable parameters, information can still be routed along different pathways, allowing for more direct influence over particular subunits. This effect is likely more pronounced in our tree architecture than in models with fully connected layers since there is greater segregation of information at each layer. The relative loss of performance in shallower networks suggests skip connections may induce destructive interference that cannot be remedied if there are not enough downstream layers for further processing. Indeed, we see that when the model is L1 regularized and has skip connections, optimal performance occurs when the architecture is neither too shallow nor too deep, but rather possesses some intermediate number of layers. This may have implications on the type of dendritic architectures we expect to see in sensory cortices.

A.3 Testing neural sampling quality by marginal statistics

The data manifold on which the model in Fig. 2.1A was trained has three embedding dimensions but is embedded in a 10-dimensional ambient space (using the terminology in Jazayeri and Ostojic (2021)). We ensured the data have non-zero values in every dimension by rotating the three dimensional manifold by $\pi/4$ around each pair of axes.

The three dimensional swiss-roll manifold is uniformly distributed along the linear dimension (i.e. the dimension orthogonal to the swiss-roll cross-section). We evaluated the marginal statistics of the model output along this dimension by comparing the cumulative distribution function (cdf) of this model output against a reference uniform cdf.

To recover the model's marginal density along the linear dimension, we first projected the network output onto the three basis vectors of the embedding space, which we calculated by transforming unit vectors by the same rotation operators as the data manifold. We then projected the model output along the third basis vector and computed the cdf along this dimension.

The empirical cdf is close to the reference cdf, indicating that the data are likely distributed according to the uniform distribution. We can quantify this using the Kolmogorov-Smirnov goodness-of-fit test, which is a nonparametric measure of the similarity between a continuous empirical sample distribution and a specified reference distribution. We applied the KS test to both the empirical distribution along the nonlinear dimension in Fig. 2.1G and along the linear dimension in Fig. A.3. For a sample size of 1000 each, the p-value along the nonlinear dimension is 0.145 while the p-value along the linear dimension is 0.125, making them statistically indistinguishable from the uniform distribution.

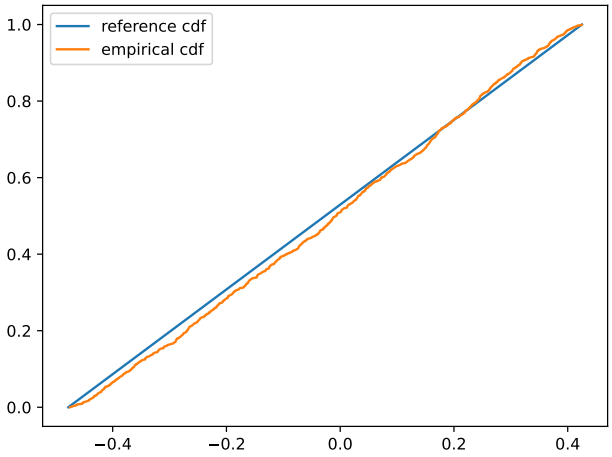


Figure A.3: Marginal statistics along the dimension orthogonal to the swiss roll cross-section.

A.4 Posterior autocorrelation

We numerically evaluated posterior samples generated by the neural circuit. We see in Fig. A.4 that the samples generated from neural sampling result in autocorrelation functions that rapidly decrease and remains steadily at zero. This indicates that the posterior samples, like the prior samples in Fig. 2.1G, are largely independent over time despite the approximations. This remains true even when the neural sampling dynamics are influenced by contextual and sensory evidence.

A.5 Stochastic neural network

To assess whether the nonlinear operations of the neural sampling process could be instantiated by a population of neurons rather than in the dendritic arbor of pyramidal neurons, we trained a Stochastic Neural Network (SNN) model, a fully connected feedforward neural network comprising neurons with linear weights, ReLU nonlinearity and somatic stochasticity. As in our dendritic model, the somatic stochasticity in each neuron is additive Gaussian, with the degree of stochasticity modulated by the phase of the global oscillation. The operation performed

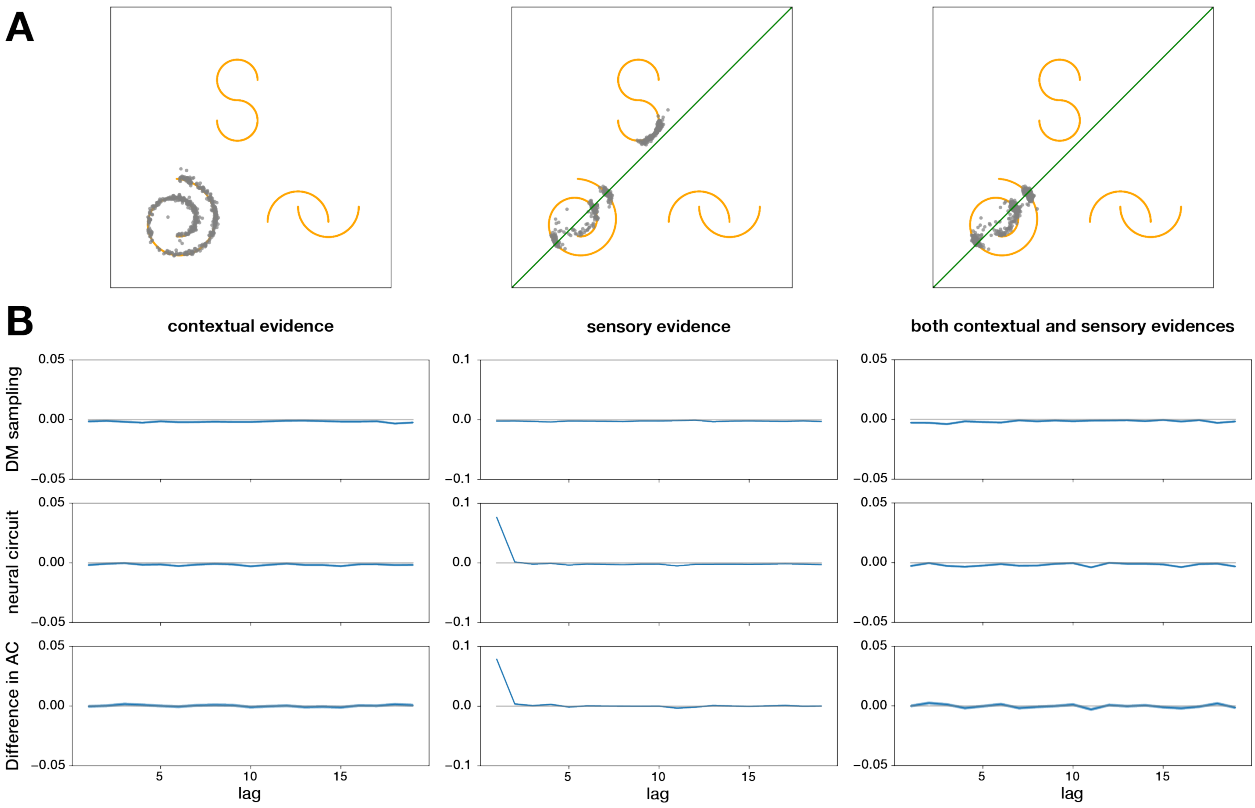


Figure A.4: Autocorrelation comparison of the posterior samples. **(A)** Posterior samples generated from the two dimensional trimodal dataset as a result of observing contextual evidence (left), sensory evidence (middle), and both forms of contextual and sensory evidence (right). The context cue for the context-conditioned posterior is the discrete context random variable c_2 indicating the swiss roll manifold. The sensory likelihood, as before, is a two dimensional linear constraint with a Gaussian-distributed uncertainty around it. **(B)** The autocorrelation functions of samples generated by the DM reverse process (top row) and neural sampling (middle row) for the different posterior distributions. The difference between the two autocorrelation functions is shown in the bottom row.

by the neurons in each hidden layer can be expressed as

$$h_l = \text{ReLU}(W_l(h_{l-1}, t)) + (1 - \beta_t)\boldsymbol{\epsilon}_t, \tag{A.1}$$

where h_l is the output of hidden layer l and W_l is the weight matrix at layer l. $\boldsymbol{\epsilon}_t$ is a sample of a Gaussian with identity covariance.

Our instantiation of this model architecture has three hidden layers, each with a width of 32 neurons. The degree of noise as given by the oscillator β_t is concatenated with the output of

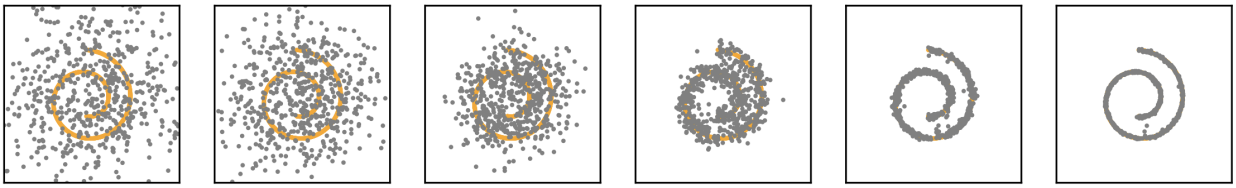


Figure A.5: DM reverse process using the SNN architecture. The samples generated by the SNN result in samples that lie on the two dimensional swiss roll manifold.

the previous hidden layer. We trained the model via backpropagation using the Adam optimizer with a learning rate of 1e-4 over 5000 epochs. We show the distribution of samples generated at intermediate points of the ascending phase below, in Fig. A.5.

A.6 More biologically realistic oscillations

In the brain, oscillations fluctuate in both amplitude and frequency and do not have the perfect sinusoid structure assumed by the model. To test the effects of such fluctuations on the circuit's ability to sample from priors, we introduced variability in the amplitudes and/or period of the β_t oscillation. The original oscillation cycled between 0 and 1 with a period of 200 timesteps (Fig. A.6A). In these variations, the amplitude and period of the oscillation were defined as random variables. The amplitude was drawn from a truncated normal distribution $A \sim \mathcal{N}(A_{\text{original}}, \sigma_A), A > 0$, i.i.d for each cycle (Fig. A.6B), where samples were obtained by rejection sampling. The period of each cycle was similarly drawn from another truncated normal distribution $T \sim \mathcal{N}(T_{\text{original}}, \sigma_T), T > 0$ (Fig. A.6C). Finally, we also considered the combined effect of simultaneous fluctuations in both amplitude and frequency (Fig. A.6D). Using the denoising-trained 2D swiss-roll model, we used the induced β_t values for circuit sampling. We assessed the effect of the generated samples by comparing the empirical distribution given by the resulting samples against the ground truth distribution, via the KL divergence. We find that the quality of the samples is largely unaffected by the deviations from the perfect sinusoid (Fig. A.6E). The KL divergence remains largely flat for a range of noise levels. This suggests that the sampling

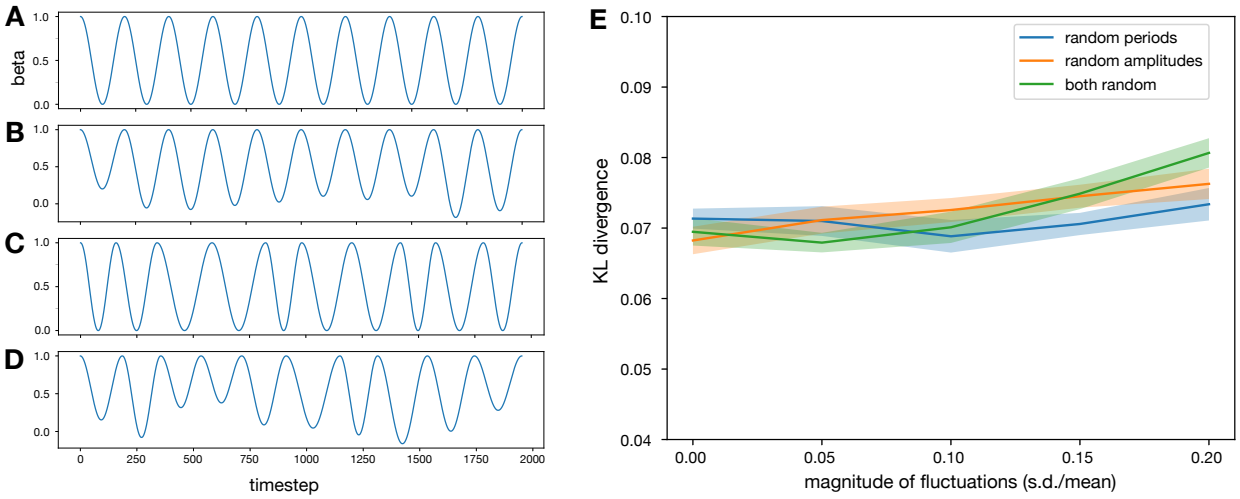


Figure A.6: (A) Reference β_t fluctuations for the simple model. **(B)** Example draw for a more naturalistic oscillations with fluctuations in amplitude ($\sigma_A = 0.15$). **(C)** Example draw for a more naturalistic oscillations with fluctuations in period ($\sigma_T = 30$). **(D)** Example draw for a more naturalistic oscillations with fluctuations in amplitude ($\sigma_A = 0.15$ and $\sigma_T = 30$). **(E)** Deviations of empirical samples from ground truth distribution for varying degrees of oscillation fluctuations. A standard deviation corresponds to the deterministic setting. KL estimates use 5000 samples per run, with shading showing SEM estimated over 20 runs.

procedure can be reasonably robust for naturalistic oscillatory signals.

A.7 MNIST DATASET

To assess whether our model scales to higher dimensional priors, we trained our circuit model on the MNIST dataset, which contains 28 by 28 pixel images of handwritten digits. We were able to train several dendritic architectures to successfully generate digits from the MNIST dataset (Fig.A.7, top two rows). They were trained by the same denoising objective, using the Adam optimizer with a learning rate of 4e-3 for 5000 epochs. The best performing architectures were two and three layer architectures with branch factors $[12, 12]$ and $[5, 5, 5]$ respectively.

We sampled a class-conditioned posterior by combining the model with a separate classifier trained on MNIST digits (Fig.A.7, top two rows). We used approximate posterior dynamics described in Section 3 and described in Supplementary section A.1. Taken together, these MNIST

results suggest that the biologically-motivated deviations from traditional diffusion models can scale to higher dimensional distributions.

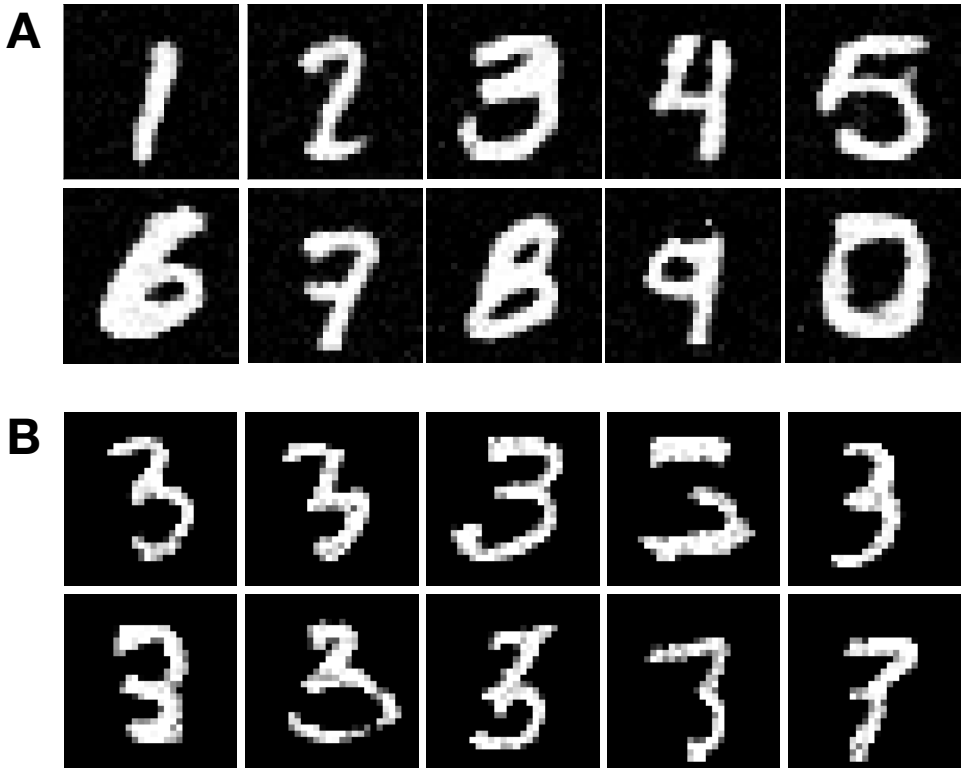


Figure A.7: Results for an an MNIST trained neural circuit. **(A)** Example prior samples, one per class. **(B)** Example samples from the posterior conditioned on class '3'.

B Appendix: Learning disentangled representations via score-based variational autoencoders

B.1 Additional results

B.1.1 Relation of MMSE solution to conditional score

Here, we provide a derivation of how the guidance score relates to the MMSE solution for the noise present in the noisy image (Eq. 3.6). For a variance-preserving noising operator $p(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$, Miyasawa/Tweedie's formula is given as

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\mathbf{x}_0] = \frac{1}{\sqrt{\bar{\alpha}_t}}[\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})].$$

We can compute the expected value of the noise in the noisy image

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\boldsymbol{\epsilon}] &= \int \boldsymbol{\epsilon}\, p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})d\mathbf{x}_0 \\ &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\mathbf{x}_0]\right].\end{aligned}$$

where on the second line we have substituted in $\boldsymbol{\epsilon} = {}^{1}/_{\sqrt{1-\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)$ via the re-parameterization trick. Now, rearranging for and substituting the expected value of $\mathbf{x}_0$ into Tweedie's formula above, we are left with

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,\mathbf{z})}[\boldsymbol{\epsilon}] &= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}[\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})]\right)\right] \\
&= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left[\mathbf{x}_t - \left(\mathbf{x}_t + (1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})\right)\right] \\
&= \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(-(1-\bar{\alpha}_t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})\right) \\
&= -\sqrt{1-\bar{\alpha}_t}\,\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{z})
\end{aligned}$$

which is the relation given by Eq. 3.6 in the main text.

B.1.2 Guiding transition operators

Here we derive the equation for how the guidance score affects the DDPM reverse transition operator, given by $p(\mathbf{x}_t|\mathbf{x}_{t+1})$. The classifier guidance equation in Eq. 3.7 can be expressed in terms of the transition operators:

$$\begin{aligned}
\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1},\mathbf{z}) &= \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1}) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t,\mathbf{x}_{t+1}) \\
&= \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_{t+1}) + \nabla_{\mathbf{x}_t}\log p(\mathbf{z}|\mathbf{x}_t),
\end{aligned}$$

where we have used the fact that since $\mathbf{x}_{t+1}$ is a noisier image than $\mathbf{x}_t$, it carries less information about $\mathbf{z}$ than $\mathbf{x}_t$, so the dependence on $\mathbf{x}_{t+1}$ can be dropped. In practice, this problem is equivalent to asking how to sample from $p(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})$ over all noise levels t. For a given noise schedule determined by $\{\alpha_t\}$, we evaluate $\mathbf{x}_t \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta, \sqrt{1-\alpha_t}I)$ by estimating the transition operator mean

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right),$$

using which we can estimate the next state in the Markov chain $\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t) + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. To sample from the $\mathbf{z}$-conditional distribution instead, this function must be dependent on $\boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z})$ rather than on $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$:

$$\begin{aligned}
\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{z}) &= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\theta,\phi}(\mathbf{x}_t, \mathbf{z})\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\left(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t)\right)\right) \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)\right) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t) \\
&= \boldsymbol{\mu}_\theta(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t).
\end{aligned}$$

Since $\mathbf{x}_{t-1}(\mathbf{x}_t, \mathbf{z}) = \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{z}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}$, we can rewrite this relation in terms of the noisy states:

$$\begin{aligned}
\mathbf{x}^*_{t-1}(\mathbf{x}_t, \mathbf{z}) &= \mathbf{x}^*_{t-1}(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla_{\mathbf{x}_t}\log q_\phi(\mathbf{z}|\mathbf{x}_t) \\
&= \mathbf{x}^*_{t-1}(\mathbf{x}_t) - \frac{1-\alpha_t}{\sqrt{\alpha_t}}\mathbf{g}_{t,\phi}.
\end{aligned}$$

B.1.3 Training and conditional generation algorithms

Below, we provide algorithms for training SAMI and for performing conditional generation using the trained networks.

B.1.4 Conditional generation as constrained optimization

As mentioned in the "Geometric Intuition" subsection, we can think of conditional generation as a constrained optimization problem, in which we have multiple constraints imposed by the inference network and the denoiser. To make this perspective more concrete, consider the generation procedure, given by Algorithm 1. We start with a sample of white noise, $\mathbf{x}_t \sim \mathcal{N}(0, I)$,

Algorithm 1: Conditional generation, given a clean guidance image $\mathbf{x}_0$, trained denoiser $\boldsymbol{\epsilon}_\theta$ and inference network f_ϕ

$(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \leftarrow f_\phi(\mathbf{x}_0)$ `// encode guidance image`
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
$\mathbf{z} \leftarrow \boldsymbol{\mu}_0 + \boldsymbol{\sigma}_0 \cdot \boldsymbol{\epsilon}$ `// sample clean image latent`

Initialize $t \leftarrow T$; $\mathbf{x}_t \sim \mathcal{N}(0, I)$
while $t \neq 0$ **do**
  $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \leftarrow f_\phi(\mathbf{x}_t)$ `// encode noisy image`
  $\log q \leftarrow -1/2\big(d_M(\mathbf{z}, \boldsymbol{\mu}_t) + \log|\boldsymbol{\sigma}_t I|\big)$ `// log posterior`
  $\mathbf{g}_t \leftarrow \nabla_{\mathbf{x}_t} \log q$ `// guidance score`
  $\boldsymbol{\mu} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\right)$
  $\Sigma \leftarrow \sqrt{1-\alpha_t} I$
  $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu} + \Sigma \mathbf{g}_t, \Sigma)$ `// guidance`
  $t \leftarrow t - 1$
end

Algorithm 2: Training, given a denoiser network $\boldsymbol{\epsilon}_\theta$ and inference network f_ϕ with randomly initialized weights

while *not converged* **do**
  $\mathbf{x}_0 \sim p(\mathbf{x}_0)$
  $t \sim [0, T]$
  $\mathbf{x}_t \sim p_{\text{fwd}}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I)$ `// noise the image`
  $\boldsymbol{\epsilon}_\mathbf{x} \leftarrow {}^{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)}/_{\sqrt{1-\bar{\alpha}_t}}$
  $\boldsymbol{\epsilon}_\mathbf{z} \sim \mathcal{N}(0, I)$
  $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0) \leftarrow f_\phi(\mathbf{x}_0)$ `// encode clean guidance image`
  $\mathbf{z} \leftarrow \boldsymbol{\mu}_0 + \boldsymbol{\sigma}_0 \cdot \boldsymbol{\epsilon}_\mathbf{z}$
  $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \leftarrow f_\phi(\mathbf{x}_t)$ `// encode noisy image`
  $\log q \leftarrow -1/2\big(d_M(\mathbf{z}, \boldsymbol{\mu}_t) + \log|\boldsymbol{\sigma}_t I|\big)$ `// log posterior`
  $\mathbf{g}_{t,\phi}(\mathbf{z}) \leftarrow \nabla_{\mathbf{x}_t} \log q$ `// guidance score`
  $\mathcal{L}_\mathbf{x} \leftarrow \|\boldsymbol{\epsilon}_\mathbf{x} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t) - \sqrt{1-\bar{\alpha}_t}\,\mathbf{g}_{t,\phi}(\mathbf{x}_t, \mathbf{z})\|^2$
  $\mathcal{L}_\mathbf{z} \leftarrow D_{KL}\left(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0)\|\mathcal{N}(0, I)\right)$
  take gradient steps on $\nabla_{\theta,\phi}\left(\mathcal{L}_\mathbf{x} + \beta\mathcal{L}_\mathbf{z}\right)$
end

from which we take gradient steps towards a region of large guided score:

$$\begin{aligned}\mathbf{x}_t &\leftarrow \mathbf{x}_t + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{z}) \\ &= \mathbf{x}_t + \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t),\end{aligned}$$

where we have decomposed the conditional score using Bayes' rule. Since the posterior q_ϕ is assumed to be conditional Gaussian, the analytic form of the log likelihood is given as

$$\log \mathscr{L}_\phi(\mathbf{x}_t; \mathbf{z}) = -\frac{1}{2}\, \mathbb{E}_{\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x}_0)} \Big[\, d_M\big(\mathbf{z}'_\phi(\mathbf{x}_0), \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big) + \log|\Sigma_\phi(\mathbf{x}_t)| + c \,\Big], \tag{B.1}$$

where $d_M = \big(\mathbf{z}'_\phi(\mathbf{x}_0) - \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big)^\top \Sigma_\phi^{-1} \big(\mathbf{z}'_\phi(\mathbf{x}_0) - \boldsymbol{\mu}_\phi(\mathbf{x}_t)\big)$ is the Mahalanobis distance between the mean estimate of the network state in latent space $\boldsymbol{\mu}_\phi(\mathbf{x}_t)$ and the latent corresponding to the target image $\mathbf{z}'(\mathbf{x}_0)$, computed under a metric that is given by the inverse of the covariance Σ_ϕ^{-1} (red arrow and concentric ellipsoids in Fig. 3.1C, $\mathcal{X}$ space).

