Extended Abstract Track

Modeling dynamic neural activity by combining naturalistic video stimuli and stimulus-independent latent factors

Editors: List of editors' names

Abstract

Understanding how the brain processes dynamic natural stimuli remains a fundamental challenge in neuroscience. Current dynamic neural encoding models either take stimuli as input but ignore shared variability in neural responses, or they model this variability by deriving latent embeddings from neural responses or behavior while ignoring the visual input. To address this gap, we propose a probabilistic model that incorporates video inputs along with stimulus-independent latent factors to capture variability in neuronal responses, predicting a joint distribution for the entire population. After training and testing our model on mouse V1 neuronal responses, we found that it outperforms video-only models in terms of log-likelihood and achieves further improvements when conditioned on responses from other neurons. Furthermore, we find that the learned latent factors strongly correlate with mouse behavior, although the model was trained without behavior data.

Keywords: Dynamics of neural representations; probabilistic neural predictive models primary visual cortex of mice; latent variable models; stimulus-independent variability

1. Introduction

Neural activity is influenced both by sensory stimuli and internal, stimulus-independent fluctuations (Stringer et al., 2019; Niell and Stryker, 2010; Reimer et al., 2016), which are often ignored by stimulus-based models. Existing approaches to neural response prediction largely fall into three main categories: (1) deterministic models that predict neural activity from visual stimuli, but neglect uncertainty and variability in responses (Wang et al., 2023a; Turishcheva et al., 2023; Sinz et al., 2018; Vystrčilová et al., 2024; Höfling et al., 2022); (2) probabilistic models that derive latent embeddings from neuronal responses to predict behavior but do not account for external sensory inputs (Schneider et al., 2023; Gokcen et al., 2022; Sussillo et al., 2016; Yu et al., 2009); and (3) models that combine task inputs and a subset of neural activity to predict the responses of conjugate neurons (Zhou and Wei, 2020; Kim et al., 2023).

To the best of our knowledge, few works exist that model the correlated variability of large neuronal populations, that also take visual stimuli as input (Bashiri et al., 2021). Specifically, none of them are designed for dynamic video-based stimuli. The flow based approach of Bashiri et al. (2021) can be computationally too expensive for modeling temporal dependencies in the latent variables. This paper addresses this gap by proposing a predictive model of dynamic neural activity as a function of naturalistic video stimuli and a temporally varying latent variable that captures correlated, stimulus-independent neural variability. We demonstrate that this model improves predictive accuracy of mouse V1 responses and implicitly learns latent variables that are correlated with behavior, which were not provided to the model during training.

