Inferring response times of perceptual decisions with Poisson variational autoencoders

Hayden R. Johnson[1,2] **Anastasia N. Krouglova**[1,2] **Hadi Vafaii**[4]
Jacob L. Yates[3,4 †] **Pedro J. Gonçalves**[1,2 †]

[1]VIB–Neuroelectronics Research Flanders, Belgium [2]KU Leuven, Belgium
[3]Herbert Wertheim School of Optometry & Vision Science, UC Berkeley, USA
[4]Redwood Center for Theoretical Neuroscience, UC Berkeley, USA
hayden.johnson@vib.be

Abstract

Many properties of perceptual decision making are well-modeled by deep neural networks. However, such architectures typically treat decisions as instantaneous readouts, overlooking the temporal dynamics of the decision process. We present an image-computable model of perceptual decision making in which choices and response times arise from efficient sensory encoding and Bayesian decoding of neural spiking activity. We use a Poisson variational autoencoder to learn unsupervised representations of visual stimuli in a population of rate-coded neurons, modeled as independent homogeneous Poisson processes. A task-optimized decoder then continually infers an approximate posterior over actions conditioned on incoming spiking activity. Combining these components with an entropy-based stopping rule yields a principled and image-computable model of perceptual decisions capable of generating trial-by-trial patterns of choices and response times. Applied to MNIST digit classification, the model reproduces key empirical signatures of perceptual decision making, including stochastic variability, right-skewed response time distributions, logarithmic scaling of response times with the number of alternatives (Hick's law), and speed–accuracy trade-offs.

1 Introduction

A central task of the brain is to process sensory information in order to guide adaptive behavior, often studied through perceptual decision-making. This process is shaped by several biological constraints. Neurons communicate through discrete action potentials (or spikes), the nervous system is subject to noise at multiple levels [1], and operates under limited metabolic resources [2, 3]. Together, these factors impose fundamental limits on both the speed and accuracy of perceptual processing [4].

Deep neural networks have emerged as powerful models of the sensory cortex, capable of performing perceptual tasks of naturalistic complexity [5, 6]. While compelling, these models typically treat decisions as instantaneous readouts, ignoring the rich temporal dynamics of the decision process. As a result, standard deep neural networks cannot account for temporal properties of perceptual decision making such as response time distributions or the speed–accuracy trade-offs.

In contrast, evidence accumulation models (e.g., drift diffusion model [7]) explicitly capture the temporal dynamics of the decision process. These models reproduce key behavioral regularities, including stochastic variability, right-skewed response time distributions, and speed–accuracy trade-offs [8]. However, they generally operate on abstract decision variables without specifying how such

† Co-senior authors.

Workshop: Data on the Brain & Mind.
39th Conference on Neural Information Processing Systems (NeurIPS 2025).

variables are derived from sensory input. Consequently, they are not image-computable (i.e., they do not compute directly on visual stimuli) and therefore cannot perform image-level perceptual tasks.

Recently, a new class of image-computable decision models has sought to fill this gap by explicitly linking sensory encoding with decision dynamics [9–13]. These models seek to combine the representational capacity of deep neural networks with temporal accumulation of evidence, providing a functional grounding for the latent decision variable in terms of task-relevant information. Doing so enables models to perform complex tasks while accounting for temporal properties of the decision process. However, existing approaches are often directly fit to human response data, or introduce temporal integration mechanisms without a clear normative justification.

In this work, we introduce $\mathcal{P}$-VAE-RT, an image-computable model for perceptual decision-making, grounded in principles of efficient coding and Bayesian evidence accumulation. Sensory representations are modeled as resource-limited encoders optimized for information transmission in a population of rate-coded neurons. Decisions are then cast as Bayesian accumulation of noisy sensory evidence (via spikes) until a criterion is reached. This formulation allows the model to perform image-level tasks while reproducing key behavioral phenomena observed in perceptual decision making. By unifying efficient sensory coding with normative accumulation of evidence, our approach provides a framework for understanding perceptual decision-making as optimal inference under resource constraints [14].

Contributions: Overall, our work makes the following contributions: **(i)** We introduce an image-computable model of perceptual decision-making that integrates efficient sensory encoding with Bayesian evidence accumulation, providing a principled link between high-dimensional sensory input and temporal decision dynamics. **(ii)** We demonstrate our model's ability to perform a complex image-level task while predicting hallmark behavioral phenomena, including trial-to-trial variability, response time distributions, and adaptive speed–accuracy trade-offs.

2 Background

Bayesian neural decoding. Previous work demonstrated that several classic response-time phenomena can be unified from the perspective of information transmission. In particular, Christie et al. [15] considers response times (RT) as the duration required for a Bayesian decoder to identify a stimulus from a population of rate-coded neurons, each modeled as a homogeneous Poisson process.

Formally, consider a set of stimuli $\mathcal{X} = \{x_1, .., x_K\}$ where each stimulus $x \in \mathcal{X}$ is sampled from a prior distribution $p(x)$. The stimulus is encoded into vector firing rates

$$\boldsymbol{\lambda} = \rho_b + \gamma\, g(x), \qquad g : \mathcal{X} \to \mathbb{R}^K,$$

where g is a one-hot encoding, γ scales the encoding magnitude, and ρ_b is a uniform baseline firing rate. In other words, each stimulus excites a unique neuron, while all others remain at baseline activity. Spikes are then generated according to a set of homogeneous Poisson processes with these rates. The decoder is modeled as a Bayesian ideal observer that infers the stimulus from cumulative spike counts at time t, denoted $\mathbf{z}_t$, by updating the posterior

$$p(x \mid \mathbf{z}_t) = \frac{p(\mathbf{z}_t \mid x)\; p(x)}{p(\mathbf{z}_t)}.$$

Under this simple one-to-one encoding, the posterior can be computed in closed-form [16]. Decoder uncertainty is quantified by the entropy $\mathcal{H}(x \mid \mathbf{z}_t)$, which decreases as spikes accumulate. Response time is defined as the first-passage time when entropy drops below a decision threshold τ.

Despite its simplicity, this framework provides a unified account for a wide range of empirical findings, including the Hick–Hyman law [17, 18], the power law of practice [19], Stroop interference [20, 21], and speed–accuracy trade-offs [4, 22]. However, extending the model to complex tasks reveals two key limitations. First, the model assumes a one-to-one neuron–stimulus mapping, incommensurate with complex encoders required for high-dimensional stimuli (e.g., images). Second, the model fails to accommodate non-identity relationships between stimuli and actions, preventing it from capturing tasks with complex or continuous outputs. In this work, we address these limitations by generalizing the framework to support complex, nonlinear stimulus encoders and by developing a task-optimized decoder capable of supporting complex actions.