Using this expression for the log likelihood, it becomes clear that we are taking gradient steps with respect to multiple constraints:

$$\begin{aligned}\mathbf{x}_t &\leftarrow \mathbf{x}_t + \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) \\ &= \mathbf{x}_t - \frac{1}{2}\nabla_{\mathbf{x}_t}\Big(\mathbb{E}_{\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\Big[d_M\left(\mathbf{z}', \boldsymbol{\mu}_\phi(\mathbf{x}_t)\right) + \log|\Sigma_\phi(\mathbf{x}_t)|\Big]\Big) + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t) \\ &= \mathbf{x}_t - \frac{1}{2}\mathbb{E}_{\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\Big[\nabla_{\mathbf{x}_t} d_M(\mathbf{z}', \boldsymbol{\mu}_\phi(\mathbf{x}_t))\Big] - \frac{1}{2}\nabla_{\mathbf{x}_t} \log|\Sigma_\phi(\mathbf{x}_t)| + \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t).\end{aligned}$$

There are three terms, each corresponding to a specific constraint during this optimization procedure. The first constraint is to land in a region of the image space such that its representation minimizes its Mahalanobis distance to the latent of the guiding image $\mathbf{z}'(\mathbf{x}_0)$. This is related to the metamer perspective (Helmholtz, 1852; Zhu et al., 1998; Portilla and Simoncelli, 2000; Freeman and Simoncelli, 2011; Mahendran and Vedaldi, 2014; Feather et al., 2023), where the goal is

to produce data samples that share the same latent representation as a target datum. Metamer generation can be formulated as a constrained optimization problem, minimizing the L2 distance between the representations of the current sample and the target, $\mathbf{x}_t \leftarrow \mathbf{x}_t - \nabla_{\mathbf{x}_t} \|\mathbf{r}_{\text{target}} - \mathbf{r}_{\mathbf{x}_t}\|^2$, where $\mathbf{r}_{\mathbf{x}_t}$ is the representation of the current sample $\mathbf{x}_t$, and $\mathbf{r}_{\text{target}}$ that the target datum. Our approach generalizes this framework by replacing the L2 distance with the log likelihood from Eq. B.1. In this view, the inference network constrains the generation process to produce samples that are metameric to the guidance image as defined by the latent representation.

The next term corresponds to taking steps towards regions of the image space that minimizes the log determinant of the covariance matrix. Geometrically, we can think of this as reducing the volume of the uncertainty ellipsoid over the features identified by the inference network. In a well-calibrated Bayesian network, the posterior uncertainty over the identified features should match the irreducible uncertainty present in the image. For SAMI, the irreducible uncertainty in the latent space comes from the noise in the noisy observations, and as such we should expect that the log determinant of the covariance to correlate strongly with the variance of the noisy image.

The final term corresponds to the denoiser, parameterized by θ, which ensures that the network is driven towards the natural image manifold, i.e. the set of highly probable images.

The region of the $\mathcal{X}$ space that we land on is at the intersection of these three signals. This perspective suggests that we can increase the diversity (i.e. the entropy) of the generated images by relaxing one or more of these constraints. For example, if the generation procedure were given only by the third term, the network would be driven to satisfy only the image prior, recovering the unconditional generation of DDPM. Alternatively, we might choose to calculate the Mahalanobis distance over only a subset of the latent dimensions, in effect making sure that $\boldsymbol{\mu}_\phi$ is close to $\mathbf{z}'$ only in those latent dimensions. This has the effect of making those latent axes "rigid", while the rest of the axes are "sloppy", since large distances in these dimensions are discounted Machta et al., 2013; Transtrum et al., 2015. Back in the pixel space, this equates to finding images that

possess a subset of the semantic attributes in the original guiding image, i.e. those that correspond to the constrained feature dimensions.

B.1.5 Architecture and training details

Parameter	Disks	CelebA 64	CelebA-HQ 256
Inference net arch.	ConvNet	Half-UNet	Half-UNet
Infnet base channels	48	64	64
Infnet channel multiplier	[2, 2]	[1, 1, 1, 1, 1, 1]	[2, 2, 2, 2, 1, 1]
Infnet nonlinearity	ReLU	ReLU	SiLU
Denoiser arch.	UNet (no attn)	UNet (no attn)	UNet (with attn, frozen)
Denoiser base channels	128	128	128
Denoiser ch. multiplier	[1, 1, 1, 1, 1, 1]	[1, 1, 1, 1, 1, 1]	[1, 1, 2, 2, 4, 4]
Num. noise levels	400	1000	1000
Latent dimensionality	3	256	512
KL weight	5e-6	1e-3	1e-8
Noise schedule	Linear	Cosine	Linear
Training set size	2000	60,000	30,000
Batch size	512	512	512
Num. epochs trained	1000	5000	70
Learning rate	6e-3	1e-3	3e-5
Timestep sampling dist.	Uniform	Uniform	Monotonically increasing
Resources (num. A100)	1	20	20
Optimizer	Adam (without weight decay)		

Table B.1: Architectural and training details for each of the three datasets mentioned in the main text.

Architectural details. We tried to use simple architectures throughout our experiments for the sake of interpretability and to emphasize the algorithmic effects as much as possible. For the disks dataset, we used a two layer ConvNet as the inference network, with kernel size 3 and stride 2. The inference networks used for the disks and CelebA-64 datasets have no biases, following Mohan et al. (2022), and the denoisers had no attention blocks. Despite the deliberate restrictions

in architecture, we saw that the model was able to learn a robust representation of the dataset. The Half-UNet architecture is a UNet where the Up blocks have been replaced by two MLP layers, both outputting a vector of dimension equal to the latent dimension. From one MLP we get the latent mean. The other outputs a vector which we apply Softplus and squaring operators; this returns the variance.

For the CelebA-HQ 256 dataset, we used an off the shelf DDPM Ho et al., 2020. We froze the denoiser's weights and trained only the inference network, which we again restricted to be a Half-UNet without attention blocks.

KL weight. We annealed the KL weight to the value given in Table B.1, starting from a number many orders of magnitude smaller and using a exponential annealing schedule.

Choosing hyperparameters. We performed an extensive grid search over the space of hyperparameters for each model, measuring model performance by the change in MSE between the MSE of an unconditional diffusion model and the MSE when the inference network was also employed. The unconditional diffusion model retained the same denoiser hyperparameters as the full SAMI. As mentioned in the main text, an informative representation should result in generated images that resemble the original guiding image, and thus a smaller MSE can be used as a measure of latent informativeness. We found that larger KL weight corresponded to more factorized representations, but the weight could only be increased up to a threshold, above which the model exhibits posterior collapse and the representation becomes uninformative. As such, we performed binary search over the KL weights until we found a value just smaller than the threshold.

Timestep sampling distribution. When training the model on the disks and CelebA 64 datasets, we sampled the noise level uniformly from 0 to T, as is common when training unconditional DDPMs. When training the inference network on top of the frozen denoiser on the CelebA-HQ

dataset, we found that oversampling the larger noise levels was more effective. This is equivalent to over-weighing the contribution of high noise level terms in the overall MSE loss rather than weighing all levels uniformly. If the inference network is unable to learn the guidance vector for all noise levels when paired with a frozen denoiser, since the guidance vector has greater relative influence at high noise levels compared to the denoising score, it is more effective to learn the guidance score for the high noise regime.

Bottleneck size. We found that the number of semantic attributes learned by the SAMI representation depended on three factors: the magnitude of the KL weight, the dimensionality of the latent representation, and the timestep sampling distribution. Relaxing the first two of these three factors increases the amount of information that can be learned by the network, but results in a less factorized and interpretable representation.

B.1.6 Reduction in variability from conditioning

As a measure of how much entropy is reduced by conditioning the diffusion process on the representation, we measured the entropy in the test set and and the entropy in conditionally generated samples. Since it is not possible to get an exact measure of entropy, we make a coarse Gaussian assumption on the distribution and measure the mean squared distance of samples from their empirical mean. We conditionally generate 4100 samples per conditioning image, which is greater than the ambient dimensionality of the image of 4096. We find that the mean distance is reduced from 6.62×10^{-2} to 2.04×10^{-3}, indicating that the representation captures much of the potential variability in the dataset. We also find a reduction in the spread of the squared distances, from a SEM value of 4.85×10^{-4} to 4.45×10^{-6}. This concentration of the squared distances around a near-zero mean value suggests that there is very little variability in conditionally generated samples.

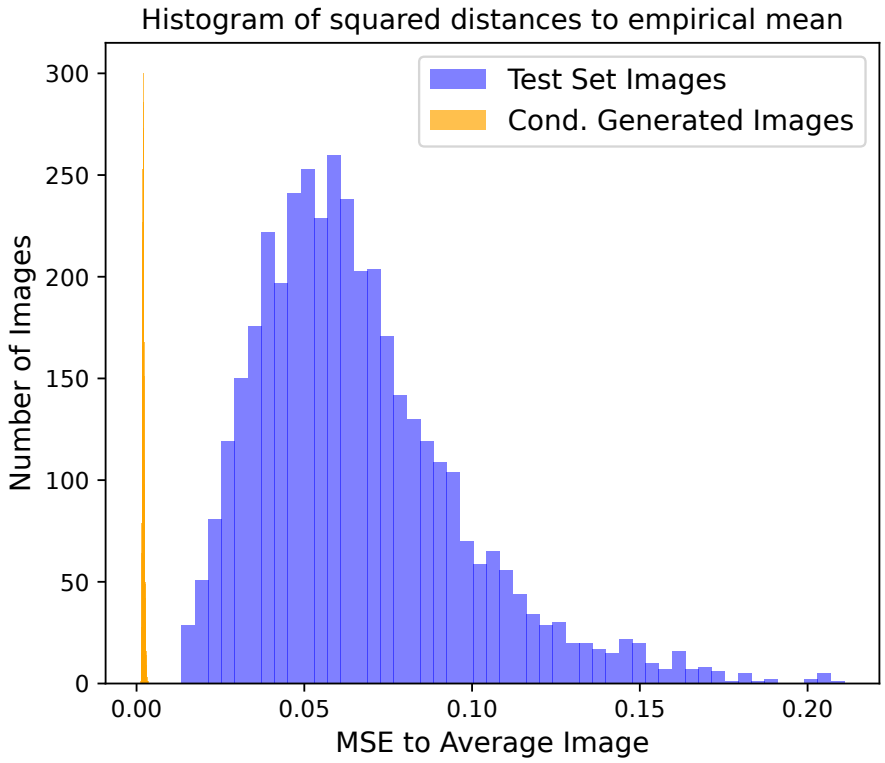


Figure B.1: Reduction in variance from conditioning on the posterior on a CelebA image.

B.1.7 Metrics

For disentanglement, we measured the degree to which SAMI's representation was disentangled when trained on CelebA 64×64. We used the Total AUROC Difference (TAD) metric, which measures the degree to which a unique attribute is detected by a unique latent representation, and the degree to which the latent axes confidently replicate the ground truth axes (Yeats et al., 2022). Among the diffusion-based representation learning methods, InfoDiff, DisDiff and DBAE are trained to exhibit disentangled representations (Wang et al., 2023; Yang et al., 2023b; Kim et al., 2024), so direct comparison against these methods are particularly interesting. We also compared our method against more standard benchmarks in the disentanglement literature, such as VAE and β-VAE.

In Table B.2, we see that though we performed little to no hyperparameter search, SAMI

performs the best or second best at both disentangling and sample quality metrics. This is likely due to the effect of training our model to produce the exact log likelihood, which requires both clean and noisy images, rather than a representation of only clean images. The coding of noisy images adds important structure to the latent space by encouraging noisy images to have the same latent value as clean images in latent dimensions that code for coarse features. This ensures that coarser features have more global coherence, a hypothesis that is borne out in Fig. 3.3F and G.

Model	TAD↑	Capt. attrs.↑	FID↓
AE	0.042 ± 0.004	1.0 ± 0.0	90.4 ± 1.8
VAE	0.000 ± 0.000	0.0 ± 0.0	94.3 ± 2.8
β-VAE	0.088 ± 0.051	1.6 ± 0.8	99.8 ± 2.4
InfoVAE	0.000 ± 0.000	0.0 ± 0.0	77.8 ± 1.6
DiffAE (Preechakul et al., 2022)	0.155 ± 0.010	2.0 ± 0.0	22.7 ± 2.1
InfoDiffusion (Wang et al., 2023)	0.299 ± 0.006	3.0 ± 0.0	22.3 ± 1.2
DisDiff (Yang et al., 2023b)	0.305 ± 0.010	-	18.2 ± 2.1
DBAE + TC (Kim et al., 2024)	0.362 ± 0.036	3.8 ± 0.8	13.4 ± 0.2
EncDiff (Yang et al., 2024)	0.638 ± 0.008	-	14.8 ± 2.3
SAMI (ours)	0.583	3.0	16.3

Table B.2: Comparisons against other generative models on TAD disentanglement metric and FID scores on 64×64 CelebA. "Capt. attrs." refers to the number of captured attributes when calculating the TAD score. Dark gray cells indicate the best, while light gray cells indicate second best.

dSprites We also trained our model on the dSprites dataset (Higgins et al., 2017), to see if the disentanglement properties we find in the synthetic disks dataset (Fig. 3.2) generalize to other synthetic benchmarks where the ground truth factors are known. Of the many disentanglement metrics, we measured Modularity (Ridgeway and Mozer, 2018) and DCI Disentanglement (Eastwood and Williams, 2018), since Locatello et al. (2019) finds that all metrics except Modularity are strongly correlated on the dSprites dataset. We measured a Modularity score of 0.882 ± 0.003 and a DCI Disentanglement score of 0.117 ± 0.0005, after initializing the measurement on 10 separate seeds. For both metrics we used 10 000 training samples and 5000 test samples.

These measures are comparable with the mean disentanglement scores produced by multiple VAE-based models, as shown in Fig. 14, top row in (Locatello et al., 2019). This suggests that our model learns a representation that disentangles ground truth factors despite the fact that our model does not have explicit regularization terms in the objective that encourage disentanglement beyond the β parameter, unlike (Chen et al., 2019; Kumar et al., 2018). Moreover, this result indicates that the disentanglement properties of our model still hold despite the fact that the multiscale inductive biases of our model are not aligned with the dSprites dataset, which compresses most of its variability within a smaller range of spatial scales than does natural datasets such as CelebA.

Indeed, we can measure the information content that exists in the dataset at different spatial scales by measuring the difference between the magnitudes of the prior and posterior scores at each noise level during the conditional generation process. The difference in these magnitudes at a particular noise level indicates the rate of change in the amount of information that is captured by the latents. Intuitively, a large magnitude suggests that a specific feature has now become "visible" to the inference network at that noise level. Conversely, small magnitudes are an indication that at that noise level there are no distinguishing features beyond the posterior mean (which is given by the unconditional denoiser).

We find that there are consistently two "bumps" in magnitude during the conditional denoising process, suggesting that most of the features in the target sprite can be found at two regions of spatial scales. Moreover, it demonstrates that our model can learn a disentangled representation, even when the distribution of the features' "characteristic noise levels" (as defined in Appendix **??**) are not spread out evenly across all noise levels, as in natural images.

B.1.8 Derivation for globally preserved semantic axes

We investigated the locality of the mapping between the axes and the semantic label by perturbing the latents of two different images along the same axis. We observed that though in

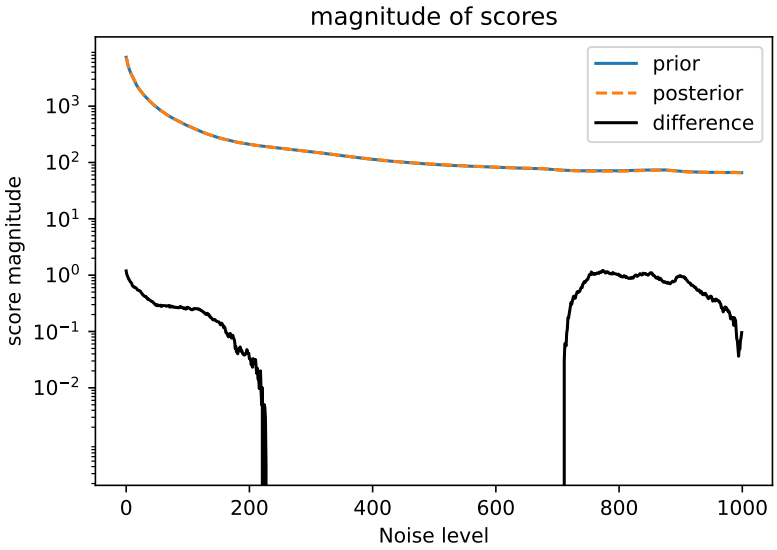


Figure B.2: Magnitude of the prior and posterior scores during a single conditional generation procedure, and their difference.

general, this mapping was not preserved across different regions of the latent space, some axes resulted in relatively similar semantic transformations irrespective of the latent.

To identify latent-semantic mappings that are relatively global, we derived a mathematical formulation that provides sufficient conditions for a mapping to be global. Suppose we have a trained conditional diffusion process that maps noisy images $\mathbf{x}_t$ to a clean image estimate $\hat{\mathbf{x}}_0$ via the use of a set of learned latents $\mathbf{z}$. This mapping is defined as a function $f_{\theta,\phi}$ that takes in $\mathbf{z}$ and $\mathbf{x}_t$ and returns a guided clean image estimate:

$$\begin{aligned} f_{\theta,\phi}(\mathbf{x}_t, \mathbf{z}) &= \mathbb{E}_{p_\theta(\mathbf{x}_0|\mathbf{x}_t,\, \mathbf{z})}\left[\mathbf{x}_0\right] \\ &= \mathbb{E}_{p_\theta(\mathbf{x}_0|\mathbf{x}_t)}\left[\mathbf{x}_0\right] + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}} \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) \\ &= \mathbf{x}_0^*(\mathbf{x}_t;\theta) + \gamma_t\, \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) \end{aligned} \tag{B.2}$$

where the analytic form of $\nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t)$ is given as the derivative of Eq. B.1. Given an initial image $\mathbf{x}_0^* = \mathbf{x}_0^*(\mathbf{x}_t, \mathbf{z}')$ that was generated by latent $\mathbf{z}'$, we define the change in the generated image in response to a perturbation to the latent along axis i as a nonlinear function $\psi(\cdot)$ that we can

define, without loss of generality, as

$$\mathbf{x}_0^* + \psi(\mathbf{x}_0^*,\ \mathbf{z},\ i) = f(\mathbf{x}_t,\ \mathbf{z} + \alpha\,\hat{\mathbf{i}})$$

where $\alpha\,\hat{\mathbf{i}}$ is a perturbation of magnitude α along a particular axis i. Our aim is to understand whether the functional form of the transformation ψ is dependent on the region of latent space that we are perturbing. In other words, if this image transformation ψ is global, it should be independent of the latent value $\mathbf{z}$, while if the transformation is more local, then the form of ψ should itself change as our test latent point $\mathbf{z}$ changes.

To make this problem tractable, let us assume we are currently in a state with large noise $\mathbf{x}_t$, where $t \approx T$. At high noise levels, the optimal estimate of the clean image that minimizes the mean squared error is the mean of the training set. This suggests that for a well-trained denoiser, we should expect the same output irrespective of the network's input:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t; \theta) = \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}\left[\mathbf{x}_0\right] \approx \mathbb{E}_{p_{\text{dataset}}(\mathbf{x}_0)}\left[\mathbf{x}_0\right] = \mathbf{c}.$$

where $\mathbf{c}$ is a constant image vector. Under this assumption, our function f is no longer dependent on the denoiser but only depends on the inference network:

$$f_\phi(\mathbf{x}_t,\ \mathbf{z}) = \mathbf{c} + \gamma_t\ \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t).$$

This means the transformation ψ is also independent of the denoiser. It is purely a function of the guidance vector:

$$\begin{aligned}
\mathbf{x}_0^* + \psi(\mathbf{x}_0^*,\ \mathbf{z},\ i) &= \mathbf{c} + \gamma_t \nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z} + \alpha\hat{\mathbf{i}}|\mathbf{x}_t) \\
&= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\left(\mathbf{z} + \alpha\,\hat{\mathbf{i}} - \boldsymbol{\mu}_\phi\right)^\top \Sigma_\phi^{-1}\left(\mathbf{z} + \alpha\,\hat{\mathbf{i}} - \boldsymbol{\mu}_\phi\right) + \log\left|\Sigma_\phi\right| - N\log(2\pi)\Big],
\end{aligned}$$

where $\mathbf{z}$ is a function of $\mathbf{x}_0$, and the means $\boldsymbol{\mu}_\phi$ and covariances Σ_ϕ are functions of the current noisy image $\mathbf{x}_t$. If we decompose the RHS of the above equation into the contribution of just $\mathbf{z}$ and the contribution of $\alpha\,\hat{\mathbf{i}}$, and denote the last two terms as the scalar s, we get:

$$\begin{aligned}
&\hat{\mathbf{x}}_0^* + \psi(\hat{\mathbf{x}}_0^*,\ \mathbf{z},\ i) \\
&= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big)^\top\Sigma_\phi^{-1}\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big) + 2\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big)^\top\Sigma_\phi^{-1}(\alpha\,\hat{\mathbf{i}}) + (\alpha\,\hat{\mathbf{i}})^\top\Sigma_\phi^{-1}(\alpha\,\hat{\mathbf{i}}) + s\ \Big] \\
&= \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big)^\top\Sigma_\phi^{-1}\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big) + 2\alpha\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big)^\top(\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii} + s\ \Big],
\end{aligned}$$

where $(\Sigma_\phi^{-1})_i$ is the i-th column of the inverse covariance matrix, and $(\Sigma_\phi^{-1})_{ii} = \sigma_i^{-2}$ is the i-th element along the diagonal of the inverse covariance matrix. We can now substitute $\hat{\mathbf{x}}_0^* = \mathbf{c} - \frac{\gamma_t}{2}\nabla_{\mathbf{x}_t}\Big[\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big)^\top\Sigma_\phi^{-1}\big(\mathbf{z}-\boldsymbol{\mu}_\phi\big) + s\ \Big]$, which yields

$$\psi(\mathbf{z},\ i) = \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}-\boldsymbol{\mu}_\phi)^\top(\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big].$$

Let's see how this transformation in $\mathcal{X}$ changes with our location $\mathbf{z}$ in latent space $\mathcal{Z}$. First suppose we use the same latent perturbation $\alpha\,\hat{\mathbf{z}}_i$ irrespective of where we are in latent space. For a given state $\mathbf{x}_t$, we compare two instances of ψ at different values of $\mathbf{z}$:

$$\begin{aligned}
\psi_1(\mathbf{z}^{(1)},\ i) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(1)}-\boldsymbol{\mu}_\phi)^\top(\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big] \\
\psi_2(\mathbf{z}^{(2)},\ i) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(2)}-\boldsymbol{\mu}_\phi)^\top(\Sigma_\phi^{-1})_i + \alpha^2\,(\Sigma_\phi^{-1})_{ii}\Big].
\end{aligned}$$

Since $\boldsymbol{\mu}_\phi$ and Σ_ϕ are functions of a common $\mathbf{x}_t$, their values are the same in both ψ_1 and ψ_2. The difference between these two transformations is therefore

$$\begin{aligned}
\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) &= \nabla_{\mathbf{x}_t}\Big[2\alpha(\mathbf{z}^{(1)}-\boldsymbol{\mu}_\phi)^\top(\Sigma_\phi^{-1})_i - 2\alpha(\mathbf{z}^{(2)}-\boldsymbol{\mu}_\phi)^\top(\Sigma_\phi^{-1})_i\Big] \\
&= 2\alpha\nabla_{\mathbf{x}_t}\Big[(\mathbf{z}^{(1)}-\mathbf{z}^{(2)})^\top(\Sigma_\phi^{-1})_i\Big].
\end{aligned}$$

We see that the difference in the transformation scales linearly with the difference in $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$, and scales with the i-th column of the inverse covariance matrix. If we assume a diagonal covariance assumption, as in our model, this is simply a unit vector scaled by the i-th element along the diagonal:

$$\begin{aligned}\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) &= 2\alpha\, \nabla_{\mathbf{x}_t}\left[(\mathbf{z}^{(1)} - \mathbf{z}^{(2)})^\top \sigma_i^{-2}\hat{\mathbf{z}}_i\right] \\ &= 2\alpha\left(\nabla_{\mathbf{x}_t}\sigma_i^{-2}\right)\left(\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\right), \end{aligned} \tag{B.3}$$

where in the second line we use the fact that only σ_i^{-2} is a function of $\mathbf{x}_t$. This means that if the clean images are mapped to similar values in this dimension (e.g. if the population variance in this dimension is very low), then the function that is applied in the data space will be very similar. The effect is also scaled by the derivative of the precision: we want to look for axes where the precision isn't affected by perturbations around $\mathbf{x}_t$:

$$\nabla_{\mathbf{x}_t}\sigma_i^{-2}(\mathbf{x}_t) = \frac{\partial}{\partial\sigma_i^2}\left(\frac{1}{\sigma_i^2}\right)\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t} = -\frac{1}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t}.$$

Plugging this into Eq. B.3 yields

$$\psi_1(\mathbf{z}^{(1)}) - \psi_2(\mathbf{z}^{(2)}) = -\frac{2\alpha}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\mathbf{x}_t}\left(\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\right).$$

Global axes are those that minimize the difference in ψ. We can find axes that satisfy three properties:

- small variance in the population of clean image latents,
- large feature uncertainty σ_i for our noisy image state $\mathbf{x}_t$,
- where the variance of the noisy latent σ_i is insensitive to changes in our noisy image $\mathbf{x}_t$.