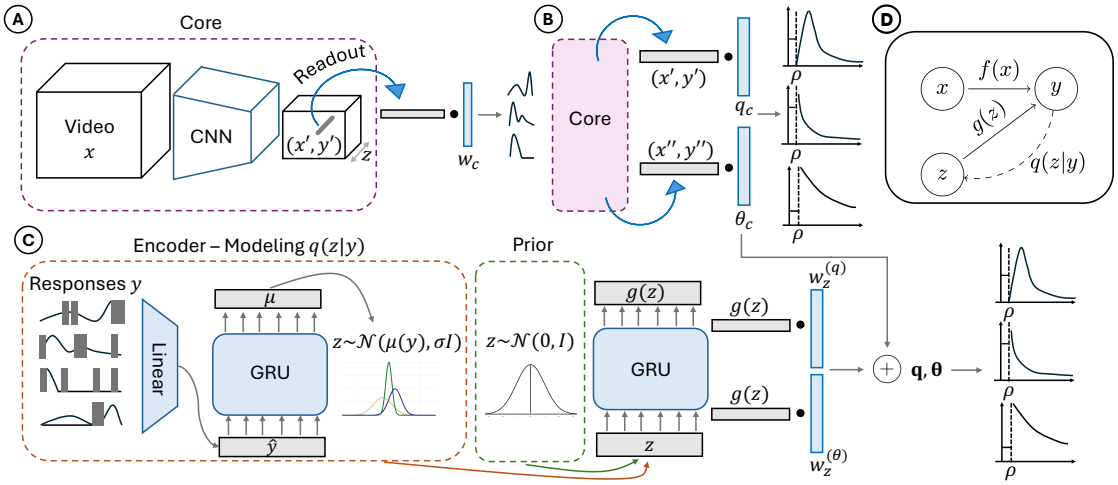

Figure 1: **A** 3D convolutional neural network-based core that extracts features of the video-input $\mathbf{x}$. For predicting a neuron's response, we take a dot product between the neuron feature vector $\mathbf{w}_c$ and core features across channel dimension z at position $(x^{'}, y^{'})$, determined by a Gaussian readout. **B** Two readouts are used to extract feature vectors at positions $(x^{'}, y^{'}), (x^{''}, y^{''})$. We calculate the parameters of the response's ZIG-distributions by computing a dot product between those vectors and the learned neuron feature vectors $\mathbf{w}_c^{(q)}, \mathbf{w}_c^{(\theta)}$. **C** The encoder processes masked neuronal responses $\mathbf{y}$, reduces their dimensionality before inputting them into a GRU to compute the posterior $q(\mathbf{z}|\mathbf{y}) \sim \mathcal{N}(\mu(\mathbf{y}), \sigma\mathbf{I})$. The decoder samples $\mathbf{z}$ from $q(\mathbf{z}|\mathbf{y})$ or $p(\mathbf{z})$, computes the dot products $g(\mathbf{z}) \cdot \mathbf{w}_z^{(q,\theta)}$, $\mathbf{w}_z^{(q,\theta)}$ are neuron specific learnt vectors, and adds them to **B**'s output. **D** Directed graphical model of video, latent and response variables. $f(\mathbf{x})$ corresponds to **B**, the dependency between $\mathbf{z}, \mathbf{y}$ is described by **C**. The dashed line indicates the variational approximation.

2. Models

Our model predicts time-varying neuronal responses (2-photon calcium traces) $\mathbf{y} \in \mathbb{R}^{N \times T}$ to a video stimulus $\mathbf{x} \in \mathbb{R}^{W \times H \times T}$, where N is the number of neurons, T the number of time points, and W and H are the width and height of one frame. The prediction additionally depends on a dynamic, stimulus-independent latent factor $\mathbf{z} \in \mathbb{R}^{k \times T}$ with latent dimension $k << N$ (Fig 1). Combining ideas from Bashiri et al. (2021) and Zhu et al. (2022), we model the distribution of neuronal responses conditioned on the stimulus and the latent factor as a Zero-Inflated-Gamma (ZIG) distribution (Wei et al., 2019):

$$p_{\text{ZIG}}\big(\mathbf{y}|\mathbf{x}, \mathbf{z}; \psi\big) = \prod_{it}^{N,T} \left(\big(1 - q_{it}(\mathbf{x}, \mathbf{z}; \psi)\big) \frac{1}{\rho} + q_{it}(\mathbf{x}, \mathbf{z}; \psi) \frac{y_{it}^{\kappa_i - 1}}{\Gamma(\kappa_i)\theta_{it}(\mathbf{x}, \mathbf{z}; \psi)^{\kappa_i}} \right), \tag{1}$$

where ρ is the mixture separation with no overlap, κ_i the shape parameter per neuron i, q_{it} the nonzero response probability, and θ_{it} the scale parameter for each neuron at time-step t. $\boldsymbol{\theta}$ and $\mathbf{q}$ are functions of the video-stimulus, the ZIG-model's parameters ψ and the latent factor while k_i is fitted once per neuron:

$$q_{it}(\mathbf{x}, \mathbf{z}, \psi) = \text{sigmoid}\left(f_{it}^{(q)}(\mathbf{x}; \psi) + \mathbf{w}_i^{(q)} \cdot g(\mathbf{z}_t; \psi) \right) \tag{2}$$

$$\theta_{it}(\mathbf{x}, \mathbf{z}, \psi) = \mathrm{ELU}\left(f_{it}^{(\theta)}(\mathbf{x}; \psi) + \mathbf{w}_i^{(\theta)} \cdot g(\mathbf{z}_t; \psi)\right) + 1. \tag{3}$$