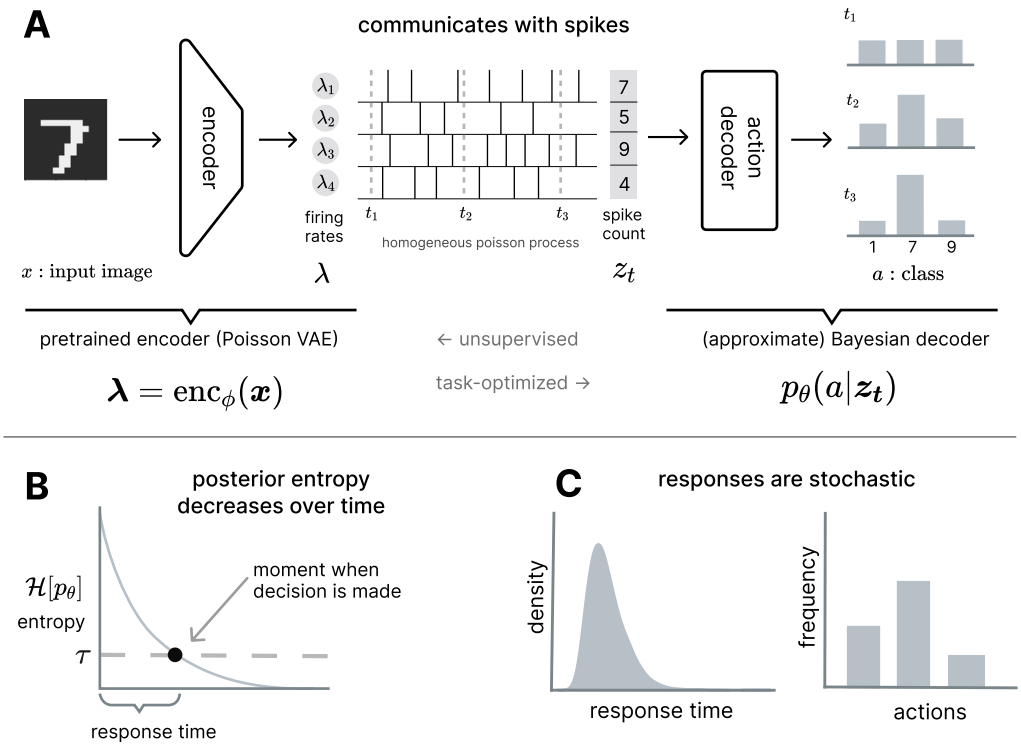

Figure 1: **(A)** $\mathcal{P}$-VAE-RT architecture. Input stimuli $\mathbf{x}$ are processed by a pretrained $\mathcal{P}$-VAE encoder, $\text{enc}_\phi(\mathbf{x})$, producing a vector of firing rates $\boldsymbol{\lambda}$. These rates generate spike trains via a set of homogeneous Poisson processes. Throughout the spike train, an approximate Bayesian decoder continually infers the posterior distribution, $p_\theta(\mathbf{a} \mid \mathbf{z}_t)$, over actions $\mathbf{a}$ based on the accumulated spike count $\mathbf{z}_t$. **(B)** Schematic of the entropy-based stopping rule. Posterior entropy $\mathcal{H}[p_\theta]$ decreases as spikes accumulate. Response times are modeled as the first passage time for the posterior to hit an entropy stopping threshold τ. **(C)** Schematic of response distributions. We generate response distributions from repeated simulation of actions and response times for a given stimulus.

Poisson variational autoencoder. Perception can be understood as the process of inferring the true state of the world from noisy and incomplete sensory measurements [23]. This perspective has directly inspired machine learning architectures such as the Helmholtz machine [24], and, more recently, variational autoencoders (VAEs; [25, 26]).

VAEs are increasingly being considered as computational models of perceptual inference in the brain. Evidence for this comes from several directions: (i) VAE representations align with the primate cortex across both ventral [27] and dorsal [28] streams; (ii) VAEs develop cortex-like topographic organization [29, 30]; and (iii) VAEs produce perceptual errors resembling those of humans [31]. Collectively, this suggests a significant degree of neural, organizational, and psychophysical alignment between VAEs and the brain.

However, standard Gaussian VAEs encode inputs into continuous latent variables that deviate from the discrete, spiking nature of biological neural codes. To address this, Vafaii et al. [32] introduced the Poisson VAE ($\mathcal{P}$-VAE), which encodes inputs into discrete spike-count variables. The $\mathcal{P}$-VAE is both rigorously grounded in the mathematics of variational inference, and is more brain-like in its representations.

Despite these advances, the $\mathcal{P}$-VAE remains a purely vision-based model. It has not yet been applied to scenarios where perceptual inference is used to guide behavior. In this work, we extend the $\mathcal{P}$-VAE framework to demonstrate that its encoding mechanism can generate spike trains that support perceptual decision-making.

3 Overview of the $\mathcal{P}$-VAE-RT

We construct our model ($\mathcal{P}$-VAE-RT) from two components, each trained independently. First, we train a Poisson variational autoencoder to learn an efficient unsupervised representation of stimuli as firing rates. These rates are then converted into spike trains via a set of homogeneous Poisson processes. The spike trains are then used as training data for a task-optimized decoder that learns to continually infer the posterior over actions, conditioned on incoming spiking activity. Combining these components with an entropy-based stopping rule yields a principled, image-computable model of perceptual decision-making that generates both choices and response times. In the following sections, we describe each component in detail and outline its role within the unified architecture.

3.1 Efficient, probabilistic encoding of stimuli

To efficiently encode visual stimuli into a vector of firing rates, we use a Poisson VAE ($\mathcal{P}$-VAE). The $\mathcal{P}$-VAE modifies a standard variational autoencoder by replacing the Gaussian prior and approximate posterior with Poisson distributions. This modification results in discrete, integer spike-count representations in the latent space of the model.

During inference, the $\mathcal{P}$-VAE encoder maps an input sample $\boldsymbol{x}$ to the rate parameters, $\boldsymbol{\lambda}(\boldsymbol{x}) = \mathrm{enc}_\phi(\boldsymbol{x})$, which are then used to construct a Poisson approximate posterior, from which spike counts $\boldsymbol{z} \sim \mathrm{Pois}(\boldsymbol{\lambda}(\boldsymbol{x}))$ are sampled. The decoder network then reconstructs the input, $\hat{\boldsymbol{x}} = \mathrm{dec}_\psi(\boldsymbol{z})$.