Though our analysis relies on the assumption that we are currently at a large noise regime, this is

also the regime in which the guidance vector has the most influence over the sampling process, since the noise-dependent factor γ_t in Eq. B.2 —which trades off between the influence of the guidance vector and the denoiser score—scales monotonically with t.

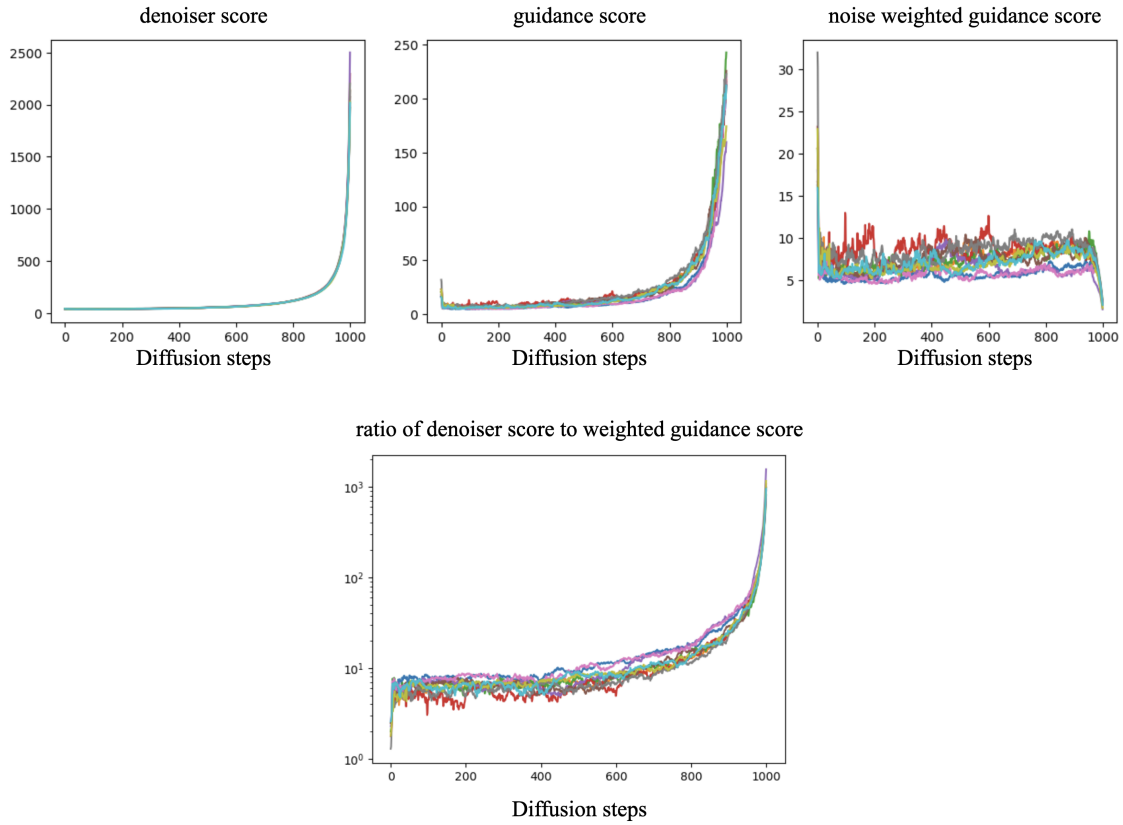


Figure B.3: Comparing the norms of the denoiser score and the guidance score. Larger (rightward) denoising steps corresponds to smaller noise level. The ratio is closest to one at the start of the diffusion process, at large noise levels.

Above, we plot the ratio between the norm of the denoiser score and the norm of the weighted guidance score. A smaller ratio indicates that the guidance score has more relative strength. We see that the ratio is smallest at large noise levels, at the start of the diffusion process.

B.1.9 Comparisons to other models

In Table B.3 we compare our model to the other diffusion-based representation learning models in the literature. Models with a † sign indicates that they use a separate neural network to

Property	**SAMI** (ours)	**DiffAE** Preechakul et al., 2022	**DRL** Mittal et al., 2023	**InfoDiff** Wang et al., 2023	**PDAE** Zhang et al., 2022b	**SODA** Hudson et al., 2024	**DisDiff** Yang et al., 2023b	**DBAE** Kim et al., 2024
Objective function	ELBO	MSE	MSE+L1 reg.	ELBO+MI reg.	ELBO	MSE	MSE+CE reg.	MSE
Guidance mechanism	Sum	Network†	Network†	Network†	Network†	Network†	Network†	Network†
Probabilistic latents	Yes	No	No	Yes	No	No	No	No
Unsupervised learning	Yes	Yes	Yes	Yes	Yes	No*	Yes	Yes
Using pre-trained DM	Both	No	No	No	Yes	No	No	No
Encodes noisy images	Yes	No	No	No	No	No	No	No
Enc. blind to noise level	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Exact guidance score calc.	Yes	No	No	No	No	No	No	No
Interpretable latent axes	Yes	No	No	Yes	No	No	No	No

Table B.3: Comparison to other diffusion-based representation learning models in the literature.

perform guidance by approximating the gradient of the log operator. In all networks other than ours, a linear projection of the latent is used to modulate the UNet denoiser's Adaptive GroupNorm layers. The asterisk (*) in the SODA column indicates that while SODA is trained in a self-supervised manner using multiple views of the original datum, their formulation allows for unsupervised training as well. DisDiff uses additional cross entropy losses to encourage representational invariance to a subset of features. However, we find that a combinatorial representation emerges automatically from training our diffusion process to use the exact log likelihood for guidance.

B.2 Proofs

B.2.1 Training on both noisy and clean images smoothens the latent space

We show below that training our inference network q_ϕ on both noisy and clean images implicitly regularizes the latent space to be smoother.

The objective (Eq. 3.10) has two components. The reconstruction term is equal to $\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}|\mathbf{x}_0}\|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t) - \beta_t^2 g_t(\mathbf{x}_t,\mathbf{z})\|^2$, where $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ is estimated by a standalone MSE based denoiser and $g_t(\mathbf{x}_t,\mathbf{z}) = \nabla_{\mathbf{x}_t}\log p_\phi(\mathbf{z}|\mathbf{x}_t)$ is the guidance vector, where $q_\phi(\mathbf{z}|\mathbf{x}_t) = \mathcal{N}(\mu_\phi(\mathbf{x}_t), \mathrm{diag}(\sigma_\phi(\mathbf{x}_t)))$ is our encoder of $\mathbf{x}_t$, and the log likelihood $\log p_\phi(\mathbf{z}|\mathbf{x}_t)$ is evaluated at $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}_0)$. Let's denote the error between the clean noise and the noise estimate as the *residual* $r(\mathbf{x}_0,\mathbf{x}_t) = \mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t)$. In this case, the reconstruction term can be written as:

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\|\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t) - \beta_t^2 g_t(\mathbf{x}_t,\mathbf{z})\|^2 \\
&= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}_0)}\|r(\mathbf{x}_0,\mathbf{x}_t) - \beta_t^2 g_t(\mathbf{x}_t,\mathbf{z})\| \\
&= \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t}\|r(\mathbf{x}_0,\mathbf{x}_t)\|^2 - 2\beta_t^2\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}\left[r(\mathbf{x}_0,\mathbf{x}_t)^\top g_t(\mathbf{x}_t,\mathbf{z})\right] + \beta_t^4\,\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t,\mathbf{z}}\|g_t(\mathbf{x}_t,\mathbf{z})\|^2 \\
&= \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3,
\end{aligned}
$$

where we have denoted each of the three terms as $\mathcal{L}_i$. Since we assume that the encoder is Gaussian, we can expand the conditional score as

$$
\log q_\phi(\mathbf{z}|\mathbf{x}_t) = -\frac{1}{2}\frac{(\mathbf{z} - \mu_\phi(\mathbf{x}_t))^2}{\sigma_\phi^2(\mathbf{x}_t)} - \log\sigma_\phi(\mathbf{x}_t) - \frac{d}{2}\log(2\pi)
$$

so now the score is

$$\nabla_{\mathbf{x}_t} \log q_\phi(\mathbf{z}|\mathbf{x}_t) = \frac{\mathbf{z} - \mu_\phi(\mathbf{x}_t)}{\sigma_\phi^2(\mathbf{x}_t)} \nabla_{\mathbf{x}_t} \mu_\phi(\mathbf{x}_t) - \frac{(\mathbf{z} - \mu_\phi(\mathbf{x}_t))^2}{\sigma_\phi^4(\mathbf{x}_t)} \nabla_{\mathbf{x}_t} \sigma_\phi^2(\mathbf{x}_t) - \frac{\nabla_{\mathbf{x}_t} \sigma_\phi^2(\mathbf{x}_t)}{2\sigma_\phi^2(\mathbf{x}_t)}$$

Let's make a simplifying assumption that $\sigma_\phi^2(\mathbf{x}_t)$ is not $\mathbf{x}_t$ dependent and is instead a constant value σ^2. In that case,

$$g_t(\mathbf{x}_t, \mathbf{z}) = \frac{\mathbf{z} - \mu_\phi(\mathbf{x}_t)}{\sigma^2} \nabla_{\mathbf{x}_t} \mu_\phi(\mathbf{x}_t)$$

Now let $J_\mu(\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \mu_\phi(\mathbf{x}_t) \in \mathbb{R}^{d \times n}$ be the encoder Jacobian.

$$g_t(\mathbf{x}_t, \mathbf{z}) = J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} (\mathbf{z} - \mu_\phi(\mathbf{x}_t))$$

where $\Sigma = \text{diag}(\sigma^2)$. Now let's assume again that $\mathbf{x}_t = \mathbf{x}_0 + \beta_t \epsilon$. We can expand $g(\mathbf{x}_t, \mathbf{z})$:

$$g_t(\mathbf{x}_t, \mathbf{z}) = g_t(\mathbf{x}_0, \mathbf{z}) + \beta_t \frac{\partial g_t}{\partial \mathbf{x}_t}\bigg|_{\mathbf{x}_t = \mathbf{x}_0} \epsilon + \frac{\beta_t^2}{2} \epsilon^\top \frac{\partial^2 g_t}{\partial \mathbf{x}_t^2}\bigg|_{\mathbf{x}_t = \mathbf{x}_0} \epsilon + O(\beta_t^3)$$

Notation: if we evaluate the Jacobian at $\mathbf{x}_0$, we will denote it as $J_\mu(\mathbf{x}_0)$. Otherwise, we will use $J_\mu = J_\mu(\mathbf{x}_t)$. The zeroth order derivative:

$$g_t(\mathbf{x}_0, \mathbf{z}) = J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} (\mathbf{z} - \mu_\phi(\mathbf{x}_0))$$

We can write this in index form:

$$g_t^{(k)}(\mathbf{x}_t, \mathbf{z}) = \sum_j \frac{\mathbf{z}_j - \mu_{\phi,k}(\mathbf{x}_0)}{\sigma_j^2} \frac{\partial \mu_{\phi,j}(\mathbf{x}_0)}{\partial \mathbf{x}_{t,k}}$$

Now, taking the derivative of $g_t^{(k)}(\mathbf{x}_t, \mathbf{z})$ with respect to $\mathbf{x}_{t,i}$, we get

$$\frac{\partial g_t^{(k)}}{\partial \mathbf{x}_{t,i}} = \sum_j \frac{1}{\sigma^2} \left[-\frac{\partial \mu_{\phi,j}}{\partial \mathbf{x}_{t,i}} \frac{\partial \mu_{\phi,j}}{\partial \mathbf{x}_{t,k}} + (\mathbf{z}_j - \mu_{\phi,j}) \frac{\partial^2 \mu_{\phi,j}}{\partial \mathbf{x}_{t,i} \partial \mathbf{x}_{t,k}} \right]$$

which in tensor form is

$$\frac{\partial g_t}{\partial \mathbf{x}_t} = -J_\mu^\top \Sigma^{-1} J_\mu + \Sigma^{-1} \sum_j (\mathbf{z}_j - \mu_{\phi,j}) H_j$$

where H_j is the Hessian matrix of $\mu_{\phi,j}$ with entries

$$H_j = \left[\frac{\partial^2 \mu_{\phi,j}}{\partial \mathbf{x}_{t,i} \partial \mathbf{x}_{t,k}} \right]_{ik}$$

so we get a Hessian weighted by the encoder residual $(\mathbf{z}_j - \mu_{\phi,j})$. Putting these derivatives back in, we get:

$$g_t(\mathbf{x}_t, \mathbf{z}) = J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} (\mathbf{z} - \mu_\phi(\mathbf{x}_0)) + \beta_t \left(-J_\mu^\top \Sigma^{-1} J_\mu + \Sigma^{-1} \sum_j (\mathbf{z}_j - \mu_{\phi,j}) H_j \right) + O(\beta_t^2).$$

Since the guidance score $g_t(\mathbf{x}_t, \mathbf{z})$ is evaluated at the clean image latent sample, i.e. $\mathbf{z} \sim \mathcal{N}(\mu_\phi(\mathbf{x}_0), \Sigma)$, we take the expectation over the distributions $q_\phi(\mathbf{z}|\mathbf{x}_0)$:

$$\mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[\mathbf{z} - \mu_\phi(\mathbf{x}_t) \right] = \mu_\phi(\mathbf{x}_0) - \mu_\phi(\mathbf{x}_t)$$

At $\mathbf{x}_t = \mathbf{x}_0$, the means are equal, so $\mathbb{E}_{\mathbf{z}|\mathbf{x}_0}[\mathbf{z} - \mu_\phi(\mathbf{x}_t)] = 0$. And at $\mathbf{x}_t = \mathbf{x}_0 + \beta_t \epsilon$,

$$\mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[\mathbf{z} - \mu_\phi(\mathbf{x}_0 + \beta_t \epsilon) \right] \approx \mu_\phi(\mathbf{x}_0) - \mu_\phi(\mathbf{x}_0) - \beta_t J_\mu \epsilon = -\beta_t J_\mu \epsilon$$

This means that the expected score is

$$\mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] = J_\mu^\top \Sigma^{-1} \mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[(\mathbf{z} - \mu_\phi(\mathbf{x}_t)) \right] \approx -J_\mu^\top \Sigma^{-1} \beta_t J_\mu \epsilon$$

Now we want to expand the Jacobian $J_\mu(\mathbf{x}_t)$ around the point $\mathbf{x}_0$:

$$J_\mu(\mathbf{x}_t) = J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j H_j + O(\beta_t^2)$$

where we use the expression for H_j that we used above. Plugging this into the expectation, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}|\mathbf{x}_0}\left[g_t(\mathbf{x}_t, \mathbf{z})\right] \\
&= -\beta_t \left(J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j H_j\right)^\top \Sigma^{-1} \left(J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j H_j\right) \\
&= -\beta_t J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) + 2\beta_t^2 \left(J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} \sum_j \epsilon_j H_j\right) - \beta_t^3 \sum_{j,k} \epsilon_j \epsilon_k H_j^\top \Sigma^{-1} H_k \\
&= -\beta_t J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) + O(\beta_t^2).
\end{aligned}$$

Now we will use these identities to evaluate $\mathcal{L}_2$ and $\mathcal{L}_3$.

Expanding loss term 2 The second term in the loss is given by

$$\mathcal{L}_2 = -2\beta_t^2 \, \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, \mathbf{z}|\mathbf{x}_0}\left[r(\mathbf{x}_0, \mathbf{x}_t)^\top g_t(\mathbf{x}_t, \mathbf{z})\right].$$

We can use the results above for $\mathbb{E}_{\mathbf{z}|\mathbf{x}_0}\left[g_t(\mathbf{x}_t, \mathbf{z})\right] \approx -\beta_t J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0)\epsilon$ to rewrite this as

$$\mathcal{L}_2 = -2\beta_t^2 \mathbb{E}_{\mathbf{x}_0, \epsilon}\left[-\beta_t r(\mathbf{x}_0, \mathbf{x}_t)^\top J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0)\epsilon\right]$$

Remember that $r(\mathbf{x}_0, \mathbf{x}_t) = \mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_0 + \beta_t\epsilon)$. If $\hat{\mathbf{x}}_0$ is a good denoiser,

$$\hat{\mathbf{x}}_0(\mathbf{x}_0 + \beta_t\epsilon) \approx \mathbb{E}\left[\mathbf{x}_0|\mathbf{x}_t\right] = \mathbf{x}_0 + \beta_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

So

$$r(\mathbf{x}_0, \mathbf{x}_t) \approx \mathbf{x}_0 - \mathbf{x}_0 - \beta_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) = -\beta_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

We can plug this back into $\mathcal{L}_2$ to get:

$$\mathcal{L}_2 = 2\beta_t^6 \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \epsilon \right]$$

When we take the expectation over ϵ, we find that

$$\mathcal{L}_2 = c \cdot \mathbb{E}_\epsilon \left[\epsilon \right] = 0,$$

so this entire term disappears at the leading order, only contributing at order $\mathcal{O}(\beta_t^3)$ or higher. However, we can show that this term disappears fully, even at higher orders. If $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ is an optimal MSE denoiser, it computes the conditional expectation. One property of MMSE estimators is their estimation error $r(\mathbf{x}_0, \mathbf{x}_t)$ are orthogonal to any function of the observation $\mathbf{x}_t$. Since the guidance score $g_t(\mathbf{x}_t, \mathbf{z})$ is a function of $\mathbf{x}_t$, after marginalizing out the contribution of $\mathbf{z}$, this means that

$$\mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t} \left[r(\mathbf{x}_0, \mathbf{x}_t)^\top \mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] \right] = 0.$$

Expanding loss term 3 The third loss term is $\mathcal{L}_3 = \beta^4 \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, \mathbf{z}} \|g_t(\mathbf{x}_t, \mathbf{z})\|^2$. From the bias-variance decomposition, we have

$$\mathcal{L}_3 = \beta_t^4 \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, \mathbf{z}|\mathbf{x}_0} \left[\|g_t(\mathbf{x}_t, \mathbf{z})\|^2 \right] = \beta_t^4 \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t} \left[\|\mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] \|^2 + \text{tr} \left(\text{var}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] \right) \right]$$

For the first (bias) term, we can use the identity for the expected guidance score:

$$\beta_t^4 \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t} \left[\left\| \mathbb{E}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] \right\|^2 \right] \approx \beta_t^6 \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\left\| J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \epsilon \right\|^2 \right]$$

where we have substituted the expectation over $\mathbf{x}_t$ with an functionally equivalent expectation over the noise ϵ. For Gaussian ϵ, we can use the identity

$$\mathbb{E}_\epsilon \left[\left\| \epsilon^\top A \right\|^2 \right] = \text{tr}\left(A^\top A\right) = \|A\|_F^2$$

So now

$$\beta_t^6 \mathbb{E}_{\mathbf{x}_0,\epsilon} \left[\left\| \epsilon^\top J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \right\|^2 \right] = \beta_t^6 \, \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \right\|_F^2$$

The second (variance) term is dependent on the variance of the score:

$$\begin{aligned}
\text{var}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] &= J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} \text{var}\left[\mathbf{z} - \mu_\phi(\mathbf{x}_0) \right] \Sigma^{-1} J_\mu(\mathbf{x}_t) \\
&= J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} \Sigma \Sigma^{-1} J_\mu(\mathbf{x}_t) \\
&= J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} J_\mu(\mathbf{x}_t)
\end{aligned}$$

Taking the trace of this gives us

$$\text{tr}\left(\text{var}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] \right) = \text{tr}\left(J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} J_\mu(\mathbf{x}_t) \right) = \| J_\mu(\mathbf{x}_t) \Sigma^{-1/2} \|_F^2$$

If we plug the Taylor expansion of $J_\mu(\mathbf{x}_t)$ (above) into this equation, we get

$$\begin{aligned}
\text{var}_{\mathbf{z}|\mathbf{x}_0} \left[g_t(\mathbf{x}_t, \mathbf{z}) \right] &= J_\mu(\mathbf{x}_t)^\top \Sigma^{-1} J_\mu(\mathbf{x}_t) \\
&= \left(J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j H_j \right)^\top \Sigma^{-1} \left(J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j H_j \right) \\
&= J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) + \beta_t \sum_j \epsilon_j \left(J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} H_j + H_j^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \right) \\
&\quad + \beta_t^2 \sum_{j,k} \epsilon_j \epsilon_k H_j^\top \Sigma^{-1} H_k + \mathcal{O}(\beta_t^3)
\end{aligned}$$

Taking the trace of this quantity up to second order terms, we get

$$\begin{aligned}\operatorname{tr}\left(\operatorname{var}_{\mathbf{z}|\mathbf{x}_0}\left[g_t(\mathbf{x}_t, \mathbf{z})\right]\right) &= \|J_\mu(\mathbf{x}_t)\Sigma^{-1/2}\|_F^2 \\ &= \|J_\mu(\mathbf{x}_0)\Sigma^{-1/2}\|_F^2 + 2\beta_t \sum_j \epsilon_j \operatorname{tr}\left(J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} H_j\right) \\ &\quad + \beta_t^2 \sum_{j,k} \epsilon_j \epsilon_k \operatorname{tr}\left(H_j^\top \Sigma^{-1} H_k\right)\end{aligned}$$

This trace term is under expectation of $\mathbf{x}_0$ and ϵ. The expectation over ϵ sets the β_t term to 0, while the β_t^2 term becomes

$$\mathbb{E}_\epsilon\left[\beta_t^2 \sum_{j,k} \epsilon_j \epsilon_k \operatorname{tr}\left(H_j^\top \Sigma^{-1} H_k\right)\right] = \beta_t^2 \sum_j \operatorname{tr}\left(H_j^\top \Sigma^{-1} H_j\right)$$

since $\mathbb{E}_\epsilon\left[\epsilon_j \epsilon_k\right] = \mathrm{Id}_{jk} = \delta_{jk}$, i.e. 1 when $j = k$ and 0 otherwise. The first equality holds because the mean is 0 and the second equality holds because the covariance is an identity matrix. Using the identity $\operatorname{tr}(A^\top A) = \sum_{j,k} A_{jk}^2 = \sum_j \|A\|_F^2$, we can simplify this term:

$$\sum_j \operatorname{tr}\left(H_j^\top \Sigma^{-1} H_j\right) = \sum_j \frac{1}{\sigma_j^2} \operatorname{tr}\left(H_j^\top H_j\right) = \sum_j \frac{1}{\sigma_j^2} \left\|H_j\right\|_F^2 .$$

So we are left with

$$\mathbb{E}_\epsilon\left[\operatorname{tr}\left(\operatorname{var}_{\mathbf{z}|\mathbf{x}_0}\left[g_t(\mathbf{x}_t, \mathbf{z})\right]\right)\right] = \|J_\mu(\mathbf{x}_0)\Sigma^{-1/2}\|_F^2 + \beta_t^2 \sum_j \frac{1}{\sigma_j^2} \left\|H_j\right\|_F^2 .$$

Putting together the bias and variance terms gives us:

$$\mathcal{L}_3 = \beta_t^4 \, \mathbb{E}_{\mathbf{x}_0} \left\|J_\mu(\mathbf{x}_0)\Sigma^{-1/2}\right\|_F^2 + \beta_t^6 \, \mathbb{E}_{\mathbf{x}_0}\left[\sum_j \frac{1}{\sigma_j^2} \left\|H_j\right\|_F^2\right] + \beta_t^6 \, \mathbb{E}_{\mathbf{x}_0} \left\|J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0)\right\|_F^2$$

Combining the losses Now, putting the second and third term back into the loss, we get

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\epsilon} \|r(\mathbf{x}_0, \mathbf{x}_t)\|^2 + \beta_t^4 \, \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0) \Sigma^{-1/2} \right\|_F^2 + \beta_t^6 \, \mathbb{E}_{\mathbf{x}_0} \left[\sum_j \frac{1}{\sigma_j^2} \left\| H_j \right\|_F^2 \right] + \beta_t^6 \, \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0)^\top \Sigma^{-1} J_\mu(\mathbf{x}_0) \right\|_F^2$$

plus higher orders. How should we interpret this loss?