We model $f_{it}^{(q)}$ and $f_{it}^{(\theta)}$ using a typical core-readout architecture (Fig 1A) where the core extracts features of the video-input, and the readout maps the relevant features from the core-output to the individual neurons (Höfling et al. (2022); Turishcheva et al. (2023); Vystrčilová et al. (2024); Sinz et al. (2018); Wang et al. (2023a)). We additionally introduce g as a one-layer recurrent neural network with gated recurrent units to capture correlations in the latents across time. The stimulus-independent latent factor is modeled with a prior distribution as a standard Gaussian independently across time $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$. We model an approximate posterior (Fig. 1C) as a Gaussian with the mean as function of the responses $\mathbf{y}$, the encoder parameters ϕ and an independent variance: $q(\mathbf{z}|\mathbf{y}; \phi) = \mathcal{N}(\mu(\mathbf{y}; \phi), \sigma\mathbf{I})$. We fit the model by maximizing its evidence lower bound (see Appendix D).

Baseline models Additionally, we train (1) a video-only ZIG-model (without latent) that maps video-stimulus $\mathbf{x}$ to dynamic responses $\mathbf{y}$, and a (2) non-probabilistic model trained with the Poisson loss (Turishcheva et al. (2023)). All models share the same hyperparameters wherever possible (Fig 1). Since we aim to use correlations between latent variables and behavioral variables (Stringer et al., 2019; Bashiri et al., 2021) as external validation for the model, we excluded behavioral data during training for all models – in contrast to previous work (Sinz et al., 2018; Wang et al., 2023a).

3. Experiments

We trained and evaluated our models on the data from the five mice of the SENSORIUM competition (Turishcheva et al., 2023). Responses from $\sim 8,000$ excitatory neurons in layers 2–5 of the primary visual cortex per mouse were recorded at 8 Hz and upsampled to 30 Hz, while the head fixed mice viewed naturalistic gray-scale videos at 30 Hz. The video input of the model has shape $(W, H, T) = (64, 36, 80)$. Behavioral variables—locomotion speed, pupil dilation and center position—were also recorded and resampled to 30 Hz.

Predictive performance When it comes to predicting neuronal responses conditioned on video-stimulus, the Poisson baseline has the best correlation performance of 0.195, whereas the latent model has the highest log-likelihood of -0.74 bits per time and neuron showing the increased capabilities of the latent model capturing full response distributions. We computed the performance of the latent model by marginalizing out the latent variable $\mathbf{z}$ via Monte-Carlo sampling (Appendix C). Since the neuronal responses are continuous and the Poisson distribution is for discrete values only, we cannot evaluate the log-likelihood of the Poisson model. The slightly lower correlation performance of the likelihood based ZIG-models could be due to a trade-off between modelling the distribution (good for likelihood) and modelling the conditional mean (good for correlation) (Lurz et al., 2023).

Latent-behavior correlation To compute correlations between the latents and behavioral variables, we used canonical correlation analysis (CCA), which finds the best linear combination of the latent variables $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)}$ that has maximal correlation with a chosen behavioral variable. Although our model has not seen any behavioral data during training the learned latents show strong correlations with behavioral data (Fig. 2), in accordance with previous works (Stringer et al., 2019; Bashiri et al., 2021; Niell and Stryker, 2010) (analysis details in Appendix B).

Table 1: Predictive performance of models

	Poisson Baseline	**Video-only ZIG**	**Latent ZIG**
Pearson Correlation	**0.195**	0.183	0.182
Log Likelihood in Bits per Neuron and Time	-	-0.98	**-0.74**

Figure 2: Canonical correlation of behavior and latent for each mouse. Error bars indicate standard error of mean of cross validation. Mice IDs are from SENSORIUM data.