The $\mathcal{P}$-VAE is trained in an unsupervised manner, using the standard Evidence Lower BOund (ELBO) objective, which assumes the following form for a Gaussian likelihood and Poisson latents:

$$\mathcal{L}_{\mathrm{PVAE}} \;=\; \underbrace{\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}\mathrm{ois}(\boldsymbol{z};\boldsymbol{\lambda}(\boldsymbol{x}))}\Big[\|\boldsymbol{x}-\mathrm{dec}_\psi(\boldsymbol{z})\|_2^2\Big]}_{\text{Reconstruction term}} \;+\; \underbrace{\sum_{i=1}^{K} r_i f(\delta r_i)}_{\text{KL term}}\,, \tag{1}$$

where $\mathrm{dec}_\psi(\cdot)$ denotes the decoder network, K is the latent dimensionality, and $f(y) = 1-y+y\log y$; $\mathbf{r}$ is the prior over firing rates, and $\delta\mathbf{r} \coloneqq \boldsymbol{\lambda} \oslash \mathbf{r}$, where $\oslash$ denotes element-wise division.

A notable consequence of incorporating Poisson latents is that the Kullback–Leibler (KL) term in the ELBO objective resembles a metabolic cost (eq. (1)) [32]. Because of this mathematical result, the overall $\mathcal{P}$-VAE objective encourages the model to faithfully reconstruct inputs while minimizing spiking activity. This establishes a theoretical connection to sparse coding [33], which was verified empirically [32, 34].

In sum, the $\mathcal{P}$-VAE provides a biologically plausible, spiking, and sparsity-promoting perceptual model, which we use as a "visual cortex" in our decision-making task.

3.2 Task-optimized neural decoding

Following Christie et al. [15], we model response time as the time required for a Bayesian decoder to infer sensory information from spikes emitted by a population of rate-coded neurons. We extend this framework by allowing the decoder to interpret non-linear encodings and learn arbitrary mappings between stimuli and actions.

Concretely, we consider the problem of inferring the posterior $p(a \mid \boldsymbol{z}_t) \propto p(\boldsymbol{z}_t \mid a)\, p(a)$, where a denotes a perceptual judgment (e.g., a class label) pertaining to stimulus $\boldsymbol{x} \in \mathbb{R}^D$. $\boldsymbol{z}_t \in \mathbb{N}^K$ is the vector of cumulative spike counts up to time t from K homogeneous Poisson neurons with rates

$$\boldsymbol{\lambda} = \mathrm{enc}_\phi(\boldsymbol{x}), \qquad \mathrm{enc}_\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^K.$$

We place no restrictions on the complexity of the encoder $\mathrm{enc}(\cdot)$, or the relation between $\boldsymbol{x}$ and a. This generalization renders the likelihood $p(\boldsymbol{z}_t \mid a)$ intractable. We therefore seek to directly approximate the posterior with a learned decoder:

$$\mathrm{dec}_\theta := p_\theta(a \mid \boldsymbol{z}_t) \approx p(a \mid \boldsymbol{z}_t), \tag{2}$$

where θ are the neural network parameters that parameterize our approximate posterior. We emphasize that dec_θ is completely distinct from the decoder used to train our $\mathcal{P}$-VAE (which learns to reconstruct

the original stimuli). Rather, this network computes an evolving posterior over actions which may be a complex function of the original stimuli (e.g., the class of an image).

To train this decoder, we construct a dataset $\mathcal{D} = (\boldsymbol{z}_t, a)$. For each stimulus, we (i) generate spike trains from the image-driven rates $\boldsymbol{\lambda}$ using a homogeneous Poisson process, (ii) discretize the spike train into time bins to form a binary event matrix of size $K \times T$, and (iii) convert this into cumulative spike counts, where entry (i, t) is the total number of spikes emitted by neuron i up to time t. Each column of this cumulative count matrix, paired with its label a, serves as a training sample. Importantly, for a homogeneous Poisson process, the likelihood depends only on the number of events in an interval, not their precise timing. Thus, the cumulative spike count vector $\boldsymbol{z}_t$ is a sufficient statistic for posterior inference [35].

In the case of a discrete action space, as seen in the subsequent example, we use a single multi-layer perceptron trained with a softmax output and cross-entropy loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(a,\mathbf{z}_t)}[-\log\, p_\theta(a \mid \mathbf{z}_t)]. \tag{3}$$

In appendix A, we provide a proof that this objective converges to the true posterior under standard realizability assumptions. For continuous actions, we replace the softmax with a parametric density (e.g., Gaussian mixture model, normalizing flow) and train the decoder using the negative log-likelihood loss.

3.3 Generating responses from $\mathcal{P}$-VAE-RT

To construct $\mathcal{P}$-VAE-RT, we combine a pretrained encoder $\text{enc}_\phi(x)$, which maps high-dimensional stimuli into firing rates, with a task-optimized decoder $p_\theta(a \mid \mathbf{z}_t)$, that continually extracts task-relevant information from resulting spike trains (see Figure 1).

A simulation proceeds as follows: the input stimulus $\boldsymbol{x}$ is first encoded into firing rates $\boldsymbol{\lambda} = \text{enc}_\phi(\boldsymbol{x})$. Spike trains are then generated from a set of homogeneous Poisson processes parameterized by these rates. At each time $t \in [0, T]$, the decoder infers the posterior distribution $p_\theta(a \mid \mathbf{z}_t)$ and its associated entropy $\mathcal{H}[p_\theta]$. Response time is defined as the earliest t for which $\mathcal{H}[p_\theta] < \tau$, where τ is a fixed entropy threshold. The selected action is given by the maximum a posteriori (MAP) estimate of the decoder at that time.

It is worth noting that the model is not trained to mimic subjects' empirically observed behavior. Instead, it is trained directly to perform a task, subject to a few guiding principles: (i) stimuli should be encoded efficiently, subject to representational constraints, (ii) information should be transmitted in a biologically plausible manner (i.e., using spikes), and (iii) task-relevant information should be decoded optimally to support action. Together, these guiding principles provide a powerful framework for understanding how temporal properties of perceptual decision arise in neural systems under biological constraints.

4 Capturing psychophysical phenomena with MNIST

We apply $\mathcal{P}$-VAE-RT to the task of handwritten digit classification using the MNIST dataset and show that it captures several hallmark regularities of human perceptual decision-making. While this serves as a concrete test case, it also illustrates the model's capacity to perform a general class of perceptual tasks.