The order β_t^4 term regularizes the Jacobian by encouraging the encoder to be contractive, basically penalizing the encoder from mapping small changes in the input space to large changes in latent space. It is weighted by $\Sigma^{-1/2}$, so it enforces the more informative (high-precision) dimensions to be more stable.

The second term has order β_t^6, and this penalizes the Frobenius norm of the Hessian, weighted by the precision. This forces the manifold of the latent space to be locally flat, with more informative (high-precision) dimensions getting a stronger penalty on their curvature. This ameliorates the "latent holes" issue in VAEs where latent interpolation fails by ensuring that linear changes in image space locally correspond to linear changes in latent space.

Extending to all noise levels One weakness of a proof based on Taylor expansions is that it only holds when the noise level (β_t) is small. However, the variance exploding noising process we assume ($\mathbf{x}_t = \mathbf{x}_0 + \beta_t \epsilon$) allows us to write noisy images in terms of slightly less noisy images. We are going to exploit this iterative definition to apply the above proof at all noise levels. Let's define an intermediate timestep that differs from timestep t by a small value δt. Due to the Markov property of the diffusion process, we can write the iterative noising process as

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \delta \mathbf{x}_t,$$

where $\delta \mathbf{x}_t \sim \mathcal{N}(0, \gamma_t^2 I)$ and we have defined the incremental noise variance as $\gamma_t^2 = \beta_t^2 - \beta_{t-1}^2$. Since δt is small, γ_t^2 is also small. We can now apply the same steps as the original proof, but instead of expanding around $\mathbf{x}_0$, we expand around the previous noisy state $\mathbf{x}_{t-1}$. This means that the Jacobian is approximated as

$$\begin{aligned} J_\mu(\mathbf{x}_t) &= J_\mu(\mathbf{x}_{t-1}) + \sum_j \delta \mathbf{x}_{t,j} H_j(\mathbf{x}_{t-1}) + O(\gamma_t^2) \\ &= J_\mu(\mathbf{x}_{t-1}) + \mathcal{H}(\mathbf{x}_{t-1}) \cdot \delta \mathbf{x}_t + O(\gamma_t^2). \end{aligned}$$

where $\mathcal{H} \cdot \delta \mathbf{x}_t$ is the Hessian tensor product with a sample of scaled white noise. Since γ_t is small, the Taylor expansions hold regardless of the size of the total noise level t. Now let's plug this into the third loss term that penalizes the curvature. In this case, we get an per-noise level expression in terms of Jacobians and Hessians evaluated at $\mathbf{x}_{t-1}$:

$$\begin{aligned} \mathcal{L}_3(t) = \gamma_t^4 \, \mathbb{E}_{\mathbf{x}_{t-1}} \left\| J_\mu(\mathbf{x}_{t-1}) \Sigma^{-1/2} \right\|_F^2 + \gamma_t^6 \, \mathbb{E}_{\mathbf{x}_{t-1}} \left[\sum_j \frac{1}{\sigma_j^2} \left\| H_j(\mathbf{x}_{t-1}) \right\|_F^2 \right] \\ + \gamma_t^6 \, \mathbb{E}_{\mathbf{x}_{t-1}} \left\| J_\mu(\mathbf{x}_{t-1})^\top \Sigma^{-1} J_\mu(\mathbf{x}_{t-1}) \right\|_F^2 . \end{aligned}$$

However, since we train over all noise levels with some distribution, e.g. $t \sim \mathcal{U}[0, \infty]$, the total regularization effect is given by the weighted average of all t dependent terms over the entire trajectory.

We can express this expected loss in terms of $\mathbf{x}_0$ by applying the Taylor expansion to the

Jacobian recursively over all noise levels, which gives us

$$\begin{aligned} J_\mu(\mathbf{x}_t) &\approx J_\mu(\mathbf{x}_{t-1}) + \mathcal{H}(\mathbf{x}_{t-1}) \cdot \delta\mathbf{x}_t \\ &= J_\mu(\mathbf{x}_{t-2}) + \mathcal{H}(\mathbf{x}_{t-2}) \cdot \delta\mathbf{x}_{t-1} + \mathcal{H}(\mathbf{x}_{t-1}) \cdot \delta\mathbf{x}_t \\ &\dots \\ &= J_\mu(\mathbf{x}_0) + \sum_{s=1}^{t} \mathcal{H}(\mathbf{x}_{s-1}) \cdot \delta\mathbf{x}_s \end{aligned}$$

In the limit of $\delta \to 0$, this sum turns into a stochastic integral with respect to the Brownian motion $\mathbf{x}_s$ of the noisy image:

$$J_\mu(\mathbf{x}_t) = J_\mu(\mathbf{x}_0) + \int_0^t \mathcal{H}(\mathbf{x}_s)\, d\mathbf{x}_s.$$

We can now plug this into the third loss term again to express it in terms of $\mathbf{x}_0$. For notational convenience, let's set $\Sigma^{-1/2} = \mathrm{Id}$.

$$\begin{aligned} \mathcal{L}_3(t) &\propto \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0)\Sigma^{-1/2} \right\|_F^2 \\ &= \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0) + \int_0^t \mathcal{H}(\mathbf{x}_s)\, d\mathbf{x}_s \right\|_F^2. \end{aligned}$$

We can expand the squared norm term using the identity $\|A+B\|^2 = \|A\|^2 + \|B\|^2 + 2\langle A, B\rangle$, which gives us

$$\mathcal{L}_3(t) = \mathbb{E}_{\mathbf{x}_0} \left\| J_\mu(\mathbf{x}_0) \right\|_F^2 + \mathbb{E}_{\mathbf{x}_0} \left\| \int_0^t \mathcal{H}(\mathbf{x}_s) d\mathbf{x}_s \right\|_F^2 + 2\mathbb{E}_{\mathbf{x}_0} \left\langle J_\mu(\mathbf{x}_0), \int_0^t \mathcal{H}(\mathbf{x}_s) d\mathbf{x}_s \right\rangle.$$

The first term is constant over time since it is not t dependent. The cross product term goes to 0 because the expectation of the martingale is 0. For the second term, we can use the Itô isometry, which allows us to rewrite the expected squared integral as the expected integral of the norm of

the Hessians.

$$\mathbb{E}_{\mathbf{x}_0}\left\|\int_0^t \mathcal{H}(\mathbf{x}_s)d\mathbf{x}_s\right\|_F^2 = \mathbb{E}_{\mathbf{x}_0}\left[\int_0^{\beta_t} \|\mathcal{H}(\mathbf{x}_s)\|_F^2\, d\beta_s^2\right].$$

If we consider all possible noise levels, the total expected loss becomes

$$\mathbb{E}_{t\sim[0,\infty]}\left[\mathcal{L}_3(t)\right] = \mathbb{E}_{\mathbf{x}_0,t}\left[\|J_\mu(\mathbf{x}_0)\|_F^2 + \int_0^{\infty} \|\mathcal{H}(\mathbf{x}_s)\|_F^2 d\beta_s^2\right].$$

Conclusion By expressing the total loss in terms of the clean image, we see that we are implicitly training the encoder to minimize the integral of the Hessian, weighted by the noise level β_t^2. This essentially encourages the network to smooth the curvature along the entire denoising *trajectory*.

B.2.2 The SAMI objective encourages disentanglement across scales

We show that when the data are generated by independent features that differ in how strongly they show up in the image, the SAMI objective prefers a representation in which each latent coordinate carries a single feature, and in which the same coordinate carries the same feature at every noise level. The coordinates switch on one at a time as the image gets cleaner, in order of how loud each feature is, so the representation grows as a nested hierarchy.

The mechanism that aligns the latent axes at a single noise level is not new. A VAE whose encoder outputs a diagonal covariance cannot represent a posterior whose coordinates are correlated, and this pushes the latent axes into alignment with the natural axes of the problem. Rolínek et al. (2019) derive this through the geometry of the decoder, and Kumar and Poole (2020) show that the uniqueness comes from the restricted variational family rather than from the prior. What is new here is what happens when this mechanism is applied across a whole family of noise levels at once. The feature that is loud enough to survive heavy noise is not the same feature that is loud enough to survive light noise, so each feature has its own noise level at which it becomes visible. We show that these levels are distinct, that they order the features, and that they give

each latent coordinate a name that does not depend on which noise level we happen to be looking at. That last point is what turns a per-level alignment result into a statement about a single stable hierarchy.

B.2.2.1 Setup

The clean image is generated from K independent features, $x = g(y_1, \dots, y_K)$, and the noisy image at signal level γ is

$$x_\gamma = \sqrt{\gamma}\, x + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_D),$$

so large γ means a clean image and small γ a heavily corrupted one. The encoder is the usual diagonal Gaussian,

$$q_\phi(z \mid x_\gamma) = \mathcal{N}\big(\mu_\phi(x_\gamma), \operatorname{diag}(\sigma^2_\phi(x_\gamma))\big),$$

and the prior is $p(z) = \mathcal{N}(0, I)$. SAMI maximizes the ELBO with the KL term weighted by β,

$$\mathcal{L}_\beta = \mathbb{E}\big[\log p_\theta(x_\gamma \mid z)\big] - \beta\, \mathbb{E}\big[D_{\mathrm{KL}}\big(q_\phi(z \mid x_\gamma) \,\|\, p(z)\big)\big]. \tag{B.4}$$

Write $\hat{y}_k = \mathbb{E}[y_k \mid x_\gamma]$ for the best estimate of feature k from the noisy image. The question is which of the many equally good codes the objective picks: if z carries the same information as $\hat{\mathbf{y}}$, then so does $A\hat{\mathbf{y}}$ for any invertible $K \times K$ matrix A, and a decoder can undo the mixing. Reconstruction quality alone therefore cannot tell these codes apart. The following assumptions specify how we distinguish between these codes.

Assumption 1. (S1) **Independent features.** The features $y_1, \dots, y_K$ are mutually independent with finite variance.

(S2) **Distinct feature energies.** Each feature has a signal energy $a_k > 0$ measuring how strongly it moves the image, $a_k = \|\partial x / \partial y_k\|^2$. The energies are distinct and separated by a fixed ratio: $a_k \geqslant (1 + \delta) a_{k+1}$ for some $\delta > 0$, so we may order them $a_1 > a_2 > \cdots > a_K$. We write

$\gamma_k \asymp 1/a_k$ for the noise level at which feature k becomes visible.

(S3) **Concentrated posterior.** At each level γ the posterior over the currently visible features is well approximated by a Gaussian whose precision matrix in the feature frame is $\gamma\,\mathrm{diag}(a_1, \ldots, a_K)$. This is the usual Laplace or Bernstein–von Mises approximation (Vaart, 1998), and it is where the assumption that features act on separate parts of the image enters.

(S4) **Linear competitors.** The codes being compared are $\mu_\phi = A\hat{\mathbf{y}}$ for invertible A, followed by any rescaling of individual coordinates. We are asking which mixing matrix the objective prefers, not searching over all possible nonlinear codes.

(S5) **Smooth dependence on the noise level.** The encoder is a single network taking both x_γ and γ as input, and its parameters vary continuously with γ.

B.2.2.2 Why a diagonal encoder picks the feature axes

The key point is that the encoder is forced to report a diagonal covariance, and a diagonal covariance cannot describe a posterior whose coordinates are correlated. In the feature frame the posterior is uncorrelated by (S3), so nothing is lost. In a rotated frame it is correlated, and the mismatch costs the objective something. The next lemma computes that cost exactly.

Lemma B.1 (Cost of a diagonal encoder in a mixed frame)**.** *Suppose the true posterior over the features is Gaussian with precision matrix P, and the code is $z = A\hat{\mathbf{y}}$. Among all diagonal-covariance Gaussian encoders, the smallest achievable gap between the ELBO and the true log likelihood is*

$$\Delta(A) = \tfrac{1}{2}\left[\sum_{i=1}^{K} \log\left(A^{-\top}PA^{-1}\right)_{ii} \;-\; \log\det\left(A^{-\top}PA^{-1}\right)\right]. \tag{B.5}$$

This quantity is always nonnegative, and it equals zero exactly when $A^{-\top}PA^{-1}$ is diagonal, i.e. when the code's axes line up with the axes in which the posterior is uncorrelated.

Proof. The ELBO differs from $\log p(x_\gamma)$ by $D_{\mathrm{KL}}\big(q_\phi(z \mid x_\gamma) \,\|\, p(z \mid x_\gamma)\big)$, so maximizing the ELBO over encoders means making this gap as small as possible. In the z coordinates the true posterior is Gaussian with precision $\tilde{P} = A^{-\top} P A^{-1}$. Writing the encoder covariance as a diagonal matrix D and matching means, the gap is

$$\tfrac{1}{2}\Big[\operatorname{tr}(\tilde{P}D) - K - \log\det(\tilde{P}D)\Big].$$

Differentiating in D_{ii} gives $\tilde{P}_{ii} - 1/D_{ii} = 0$, so the best diagonal choice is $D_{ii} = 1/\tilde{P}_{ii}$, and substituting back leaves equation B.5. Nonnegativity and the equality condition are Hadamard's inequality, $\prod_i \tilde{P}_{ii} \geqslant \det \tilde{P}$, with equality exactly when $\tilde{P}$ is diagonal. □

Two things are worth noticing. First, the cost in equation B.5 falls on the reconstruction side of the objective, not on the KL side: a mixed code makes the encoder a worse description of the true posterior, so it reconstructs worse. This is why β does not enter the lemma. Second, the cost depends on A only through how far A is from diagonalizing the posterior.

Proposition B.2 (Alignment at a fixed noise level)**.** *Under Assumption 1, at any noise level γ the encoders maximizing equation B.4 are exactly those whose mixing matrix A is a signed permutation restricted to the currently visible features. Each visible feature is therefore carried by one coordinate, up to reordering the coordinates, flipping their signs, and rescaling them.*

Proof. By (S3) the posterior precision in the feature frame is $P = \gamma \operatorname{diag}(a_k)$. By Lemma B.1 the objective is maximized by any A making $A^{-\top} P A^{-1}$ diagonal. Restricting to the rotations within (S4) (rescalings are absorbed by the coordinatewise freedom), $RPR^\top$ is diagonal precisely when R maps eigenvectors of P to coordinate axes. Since the a_k are distinct by (S2), the eigenvalues of P are distinct, its eigenvectors are the feature axes and are unique up to sign, and so R must be a signed permutation. □

Remark 1 (Distinct energies are doing the work). If two features had equal energy, the posterior precision would have a repeated eigenvalue, any rotation within that two-dimensional subspace would leave equation B.5 at zero, and those two features would not be separated. This is the same condition under which PCA has a well-defined set of principal axes. This is why we need Assumption (S2): it is what makes the feature frame identifiable at all. It is also, as the next subsection shows, exactly what orders the features across noise levels.

Remark 2 (Where β acts). Since β does not affect which frame is preferred, its role is to set how much information a coordinate must carry before it is worth keeping. A larger β charges more per unit of encoded information, so a feature must be correspondingly louder before the reconstruction gain outweighs the cost. In the language of the next subsection, β sets the activation level η and hence shifts all the thresholds γ_k together, without changing their order.

B.2.2.3 From one noise level to a hierarchy

Proposition B.2 applies at each noise level separately, and it determines the code only up to reordering the coordinates. That reordering is free to differ from one level to the next, meaning that an encoder may assign a particular feature to one coordinate at high noise and to another coordinate at low noise. This satisfies Proposition B.2 everywhere but does not result in a representation with a consistent hierarchy. To get *hierarchical* disentanglement, we use the assumption that each feature becomes visible at its own noise level.

Lemma B.3 (Features switch on one at a time and stay on). *For $\gamma \leqslant \gamma'$, the noisy image x_γ can be obtained from $x_{\gamma'}$ by adding further noise, so no feature can be more visible at γ than at γ': the information $I(y_k; x_\gamma)$ is nondecreasing in γ. Fixing an activation level $\eta > 0$ and calling feature k active when $I(y_k; x_\gamma) \geqslant \eta$, the set of active features only grows as γ increases, and under (S2) and (S3) the features become active one at a time in the order $y_1, y_2, \ldots, y_K$, feature k switching on at $\gamma \asymp \gamma_k$.*

Proof. Monotonicity is the data-processing inequality applied to the extra noise. A nondecreasing function crosses the level η at most once, so once a feature is active it stays active and the active sets are nested. By (S3) the information $I(y_k; x_\gamma)$ depends on γ only through the product γa_k, so the crossing happens at $\gamma \asymp 1/a_k = \gamma_k$; the ratio gap in (S2) keeps the crossings separated, giving the stated order. □

Lemma B.4 (The assignment cannot drift)**.** *Let $\phi^\star(\gamma)$ be an optimal encoder at level γ, and for a feature k let $\pi_k(\gamma)$ be the index of the coordinate carrying it. Then π_k is constant on the whole range of levels where feature k is active.*

Proof. By Proposition B.2 the mixing matrix at each level is a signed permutation on the active features, so $\pi_k(\gamma)$ is well defined there. There are finitely many signed permutations, and distinct ones are separated by a fixed distance, so as γ varies the mixing matrix cannot pass from one to another without leaving the set of optima in between. By (S5) the encoder varies continuously with γ, so it cannot jump; hence π_k is locally constant, and being locally constant on an interval it is constant. □

Remark 3 (Why continuity is needed, and why sharing parameters is not enough). A single network taking γ as an input is perfectly capable of permuting its own output coordinates as a function of γ, and doing so costs nothing: the objective at each level is unchanged, so such an encoder is just as optimal. Parameter sharing alone therefore does not give consistency. What rules the drifting solution out is that the optimal assignments form a finite, separated set while the encoder moves continuously, so it has nowhere to drift to. This is the one place (S5) is used, and it cannot be dropped.

Lemma B.5 (Switch-on levels identify the coordinates)**.** *Each coordinate carries no information below one particular noise level and information above it, and these switch-on levels are the distinct values $\gamma_1, \dots, \gamma_K$. Reading off which level a coordinate switches on at therefore identifies which feature it carries, and the answer does not depend on the level at which we inspect the code.*

Proof. By Lemma B.4 the coordinate carrying feature k is the same throughout that feature's active range, so its information curve is that of y_k, which by Lemma B.3 crosses η exactly once, at γ_k. Rescaling a coordinate does not change how much information it carries, so the crossing point is unaffected by the remaining freedom in Proposition B.2. The thresholds are distinct by (S2), so the map from coordinates to switch-on levels is one-to-one. □

Theorem B.6 (Hierarchical disentanglement across scales)**.** *Under Assumption 1, an encoder maximizing the SAMI objective across the family of noise levels satisfies:*

(i) Alignment. *At every noise level, each visible feature is carried by a single latent coordinate, up to reordering, sign, and rescaling.*

(ii) Stability. *A given coordinate carries the same feature at every level where that feature is visible.*

(iii) Hierarchy. *As the image gets cleaner the coordinates switch on one at a time, loudest feature first, and the coordinates active at one level are a subset of those active at every cleaner level.*

Proof. (i) is Proposition B.2 applied at each level, which is available because the cost in Lemma B.1 is incurred level by level and is minimized simultaneously at every level by the aligned code. (ii) is Lemma B.4. (iii) combines Lemma B.3, which nests and orders the features, with (ii) and Lemma C.4, which attach each feature to a fixed coordinate throughout its active range. □

Remark 4 (What is new). Part (i) is the known effect of a diagonal encoder, in the form given by Rolínek et al. (2019) and Kumar and Poole (2020), restated here for the noisy observation x_γ. Parts (ii) and (iii) are the novel contribution: we show that the assignment of feature to latent coordinate does not drift as the encoder varies continuously with γ, and given this fixed assignment, observing when the coordinate becomes informative as we increase the SNR tells us which feature it is representing. Both aspects make the representation disentangled.

C | Appendix: A rate-distortion account of InfoMax

C.1 Derivation of the Barber-Agakov bound

Given a fixed data distribution $p(x)$ and an encoder $p(z|x)$, the mutual information (MI) is defined as

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &= H(X) + \mathbb{E}_{p(x,z)} \left[\log p(x|z)\right] . \end{aligned}$$

Let $q(x|z)$ be a variational approximation to $p(x|z)$. The MI can be written in terms of $q(x|z)$:

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

Defining the variational approximation to MI as $\tilde{I}(X;Z) := H(X) + \mathbb{E}_{p(x,z)} \left[\log q(x|z)\right]$, we get

$$I(X;Z) = \tilde{I}(X;Z) + D_{\mathrm{KL}} \left(p(x|z) \| q(x|z)\right) .$$

Since $D_{\mathrm{KL}}\left(p(x|z)\|q(x|z)\right) \geq 0$, the variational approximation of the MI forms a lower bound on the true MI,

$$\tilde{I}(X;Z) \leq I(X;Z),$$

where equality occurs when $q(x|z) = p(x|z)$, such that $D_{\mathrm{KL}}\left(p(x|z)\|q(x|z)\right) = 0$.

C.2 Marginal distortion

Here we show that the mean squared error $\mathrm{mmse}(\gamma|Z) := \mathbb{E}_{x,x_\gamma,z}\left[\left\|x - \hat{x}(x_\gamma, \gamma, z)\right\|^2\right]$ is equal to the rate of change of the Z-conditional entropy of X_γ with respect to the SNR γ, given by $\partial_\gamma h(X_\gamma|Z)$. We also show that this is equal to $-\partial_\gamma h(X|X_\gamma, Z)$.