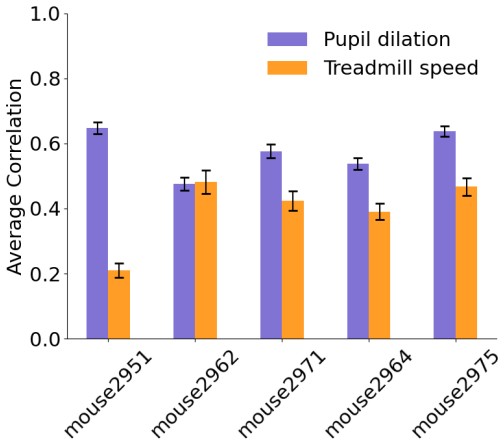

Table 2: Average canonical correlation for behavior and latent with standard error of mean

	Latent Model
Pupil Dilation	0.57 ± 0.009
Treadmill Speed	0.40 ± 0.014

Table 3: Correlation for Neuron Conditioning

	Default Latent Model	**Forecasting Latent Model**
Conditioning on half of neurons	**0.229**	0.226
Forecasting y_t given $y_{1,\ldots,t-1}$	0.233	**0.234**

Neuron Conditioning We assessed the predictive correlation of the model by providing half of the responses as input, with the latent model achieving a correlation of 0.229, outperforming the Poisson Baseline's 0.195. This indicates that the latent model has learned to capture relevant information from neuronal responses. Additionally, we analyzed neuron responses at time t, using data only untile time $t-1$. The latent model outperformed baselines, indicating its capacity for accurately forecasting neuron responses. The forecasting-trained model (Appendix E) did not notably increase forecasting performance compared to the default latent model.

4. Discussion

In this work we showed that adding latent factors to a video encoding model enables the prediction of a joint dynamic response distribution and captures biological variables by implicitly learning correlations between latent factors and behavior. Future work could investigate modeling the latent state further, such as experimenting with the number of neurons needed, the optimal dimensionality k, and distributional assumptions in the generative model, potentially learning latent variables with temporal dependencies.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data, 2024. URL https://arxiv.org/abs/2311.00136.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva Dyer. A unified, scalable framework for neural population decoding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44937–44956. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8ca113d122584f12a6727341aaf58887-Paper-Conference.pdf.

Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15801–15815. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/84a529a92de322be42dd3365afd54f91-Paper.pdf.

Martin Bjerke, Lukas Schott, Kristopher T. Jensen, Claudia Battistin, David A. Klindt, and Benjamin A. Dunn. Understanding neural coding on latent manifolds by sharing features and dividing ensembles, 2023. URL https://arxiv.org/abs/2210.03155.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 1537-274X. doi: 10.1080/01621459.2017.1285773. URL http://dx.doi.org/10.1080/01621459.2017.1285773.

Eloy Geenjaar, Donghyun Kim, Riyasat Ohib, Marlena Duda, Amrit Kashyap, Sergey Plis, and Vince Calhoun. Learning low-dimensional dynamics from whole-brain data improves task capture, 2023. URL https://arxiv.org/abs/2305.14369.

Evren Gokcen, Anna I Jasper, João D Semedo, Amin Zandvakili, Adam Kohn, Christian K Machens, and Byron M Yu. Disentangling the flow of signals between populations of neurons. *Nature Computational Science*, 2(8):512–525, August 2022.

Larissa Höfling, Klaudia P. Szatko, Christian Behrens, Yongrong Qiu, David A. Klindt, Zachary Jessen, Gregory W. Schwartz, Matthias Bethge, Philipp Berens, Katrin Franke, Alexander S. Ecker, and Thomas Euler. A chromatic feature detector in the retina signals visual context changes. *bioRxiv*, 2022. doi: 10.1101/2022.11.30.518492. URL https://www.biorxiv.org/content/early/2022/12/01/2022.11.30.518492.

Kristopher Jensen, Ta-Chu Kao, Jasmine Stone, and Guillaume Hennequin. Scalable bayesian gpfa with automatic relevance determination and discrete noise models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10613–10626. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/58238e9ae2dd305d79c2ebc8c1883422-Paper.pdf.