To evaluate behavior, we report response times in arbitrary units (AU). These can be mapped to real time (e.g., seconds) by specifying a time constant, which in practice can be estimated from empirical data. For the purposes of this section, however, we emphasize the model's ability to reproduce well-established psychophysical trends, which remain invariant under rescaling of the time axis.

Unless otherwise noted, we use an entropy stopping threshold of $\tau = 0.5$ (except when systematically varied to study the speed–accuracy trade-off) and a latent dimension of 128. Additional analyses on the effect of varying the latent dimensions are provided in appendix B and further details on the training procedure can be found in Appendix C.

Response distributions: Human response times are among the most extensively studied measures in behavioral psychology, providing a unique window into the processes that underlie perceptual decision-making. A key reason is that response time distributions, together with behavioral outcomes,

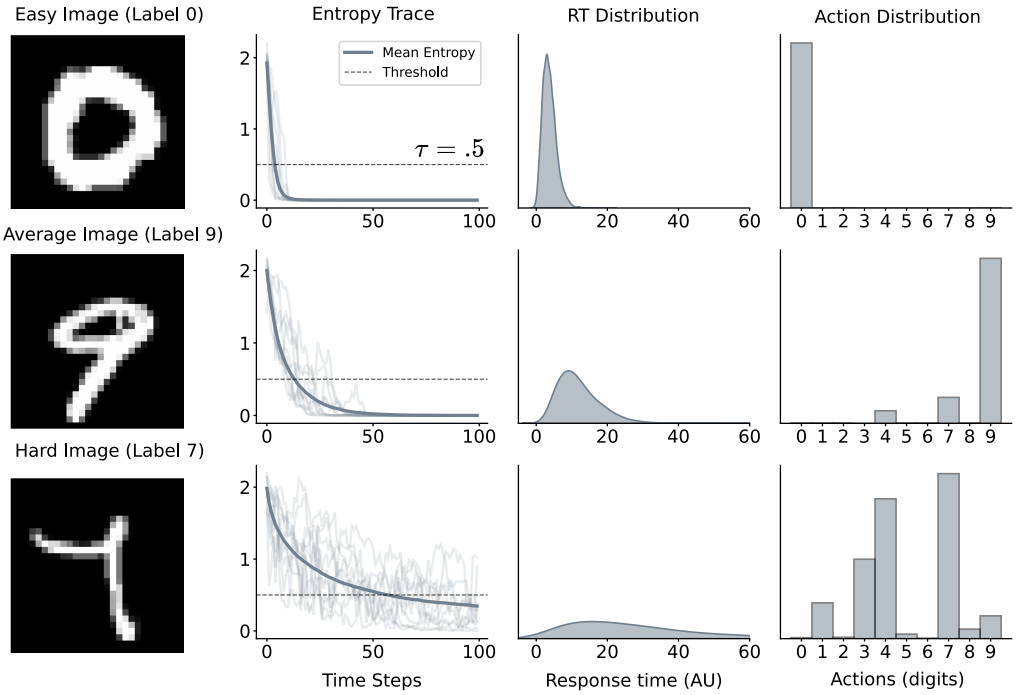

Figure 2: **Response distributions depend on task difficulty.** Response distributions for stimuli of varying difficulty. Each row corresponds to a digit at a given difficulty level, while each column highlights a property of the response distribution across repeated trials with a fixed stimulus. Easy stimuli are characterized by rapidly decreasing entropy, strongly skewed distributions, and low variance in the action distribution. Difficult stimuli, by contrast, exhibit slower entropy reduction, more symmetric distributions, and greater variability across actions.

reliably exhibit several regularities. We highlight three core phenomena and show that our model reproduces them: (i) decisions are stochastic, i.e., presenting the same stimulus can yield different responses and response times [36, 37]; (ii) response time distributions are typically right-skewed; and (iii) the degree of skewness decreases as task difficulty increases [8, 38].

Our model successfully reproduces all three effects (Figure 2). To illustrate this, we selected three images that differ in classification difficulty for human observers. The easy image is quickly and reliably classified as a 0, producing a strongly right-skewed response-time distribution with minimal variance across actions. The average difficulty image is labeled as a 9, but its white pixels also resemble a 4 or a 7, leading to slightly slower response times, reduced skew in the response-time distribution, and increased variance across actions. The hard image closely resembles both a 4 and a 7, creating ambiguity that generates a bimodal action distribution, substantially slower response times, and a broad response time distribution.

Reproducing the Hick–Hyman law. One of the earliest applications of information theory to human behavior is the Hick–Hyman law, which states that response time increases linearly with the information content of a stimulus. Given a uniform prior over stimuli, this implies a logarithmic relationship between response time and the number of possible choices. This effect was first demonstrated by William Hick in a classic experiment where subjects responded to a variable number of stimuli [17], and further developed by Ray Hyman [18]. Together, these studies suggest that response times scale with stimulus uncertainty, consistent with the idea that human decision-making proceeds at an approximately constant rate of information processing.

Our model reproduces this trend, when the decoder is trained to classify among varying numbers of digit classes. As shown in Figure 3, response times increase systematically with the number of alternatives, in agreement with the classic behavioral data of Hick [17].

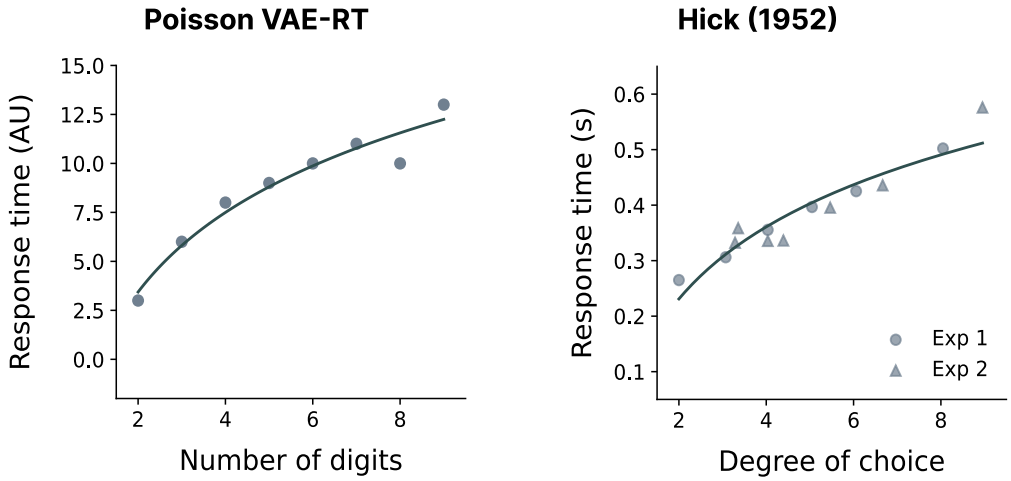

Figure 3: **Hick's Law**. (*Left*) Mean response time from the $\mathcal{P}$-VAE-RT where the decoder is trained to classify among a varying number of MNIST digits. RT increases monotonically with the number of alternatives with an approximately logarithmic trend. (*Right*) Human response times replotted from Hick [17], shown in seconds.