Given a noising process $x_\gamma = \sqrt{\gamma}x + \epsilon$ where γ is the SNR and $\epsilon \sim \mathcal{N}(0, I)$, the I-MMSE relation tells us that

$$\partial_\gamma I(X;X_\gamma|Z) = \frac{1}{2}\mathrm{mmse}(\gamma|Z) \tag{1}$$

where $\mathrm{mmse}(\gamma|Z) = \mathbb{E}\|x - \hat{x}(x_\gamma, \gamma, z)\|^2$. From the definition of mutual information:

$$\partial_\gamma I(X;X_\gamma|Z) = \partial_\gamma\left(h(X_\gamma|Z) - h(X_\gamma|X,Z)\right) = \partial_\gamma\left(h(X_\gamma|Z) - h(X_\gamma|X)\right)$$

where the second equality is from the Markov chain relation $z - x - x_\gamma$. The second term disappears because it is a constant; it has no dependency on γ for all x:

$$h(X_\gamma|X) = \frac{d}{2}\log(2\pi e) \quad \forall x.$$

Thus we know that $\partial_\gamma I(X;X_\gamma|Z) = \partial_\gamma h(X_\gamma|Z)$. Combining this with Eq. 1, we get

$$\frac{1}{2}\mathrm{mmse}(\gamma|Z) = \partial_\gamma h(X_\gamma|Z).$$

By decomposing the mutual information as entropies of X rather than X_γ,

$$\partial_\gamma I(X; X_\gamma|Z) = \partial_\gamma \left(h(X|Z) - h(X|X_\gamma, Z)\right) = -\partial_\gamma h(X|X_\gamma, Z),$$

the mmse is also equal to

$$\frac{1}{2}\text{mmse}(\gamma|Z) = -\partial_\gamma h(X|X_\gamma, Z).$$

C.3 Relation of marginal mutual information to Fisher information

Let X be any random variable with a finite variance with a density $f(x)$. Let ϵ be an independent normally distributed random variable with zero mean and unit variance, and suppose $X_\gamma = X + \sqrt{\gamma}\epsilon$. We want to show that

$$R_\gamma = \partial_\gamma I(Z; X_\gamma) = \frac{1}{2}\,\mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|\nabla_{x_\gamma} \log p(z \mid x_\gamma)\right\|^2\right] \geqslant 0.$$

de Bruijn's identity connects the entropy to Fisher information:

$$\frac{\partial}{\partial\gamma} h_e(X_\gamma) = \frac{1}{2} J(X_\gamma).$$

where h_e is the differential entropy to base e and J is the Fisher information. The proof can be found in Theorem 17.7.2 of Cover and Thomas (2006).

Meanwhile, given conditioning variable Z, conditional de Bruijn's identity (Lemma C.1) is

$$\frac{\partial}{\partial\gamma} h_e(X_\gamma|Z) = \frac{1}{2} J(X_\gamma|Z).$$

The marginal mutual information is

$$\partial_\gamma I(Z;X_\gamma) = \partial_\gamma h(X_\gamma) - \partial_\gamma h(X_\gamma|Z) = \frac{1}{2}\big(\, J(X_\gamma) - J(X_\gamma|Z)\,\big) = \frac{1}{2}J(Z \mid X_\gamma).$$

Using the definition of the Fisher information,

$$R_\gamma = J(Z \mid X_\gamma) = \mathbb{E}_{p(x,x_\gamma,z)}\Big[\, \big\|\nabla_{x_\gamma} \log p(z|x_\gamma)\big\|^2 \,\Big].$$

Lemma C.1 (Conditional de Bruijn's identity)**.** *Let X be any random variable with a finite variance with a density $f(x)$. Let ϵ be an independent normally distributed random variable with zero mean and unit variance. Let W be a random variable that we wish to condition on. Suppose $Y_\gamma = X + \sqrt{\gamma}\epsilon$. The conditional de Bruijn's identity is given by*

$$\frac{\partial}{\partial\gamma} h_e(Y_\gamma|W) = \frac{1}{2}J(Y_\gamma|W). \tag{C.1}$$

Proof. The density of Y_γ is

$$g_\gamma(y) = \int_{-\infty}^{\infty} f(x)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx.$$

Since ϵ is independent of everything, conditioning on $W = w$ only affects the distribution of X, but not the additive Gaussian noise. Hence the conditional density of Y_γ given W is

$$g_\gamma(y|w) = \int_{-\infty}^{\infty} f(x|w)\frac{1}{\sqrt{2\pi\gamma}}\exp\left(-\frac{(y-x)^2}{2\gamma}\right)dx.$$

Likewise, the joint distribution is defined as

$$\begin{aligned} g_\gamma(y,w) &= f_W(w) g_\gamma(y|w) \\ &= f_W(w) \int_{-\infty}^{\infty} f(x|w) \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(y-x)^2}{2\gamma}\right) dx \\ &= \int_{-\infty}^{\infty} f(x,w) \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(y-x)^2}{2\gamma}\right) dx. \end{aligned}$$

Since the identities in de Bruijn's identity (derivatives with respect to γ and y) do not affect $f(x)$, we can simply replace $f(x)$ with $f(x|w)$ to get similar identities for $g_\gamma(y|w)$:

$$\begin{aligned} \frac{\partial}{\partial\gamma} g_\gamma(y|w) &= \int_{-\infty}^{\infty} f(x|w) \left[-\frac{1}{2\gamma}\frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2\gamma}\right) + \frac{(y-x)^2}{2\gamma^2}\frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(y-x)^2}{2\gamma}\right)\right] dx \\ \frac{\partial}{\partial y} g_\gamma(y|w) &= \int_{-\infty}^{\infty} f(x|w) \frac{1}{\sqrt{2\pi t}} \left(-\frac{y-x}{\gamma}\right) \exp\left(-\frac{(y-x)^2}{2\gamma}\right) dx \\ \frac{\partial^2}{\partial y^2} g_\gamma(y|w) &= \int_{-\infty}^{\infty} f(x|w) \frac{1}{\sqrt{2\pi\gamma}} \left[-\frac{1}{\gamma} \exp\left(-\frac{(y-x)^2}{2\gamma}\right) + \frac{(y-x)^2}{\gamma^2} \exp\left(-\frac{(y-x)^2}{2\gamma}\right)\right] dx \end{aligned}$$

Thus we can see that

$$\frac{\partial}{\partial\gamma} g_\gamma(y|w) = \frac{1}{2}\frac{\partial^2}{\partial y^2} g_\gamma(y|w).$$

Now let's use this to get de Bruijn's identity for $\frac{\partial}{\partial\gamma} h_e(Y_\gamma|W)$. We can define $h_e(Y_\gamma|W)$ as

$$\begin{aligned} h_e(Y_\gamma|W) &= -\int_{-\infty}^{\infty} g_\gamma(y,w) \ln g_\gamma(y|w) dy \\ &= -\int_{-\infty}^{\infty} f(w) \int_{-\infty}^{\infty} g_\gamma(y|w) \ln g_\gamma(y|w)\, dy\, dw \end{aligned}$$

Differentiating, we get

$$\begin{aligned}
\frac{\partial}{\partial\gamma}h_e(Y_\gamma|W) &= -\frac{\partial}{\partial\gamma}\int_{-\infty}^{\infty} f(w)h(Y_\gamma|W=w)\,dy\,dw \\
&= -\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial\gamma}\ln g_\gamma(y|w) + g_\gamma(y|w)\frac{1}{g_\gamma(y|w)}\frac{\partial g_\gamma(y|w)}{\partial y}dy\,dw \\
&= -\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial t}\ln g_\gamma(y|w) + \frac{\partial g_\gamma(y|w)}{\partial t}dy\,dw \\
&= -\int_{-\infty}^{\infty} f(w)\frac{\partial}{\partial t}\int_{-\infty}^{\infty} g_\gamma(y|w)\,dy\,dw - \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial g_\gamma(y|w)}{\partial t}\ln g_\gamma(y|w)\,dy\,dw
\end{aligned}$$

The first term goes to zero since the inner integrand $\int g_\gamma(y|w)dy = 1$. We can use the identity above to get

$$\frac{\partial}{\partial t}h_e(Y_\gamma|W) = -\frac{1}{2}\int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty}\frac{\partial^2}{\partial y^2}g_\gamma(y|w)\ln g_\gamma(y|w)\,dy\,dw$$

Let's do integration by parts, just like before.

$$\frac{\partial}{\partial t}h_e(Y_\gamma|W) = \int_{-\infty}^{\infty} f(w)\left(\frac{1}{2}\left[\frac{\partial}{\partial y}g_\gamma(y|w)\ln g_\gamma(y|w)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty}\left[\frac{\partial}{\partial y}g_\gamma(y|w)\right]^2\frac{1}{g_\gamma(y|w)}dy\right)dw.$$

Since the definition of $J(X|W)$ is

$$\begin{aligned}
J(X|W) &= \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty} f(x|w)\left[\frac{\partial}{\partial x}\ln f(x|w)\right]^2 dx\,dw \\
&= \int_{-\infty}^{\infty} f(w)\int_{-\infty}^{\infty} f(x|w)\left[\frac{1}{f(x|w)}\frac{\partial}{\partial x}f(x|w)\right]^2 dx\,dw,
\end{aligned}$$

the second term is the conditional Fisher information $\frac{1}{2}J(Y_\gamma|W)$. All that remains is to show that the first term is zero. We can rewrite the first term as

$$\frac{\partial}{\partial y}g_\gamma(y|w)\ln g_\gamma(y|w) = \left[\frac{1}{\sqrt{g_\gamma(y|w)}}\frac{\partial g_\gamma(y|w)}{\partial y}\right]\left[2\sqrt{g_\gamma(y|w)}\ln\sqrt{g_\gamma(y|w)}\right]$$

This goes to zero, just like for $\frac{\partial}{\partial y} g_\gamma(y) \ln g_\gamma(y)$. Thus

$$\frac{\partial}{\partial \gamma} h_e(Y_\gamma|W) = \frac{1}{2} J(Y_\gamma|W).$$

□

C.4 Rate as a dot product

We want to show that the norm-based form of the marginal rate (Eq. 4.9) is equivalent to the dot-product form of the marginal rate (Eq. 4.10):

$$R_\gamma := \partial_\gamma I(Z;X_\gamma) = \frac{1}{2}\mathbb{E}_{p(x,x_\gamma,z)}\left[\left\|\nabla_{x_\gamma} \log p(z \mid x_\gamma)\right\|^2\right] = \mathbb{E}_{p(x,x_\gamma,z)}\left[\nabla_{x_\gamma} \log p(z \mid x_\gamma) \cdot \partial_\gamma x_\gamma\right].$$

Since $I(Z;X_\gamma) = H(Z) - H(Z|X_\gamma)$ and $H(Z)$ is independent of γ, we have

$$\begin{aligned}
\partial_\gamma I(Z;X_\gamma) &= \partial_\gamma \mathbb{E}\left[\log p(z|x_\gamma)\right] \\
&= \mathbb{E}\left[\frac{d}{d\gamma} \log p(z \mid x_\gamma)\right] \\
&= \mathbb{E}\left[\partial_\gamma \log p(z|x_\gamma) + \nabla_{x_\gamma} \log p(z|x_\gamma) \cdot \partial_\gamma x_\gamma\right]
\end{aligned}$$

The first term can be written as

$$\begin{aligned}
\mathbb{E}_{p(z,x_\gamma)}\left[\partial_\gamma \log p(z|x_\gamma)\right] &= \mathbb{E}_{p(x_\gamma)}\left[\int p(z|x_\gamma)\partial_\gamma \log p(z|x_\gamma)\, dz\right] \\
&= \mathbb{E}_{p(x_\gamma)}\left[\int \partial_\gamma p(z|x_\gamma)\, dz\right] \\
&= \mathbb{E}_{p(x_\gamma)}\left[\partial_\gamma \int p(z|x_\gamma)\, dz\right].
\end{aligned}$$

where in the second line we have used the identity $p_\gamma \partial_\gamma \log p_\gamma = \partial_\gamma p_\gamma$. Since $\int p(z|x_\gamma)\, dz = 1$, the

entire expectation goes to 0. Thus we are left with

$$\partial_\gamma I(Z; X_\gamma) = \mathbb{E}_{p(x,z,x_\gamma)} \left[\nabla_{x_\gamma} \log p(z|x_\gamma) \cdot \partial_\gamma x_\gamma \right].$$

C.5 Hierarchical disentanglement from noise-level structure

Setup. Suppose an image $x \in \mathbb{R}^D$ is generated from K independent latent *features* $\mathbf{y} = (y_1, \ldots, y_K)$ through a smooth injective map $x = f(\mathbf{y})$ that may otherwise be arbitrarily nonlinear. Each feature is standardized to unit variance. We never observe $\mathbf{y}$ directly, but only through the noised observation

$$x_\gamma \;=\; \sqrt{\gamma}\, f(\mathbf{y}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_D),$$

where γ is the signal-to-noise ratio (SNR): $\gamma \to \infty$ is a clean image and $\gamma \to 0$ is one buried in noise.

We build the theorem from the following two quantities. Denoting $J(\mathbf{y}) = \partial f / \partial \mathbf{y} \in \mathbb{R}^{D \times K}$ for the Jacobian and $J_k(\mathbf{y})$ for its k-th column, define the *energy* of feature y_k and its *SNR threshold* as

$$a_k(\mathbf{y}) \;:=\; \|J_k(\mathbf{y})\|^2, \qquad \gamma_k \;\asymp\; 1/a_k.$$

The energy a_k measures how much the image moves when y_k is perturbed, and γ_k is the noise level at which that motion first rises above the unit noise floor. Below γ_k the observation is *essentially uninformative* about y_k, carrying only $o(\gamma\, a_k)$ nats (Lemma C.4); above it, information about y_k accumulates. We assume the SNR thresholds are *distinct and ordered*, $\gamma_1 > \cdots > \gamma_K$. Equivalently, since $\gamma_k \asymp 1/a_k$, the energies run the *opposite* way, $a_1 < \cdots < a_K$: a feature with a large threshold has small energy. As γ increases from 0 we cross γ_K first, so features become detectable one at a time in the order $y_K, \ldots, y_1$.

Our goal is to show that a conditional Gaussian encoder with diagonal covariance,

$$p_\phi(z \mid x_\gamma) = \mathcal{N}\big(\mu_\phi(x_\gamma),\, \mathrm{diag}(\sigma_\phi(x_\gamma))\big),$$

trained on the InfoMax objective, learns latents $\mathbf{z} = (z_1, \ldots, z_M)$ that are disentangled according to the SNR threshold: the encoder assigns each feature to a unique latent and maintains that bijective map across all noise levels.

Assumption 2.

(A1) There exist features $\mathbf{y} = (y_1, \ldots, y_K)$ that are mutually independent, with smooth density, standardized so that $\mathbb{E}[y_k] = 0$, $\mathrm{Var}(y_k) = 1$.

(A2) $f : \mathbb{R}^K \to \mathbb{R}^D$ is smooth and injective, and its Jacobian $J(\mathbf{y}) = \partial f / \partial \mathbf{y} \in \mathbb{R}^{D \times K}$ has full column rank.

(A3) *(Separability)* $J(\mathbf{y})^\top J(\mathbf{y})$ is diagonal for every $\mathbf{y}$; that is, $J_k(\mathbf{y})^\top J_j(\mathbf{y}) = 0$ for $k \neq j$, with diagonal entries $a_k(\mathbf{y}) = \|J_k(\mathbf{y})\|^2$.

(A4) *(Distinct ordered energies)* The energies occupy non-overlapping bands, $a_k(\mathbf{y}) < a_{k+1}(\mathbf{y})$, so the SNR thresholds $\gamma_k \asymp 1/a_k$ are distinct and *oppositely* ordered, $\gamma_1 > \gamma_2 > \cdots > \gamma_K$.

(A5) *(Bounded, weakly nonlinear energies)* Over the support of the prior $p(\mathbf{y})$, the energies $a_k(\mathbf{y})$ are (i) *uniformly bounded*, with fixed constants $0 < a_{\min} \leqslant a_k(\mathbf{y}) \leqslant a_{\max} < \infty$ for all $\mathbf{y}$; and (ii) *slowly varying*, the curvatures $\partial^2 f / \partial y_k^2$ being small relative to the unit spread of y_k, so each column $J_k(\mathbf{y})$ changes little as $\mathbf{y}$ ranges over the prior.

(A6) *(Regularity of generative posterior)* The Bayesian posterior $p(\mathbf{y} \mid x_\gamma)$ induced by the generative model concentrates at high SNR in a local Gaussian about the maximum-likelihood estimate. We assume f is proper.

Theorem C.2 (Disentangled hierarchical representation)**.** *Under Assumption 2, the maximally informative code for x_γ is the assignment $z_i = \hat{y}_{k(i)}$ of one latent per feature, and this assignment is preserved across noise levels, the latents activating in the order $y_K, \dots, y_1$ as γ increases. The assignment holds up to a permutation and a per-coordinate reparametrization.*

Proof sketch. The claim rests on two facts: because each feature has a distinct noise threshold, the optimal code assigns each feature to a single latent; and because information only grows with γ, that assignment is stable as the image gets cleaner. The five steps below outline this argument and point to the relevant lemma.

Step 1 (visibility is energy). To detect y_k we must notice its effect on the image. Perturbing y_k by a small δ shifts the noised mean by $\sqrt{\gamma}\,\delta\,J_k(\mathbf{y})$; against unit-variance noise this is detectable only when the *effective per-feature SNR*

$$s_k(\gamma) \;=\; \gamma\, a_k(\mathbf{y}) \tag{C.2}$$

exceeds one. This is made precise by Lemma C.3.

Step 2 (features appear one at a time). Feature k is invisible while $s_k \ll 1$ and visible once $s_k \gtrsim 1$, i.e. once $\gamma \gtrsim \gamma_k = 1/a_k$. "Loud" features (large a_k, small γ_k) survive heavy noise, while "faint" features (small a_k, large γ_k) are only visible in a clean image. Distinct energies correspond to distinct thresholds, so as we lower the noise the features appear in a fixed order. This monotonicity is given by Lemma C.4(ii)-(iii).

Step 3 (the order labels the features). Because the features switch on at *distinct* noise levels, the crossing level itself identifies which feature has appeared: γ acts as a label sorting the features. This step is laid out in Lemma C.6.

Step 4 (the best code assigns latents to features). Observe a single latent z_i as we sweep γ. If z_i carries one feature it has a *single onset*, defined as a latent that is silent below γ_k and informative above. If the latent mixes two features with different thresholds it has *two* onsets. Among informative

codes we show the noise-level structure forces a bijection $z_i \leftrightarrow y_k$ up to relabeling. This step is given by Lemma C.6.

Step 5 (the arrangement is stable across noise levels). This assignment persist across noise levels for two reasons: features appear in a fixed order set by their energy, and since information is monotone in γ, once a feature is visible it stays visible. The *active set* of features therefore grows in a *nested* fashion, with a loud feature appearing low in the hierarchy and a faint feature appearing high. This result comes from combining Lemma C.4(i) with (A4).

The three lemmas below examine a single object, the per-feature threshold γ_k, from three angles. Lemma C.3 has to do with *precision*: we use the Cramér-Rao bound to determine γ_k. Lemma C.4 approaches this from an *information* perspective: the I-MMSE relation recovers the same γ_k and shows that information is monotone. Lemma C.6 is the *labeling* step: distinct thresholds turn onsets into a coordinate frame.

Lemma C.3 (Fisher information and the SNR threshold)**.** *The Fisher information that x_γ carries about* $\mathbf{y}$ *is*

$$\mathcal{I}_\gamma(\mathbf{y}) = \gamma\, J(\mathbf{y})^\top J(\mathbf{y}), \tag{C.3}$$

which under (A3) *is diagonal with entries $\gamma\, a_k(\mathbf{y})$. Any unbiased estimator of y_k from x_γ therefore has variance at least $(\gamma\, a_k)^{-1}$. Since y_k carries unit prior variance, x_γ is essentially uninformative about y_k while $\gamma\, a_k \lesssim 1$ and becomes informative once $\gamma\, a_k \gtrsim 1$, fixing the SNR threshold $\gamma_k \asymp 1/a_k$.*

Proof. Since $x_\gamma \mid \mathbf{y} \sim \mathcal{N}(\sqrt{\gamma}\, f(\mathbf{y}), I_D)$, the score is $\nabla_{\mathbf{y}} \log p(x_\gamma \mid \mathbf{y}) = \sqrt{\gamma}\, J(\mathbf{y})^\top \big(x_\gamma - \sqrt{\gamma}\, f(\mathbf{y})\big) = \sqrt{\gamma}\, J(\mathbf{y})^\top \epsilon$. The covariance of the score is the Fisher information, $\gamma\, J^\top J$, giving us equation C.3. (A3) makes it diagonal.

By Cramér-Rao, any unbiased estimate of y_k from x_γ has standard deviation at least $(\gamma a_k)^{-1/2}$, the inverse square root of its Fisher information. The observation becomes informative about y_k only once this variance falls below the unit feature variance (Recall (A1)), which occurs once $\gamma a_k \gtrsim 1$. The crossover at $\gamma a_k \asymp 1$ gives us the SNR threshold $\gamma_k \asymp 1/a_k$. □

Lemma C.4 (Monotone per-feature information). *Denote* $\mathbf{y}_{-k} = (y_j)_{j\neq k}$ *for the features in* $\mathbf{y}$ *other than* y_k.

(i) For $\gamma < \gamma'$ *the chain* $\mathbf{y} \to x_{\gamma'} \to x_\gamma$ *is Markov; hence* $\gamma \mapsto I(y_k; x_\gamma)$ *is non-decreasing.*

(ii) Let $\mathrm{mmse}(\gamma) = \mathbb{E}\|f(\mathbf{y}) - \mathbb{E}[f(\mathbf{y}) \mid x_\gamma]\|^2$, *and let* $\mathrm{mmse}(\gamma \mid y_k) = \mathbb{E}\|f(\mathbf{y}) - \mathbb{E}[f(\mathbf{y}) \mid x_\gamma, y_k]\|^2$ *be the residual error when* y_k *is known. Then, with no assumption on* f *beyond injectivity,*

$$\frac{d}{d\gamma} I(y_k; x_\gamma) \;=\; \tfrac{1}{2}\,\mathrm{mmse}_k(\gamma), \qquad \mathrm{mmse}_k(\gamma) := \mathrm{mmse}(\gamma) - \mathrm{mmse}(\gamma \mid y_k) \;\geqslant\; 0.$$

The per-feature rate of change in information is thus exactly half the error that knowing y_k *removes, and its non-negativity recovers the monotonicity of (i).*

(iii) The onset of y_k *is at* γ_k. *As* $\gamma \to 0$,

$$I(y_k; x_\gamma) = \tfrac{1}{2}\, c_k\, \gamma + o(\gamma), \qquad c_k := \mathrm{mmse}_k(0) = \mathrm{tr}\,\mathrm{Cov}\big(\mathbb{E}[f(\mathbf{y}) \mid y_k]\big),$$

with $c_k \asymp a_k$ *under* (A5). *Thus* $I(y_k; x_\gamma) = O(\gamma\, a_k)$ *until* $\gamma \gtrsim \gamma_k \asymp 1/a_k$, *recovering the threshold of Lemma C.3.*

Proof. (i) Degrading a cleaner observation yields a noisier one: $x_\gamma \overset{d}{=} \sqrt{\gamma/\gamma'}\, x_{\gamma'} + \tilde{\epsilon}$ with $\tilde{\epsilon} \sim \mathcal{N}\big(0, (1 - \gamma/\gamma')I_D\big)$ independent of $x_{\gamma'}$, so $\mathbf{y} \to x_{\gamma'} \to x_\gamma$ is Markov and the data-processing inequality (Cover and Thomas, 2006) gives $I(y_k; x_\gamma) \leqslant I(y_k; x_{\gamma'})$.