TD Kim, TZ Luo, T Can, K Krishnamurthy, JW Pillow, and CD Brody. Flow-field inference from neural data using deep recurrent networks. *bioRxiv*, 2023. doi: 10.1101/2023.11.14.567136. URL https://doi.org/10.1101/2023.11.14.567136. PMID: 38014290; PMCID: PMC10680687.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Konstantin-Klemens Lurz, Mohammad Bashiri, Edgar Y. Walker, and Fabian H. Sinz. Bayesian oracle for bounding information gain in neural encoding models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iYC5hOMqUg.

Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

Jacob Reimer, Matthew J McGinley, Yang Liu, Charles Rodenkirch, Qi Wang, David A McCormick, and Andreas S Tolias. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, 7(1):13289, November 2016.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023.

K Seeliger, L Ambrogioni, Y Güçlütürk, LM van den Bulk, U Güçlü, and MAJ van Gerven. End-to-end neural system identification with neural information flow. *PLoS Comput Biol*, 17(2):e1008558, Feb 2021. doi: 10.1371/journal.pcbi.1008558.

Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/9d684c589d67031a627ad33d59db65e5-Paper.pdf.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu B Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):255, 2019. doi: 10.1126/science.aav7893.

David Sussillo, Rafal Jozefowicz, L. F. Abbott, and Chethan Pandarinath. Lfads - latent factor analysis via dynamical systems, 2016. URL https://arxiv.org/abs/1608.06315.

Polina Turishcheva, Paul G. Fahey, Laura Hansel, Rachel Froebe, Kayla Ponder, Michaela Vystrčilová, Konstantin F. Willeke, Mohammad Bashiri, Eric Wang, Zhiwei Ding, Andreas S. Tolias, Fabian H. Sinz, and Alexander S. Ecker. The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *arXiv*, preprint, 2023. doi: 10.48550/arXiv.2305.19654. URL https://doi.org/10.48550/arXiv.2305.19654.

Polina Turishcheva, Max Burg, Fabian H Sinz, and Alexander Ecker. Reproducibility of predictive networks for mouse visual cortex. *arXiv preprint arXiv:2406.12625*, 2024.

Michaela Vystrčilová, Shashwat Sridhar, Max F. Burg, Tim Gollisch, and Alexander S. Ecker. Convolutional neural network models of the primate retina reveal adaptation to natural stimulus statistics. *bioRxiv*, 2024. doi: 10.1101/2024.03.06.583740. URL https://www.biorxiv.org/content/early/2024/03/09/2024.03.06.583740.

EY Wang, PG Fahey, K Ponder, Z Ding, A Chang, T Muhammad, S Patel, Z Ding, D Tran, J Fu, S Papadopoulos, K Franke, AS Ecker, J Reimer, X Pitkow, FH Sinz, and AS Tolias. Towards a foundation model of the mouse visual cortex. *bioRxiv*, 2023a. doi: 10.1101/2023.03.21.533548. URL https://doi.org/10.1101/2023.03.21.533548. PMID: 36993435; PMCID: PMC10055288.

JH Wang, D Tsin, and TA Engel. Predictive variational autoencoder for learning robust representations of time-series data. *arXiv preprint arXiv:2312.06932*, 2023b. PMID: 38168462; PMCID: PMC10760197.

Yule Wang, Zijing Wu, Chengrui Li, and Anqi Wu. Extraction and recovery of spatio-temporal structure in latent dynamics alignment with diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 38988–39005. Curran Associates, Inc., 2023c. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7abbcb05a5d55157ede410bb718e32d7-Paper-Conference.pdf.

Xue-Xin Wei, Ding Zhou, Andres Grosmark, Zaki Ajabi, Fraser Sparks, Pengcheng Zhou, Mark Brandon, Attila Losonczy, and Liam Paninski. A zero-inflated gamma model for post-deconvolved calcium imaging traces. May 2019. doi: 10.1101/637652. URL http://dx.doi.org/10.1101/637652.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009. doi: 10.1152/jn.90941.2008. Epub 2009 Apr 8; Erratum in: Journal of Neurophysiology. 2009 Sep;102(3):2008.

Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7234–7247. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/510f2318f324cf07fce24c3a4b89c771-Paper.pdf.