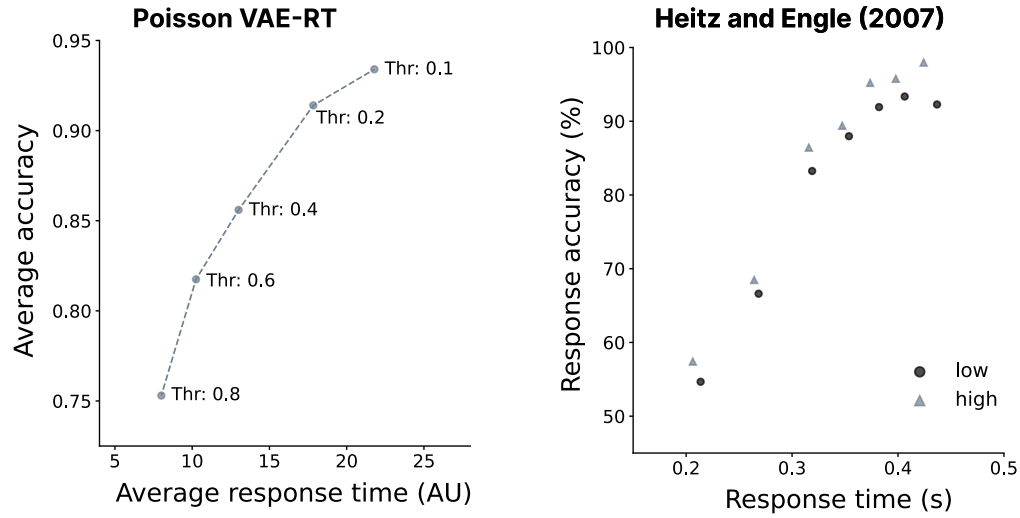

Figure 4: **Speed-accuracy trade-off.** (*Left*) Model accuracy versus average RT as the entropy threshold is swept ($\tau \in \{0.1, 0.2, 0.4, 0.6, 0.8\}$). Lower τ values produce longer RTs and higher accuracy. Points represent averages across images and trials. (*Right*) Human response time data from a working memory task under low versus high time pressure (Heitz & Engle [39]).

Speed-accuracy trade-off: Biological organisms must often act under variable time constraints, balancing the need for rapid responses against the benefits of accuracy. This balance is formalized by the speed–accuracy trade-off (SAT), whereby faster decisions typically yield higher error rates, while slower decisions improve accuracy at the cost of timeliness. The phenomenon has been documented across a wide range of species and tasks, including perceptual and cognitive decisions in humans [4, 40, 41], macaques [42], rats [43], and even insects such as bees [44].

Our model reproduces this trade-off, when the entropy stopping threshold τ is modulated. As illustrated in Figure 4, higher thresholds yield faster responses at the cost of accuracy, whereas lower thresholds prolong evidence accumulation and improve performance. This mirrors behavioral trends observed in working-memory experiments with varying levels of time pressure [39]. By linking stopping rules to decision performance, our framework provides a principled account of how organisms can tune their behavior to the demands of the environment.

5 Discussion

In this work, we introduced $\mathcal{P}$-VAE-RT, an image-computable model of perceptual decision-making grounded in the principles of efficient coding and Bayesian evidence accumulation. The model employs an unsupervised representation from a Poisson variational autoencoder to map high-dimensional stimuli into a population of rate-coded neurons, modeled as a set of independent homogeneous Poisson processes. These rates generate spike trains that drive a task-optimized decoder that computes an approximate Bayesian posterior over actions throughout the spike train. As spikes accumulate, posterior entropy decreases until reaching a preset threshold τ, at which point the model commits to an action and reports a response time.

We demonstrated the framework on the task of MNIST digit classification and showed that it reproduces classic psychophysical regularities. These include response variability, right-skewed response-time distributions, the logarithmic increase in response times with the number of alternatives (Hick's law), and speed–accuracy trade-offs.

We believe this approach makes several contributions to the study of perceptual decision-making. From a pragmatic perspective, it offers a highly flexible framework for modeling perceptual decisions with minimal constraints on stimulus complexity or the space of potential actions. We see this flexibility as an important step toward investigating decision-making in more complex and naturalistic contexts. At a conceptual level, our model provides a principled link between efficient sensory coding and evidence accumulation, highlighting how the variability and resource limitations of neural population activity can give rise to variability in behavior. From a broader perspective, this work underscores the value of response times as a meaningful axis for comparing artificial and biological neural networks, extending alignment efforts beyond static measures such as accuracy or representational similarity.

Limitations and future work. While we view these results as promising, several limitations remain to be addressed in future work. First, we have not yet performed a direct quantitative comparison to human behavioral data, an essential step in assessing the model's explanatory power. Applying the framework to empirical datasets requires estimating several free parameters, including the time constant, entropy threshold, and neural population size. Future work should explore how these parameters can be constrained by normative principles or efficiently fit to behavioral measurements.

Second, the current formulation assumes a homogeneous Poisson population, an idealization that abstracts away richer neural dynamics. In particular, inhomogeneous Poisson processes are more biologically realistic but also invalidate the property that cumulative spike counts are sufficient statistics for posterior inference. Addressing this would likely require more expressive decoding architectures, for example, incorporating recurrence or state-space formulations.

Third, the present architecture is restricted to perceptual decisions on static images and thus cannot capture tasks with temporally varying stimuli (e.g., the random-dot motion paradigm). Extending the model to handle dynamic evidence would broaden its scope to a wider range of decision-making contexts and allow comparison to several classic experiments connecting neural and behavioral data.

Conclusion. Our findings reveal that image-computable models can capture key dynamics of perceptual decision making by linking neural activity to choices and response times. This suggests a promising direction toward more expressive models of biological neural computation.

Acknowledgments and Funding

We thank Thomas Christie for his thoughtful comments on the manuscript and for the conversations that helped shape and develop this line of work. We also thank Jonathan Pillow, Chris Summerfield, Fabian Sinz, Srini Turaga, Leyla Isik, and Janne Lappalainen for their valuable feedback and discussions during the early development of this project at the Cajal NeuroAI summer course. Hayden R. Johnson was supported by an FWO grant (G053624N). Anastasia N. Krouglova was supported by an FWO grant (G097022N).