(ii) Since f is injective, $\mathbf{y} \leftrightarrow f(\mathbf{y})$ is a bijection, and given y_k so is $\mathbf{y}_{-k} \leftrightarrow f(\mathbf{y})$:

$$I(\mathbf{y}; x_\gamma) = I(f(\mathbf{y}); x_\gamma), \quad I(\mathbf{y}_{-k}; x_\gamma \mid y_k) = I(f(\mathbf{y}); x_\gamma \mid y_k).$$

Both are governed by the (conditional) I-MMSE relation (Guo et al., 2005),

$$\frac{d}{d\gamma} I(\mathbf{y}; x_\gamma) = \tfrac{1}{2}\,\mathrm{mmse}(\gamma), \qquad \frac{d}{d\gamma} I(\mathbf{y}_{-k}; x_\gamma \mid y_k) = \tfrac{1}{2}\,\mathrm{mmse}(\gamma \mid y_k).$$

They are related by the chain rule for mutual information,

$$I(\mathbf{y}; x_\gamma) = I(y_k; x_\gamma) + I(\mathbf{y}_{-k}; x_\gamma \mid y_k),$$

so differentiating and subtracting gives

$$\frac{d}{d\gamma} I(y_k; x_\gamma) \;=\; \frac{1}{2}\big(\mathrm{mmse}(\gamma) - \mathrm{mmse}(\gamma \mid y_k)\big) \;:=\; \frac{1}{2}\mathrm{mmse}_k(\gamma).$$

where we have defined $\mathrm{mmse}_k(\gamma)$ as the mmse error removed by knowing y_k. This can essentially be interpreted as the amount of residual uncertainty in f that y_k explains at SNR γ. We know that $\mathrm{mmse}(\gamma|y_k) \leq \mathrm{mmse}(\gamma)$ because $\mathbb{E}[f|x_\gamma]$ is a suboptimal estimator when we also know y_k. Hence $\mathrm{mmse}_k(\gamma) \geq 0$. This independently reproves (i) that $I(y_k; x_\gamma)$ is non-decreasing.
(iii) At $\gamma = 0$ (high noise) the observation $x_0 = \epsilon$ is independent of $\mathbf{y}$, so the estimators reduce to prior means and the errors to prior variances:

$$\mathrm{mmse}(0) = \mathrm{tr}\,\mathrm{Cov}\big(f(\mathbf{y})\big), \qquad \mathrm{mmse}(0 \mid y_k) = \mathbb{E}\Big[\,\mathrm{tr}\,\mathrm{Cov}\big(f(\mathbf{y}) \mid y_k\big)\Big].$$

The law of total variance, $\mathrm{Cov}(f) = \mathbb{E}[\mathrm{Cov}(f \mid y_k)] + \mathrm{Cov}(\mathbb{E}[f \mid y_k])$, gives

$$c_k := \mathrm{mmse}_k(0) = \mathrm{mmse}(0) - \mathrm{mmse}(0 \mid y_k) = \mathrm{tr}\,\mathrm{Cov}\big(\mathbb{E}[f(\mathbf{y}) \mid y_k]\big),$$

the variance of the k-th component of f explained by y_k. To first order, a unit change in y_k moves the image along the Jacobian J_k, so $\mathbb{E}[f \mid y_k] - \mathbb{E}[f] \approx J_k\, y_k$. Thus

$$\mathrm{Cov}\left(\mathbb{E}[f \mid y_k]\right) \approx J_k J_k^\top \mathrm{Var}(y_k) = J_k J_k^\top$$

where in the equality we have used the unit variance assumption (A1). This means that

$$c_k \approx \operatorname{tr}\left(J_k J_k^\top\right) = \|J_k\|^2 = a_k.$$

where we have used assumption (A5)(ii) that J_k varies slowly. Thus $c_k \asymp a_k$.

Integrating the relation from (ii) from 0, and using $I(y_k; x_0) = 0$, we get

$$I(y_k; x_\gamma) = \tfrac{1}{2}\int_0^\gamma \operatorname{mmse}_k(s)\, ds = \tfrac{1}{2}\, c_k\, \gamma + \tfrac{1}{2}\int_0^\gamma \big(\operatorname{mmse}_k(s) - c_k\big)\, ds = \tfrac{1}{2}\, c_k\, \gamma + o(\gamma).$$

The first term is linear and the remainder is $o(\gamma)$ since $\operatorname{mmse}_k(s) \to c_k$ as $s \to 0$ (A5).

The information grows linearly at rate $\frac{1}{2}c_k \asymp \frac{1}{2}a_k$. This means it is negligible while $\gamma a_k \ll 1$ and order one once $\gamma a_k \asymp 1$, i.e. once $\gamma \asymp 1/a_k = \gamma_k$. (At $\gamma = \gamma_k$ the leading term is $\frac{1}{2}a_k \cdot (1/a_k) = \frac{1}{2}$ nat.) This low-SNR slope, extrapolated to where I reaches $O(1)$, places the onset at the same $\gamma_k \asymp 1/a_k$ as the high-SNR Cramér-Rao bound of Lemma C.3; the two regimes meet at the SNR threshold. □

Proof. (i) We can show the Markov relation by writing x_γ in terms of $x_{\gamma'}$: $x_\gamma \overset{d}{=} \sqrt{\gamma/\gamma'}\, x_{\gamma'} + \tilde{\epsilon}$ with $\tilde{\epsilon} \sim \mathcal{N}\big(0, (1-\gamma/\gamma')I_D\big)$ independent of $x_{\gamma'}$. By the data-processing inequality (Cover and Thomas, 2006) we get $I(y_k; x_\gamma) \leqslant I(y_k; x_{\gamma'})$.

(ii) The scalar identity is given by Guo et al. (2005); under (A3) the orthogonal columns decouple the per-feature estimation errors to leading order, so $\operatorname{mmse}(\gamma) = \sum_k \operatorname{mmse}_k(\gamma)$ and the derivative can be split into per-feature terms.

(iii) By part (ii), the I-MMSE relation gives $\frac{d}{d\gamma} I(y_k; x_\gamma) = \frac{1}{2}\operatorname{mmse}_k(\gamma)$, and we will use this relation to analyze the behavior of the information near $\gamma = 0$. At $\gamma = 0$ the observation $x_0 = \epsilon$ is pure noise, independent of $\mathbf{y}$, so the MMSE estimator of feature k's contribution to the image is its

prior mean, and the residual error is the full prior variance,

$$c_k \;:=\; \mathrm{mmse}_k(0) \;=\; \mathrm{Var}\big(\text{contribution of } y_k \text{ to } f(\mathbf{y})\big).$$

To first order, a unit perturbation of y_k moves the image along $J_k(\mathbf{y})$, so this variance scales as $\|J_k\|^2 = a_k$, giving $c_k \asymp a_k$ under (A5). Integrating the I-MMSE relation from 0 to γ, and using $I(y_k; x_0) = 0$,

$$I(y_k; x_\gamma) \;=\; \tfrac{1}{2}\int_0^\gamma \mathrm{mmse}_k(s)\,ds \;=\; \tfrac{1}{2}\,c_k\,\gamma + \tfrac{1}{2}\int_0^\gamma \big(\mathrm{mmse}_k(s) - c_k\big)\,ds \;=\; \tfrac{1}{2}\,c_k\,\gamma + o(\gamma),$$

where the remainder is $o(\gamma)$ because the integrand vanishes as $s \to 0$, by the regularity of mmse_k near the origin (A5).

This indicates that as we move away from $\gamma = 0$, the growth in information is linear of order $O(\gamma\, a_k)$. This is negligible while $\gamma\, a_k \ll 1$ and of order one once $\gamma\, a_k \asymp 1$, i.e. once $\gamma \asymp 1/a_k = \gamma_k$. (At $\gamma = \gamma_k$ the leading term is $\frac{1}{2}a_k \cdot (1/a_k) = \frac{1}{2}$ nat, so roughly half a nat has accumulated by the SNR threshold.) Strictly, this is an expansion around $\gamma = 0$, so our statement above extrapolates the low-SNR slope out to where the information reaches $O(1)$. This extrapolation agrees with the high-information estimate of Lemma C.3, which places the onset at the same SNR $\gamma_k \asymp 1/a_k$; both the Cramér-Rao bound and the I-MMSE slope perspective approach the threshold from opposite regimes. □

Definition C.5 (Onset)**.** Let $z = \psi(x_\gamma)$ be any scalar statistic of the observation. Its information $\gamma \mapsto I(z; x_\gamma)$ is non-decreasing (data processing, Lemma C.4(i)). Say z has an *onset* at γ_k if $I(z; x_\gamma)$ crosses from $o(1)$ to $\Omega(1)$ as γ passes γ_k, where $o(1)$ and $\Omega(1)$ are taken in the effective-SNR sense ($s_k = \gamma a_k \to 0$ and $\to \infty$ respectively). Onsets are counted only at the threshold set $\{\gamma_1, \ldots, \gamma_K\}$.

Lemma C.6 (Distinct thresholds determine the basis)**.** *Suppose a code is aligned to the feature basis: each coordinate z_i is an invertible reparametrization of a single feature $y_{k(i)}$, for some assignment*

$k(\cdot)$. *Then*

(i) each coordinate becomes informative at the threshold of the feature it carries, and

(ii) the distinctness of these thresholds (A4) *recovers the assignment* $k(\cdot)$ *uniquely.*

Proof. (i) *Each feature has a single onset at its own SNR threshold.* By separability (A3) and Lemma C.4(ii), $I(y_k; x_\gamma)$ depends on the noise level only through the feature's effective SNR $s_k = \gamma a_k$. It is a fixed increasing function of s_k that vanishes at $s_k = 0$, is $o(1)$ for $s_k \ll 1$ and is $\Omega(1)$ for $s_k \gtrsim 1$. Such a curve crosses from $o(1)$ to $\Omega(1)$ exactly once, at $s_k \asymp 1$, i.e. $\gamma \asymp \gamma_k$. Because the other thresholds γ_j $(j \neq k)$ never enter this curve, it has no second crossing. Hence y_k has a single onset at γ_k.

Each aligned coordinate inherits its feature's onset. By hypothesis, latent z_i is an invertible reparametrization of the single feature $y_{k(i)}$ and mutual information is invariant under such reparamaterization, so

$$I(z_i; x_\gamma) = I(y_{k(i)}; x_\gamma).$$

Latent z_i thus has the same information rate as $y_{k(i)}$ and, by Step 1, a single onset at $\gamma_{k(i)}$.
(ii) The thresholds $\gamma_1 > \cdots > \gamma_K$ are distinct (A4), so the map $i \mapsto \gamma_{k(i)}$ is injective. Reading each coordinate's onset level off against $\{\gamma_k\}$ therefore uniquely recovers the assignment $k(\cdot)$. □

Assembling the theorem. The proof has three steps. *Concentration*: at high information the posterior over the active features is Gaussian and, under separability, diagonal in the feature frame — but only up to a rotation within the active block, which sufficiency alone cannot remove. *De-rotation*: the diagonal encoder, under InfoMax, is exact only in the frame that diagonalizes the posterior, and distinct energies make that frame unique — so it lands on the feature frame, and Lemma C.6 labels the axes. *Nesting*: monotone information grows the active set in threshold order.

Proof. Concentration. By Lemma C.3 the posterior precision is $\mathcal{I}_\gamma(\mathbf{y}) = \gamma \operatorname{diag}(a_k)$. Restrict to the active features $A(\gamma) = \{k : \gamma > \gamma_k\}$; inactive features carry $o(1)$ information (Lemma C.4(iii)) and

do not concentrate. Let $\hat{\mathbf{y}}$ denote the maximum-likelihood estimate of $\mathbf{y}$ from x_γ, i.e. the minimizer of $\|x_\gamma - \sqrt{\gamma} f(\mathbf{y})\|^2$, and $\hat{y}_k$ its k-th coordinate.

As γ increases the posterior sharpens, and by Bernstein-von Mises theorem (Vaart, 1998) (otherwise known as *local asymptotic normality* in the frequentist literature) the posterior over the active features $\mathbf{y}_{A(\gamma)}$ converges, to leading order $1/\gamma$, to a normal distribution about $\hat{\mathbf{y}}$ with covariance $\mathcal{I}_\gamma^{-1}$. Diagonality of $\mathcal{I}_\gamma$ (A3) makes this covariance diagonal, so the posterior factorizes over features,

$$p\big(\mathbf{y}_{A(\gamma)} \mid x_\gamma\big) \;\approx\; \prod_{k \in A(\gamma)} \mathcal{N}\big(\hat{y}_k,\, (\gamma a_k)^{-1}\big), \qquad A(\gamma) = \{k : \gamma > \gamma_k\}.$$

Because this depends on x_γ only through the centers $\hat{y}_k$ (the widths being fixed), the vector $(\hat{y}_k)_{k \in A(\gamma)}$ is a minimal sufficient statistic for $\mathbf{y}_{A(\gamma)}$. Moreover, since the posterior factorizes, this statistic is coordinate-wise: each $\hat{y}_k$ is sufficient for its own feature y_k alone. Hence the maximally informative code's coordinates are re-parametrizations of the active features $\{y_k\}_{k \in A(\gamma)}$. This is determined, however, only up to a rotation *within* the active block, since $I(z; x_\gamma)$ is invariant under any invertible reparametrization of z. This residual rotation is removed in the de-rotation step.

De-rotation. Concentration determined the code only up to a rotation within the active block; we show the diagonal encoder, under InfoMax, selects the feature basis out of this rotation family. By the *feature basis* we mean the coordinate axes aligned with $y_1, \ldots, y_K$; under (A3) this is exactly the eigenbasis of the posterior covariance $\mathcal{I}_\gamma^{-1} = \Lambda := \mathrm{diag}\big((\gamma a_k)^{-1}\big)$.

Information alone cannot make this selection. For any rotation R of the active block, $R\hat{\mathbf{y}}$ is an equally valid sufficient statistic and $I(Rz; x_\gamma) = I(z; x_\gamma)$, so InfoMax evaluated on a fixed code is blind to R. The tie is broken instead by the factorial *constraint* on the encoder.

The encoder is restricted to diagonal covariance, $p_\phi(z \mid x_\gamma) = \mathcal{N}\big(\mu_\phi, \mathrm{diag}(\sigma_\phi)\big)$, so its coordinates are conditionally independent given x_γ. A mean-field code matches the posterior exactly

only when the posterior is itself diagonal in the code coordinates; otherwise it approximates a correlated Gaussian by a factorized one and pays a gap equal to the posterior's total correlation. In a basis where the active block has covariance Σ, the information among the coordinates that a factorized code must discard is given by

$$\Delta(\Sigma) \;=\; \min_{q \text{ factorized}} D_{\mathrm{KL}}\big(p \,\|\, q\big) \;=\; \tfrac{1}{2}\left(\sum_k \log \Sigma_{kk} - \log\det \Sigma\right) \;\geqslant\; 0.$$

By Hadamard's inequality, $\Delta(\Sigma) = 0$ iff Σ is diagonal.

Applying this to the rotation family R gives covariance $R\Lambda R^\top$ with fixed $\det(R\Lambda R^\top) = \det\Lambda$, so

$$\Delta\big(R\Lambda R^\top\big) \;=\; \tfrac{1}{2}\log \frac{\prod_k \big(R\Lambda R^\top\big)_{kk}}{\det\Lambda}.$$

By contrast, in the feature basis $R = I$, the covariance Λ is already diagonal so $\Delta = 0$, meaning that a diagonal encoder is exact and InfoMax-optimal. For any $R \neq I$, $R\Lambda R^\top$ is non-diagonal because the eigenvalues $(\gamma a_k)^{-1}$ are distinct (A4), so $\Delta(R\Lambda R^\top) > 0$ and the rotated basis is suboptimal. The only rotations preserving diagonality are permutations and sign flips.

The InfoMax-optimal diagonal encoder thus selects the unique basis that diagonalizes the posterior, i.e. the feature basis, up to a permutation and a sign. Lemma C.6 fixes the permutation by matching each coordinate's onset to a threshold γ_k.

Remark: the mechanism uses the finite posterior width $(\gamma a_k)^{-1}$ and the distinctness of those widths (A4): in the deterministic $\sigma \to 0$ limit the gap vanishes and rotations are free again. The alignment is thus *derived* from the diagonal-encoder InfoMax objective, not imposed, and holds to leading order in $1/\gamma$, matching concentration. Without the diagonal constraint the rotation survives.

Nesting: As γ increases, the active set $A(\gamma) = \{k : \gamma > \gamma_k\}$ only gains elements, and by the monotonicity of $I(y_k; x_\gamma)$ (Lemma C.4(i)) a feature stays informative once past its threshold. The active set is therefore nested, $A(\gamma) \subseteq A(\gamma')$ for $\gamma < \gamma'$, and grows one feature at a time in the

order set by (A4), $\gamma_1 > \cdots > \gamma_K$: features are represented as $y_K, \ldots, y_1$ as the image possesses less noise. This nested, ordered activation is the hierarchy. □

C.6 Rate-optimal likelihood bias is orthogonal to the prior drift

Let the combined inference drive be $u = s + g$, where $s \in \mathbb{R}^d$ is the attractor drift from the prior population and $g \in \mathbb{R}^d$ is the additive bias from the likelihood population. Suppose the coding rate increases with the drive power, $R = \phi\big(\|u\|^2\big)$ with $\phi' > 0$. Let $g_\perp$ be the component in g orthogonal to s such that $\langle g_\perp, s\rangle = 0$. We can thus write $g = \alpha\, s + g_\perp$. Suppose the bias does not attenuate the prior drift, so that $\langle s, g\rangle \geqslant 0$. Then for any fixed orthogonal component $g_\perp$, the rate is minimized at $\alpha = 0$, so the rate-optimal bias satisfies $\langle s, g\rangle = 0$.

Proof. Since $\alpha\, s \parallel s$ and $g_\perp \perp s$, the Pythagorean identity gives

$$\|u\|^2 = \|(1+\alpha)\, s\|^2 + \|g_\perp\|^2 = (1+\alpha)^2\|s\|^2 + \|g_\perp\|^2.$$

The condition $\langle s, g\rangle = \alpha\|s\|^2 \geqslant 0$ restricts $\alpha \geqslant 0$. On this range $\|u\|^2$ is strictly increasing in α, and since $\phi' > 0$ so is R. Hence R is minimized at $\alpha = 0$, giving $\langle s, g\rangle = 0$ and $u = s + g_\perp$. □

What this implies is that under a shared power budget on the combined drive, the norm form $\|s\|^2 + \|g\|^2$ and the dot-product form $2\langle s, g\rangle$ are not independent constraints but the two terms of a single quadratic. Any prior-parallel component of the likelihood bias inflates the combined power through the cross term while conveying nothing the prior population could not produce by gain modulation alone. A code that penalizes the total power of the combined code thus implicitly orthogonalizes the likelihood bias and the prior drift.

D | Appendix: Hierarchical representations of prior knowledge

D.1 Marginal rate

Here we derive the identity

$$\partial_\eta I(X; Y_\eta|Z) = \mathbb{E}_{x,y,y_\eta,z}\left[r_\eta^2 \left\|\nabla_{y_\eta} \log p(x|y_\eta, z)\right\|^2\right] := \mathbb{E}_{x,y,y_\eta,z}\left[r_\eta^2 \left\|\mathbf{g}_{x|y_\eta,z}\right\|^2\right].$$

We can express the marginal rate in terms of the difference of the unconditional and conditional MMSEs using the I-MMSE relation:

$$\begin{aligned}
\partial_\eta I(X; Y|Z) &= \mathbb{E}_{p(x,y,y_\eta,z)}\left[\left\|y - \left(\hat{y}(y_\eta, \eta) - r_\eta \mathbf{g}_{z|y_\eta}\right)\right\|^2 - \left\|y - \left(\hat{y}(y_\eta, \eta) - r_\eta \mathbf{g}_{z|y_\eta} - r_\eta \mathbf{g}_{x|y_\eta,z}\right)\right\|^2\right] \\
&= \mathbb{E}_{p(x,y,y_\eta,z)}\left[\mathbf{u}^\top \mathbf{u} - \|\mathbf{u} + r_\eta \mathbf{g}_{x|z}\|^2\right] \\
&= \mathbb{E}_{p(x,y,y_\eta,z)}\left[-2r_\eta \mathbf{u}^\top \mathbf{g}_{x|z} - r_\eta^2 \mathbf{g}_{x|z}^\top \mathbf{g}_{x|z}\right]
\end{aligned}$$

where $\mathbf{g}_{z|y_\eta}(y_\eta, z) = \nabla_{y_\eta} \log p(z|y_\eta)$, $\mathbf{g}_{x|y_\eta,z}(y_\eta, x, z) = \nabla_{y_\eta} \log p(x|y_\eta, z)$ and $\mathbf{u} = y - \hat{y}(y_\eta, \eta, z)$. Taking the expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{u}|y_\eta, \eta, z, x\right] &= \mathbb{E}\left[y|y_\eta, \eta, z, x\right] - \hat{y}(y_\eta, \eta, z) \\
&= \hat{y}(y_\eta, \eta, z, x) - \hat{y}(y_\eta, \eta, z) \\
&= \left(\hat{y}(y_\eta, \eta) - r_\eta \mathbf{g}_z(y_\eta, z) - r_\eta \mathbf{g}_{x|z}(y_\eta, z, x)\right) - \left(\hat{y}(y_\eta, \eta) - r_\eta \mathbf{g}_z(y_\eta, z)\right) \\
&= -r_\eta \mathbf{g}_{x|z}(y_\eta, z, x)
\end{aligned}$$

We can plug this back into our definition of $\partial_\eta I(X; Y_\eta|Z)$:

$$\begin{aligned}
\partial_\eta I(X; Y_\eta|Z) &= \mathbb{E}_{p(x,y,y_\eta,z)}\left[-2r_\eta \mathbf{u}^\top \mathbf{g}_{x|y_\eta,z} - r_\eta^2 \mathbf{g}_{x|y_\eta,z}^\top \mathbf{g}_{x|y_\eta,z}\right] \\
&= \mathbb{E}\left[-2r_\eta \mathbb{E}\left[\mathbf{u} \mid y_\eta, \eta, z, x\right]^\top \mathbf{g}_{x|z} - r_\eta^2 \mathbf{g}_{x|y_\eta,z}^\top \mathbf{g}_{x|y_\eta,z}\right] \\
&= \mathbb{E}_{p(x,y,y_\eta,z)}\left[r_\eta^2 \left\|\mathbf{g}_{x|y_\eta,z}\right\|^2\right]
\end{aligned}$$

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). "A Learning Algorithm for Boltzmann Machines". In: *Cognitive Science* 9.1, pp. 147–169. ISSN: 1551-6709. DOI: 10.1207/s15516709cog0901_7.

Adams, W. J. (Aug. 2007). "A Common Light-Prior for Visual Search, Shape, and Reflectance Judgments". In: *Journal of Vision* 7.11, p. 11. ISSN: 1534-7362. DOI: 10.1167/7.11.11.

Adams, W. J., Graf, E. W., and Ernst, M. O. (Oct. 2004). "Experience Can Change the 'light-from-above' Prior". In: *Nature Neuroscience* 7.10, pp. 1057–1058. ISSN: 1546-1726. DOI: 10.1038/nn1312.

Aitchison, L., Jegminat, J., Menendez, J. A., Pfister, J.-P., Pouget, A., and Latham, P. E. (Apr. 2021). "Synaptic Plasticity as Bayesian Inference". In: *Nature Neuroscience* 24.4, pp. 565–571.

Aitchison, L. and Lengyel, M. (Dec. 2016). "The Hamiltonian Brain: Efficient Probabilistic Inference with Excitatory-Inhibitory Neural Circuit Dynamics". In: *PLOS Computational Biology* 12.12. Ed. by Kording, K. P., e1005186.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (July 2018). "Fixing a Broken ELBO". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 159–168.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (Oct. 2019). *Deep Variational Information Bottleneck*.

Angeletos Chrysaitis, N. and Seriès, P. (Feb. 2023). "10 Years of Bayesian Theories of Autism: A Comprehensive Review". In: *Neuroscience & Biobehavioral Reviews* 145, p. 105022.

Archer, E., Park, I. M., and Pillow, J. W. (May 2013). "Bayesian and Quasi-Bayesian Estimators for Mutual Information from Discrete Data". In: *Entropy* 15.5, pp. 1738–1755. ISSN: 1099-4300. DOI: 10.3390/e15051738.

Atick, J. J. and Redlich, A. N. (Sept. 1990). "Towards a Theory of Early Visual Processing". In: *Neural Computation* 2.3, pp. 308–320. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1990.2.3.308.

— (Mar. 1992). "What Does the Retina Know about Natural Scenes?" In: *Neural Computation* 4.2, pp. 196–210. ISSN: 0899-7667. DOI: 10.1162/neco.1992.4.2.196.