References

[1] A Aldo Faisal et al. "Noise in the nervous system". In: *Nature reviews neuroscience* 9.4 (2008), pp. 292–303.

[2] Simon B Laughlin. "Energy as a constraint on the coding and processing of sensory information". In: *Current opinion in neurobiology* 11.4 (2001), pp. 475–480.

[3] James V Stone. "Principles of neural information theory". In: *Computational Neuroscience and Metabolic Efficiency* (2018).

[4] Richard P Heitz. "The speed-accuracy tradeoff: history, physiology, methodology, and behavior". In: *Frontiers in neuroscience* 8 (2014), p. 150.

[5] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[6] Alexander JE Kell et al. "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy". In: *Neuron* 98.3 (2018), pp. 630–644.

[7] Roger Ratcliff. "A theory of memory retrieval." In: *Psychological review* 85.2 (1978), p. 59.

[8] Nathan J Evans, Eric-Jan Wagenmakers, et al. "Evidence accumulation models: Current limitations and future directions". In: *The Quantitative Methods for Psychology* 16.2 (2020), pp. 73–90.

[9] Bo Chen and Pietro Perona. "Seeing into darkness: Scotopic visual recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3826–3835.

[10] Lore Goetschalckx et al. "Computing a human-like reaction time metric from stable recurrent vision models". In: *Advances in neural information processing systems* 36 (2023), pp. 14338–14365.

[11] Yu-Ang Cheng et al. "RTify: Aligning deep neural networks with human behavioral decisions". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 130485–130510.

[12] Farshad Rafiei et al. "The neural network RTNet exhibits the signatures of human perceptual decision-making". In: *Nature Human Behaviour* 8.9 (2024), pp. 1752–1770. DOI: 10.1038/s41562-024-01914-8.

[13] Paul I Jaffe et al. "An image-computable model of speeded decision-making". In: *eLife* 13 (2025), RP98351.

[14] Rahul Bhui et al. "Resource-rational decision making". In: *Current Opinion in Behavioral Sciences* 41 (2021), pp. 15–21.

[15] S Thomas Christie et al. "Information-Theoretic Neural Decoding Reproduces Several Laws of Human Behavior". In: *Open Mind* 7 (2023), pp. 675–690.

[16] Scott Thomas Christie. *Information-theoretic bounded rationality: timing laws and cognitive costs emerge from rational bounds on information coding and transmission*. University of Minnesota, 2019.

[17] William E Hick. "On the rate of gain of information". In: *Quarterly Journal of experimental psychology* 4.1 (1952), pp. 11–26.

[18] Ray Hyman. "Stimulus information as a determinant of reaction time." In: *Journal of experimental psychology* 45.3 (1953), p. 188.

[19] Allen Newell and Paul S Rosenbloom. "Mechanisms of skill acquisition and the law of practice". In: *Cognitive skills and their acquisition*. Psychology Press, 2013, pp. 1–55.

[20] Colin M MacLeod. "Half a century of research on the Stroop effect: an integrative review." In: *Psychological bulletin* 109.2 (1991), p. 163.

[21] J Ridley Stroop. "Studies of interference in serial verbal reactions." In: *Journal of experimental psychology* 18.6 (1935), p. 643.

[22] Wayne A Wickelgren. "Speed-accuracy tradeoff and information processing dynamics". In: *Acta psychologica* 41.1 (1977), pp. 67–85.

[23] Hermann Von Helmholtz. *Handbuch der physiologischen Optik*. Vol. 9. L. Voss, 1867.

[24] Peter Dayan et al. "The helmholtz machine". In: *Neural computation* 7.5 (1995), pp. 889–904.

[25] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[26] Danilo Jimenez Rezende et al. "Stochastic backpropagation and approximate inference in deep generative models". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1278–1286. URL:

[27] Irina Higgins et al. "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons". In: *Nature Communications* 12.1 (2021), p. 6456. DOI: 10.1038/s41467-021-26751-5.

[28] Hadi Vafaii et al. "Hierarchical VAEs provide a normative account of motion processing in the primate brain". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL:

[29] T. Anderson Keller et al. "Modeling Category-Selective Cortical Regions with Topographic Variational Autoencoders". In: *SVRHM 2021 Workshop @ NeurIPS*. 2021. URL:

[30] T. Anderson Keller and Max Welling. "Topographic VAEs learn Equivariant Capsules". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 28585–28597. URL:

[31] Katherine R Storrs et al. "Unsupervised learning predicts human perception and misperception of gloss". In: *Nature Human Behaviour* 5.10 (2021), pp. 1402–1417. DOI: 10.1038/s41562-021-01097-6.

[32] Hadi Vafaii et al. "Poisson variational autoencoder". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 44871–44906.

[33] Bruno A Olshausen and David J Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583 (1996), pp. 607–609. DOI: 10.1038/381607a0.

[34] Hadi Vafaii et al. *Brain-like variational inference*. 2025. arXiv: 2410.19315 [cs.AI].

[35] John Frank Charles Kingman. *Poisson processes*. Vol. 3. Clarendon Press, 1992.

[36] Jeffrey M Beck et al. "Not noisy, just wrong: the role of suboptimal inference in behavioral variability". In: *Neuron* 74.1 (2012), pp. 30–39.

[37] Alfonso Renart and Christian K Machens. "Variability in neural activity and behavior". In: *Current opinion in neurobiology* 25 (2014), pp. 211–220.

[38] Birte U Forstmann et al. "Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions". In: *Annual review of psychology* 67.1 (2016), pp. 641–666.

[39] Richard P Heitz and Randall W Engle. "Focusing the spotlight: individual differences in visual attention control." In: *Journal of Experimental Psychology: General* 136.2 (2007), p. 217.

[40] Rafal Bogacz et al. "Do humans produce the speed–accuracy trade-off that maximizes reward rate?" In: *Quarterly journal of experimental psychology* 63.5 (2010), pp. 863–891.

[41] Jan Drugowitsch et al. "Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making". In: *elife* 4 (2015), e06678.

[42] Timothy Hanks et al. "A neural mechanism of speed-accuracy tradeoff in macaque area LIP". In: *Elife* 3 (2014), e02260.

[43] André G Mendonça et al. "The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs". In: *Nature communications* 11.1 (2020), p. 2757.

[44] Lars Chittka et al. "Bees trade off foraging speed for accuracy". In: *Nature* 424.6947 (2003), pp. 388–388.

[45] Edmund T Rolls and Martin J Tovee. "Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex". In: *Journal of neurophysiology* 73.2 (1995), pp. 713–726.