Attneave, F. (1954). "Some Informational Aspects of Visual Perception". In: *Psychological Review* 61.3, pp. 183–193. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0054663.

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. (Feb. 2021). "Nonlinear Transform Coding". In: *IEEE Journal of Selected Topics in Signal Processing* 15.2, pp. 339–353. ISSN: 1941-0484. DOI: 10.1109/JSTSP.2020.3034501.

Ballé, J., Laparra, V., and Simoncelli, E. P. (Mar. 2017). "End-to-End Optimized Image Compression". In: *arXiv:1611.01704 [cs, math]*.

Bar, M. (Aug. 2004). "Visual Objects in Context". In: *Nature Reviews Neuroscience* 5.8, pp. 617–629. ISSN: 1471-0048. DOI: 10.1038/nrn1476.

Barber, D. and Agakov, F. (2003). "The IM Algorithm : A Variational Approach to Information Maximization". In: *Advances in Neural Information Processing Systems*. Vol. 16. MIT Press.

Barlow, H. B. (1961). "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication*. Ed. by Rosenblith, W. A. The MIT Press, pp. 216–234. ISBN: 978-0-262-51842-0. DOI: 10.7551/mitpress/9780262518420.003.0013.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. (Aug. 2021). *MINE: Mutual Information Neural Estimation*. DOI: 10.48550/arXiv.1801.04062.

Bengio, Y., Courville, A., and Vincent, P. (Apr. 2014). *Representation Learning: A Review and New Perspectives*.

Bergen, J. R. and Adelson, E. H. (May 1988). "Early Vision and Texture Perception". In: *Nature* 333.6171, pp. 363–364. ISSN: 1476-4687. DOI: 10.1038/333363a0.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (Jan. 2011). "Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment". In: *Science* 331.6013, pp. 83–87.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., and Warland, D. (June 1991). "Reading a Neural Code". In: *Science* 252.5014, pp. 1854–1857. DOI: 10.1126/science.2063199.

Bredenberg, C., Lyo, B., Simoncelli, E., and Savin, C. (2021). "Impression Learning: Online Representation Learning with Synaptic Plasticity". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 11717–11729.

Bredenberg, C., Simoncelli, E., and Savin, C. (2020). "Learning Efficient Task-Dependent Representations with Synaptic Plasticity". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 15714–15724.

Brown, B. C. A., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. (Mar. 2023). *Verifying the Union of Manifolds Hypothesis for Image Data*.

Burda, Y., Grosse, R., and Salakhutdinov, R. (Nov. 2016). *Importance Weighted Autoencoders*. en. arXiv:1509.00519 [cs]. DOI: 10.48550/arXiv.1509.00519.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (Apr. 2018). "Understanding Disentangling in β-VAE". In: *arXiv:1804.03599 [cs, stat]*.

Carlsson, G., Ishkhanov, T., De Silva, V., and Zomorodian, A. (Jan. 2008). "On the Local Behavior of Spaces of Natural Images". In: *International Journal of Computer Vision* 76.1, pp. 1–12.

Chapman, B. and Bonhoeffer, T. (Mar. 1998). "Overrepresentation of Horizontal and Vertical Orientation Preferences in Developing Ferret Area 17". In: *Proceedings of the National Academy of Sciences* 95.5, pp. 2609–2614. DOI: 10.1073/pnas.95.5.2609.

Chatterjee, S. and Diaconis, P. (Apr. 2018). "The Sample Size Required in Importance Sampling". In: *The Annals of Applied Probability* 28.2, pp. 1099–1135. ISSN: 1050-5164, 2168-8737. DOI: 10.1214/17-AAP1326.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (Apr. 2019). *Isolating Sources of Disentanglement in Variational Autoencoders*.

Chen, S., Jiang, L., Rao, R. P., and Shea-Brown, E. (2023). "Expressive Probabilistic Sampling in Recurrent Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., pp. 34981–35005. DOI: 10.52202/075280-1520.

Cheng, W., Darnell, G., Ramachandran, S., and Crawford, L. (July 2020). *Generalizing Variational Autoencoders with Hierarchical Empirical Bayes*.

Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., and Lu, J. (Sept. 2024). *What Does Guidance Do? A Fine-Grained Analysis in a Simple Setting*.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. 2nd ed. Hoboken, N.J: Wiley-Interscience. ISBN: 978-0-471-24195-9.

Cremer, C., Li, X., and Duvenaud, D. (May 2018). *Inference Suboptimality in Variational Autoencoders*. en. arXiv:1801.03558 [cs]. DOI: 10.48550/arXiv.1801.03558.

Csikor, F., Meszéna, B., Szabó, B., and Orbán, G. (June 2022). *Top-down Inference in an Early Visual Cortex Inspired Hierarchical Variational Autoencoder*.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. (2018). "Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models". In: *Journal of Machine Learning Research* 19.41, pp. 1–42. ISSN: 1533-7928.

Dai, B. and Wipf, D. (Sept. 2018). "Diagnosing and Enhancing VAE Models". In: *International Conference on Learning Representations*.

Darlington, T. R., Beck, J. M., and Lisberger, S. G. (Oct. 2018). "Neural Implementation of Bayesian Inference in a Sensorimotor Behavior". In: *Nature Neuroscience* 21.10, pp. 1442–1451.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (Sept. 1995). "The Helmholtz Machine". In: *Neural Computation* 7.5, pp. 889–904.

Deng, L. (2012). "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

Dhariwal, P. and Nichol, A. (2021). "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 8780–8794.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (Feb. 2012). "How Does the Brain Solve Visual Object Recognition?" In: *Neuron* 73.3, pp. 415–434. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2012.01.010.

Doi, E., Gauthier, J. L., Field, G. D., Shlens, J., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., Mathieson, K., Gunning, D. E., Litke, A. M., Paninski, L., Chichilnisky, E. J., and Simoncelli, E. P. (Nov. 2012). "Efficient Coding of Spatial Information in the Primate Retina". In: *Journal of Neuroscience* 32.46, pp. 16256–16264. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.4036-12.2012.

Downing, C. J. (May 1988). "Expectancy and Visual-Spatial Attention: Effects on Perceptual Quality". In: *Journal of Experimental Psychology. Human Perception and Performance* 14.2, pp. 188–202. ISSN: 0096-1523. DOI: 10.1037//0096-1523.14.2.188.

Durrant, T., Kang, Y., Wolpert, D., and Lengyel, M. (2023). "Here, there, but not everywhere: multimodal beliefs captured by a population code during navigation". In: *Bernstein conference proceedings*. DOI: 10.12751/nncn.bc2023.296.

Eastwood, C. and Williams, C. K. I. (Feb. 2018). "A Framework for the Quantitative Evaluation of Disentangled Representations". In: *International Conference on Learning Representations*.

Echeveste, R., Aitchison, L., Hennequin, G., and Lengyel, M. (Sept. 2020). "Cortical-like Dynamics in Recurrent Circuits Optimized for Sampling-Based Probabilistic Inference". In: *Nature Neuroscience* 23.9, pp. 1138–1149.

Ernst, M. O. and Banks, M. S. (Jan. 2002). “Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion”. In: *Nature* 415.6870, pp. 429–433. ISSN: 1476-4687. DOI: 10.1038/415429a.

Falorsi, L., Haan, P. de, Davidson, T. R., Cao, N. D., Weiler, M., Forré, P., and Cohen, T. S. (July 2018). *Explorations in Homeomorphic Variational Auto-Encoding*. DOI: 10.48550/arXiv.1807.04689.

Fang, C., Cai, X., and Lu, H. D. (Apr. 2022). “Orientation Anisotropies in Macaque Visual Areas”. In: *Proceedings of the National Academy of Sciences* 119.15, e2113407119. DOI: 10.1073/pnas.2113407119.

Feather, J., Leclerc, G., Mądry, A., and McDermott, J. H. (Nov. 2023). “Model metamers reveal divergent invariances between biological and artificial neural networks”. en. In: *Nature Neuroscience* 26.11. Publisher: Nature Publishing Group, pp. 2017–2034. ISSN: 1546-1726. DOI: 10.1038/s41593-023-01442-0.

Fefferman, C., Mitter, S., and Narayanan, H. (Feb. 2016). “Testing the Manifold Hypothesis”. In: *Journal of the American Mathematical Society* 29.4, pp. 983–1049.

Felleman, D. J. and Van Essen, D. C. (1991). “Distributed Hierarchical Processing in the Primate Cerebral Cortex”. In: *Cerebral Cortex* 1.1, pp. 1–47. ISSN: 1047-3211. DOI: 10.1093/cercor/1.1.1-a.

Field, D. J. (Dec. 1987). “Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells”. In: *Journal of the Optical Society of America. A, Optics and Image Science* 4.12, pp. 2379–2394. ISSN: 0740-3232. DOI: 10.1364/josaa.4.002379.

Fischer, B. J. and Peña, J. L. (Aug. 2011). “Owl’s Behavior and Neural Representation Predicted by Bayesian Inference”. In: *Nature Neuroscience* 14.8, pp. 1061–1066.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (Mar. 2010). “Statistically Optimal Perception and Learning: From Behavior to Neural Representations”. In: *Trends in cognitive sciences* 14.3, pp. 119–130.

Franzese, G., Bounoua, M., and Michiardi, P. (May 2024). *MINDE: Mutual Information Neural Diffusion Estimation*. DOI: 10.48550/arXiv.2310.09031.

Freeman, J. and Simoncelli, E. P. (Sept. 2011). "Metamers of the ventral stream". en. In: *Nature Neuroscience* 14.9. Publisher: Nature Publishing Group, pp. 1195–1201. ISSN: 1546-1726. DOI: 10.1038/nn.2889.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (July 2013). "A Functional and Perceptual Signature of the Second Visual Area in Primates". In: *Nature Neuroscience* 16.7, pp. 974–981. ISSN: 1546-1726. DOI: 10.1038/nn.3402.

Fuest, M., Ma, P., Gui, M., Schusterbauer, J., Hu, V. T., and Ommer, B. (June 2024). *Diffusion Models and Representation Learning: A Survey*. en. arXiv:2407.00783 [cs]. DOI: 10.48550/arXiv.2407.00783.

Fukushima, K. (Apr. 1980). "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 0340-1200, 1432-0770. DOI: 10.1007/BF00344251.

Ganguli, D. and Simoncelli, E. P. (2010). "Implicit Encoding of Prior Probabilities in Optimal Neural Populations". In: p. 9.

— (Oct. 2014). "Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations". In: *Neural Computation* 26.10, pp. 2103–2134.

Geisler, W. S. and Perry, J. S. (Jan. 2009). "Contour Statistics in Natural Images: Grouping across Occlusions". In: *Visual Neuroscience* 26.1, pp. 109–121. ISSN: 0952-5238, 1469-8714. DOI: 10.1017/S0952523808080875.

Geman, S. and Geman, D. (Nov. 1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, pp. 721–741. ISSN: 1939-3539. DOI: 10.1109/TPAMI.1984.4767596.

Gerardin, P., Kourtzi, Z., and Mamassian, P. (Sept. 2010). "Prior Knowledge of Illumination for 3D Perception in the Human Brain". In: *Proceedings of the National Academy of Sciences* 107.37, pp. 16309–16314.

Gershman, S. and Goodman, N. D. (2014). "Amortized Inference in Probabilistic Reasoning". In: *Cognitive Science*.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (July 2011). "Cardinal Rules: Visual Orientation Perception Reflects Knowledge of Environmental Statistics". In: *Nature Neuroscience* 14.7, pp. 926–932.

Glazunov, M. and Zarras, A. (July 2023). "Vacant Holes for Unsupervised Detection of the Outliers in Compact Latent Representation". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. PMLR, pp. 701–711.

Gold, J. I. and Shadlen, M. N. (July 2007). "The Neural Basis of Decision Making". In: *Annual Review of Neuroscience* 30.1, pp. 535–574. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev.neuro.29.051605.113038.

Goldreich, D. and Peterson, M. A. (2012). "A Bayesian Observer Replicates Convexity Context Effects in Figure–Ground Perception". In: *Seeing and Perceiving* 25.3-4, pp. 365–395. ISSN: 1878-4755, 1878-4763. DOI: 10.1163/187847612X634445.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (June 2014). *Generative Adversarial Networks*. arXiv:1406.2661 [stat]. DOI: 10.48550/arXiv.1406.2661.

Gorban, A. N. and Tyukin, I. Y. (Apr. 2018). "Blessing of Dimensionality: Mathematical Foundations of the Statistical Physics of Data". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 376.2118, p. 20170237.

Guo, D., Shamai, S., and Verdu, S. (Apr. 2005). "Mutual Information and Minimum Mean-Square Error in Gaussian Channels". In: *IEEE Transactions on Information Theory* 51.4, pp. 1261–1282. ISSN: 1557-9654. DOI: 10.1109/TIT.2005.844072.

Guth, F., Kadkhodaie, Z., and Simoncelli, E. P. (June 2025). *Learning Normalized Image Densities via Dual Score Matching*. DOI: 10.48550/arXiv.2506.05310.

Hahn, M. and Wei, X.-X. (Dec. 2022). *A Unifying Theory Explains Seemingly Contradicting Biases in Perceptual Estimation*.

Haijiang, Q., Saunders, J. A., Stone, R. W., and Backus, B. T. (Jan. 2006). "Demonstration of Cue Recruitment: Change in Visual Appearance by Means of Pavlovian Conditioning". In: *Proceedings of the National Academy of Sciences* 103.2, pp. 483–488. DOI: 10.1073/pnas.0506728103.

He, K., Zhang, X., Ren, S., and Sun, J. (Dec. 2015). *Deep Residual Learning for Image Recognition*.

Heeger, D. J. and Bergen, J. R. (Sept. 1995). "Pyramid-Based Texture Analysis/Synthesis". In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '95. New York, NY, USA: Association for Computing Machinery, pp. 229–238. ISBN: 978-0-89791-701-8. DOI: 10.1145/218380.218446.

Helmholtz, H. (Jan. 1852). "LXXXI. On the theory of compound colours". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.28. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786445208647175, pp. 519–534. ISSN: 1941-5982. DOI: 10.1080/14786445208647175.

Hénaff, O. J., Goris, R. L. T., and Simoncelli, E. P. (June 2019). "Perceptual Straightening of Natural Videos". In: *Nature Neuroscience* 22.6, pp. 984–991.

Hennequin, G., Aitchison, L., and Lengyel, M. (Apr. 2014). *Fast Sampling for Bayesian Inference in Neural Circuits*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (Feb. 2017). "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (May 1995). “The "Wake-Sleep" Algorithm for Unsupervised Neural Networks”. In: *Science* 268.5214, pp. 1158–1161.

Ho, J., Jain, A., and Abbeel, P. (2020). “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 6840–6851.

Ho, J. and Salimans, T. (Dec. 2021). “Classifier-Free Diffusion Guidance”. In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hoffman, M. D. and Johnson, M. J. (2016). “ELBO Surgery: Yet Another Way to Carve up the Variational Evidence Lower Bound”. In.

Houlsby, N. M., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., and Lengyel, M. (Nov. 2013). “Cognitive Tomography Reveals Complex, Task-Independent Mental Representations”. In: *Current Biology* 23.21, pp. 2169–2175.

Hoyer, P. and Hyvärinen, A. (2002). “Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior”. In: *Advances in Neural Information Processing Systems*. Vol. 15. MIT Press.

Hubel, D. H. and Wiesel, T. N. (1962). “Receptive Fields, Binocular Interaction and Functional Architecture in the Cat’s Visual Cortex”. In: *The Journal of Physiology* 160.1, pp. 106–154. ISSN: 1469-7793. DOI: 10.1113/jphysiol.1962.sp006837.

— (Mar. 1968). “Receptive Fields and Functional Architecture of Monkey Striate Cortex”. In: *The Journal of Physiology* 195.1, pp. 215–243. ISSN: 0022-3751. DOI: 10.1113/jphysiol.1968.sp008455.

Hudson, D. A., Zoran, D., Malinowski, M., Lampinen, A. K., Jaegle, A., McClelland, J. L., Matthey, L., Hill, F., and Lerchner, A. (June 2024). “SODA: Bottleneck Diffusion Models for Representation Learning”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, pp. 23115–23127. ISBN: 979-8-3503-5300-6. DOI: 10.1109/CVPR52733.2024.02181.

Hyvarinen, A. and Morioka, H. (May 2016). *Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA*. DOI: 10.48550/arXiv.1605.06336.

— (Apr. 2017). "Nonlinear ICA of Temporally Dependent Stationary Sources". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 460–469.

Hyvärinen, A. (Dec. 1997). "New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit". In: *Neural Information Processing Systems*.

— (2005). "Estimation of Non-Normalized Statistical Models by Score Matching". In: p. 15.

Hyvärinen, A. and Pajunen, P. (Apr. 1999). "Nonlinear Independent Component Analysis: Existence and Uniqueness Results". In: *Neural Networks* 12.3, pp. 429–439. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(98)00140-3.

Ifriqi, T. B., Astolfi, P., Hall, M., Askari-Hemmat, R., Benchetrit, Y., Havasi, M., Muckley, M., Alahari, K., Romero-Soriano, A., Verbeek, J., and Drozdzal, M. (Jan. 2025). *On Improved Conditioning Mechanisms and Pre-training Strategies for Diffusion Models*. en. arXiv:2411.03177 [cs]. DOI: 10.48550/arXiv.2411.03177.

Jacobs, R. A. (Oct. 1999). "Optimal Integration of Texture and Motion Cues to Depth". In: *Vision Research* 39.21, pp. 3621–3629. ISSN: 00426989. DOI: 10.1016/S0042-6989(99)00088-7.

Jadi, M. P., Behabadi, B. F., Poleg-Polsky, A., Schiller, J., and Mel, B. W. (May 2014). "An Augmented Two-Layer Model Captures Nonlinear Analog Spatial Integration Effects in Pyramidal Neuron Dendrites". In: *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers* 102.5, p. 1.

Jazayeri, M. and Ostojic, S. (Oct. 2021). "Interpreting Neural Computations by Examining Intrinsic and Embedding Dimensionality of Neural Activity". In: *Current Opinion in Neurobiology*. Computational Neuroscience 70, pp. 113–120.

Jones, J. P. and Palmer, L. A. (Dec. 1987). "The Two-Dimensional Spatial Structure of Simple Receptive Fields in Cat Striate Cortex". In: *Journal of Neurophysiology* 58.6, pp. 1187–1211. ISSN: 0022-3077. DOI: 10.1152/jn.1987.58.6.1187.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. (May 2024). "Generalization in diffusion models arises from geometry-adaptive harmonic representation". In: *Int'l Conf on Learning Representations (ICLR)*. Vol. 12. Vienna, Austria.

Kadkhodaie, Z. and Simoncelli, E. (2021). "Stochastic Solutions for Linear Inverse Problems Using the Prior Implicit in a Denoiser". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 13242–13254.

Kadkhodaie, Z. and Simoncelli, E. P. (2020a). *Solving Linear Inverse Problems Using the Prior Implicit in a Denoiser*.

— (2020b). *Solving Linear Inverse Problems Using the Prior Implicit in a Denoiser*.

Kaiser, T., Adaloglou, N., and Kollmann, M. (Nov. 2024). *The Unreasonable Effectiveness of Guidance for Diffusion Models*. arXiv:2411.10257 [cs] version: 1. DOI: 10.48550/arXiv.2411.10257.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (Feb. 2018). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. arXiv:1710.10196 [cs]. DOI: 10.48550/arXiv.1710.10196.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., and Shenoy, K. V. (Mar. 2014). "Cortical Activity in the Null Space: Permitting Preparation without Movement". In: *Nature Neuroscience* 17.3, pp. 440–448. ISSN: 1546-1726. DOI: 10.1038/nn.3643.

Kersten, D. and Yuille, A. (Apr. 2003). "Bayesian Models of Object Perception". In: *Current Opinion in Neurobiology* 13.2, pp. 150–158.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (June 2020). "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2207–2217.

Kim, H. and Mnih, A. (July 2019). *Disentangling by Factorising*.

Kim, Y., Lee, K., Park, M., Na, B., and Moon, I.-c. (Oct. 2024). "Diffusion Bridge AutoEncoders for Unsupervised Representation Learning". In: *The Thirteenth International Conference on Learning Representations*.

Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). "Variational Diffusion Models". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 21696–21707.

Kingma, D. P. and Gao, R. (Sept. 2023). *Understanding Diffusion Objectives as the ELBO with Simple Data Augmentation*.

Kingma, D. P. and Welling, M. (2013). "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]*.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. (Mar. 2021). *Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding*. DOI: 10.48550/arXiv.2007.10930.

Klushyn, A., Chen, N., Kurle, R., Cseke, B., and van der Smagt, P. (2019). "Learning Hierarchical Priors in VAEs". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Knill, D. C. (June 1998). "Discrimination of Planar Surface Slant from Texture: Human and Ideal Observers Compared". In: *Vision Research* 38.11, pp. 1683–1711. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(97)00325-8.

— (June 2007). "Learning Bayesian Priors for Depth Perception". In: *Journal of Vision* 7.8, p. 13. ISSN: 1534-7362. DOI: 10.1167/7.8.13.

Knill, D. C. and Richards, W., eds. (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press. ISBN: 978-0-521-46109-2. DOI: 10.1017/CBO9780511984037.

Kong, X., Brekelmans, R., and Steeg, G. V. (Feb. 2023). *Information-Theoretic Diffusion*.

Kong, X., Liu, O., Li, H., Yogatama, D., and Steeg, G. V. (May 2024). *Interpretable Diffusion via Information Decomposition*.

Kraskov, A., Stögbauer, H., and Grassberger, P. (June 2004). "Estimating Mutual Information". In: *Physical Review E* 69.6, p. 066138. DOI: 10.1103/PhysRevE.69.066138.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (May 2017). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Communications of the ACM* 60.6, pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386.

Kuchibhotla, K. V., Gill, J. V., Lindsay, G. W., Papadoyannis, E. S., Field, R. E., Sten, T. A. H., Miller, K. D., and Froemke, R. C. (Jan. 2017). "Parallel Processing by Cortical Inhibition Enables Context-Dependent Behavior". In: *Nature Neuroscience* 20.1, pp. 62–71.

Kumar, A. and Poole, B. (Nov. 2020). "On Implicit Regularization in $$\beta$$-VAEs". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 5480–5490.

Kumar, A., Sattigeri, P., and Balakrishnan, A. (Dec. 2018). *Variational Inference of Disentangled Latent Concepts from Unlabeled Observations*. arXiv:1711.00848 [cs]. DOI: 10.48550/arXiv.1711.00848.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Priol, R. L., Lacoste, A., and Lacoste-Julien, S. (June 2022). "Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA". In: *Proceedings of the First Conference on Causal Learning and Reasoning*. PMLR, pp. 428–484.

Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. (Feb. 1995). "Measurement and Modeling of Depth Cue Combination: In Defense of Weak Fusion". In: *Vision Research* 35.3, pp. 389–412. ISSN: 0042-6989. DOI: 10.1016/0042-6989(94)00176-M.

Langer, M. S. and Bülthoff, H. H. (2001). "A Prior for Global Convexity in Local Shape-from-Shading". In: *Perception* 30.4, pp. 403–410. ISSN: 0301-0066. DOI: 10.1068/p3178.

Laquitaine, S., Azeglio, S., Paris, C., Ferrari, U., and Chalk, M. (Oct. 2025). "Decomposing Stimulus-Specific Sensory Neural Information via Diffusion Models". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Laughlin, S. (Oct. 1981). "A Simple Coding Procedure Enhances a Neuron's Information Capacity". In: *Zeitschrift für Naturforschung C* 36.9-10, pp. 910–912. ISSN: 1865-7125, 0939-5075. DOI: 10.1515/znc-1981-9-1040.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (Nov. 1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791.

Lee, T. S. and Mumford, D. (July 2003). "Hierarchical Bayesian Inference in the Visual Cortex". In: *Journal of the Optical Society of America A* 20.7, p. 1434.

Li, B., Peterson, M. R., and Freeman, R. D. (July 2003). "Oblique Effect: A Neural Basis in the Visual Cortex". In: *Journal of Neurophysiology* 90.1, pp. 204–217.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). "Visualizing the Loss Landscape of Neural Nets". In.

Li, R., Peng, X., and Lin, C. (Oct. 2021). *On the Latent Holes of VAEs for Text Generation*. DOI: 10.48550/arXiv.2110.03318.

Liang, Q., Liu, Z., Ostrow, M., and Fiete, I. (Aug. 2024). *How Diffusion Models Learn to Factorize and Compose*.

Liang, Q., Qian, D., Ziyin, L., and Fiete, I. (May 2025). *Compositional Generalization via Forced Rendering of Disentangled Latents*. DOI: 10.48550/arXiv.2501.18797.

Lin, H. W., Tegmark, M., and Rolnick, D. (Sept. 2017). "Why Does Deep and Cheap Learning Work so Well?" In: *Journal of Statistical Physics* 168.6, pp. 1223–1247. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-017-1836-5.

Linsker, R. (1988). "Self-organization in a perceptual network". In: *Computer* 21.3, pp. 105–117. DOI: 10.1109/2.36.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (May 2019). "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR, pp. 4114–4124.

Locatello, F., Locatello, F., Bauer, S., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2020). "A Sober Look at the Unsupervised Learning of Disentangled Representations and Their Evaluation". In.

Lyo, B. S. H. and Savin, C. (Oct. 2023). *Complex Priors and Flexible Inference in Recurrent Circuits with Dendritic Nonlinearities*.

Lyo, B. S. H., Simoncelli, E. P., and Savin, C. (Dec. 2025). *Disentangled Representations via Score-Based Variational Autoencoders*. DOI: 10.48550/arXiv.2512.17127.

Ma, W. J. (Oct. 2012). "Organizing Probabilistic Models of Perception". In: *Trends in Cognitive Sciences* 16.10, pp. 511–518.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (Nov. 2006). "Bayesian Inference with Probabilistic Population Codes". In: *Nature Neuroscience* 9.11, pp. 1432–1438.

Machta, B. B., Chachra, R., Transtrum, M. K., and Sethna, J. P. (Nov. 2013). "Parameter Space Compression Underlies Emergent Theories and Predictive Models". In: *Science* 342.6158, pp. 604–607.

Mahendran, A. and Vedaldi, A. (Nov. 2014). *Understanding Deep Image Representations by Inverting Them*. en. arXiv:1412.0035 [cs]. DOI: 10.48550/arXiv.1412.0035.

Mallat, S. (July 1989). "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7, pp. 674–693. ISSN: 1939-3539. DOI: 10.1109/34.192463.

Mamassian, P., Landy, M., and Maloney, L. T. (Mar. 2002). "Bayesian Modelling of Visual Perception". In: *Probabilistic Models of the Brain: Perception and Neural Function*. The MIT Press, pp. 13–36.

Mamassian, P. and Landy, M. S. (Sept. 1998). "Observer Biases in the 3D Interpretation of Line Drawings". In: *Vision Research* 38.18, pp. 2817–2832. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(97)00438-0.

Manduchi, L., Vandenhirtz, M., Ryser, A., and Vogt, J. (Dec. 2023). "Tree Variational Autoencoders". In: *Advances in Neural Information Processing Systems* 36, pp. 54952–54986.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (Nov. 2013). "Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex". In: *Nature* 503.7474, pp. 78–84. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12742.

Mareschal, I., Calder, A. J., and Clifford, C. W. (Apr. 2013). "Humans Have an Expectation That Gaze Is Directed Toward Them". In: *Current Biology* 23.8, pp. 717–721. ISSN: 0960-9822. DOI: 10.1016/j.cub.2013.03.030.

Masset, P., Zavatone-Veth, J. A., Connor, J. P., Murthy, V. N., and Pehlevan, C. (2022). "Natural Gradient Enables Fast Sampling in Spiking Neural Networks". In: *Advances in neural information processing systems* 35, pp. 22018–22034.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. (June 2019). *Disentangling Disentanglement in Variational Autoencoders*.

Metzler, C. A., Mousavi, A., Heckel, R., and Baraniuk, R. G. (July 2020). *Unsupervised Learning with Stein's Unbiased Risk Estimator*. DOI: 10.48550/arXiv.1805.10531.

Mittal, S., Abstreiter, K., Bauer, S., Schölkopf, B., and Mehrjou, A. (July 2023). "Diffusion Based Representation Learning". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 24963–24982.

Miyasawa, K. et al. (1961). "An empirical Bayes estimator of the mean of a normal population". In: *Bull. Inst. Internat. Statist* 38.181-188, pp. 1–2.

Mohan, S., Kadkhodaie, Z., Simoncelli, E. P., and Fernandez-Granda, C. (Mar. 2022). "Robust And Interpretable Blind Image Denoising Via Bias-Free Convolutional Neural Networks". In: *International Conference on Learning Representations*.

Mumford, D. (Jan. 1992). "On the Computational Architecture of the Neocortex". In: *Biological Cybernetics* 66.3, pp. 241–251. ISSN: 1432-0770. DOI: 10.1007/BF00198477.

Murayama, M., Pérez-Garci, E., Nevian, T., Bock, T., Senn, W., and Larkum, M. E. (Feb. 2009). "Dendritic Encoding of Sensory Stimuli Controlled by Deep Cortical Interneurons". In: *Nature* 457.7233, pp. 1137–1141.

Neal, R. M. (July 1992). "Connectionist Learning of Belief Networks". In: *Artificial Intelligence* 56.1, pp. 71–113. ISSN: 00043702. DOI: 10.1016/0004-3702(92)90065-6.

— (Dec. 1996). "Sampling from Multimodal Distributions Using Tempered Transitions". In: *Statistics and Computing* 6.4, pp. 353–366.

Nemenman, I., Shafee, F., and Bialek, W. (2001). "Entropy and Inference, Revisited". In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (Nov. 2010). "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization". In: *IEEE Transactions on Information Theory* 56.11, pp. 5847–5861. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.2010.2068870.

Nichol, A. Q. and Dhariwal, P. (July 2021). "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, pp. 8162–8171.

Niu, X., Savin, C., and Simoncelli, E. P. (Dec. 2024). "Learning predictable and robust neural representations by straightening image sequences". en. In: *Advances in Neural Information Processing Systems* 37, pp. 40316–40335.

Olshausen, B. A. and Field, D. J. (1996). "Natural Image Statistics and Efficient Coding". In.

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (Oct. 2016). "Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex". In: *Neuron* 92.2, pp. 530–543.

Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (June 2015). "Predictive Information in a Sensory Population". In: *Proceedings of the National Academy of Sciences* 112.22, pp. 6908–6913.

Paninski, L. (June 2003). "Estimation of Entropy and Mutual Information". In: *Neural Computation* 15.6, pp. 1191–1253. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976603321780272.

Pasupathy, A. and Connor, C. E. (Nov. 2001). "Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation". In: *Journal of Neurophysiology* 86.5, pp. 2505–2519. ISSN: 0022-3077, 1522-1598. DOI: 10.1152/jn.2001.86.5.2505.

Pearl, J. (2009). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Rev. 2. ed., transferred to digital printing. The Morgan Kaufmann Series in Representation and Reasoning. San Francisco, Calif: Morgan Kaufmann.

Peters, B., DiCarlo, J. J., Gureckis, T., Isik, L., Tenenbaum, J., Konkle, T., Naselaris, T., Stachenfeld, K., Tavares, Z., Tsao, D., Yildirim, I., and Kriegeskorte, N. (2024). "How Does the Primate Brain Combine Generative and Discriminative Computations in Vision?" In.

Poirazi, P., Brannon, T., and Mel, B. W. (Mar. 2003). "Pyramidal Neuron as Two-Layer Neural Network". In: *Neuron* 37.6, pp. 989–999.

Polsky, A., Mel, B. W., and Schiller, J. (June 2004). "Computational Subunits in Thin Dendrites of Pyramidal Cells". In: *Nature Neuroscience* 7.6, pp. 621–627.

Poole, B., Ozair, S., Oord, A. van den, Alemi, A. A., and Tucker, G. (May 2019). *On Variational Bounds of Mutual Information*. https://arxiv.org/abs/1905.06922v1.

Portilla, J. and Simoncelli, E. P. (Oct. 2000). "A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients". en. In: *International Journal of Computer Vision* 40.1, pp. 49–70. ISSN: 1573-1405. DOI: 10.1023/A:1026553619983.

Posner, M. I. (Feb. 1980). "Orienting of Attention". In: *Quarterly Journal of Experimental Psychology* 32.1, pp. 3–25. ISSN: 0033-555X. DOI: 10.1080/00335558008248231.

Preechakul, K., Chatthee, N., Wizadwongsa, S., and Suwajanakorn, S. (2022). "Diffusion Autoencoders: Toward a Meaningful and Decodable Representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629.

Qiang, B., Song, Y., Xu, M., Gong, J., Gao, B., Zhou, H., Ma, W.-Y., and Lan, Y. (July 2023). "Coarse-to-Fine: A Hierarchical Diffusion Model for Molecule Generation in 3D". In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, pp. 28277–28299.

Raiguel, S., Vogels, R., Mysore, S. G., and Orban, G. A. (June 2006). "Learning to See the Difference Specifically Alters the Most Informative V4 Neurons". In: *Journal of Neuroscience* 26.24, pp. 6589–6602. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.0457-06.2006.

Ramachandran, V. S. (Aug. 1988). "Perceiving Shape from Shading". In: *Scientific American*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (July 2021). "Zero-Shot Text-to-Image Generation". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, pp. 8821–8831.

Rao, R. P. N. and Ballard, D. H. (Jan. 1999). "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects". In: *Nature Neuroscience* 2.1, pp. 79–87. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/4580.

Raphan, M. and Simoncelli, E. P. (Feb. 2011). "Least Squares Estimation Without Priors or Supervision". In: *Neural Computation* 23.2, pp. 374–420.

Reichert, D. P., Series, P., and Storkey, A. J. (July 2011). "Unifying Low-Level Mechanistic and High-Level Bayesian Explanations of Bistable Perceptions: Neuronal Adaptation for Cortical Inference". In: *BMC Neuroscience* 12.1, P320.

Rezende, D. J. and Mohamed, S. (July 2015). "Variational inference with normalizing flows". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, pp. 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (Jan. 2014). *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*. https://arxiv.org/abs/1401.4082v3.

Rezende, D. J. and Viola, F. (Oct. 2018). *Taming VAEs*.

Ridgeway, K. and Mozer, M. C. (2018). "Learning Deep Disentangled Embeddings With the F-Statistic Loss". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc.

Riesenhuber, M. and Poggio, T. (Nov. 1999). "Hierarchical Models of Object Recognition in Cortex". In: *Nature Neuroscience* 2.11, pp. 1019–1025. ISSN: 1546-1726. DOI: 10.1038/14819.

Robbins, H. (Jan. 1956). “An Empirical Bayes Approach to Statistics”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 3.1. University of California Press, pp. 157–164.

Rolínek, M., Zietlow, D., and Martius, G. (June 2019). “Variational Autoencoders Pursue PCA Directions (by Accident)”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, pp. 12398–12407. ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.01269.

Roux, L. and Buzsáki, G. (Jan. 2015). “Tasks for Inhibitory Interneurons in Intact Brain Circuits”. In: *Neuropharmacology*. GABAergic Signaling in Health and Disease 88, pp. 10–23.

Rullán Buxó, C. and Savin, C. (2021). “A Sampling-Based Circuit for Optimal Decision Making”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 14163–14175.

Sachdeva, V., Mora, T., Walczak, A. M., and Palmer, S. (May 2020). *Optimal Prediction with Resource Constraints Using the Information Bottleneck*. Preprint. Biophysics.

Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., and Weber, R. M. (May 2024). *CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling*. en. arXiv:2310.17347 [cs]. DOI: 10.48550/arXiv.2310.17347.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (Dec. 2022). “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 36479–36494.

Savin, C., Dayan, P., and Lengyel, M. (Feb. 2014). “Optimal Recall from Bounded Metaplastic Synapses: Predicting Functional Adaptations in Hippocampal Area CA3”. In: *PLoS Computational Biology* 10.2. Ed. by Sporns, O., e1003489.

Savin, C. and Denève, S. (2014). "Spatio-Temporal Representations of Uncertainty in Spiking Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc.

Schoups, A., Vogels, R., Qian, N., and Orban, G. (Aug. 2001). "Practising Orientation Identification Improves Orientation Coding in V1 Neurons". In: *Nature* 412.6846, pp. 549–553. ISSN: 1476-4687. DOI: 10.1038/35087601.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (Jan. 2020). *Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?* DOI: 10.1101/407007.

Schwartz, O., Hsu, A., and Dayan, P. (July 2007). "Space and Time in Visual Context". In: *Nature Reviews Neuroscience* 8.7, pp. 522–535. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn2155.

Schwartz, O. and Simoncelli, E. P. (Aug. 2001). "Natural Signal Statistics and Sensory Gain Control". In: *Nature Neuroscience* 4.8, pp. 819–825.

Sclocchi, A., Favero, A., and Wyart, M. (Jan. 2025). "A Phase Transition in Diffusion Models Reveals the Hierarchical Nature of Data". In: *Proceedings of the National Academy of Sciences* 122.1, e2408799121. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2408799121.

Sekuler, R. and Ball, K. (Oct. 1977). "Mental Set Alters Visibility of Moving Targets". In: *Science* 198.4312, pp. 60–62. DOI: 10.1126/science.897682.

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., and Kohn, A. (Apr. 2019). "Cortical Areas Interact through a Communication Subspace". In: *Neuron* 102.1, 249–259.e4. ISSN: 08966273. DOI: 10.1016/j.neuron.2019.01.026.

Sengpiel, F., Stawinski, P., and Bonhoeffer, T. (Aug. 1999). "Influence of Experience on Orientation Maps in Cat Visual Cortex". In: *Nature Neuroscience* 2.8, pp. 727–732. ISSN: 1546-1726. DOI: 10.1038/11192.

Seriès, P. and Seitz, A. R. (2013). "Learning What to Expect (in Visual Perception)". In: *Frontiers in Human Neuroscience* 7.

Serre, T., Oliva, A., and Poggio, T. (Apr. 2007). "A Feedforward Architecture Accounts for Rapid Categorization". In: *Proceedings of the National Academy of Sciences* 104.15, pp. 6424–6429. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0700622104.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. ISSN: 1538-7305. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Sikka, H., Zhong, W., Yin, J., and Pehlevan, C. (Dec. 2019). "A Closer Look at Disentangling in β-VAE". In: *arXiv:1912.05127 [cs, stat]*.

Simon, H. A. (1962). "The Architecture of Complexity". In: *The Roots of Logistics: A Reader of Classical Contributions to the History and Conceptual Foundations of the Science of Logistics*. Ed. by Klaus, P. and Müller, S. Berlin, Heidelberg: Springer, pp. 335–361. ISBN: 978-3-642-27922-5. DOI: 10.1007/978-3-642-27922-5_23.

Simoncelli, E. and Freeman, W. (1995). "The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation". In: *Proceedings., International Conference on Image Processing*. Vol. 3. Washington, DC, USA: IEEE Comput. Soc. Press, pp. 444–447. ISBN: 978-0-7803-3122-8. DOI: 10.1109/ICIP.1995.537667.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (June 2015a). "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, pp. 2256–2265.

— (June 2015b). "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, pp. 2256–2265.

Sønderby, C. K., Raiko, T., Maaløe, L., Sø nderby, S. r. K., and Winther, O. (2016). "Ladder Variational Autoencoders". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Song, J. and Ermon, S. (Mar. 2020). *Understanding the Limitations of Variational Mutual Information Estimators*. DOI: 10.48550/arXiv.1910.06222.

Song, Y. and Ermon, S. (2019). "Generative Modeling by Estimating Gradients of the Data Distribution". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Stein, C. M. (Nov. 1981). "Estimation of the Mean of a Multivariate Normal Distribution". In: *The Annals of Statistics* 9.6, pp. 1135–1151. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176345632.

Sterzer, P., Frith, C., and Petrovic, P. (Aug. 2008). "Believing Is Seeing: Expectations Alter Visual Awareness". In: *Current Biology* 18.16, R697–R698. ISSN: 0960-9822. DOI: 10.1016/j.cub.2008.06.021.

Stocker, A. A. and Simoncelli, E. P. (Apr. 2006). "Noise Characteristics and Prior Expectations in Human Visual Speed Perception". In: *Nature Neuroscience* 9.4, pp. 578–585.

Theis, L., Oord, A. van den, and Bethge, M. (Apr. 2016). "A Note on the Evaluation of Generative Models". In: *International Conference on Machine Learning*.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). *The information bottleneck method*.

Tishby, N. and Zaslavsky, N. (Mar. 2015). *Deep Learning and the Information Bottleneck Principle*. DOI: 10.48550/arXiv.1503.02406.

Transtrum, M. K., Machta, B. B., Brown, K. S., Daniels, B. C., Myers, C. R., and Sethna, J. P. (July 2015). "Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond". In: *The Journal of Chemical Physics* 143.1, p. 010901.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., and Chun, M. M. (Aug. 2010). "Implicit Perceptual Anticipation Triggered by Statistical Learning". In: *The Journal of Neuroscience* 30.33, pp. 11177–11187. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.0858-10.2010.

Turner, R. E., Diaconu, C.-D., Markou, S., Shysheya, A., Foong, A. Y. K., and Mlodozeniec, B. (Feb. 2024). *Denoising Diffusion Probabilistic Models in Six Simple Steps*.

Ujfalussy, B. B., Makara, J. K., Lengyel, M., and Branco, T. (Nov. 2018). “Global and Multiplexed Dendritic Computations under In Vivo-like Conditions”. In: *Neuron* 100.3, 579–592.e5.

Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vafaii, H., Yates, J. L., and Butts, D. A. (Sept. 2023). *Hierarchical VAEs Provide a Normative Account of Motion Processing in the Primate Brain*. Preprint. Neuroscience.

Vahdat, A. and Kautz, J. (Jan. 2021). “NVAE: A Deep Hierarchical Variational Autoencoder”. In: *arXiv:2007.03898 [cs, stat]*.

van den Oord, A., Li, Y., and Vinyals, O. (July 2018). “Representation Learning with Contrastive Predictive Coding”. In.

van Hateren, J. H. (1992). “A Theory of Maximizing Sensory Information”. In: *Biological Cybernetics* 68.1, pp. 23–29. ISSN: 0340-1200. DOI: 10.1007/BF00203134.

Vincent, P. (July 2011). “A Connection Between Score Matching and Denoising Autoencoders”. In: *Neural Computation* 23.7, pp. 1661–1674.

Wainwright, M. J. and Simoncelli, E. (1999). “Scale Mixtures of Gaussians and the Statistics of Natural Images”. In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press.

Wang, G., Ding, S., and Yunokuchi, K. (Dec. 2003). “Representation of Cardinal Contour Overlaps Less With Representation of Nearby Angles in Cat Visual Cortex”. In: *Journal of Neurophysiology* 90.6, pp. 3912–3920. ISSN: 0022-3077. DOI: 10.1152/jn.00219.2003.

Wang, Y., Schiff, Y., Gokaslan, A., Pan, W., Wang, F., De Sa, C., and Kuleshov, V. (June 2023). *InfoDiffusion: Representation Learning Using Information Maximizing Diffusion Models*.

Wehenkel, A. and Louppe, G. (June 2021). *Diffusion Priors In Variational Autoencoders*.

Wei, X.-x. and Stocker, A. A. (2012). “Efficient Coding Provides a Direct Link between Prior and Likelihood in Perceptual Bayesian Inference”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.

Wei, X.-X. and Stocker, A. A. (Oct. 2015). "A Bayesian Observer Model Constrained by Efficient Coding Can Explain 'anti-Bayesian' Percepts". In: *Nature Neuroscience* 18.10, pp. 1509–1517. ISSN: 1546-1726. DOI: 10.1038/nn.4105.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (June 2002). "Motion Illusions as Optimal Percepts". In: *Nature Neuroscience* 5.6, pp. 598–604. ISSN: 1546-1726. DOI: 10.1038/nn0602-858.

Whiteley, L. and Sahani, M. (2012). "Attention in a Bayesian Framework". In: *Frontiers in Human Neuroscience* 6.

Wilson, K. G. (Oct. 1975). "The Renormalization Group: Critical Phenomena and the Kondo Problem". In: *Reviews of Modern Physics* 47.4, pp. 773–840. DOI: 10.1103/RevModPhys.47.773.

Womelsdorf, T., Valiante, T. A., Sahin, N. T., Miller, K. J., and Tiesinga, P. (Aug. 2014). "Dynamic Circuit Motifs Underlying Rhythmic Gain Control, Gating and Integration". In: *Nature Neuroscience* 17.8, pp. 1031–1039.

Wyner, A. and Ziv, J. (Jan. 1976). "The Rate-Distortion Function for Source Coding with Side Information at the Decoder". In: *IEEE Transactions on Information Theory* 22.1, pp. 1–10. ISSN: 1557-9654. DOI: 10.1109/TIT.1976.1055508.

Xu, P., Cheung, J. C. K., and Cao, Y. (Aug. 2020). *On Variational Learning of Controllable Representations for Text without Supervision*. DOI: 10.48550/arXiv.1905.11975.

Xu, X., Collins, C. E., Khaytin, I., Kaas, J. H., and Casagrande, V. A. (Nov. 2006). "Unequal Representation of Cardinal vs. Oblique Orientations in the Middle Temporal Visual Area". In: *Proceedings of the National Academy of Sciences* 103.46, pp. 17490–17495. DOI: 10.1073/pnas.0608502103.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (June 2014). "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex". In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1403112111.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (Oct. 2023a). *Diffusion Models: A Comprehensive Survey of Methods and Applications*.

Yang, T., Lan, C., Lu, Y., and zheng, N. (June 2024). *Diffusion Model with Cross Attention as an Inductive Bias for Disentanglement*. DOI: 10.48550/arXiv.2402.09712.

Yang, T., Wang, Y., Lu, Y., and Zheng, N. (Nov. 2023b). "DisDiff: Unsupervised Disentanglement of Diffusion Probabilistic Models". In: *Thirty-Seventh Conference on Neural Information Processing Systems*.

Yeats, E., Liu, F., Womble, D., and Li, H. (2022). "NashAE: Disentangling Representations Through Adversarial Covariance Minimization". In: *Computer Vision – ECCV 2022*. Ed. by Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. Vol. 13687. Cham: Springer Nature Switzerland, pp. 36–51. ISBN: 978-3-031-19811-3 978-3-031-19812-0. DOI: 10.1007/978-3-031-19812-0_3.

Yu, J., Zhu, H., Jiang, L., Loy, C. C., Cai, W., and Wu, W. (Mar. 2023). *CelebV-Text: A Large-Scale Facial Text-Video Dataset*. arXiv:2303.14717 [cs]. DOI: 10.48550/arXiv.2303.14717.

Yu, L., Shi, X., Kong, X., Jia, T., and Steeg, G. V. (Nov. 2025). *MMG: Mutual Information Estimation via the MMSE Gap in Diffusion*. DOI: 10.48550/arXiv.2509.20609.

Zemel, R. and Dayan, P. (1998). "Distributional Population Codes and Multiple Motion Models". In: *Advances in Neural Information Processing Systems*. Vol. 11. MIT Press.

Zhang, J., Bai, J., Lin, C., Wang, Y., and Rong, W. (Dec. 2022a). "Improving Variational Autoencoders with Density Gap-based Regularization". In: *Advances in Neural Information Processing Systems* 35, pp. 19470–19483.

Zhang, Z., Zhao, Z., and Lin, Z. (Dec. 2022b). "Unsupervised Representation Learning from Pretrained Diffusion Probabilistic Models". en. In: *Advances in Neural Information Processing Systems* 35, pp. 22117–22130.

Zhao, S., Song, J., and Ermon, S. (May 2018). *InfoVAE: Information Maximizing Variational Autoencoders*.

Zhu, S. C., Wu, Y., and Mumford, D. (Mar. 1998). “Filters, Random Fields and Maximum Entropy (FRAME): Towards a Unified Theory for Texture Modeling”. en. In: *International Journal of Computer Vision* 27.2, pp. 107–126. ISSN: 1573-1405. DOI: 10.1023/A:1007925832420.