A Proof of convergence

Let $\{(a_i, \mathbf{z}_{t,i})\}_{i=1}^N$ be i.i.d. samples from the true joint $p(a)\,p(\mathbf{z}_t \mid a)$, where $p(a)$ is the (fixed) prior over a and $p(\mathbf{z}_t \mid a)$ is the likelihood of a. We fit a conditional model $p_\theta(a \mid \mathbf{z}_t)$ by maximum likelihood.

Maximizing the product $\prod_{i=1}^N p_\theta(a_i \mid \mathbf{z}_{t,i})$ with respect to θ is equivalent to maximizing the average log-likelihood

$$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(a_i \mid \mathbf{z}_{t,i}). \tag{4}$$

By the strong law of large numbers, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(a_i \mid \mathbf{z}_{t,i}) \xrightarrow{\text{a.s.}} \mathbb{E}_{p(a)\,p(z_t|a)}[\log p_\theta(a \mid \mathbf{z}_t)]\,. \tag{5}$$

Let $p(\mathbf{z}_t) = \int p(a)\,p(\mathbf{z}_t \mid a)\,da$ be the induced marginal of $\mathbf{z}_t$. Consider the KL divergence between the two joint distributions on $(a, \mathbf{z}_t)$:

$$D_{\mathrm{KL}}\Big(p(a)\,p(\mathbf{z}_t \mid a)\,\Big\|\,p(\mathbf{z}_t)\,p_\theta(a \mid \mathbf{z}_t)\Big) = \mathbb{E}_{p(a)\,p(\mathbf{z}_t|a)}\left[\log \frac{p(a)\,p(\mathbf{z}_t \mid a)}{p(\mathbf{z}_t)\,p_\theta(a \mid \mathbf{z}_t)}\right]. \tag{6}$$

Rearranging (6) yields

$$\mathbb{E}_{p(a)\,p(\mathbf{z}_t|a)}[\log p_\theta(a \mid \mathbf{z}_t)] = -\,D_{\mathrm{KL}}\Big(p(a)\,p(\mathbf{z}_t \mid a)\,\Big\|\,p(\mathbf{z}_t)\,p_\theta(a \mid \mathbf{z}_t)\Big) + \text{const.}, \tag{7}$$

where the constant does not depend on θ. Thus, maximizing the expected log-likelihood is equivalent to minimizing the KL divergence in (6).

The KL divergence in (6) is minimized (to 0) if and only if the two joint distributions are equal almost everywhere:

$$p(a)\,p(\mathbf{z}_t \mid a) \;=\; p(z_t)\,p_\theta(a \mid z_t) \qquad \text{a.e.} \tag{8}$$

Solving (8) for $p_\theta(a \mid \mathbf{z}_t)$ gives

$$p_\theta(a \mid z_t) \;=\; \frac{p(a)\,p(\mathbf{z}_t \mid a)}{p(\mathbf{z}_t)} \;=\; p(a \mid \mathbf{z}_t). \tag{9}$$

Combining (5) and (7)–(9), we conclude that any maximizer $\theta^\star$ of the limiting objective satisfies $p_{\theta^\star}(a \mid \mathbf{z}_t) = p(a \mid \mathbf{z}_t)$ almost everywhere. Consequently, under standard regularity (realizability of the true posterior within the model class and optimization that attains the global maximum), the MLE estimator satisfies

$$p_{\hat{\theta}_N}(a \mid \mathbf{z}_t) \;\longrightarrow\; p(a \mid \mathbf{z}_t) \quad \text{as } N \to \infty.$$

□

B Varying the latent dimension

The latent dimensionality is a core architectural hyperparameter that is fixed prior to training and constrains the model's representational capacity. In our main experiments, we set the latent dimensionality to 128, following Vafaii et al. [32]. Here, we systematically vary this dimensionality to examine its influence on three aspects of model behavior, which we investigate in turn:

(i) the quality of $\mathcal{P}$-VAE reconstructions,

(ii) the rate of evidence accumulation, and

(iii) the sparsity of the learned latent representation.

B.1 Reconstruction of the MNIST dataset

To assess how latent dimensionality affects reconstruction fidelity, we varied the number of latent dimensions in $\mathcal{P}$-VAE on a logarithmic scale from 2 to 512, holding other training hyperparameters constant. As expected, reconstruction quality improves monotonically with dimensionality (Figure 5), reflecting increased representational capacity. At very low dimensions, the bottleneck severely restricts the encoder's ability to preserve image structure, whereas higher-dimensional latents enable accurate reconstructions.

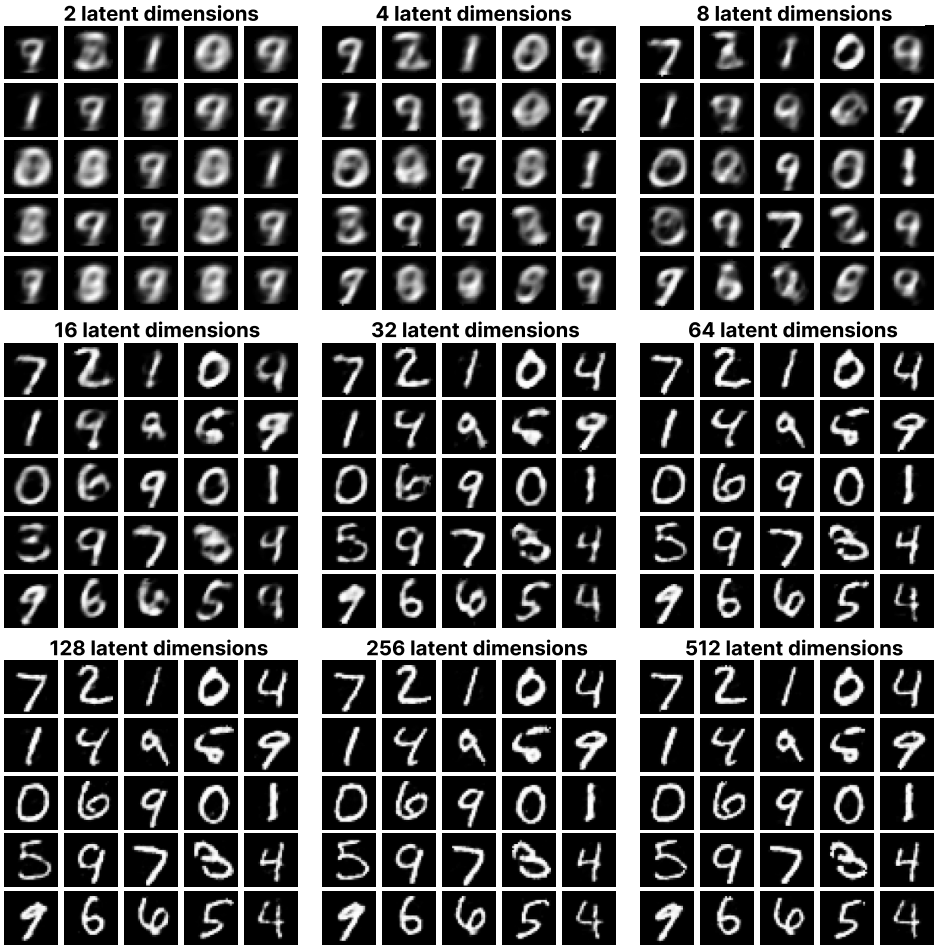

Figure 5: **Reconstruction quality of the MNIST dataset.** The reconstruction of 25 MNIST images from the test set for $\mathcal{P}$-VAE with varying latent dimensions. Reconstruction improves as latent dimension increases, indicating increased capacity.

B.2 Rate of evidence accumulation

We next analyzed how latent dimensionality influences the rate of evidence accumulation. We quantified this effect using the mean posterior entropy $\mathcal{H}[p_\theta]$ across images and trials in $\mathcal{P}$-VAE-RT. Posterior entropy decreases stochastically within trials and, when averaged, yields smooth traces characterizing the accumulation rate. Consistent with Christie et al. [15], higher latent dimensionality accelerates entropy reduction – interpreted as increased signal power – while lower-dimensional spaces constrain evidence accumulation, producing slower response times (Figure 6).

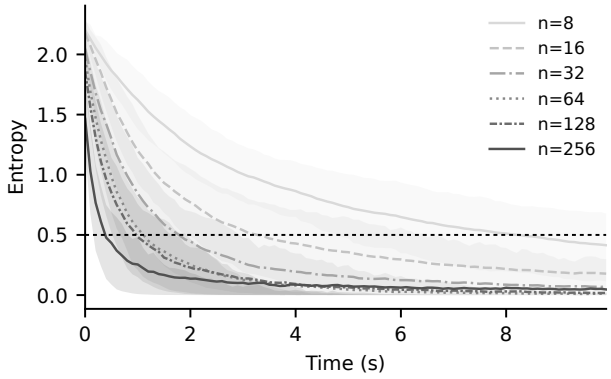

Figure 6: Mean entropy traces for n-dimensional latent representations in $\mathcal{P}$-VAE-RT across 1000 images and 100 trials. Decreasing the dimensionality of the latent reduces the rate of evidence accumulation, leading to slower response times.

B.3 Sparsity of learned codes

Inspired by the relation to sparse coding, we examine the sparsity of the learned representation as a function of latent dimensionality. First, sparsity was quantified as the proportion of neurons whose responses exceeded a threshold (over 1000 images):

$$\text{sparsity } \% = \frac{\#\{\text{neurons with response} > \Psi\}}{\text{total neurons}} \times 100, \tag{10}$$

where Ψ is a tunable threshold. Across thresholds, sparsity consistently increased with latent dimensionality – an example for $\Psi = 1$ is shown in Figure 7 (left).

To validate our results, we additionally quantified the lifetime sparsity using the Treves–Rolls sparseness index [45], which captures the distributional shape of firing rates r:

$$\alpha = \frac{\left(\frac{1}{N}\sum_i r_i\right)^2}{\frac{1}{N}\sum_i r_i^2}, \qquad \text{TR sparsity } \% = \frac{1-\alpha}{1-1/N}, \tag{11}$$

where N is the number of images and r_i denotes the mean firing rate of a neuron in response to image i.

Following the analysis in Christie et al. [15], we varied the number of latent dimensions (i.e., neurons) from 2 to 2048 on a logarithmic scale. Both metrics revealed the same overall trend: sparsity increases with latent dimensionality.

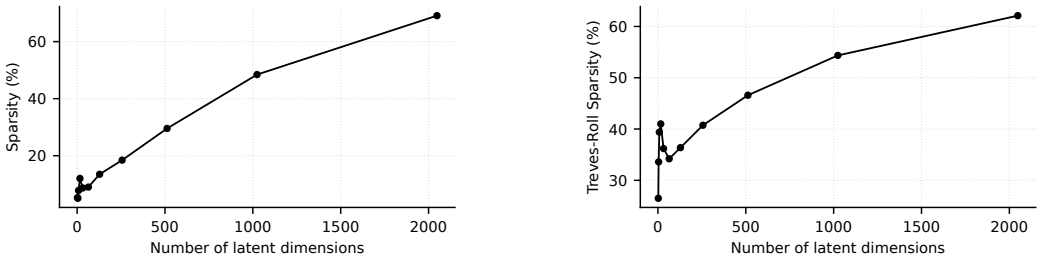

Figure 7: **Sparsity over latent dimensions.** (*Left*) Sparsity as a proportion of neurons exceeding threshold $\Psi = 1$. (*Right*) Treves-Rolls Sparseness index.

C Training details

All training can be completed in a few hours under modest compute resources.

C.1 Software and hardware

Programming language: Python 3.10. Core ML package: PyTorch 2.5.1. Analysis packages: NumPy, Matplotlib, Pandas, scikit-learn. Hardware: 2024 MacBook Pro with Apple M4 Max.

C.2 Poisson VAE

Dataset: MNIST; 28×28 grayscale images flattened.

Architecture:
Latent dimensionality: 128
Encoder, $\text{enc}_\phi(x)$: 784 $\to$ 128 (linear).
Decoder, $\text{dec}_\psi(z)$: 128 $\to$ 784 (linear) with Gaussian output.

Training: ELBO objective; Adam optimizer with learning rate 1e-3; 50 epochs.

C.3 Task-optimized decoder

Spike processing: For each stimulus, we generate spike trains from the image-driven rates $\lambda = \text{enc}_\phi(x)$ using independent homogeneous Poisson processes. The spike trains are discretized into $T = 100$ time bins to form a binary event matrix (neurons $\times$ time). This is further converted into cumulative spike counts, where entry (i, t) denotes the total number of spikes emitted by neuron i up to time t. Each cumulative count vector $\boldsymbol{z}_t$, paired with its label a, serves as a supervised training sample.

Architecture: multilayer perceptron 128 $\to$ 64 $\to$ 32 $\to$ output size; ReLU activations in hidden layers; final layer produces class logits.

Training: cross-entropy loss; Adam optimizer with learning rate 1e-3; batch size 256; 100 epochs.