Feng Zhu, Harrison A Grier, Raghav Tandon, Changjia Cai, Anjali Agarwal, Andrea Giovannucci, Matthew T Kaufman, and Chethan Pandarinath. A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with subframe temporal resolution. *Nature neuroscience*, 25(12):1724–1734, 2022.

Appendix A. Related Works Overview

We give in 4 an overview of related work, which either predicts neuron responses for given natural stimuli, uses latent representations for encoding neuron responses or takes behavior of the animal into account.

Appendix B. Behavior Analysis

For the pupil dilation and the treadmill speed of each mouse we performed one CCA analysis each. The CCA analysis was done with 5-fold cross validation. We split the recording time with a 80/20 ratio. This was repeated on five different seeds. The correlations were computed between the CCA combination $\sum_i w_i^{(cca)} \mathbf{z}^{(i)}$ on the test time, where $w_i^{(cca)}$ are the CCA weights and $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k)} \in \mathbb{R}^T$ are the latent variables. For two selected videos and mice we plotted their normalized pupil dilation and treadmill speed against the corresponding normalized CCA combination of the latent over whole videos time, which correspond to ~ 300 time points (3).

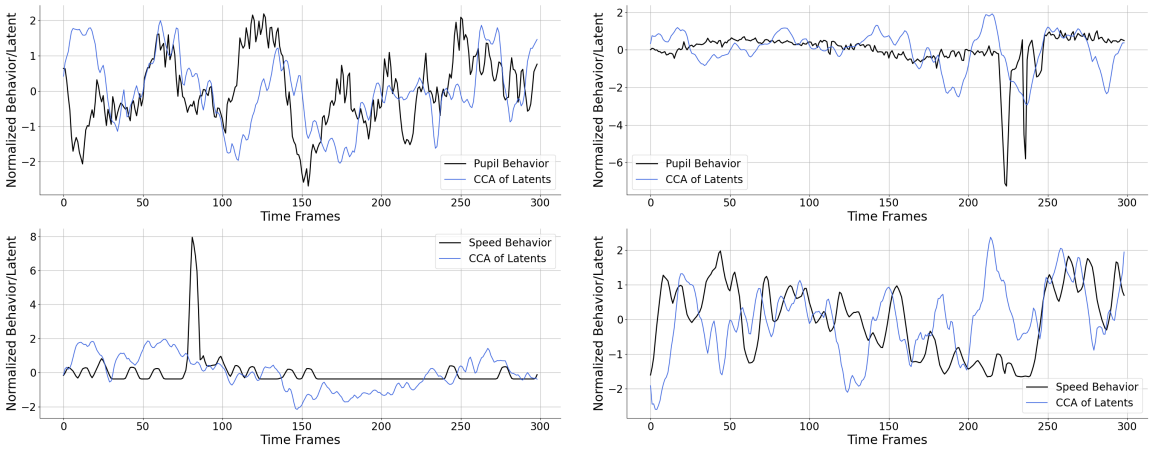

Figure 3: Normalized behavior and CCA combination of latent factors during two selected videos

Appendix C. Training and Marginalization

Training In order to train our model, we maximize $p_{\text{ZIG}}(\mathbf{y}|\mathbf{x})$ via its evidence lower bound (ELBO) via variational inference (Kingma and Welling, 2022; Blei et al., 2017):

$$\log p_{\text{ZIG}}(\mathbf{y}|\mathbf{x}) \geq \langle \log p(\mathbf{y}|\mathbf{z}, \mathbf{x}) \rangle_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{y})} + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{y})|p(\mathbf{z})), \tag{4}$$

where $\langle \cdot \rangle$ represents expected value and D_{KL} Kullback-Leibler divergence.
Marginalization The marginalized performance of the latent model is obtained by calcu-

lating:

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}, \mathbf{z}|\mathbf{x})\, d\mathbf{z} \tag{5}$$

$$= \int p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{z}|\mathbf{x})\, d\mathbf{z} \tag{6}$$

$$= \int p(\mathbf{y}|\mathbf{x}, \mathbf{z}) p(\mathbf{z})\, d\mathbf{z} \tag{7}$$

$$= \frac{1}{L} \sum_{l}^{L} p(\mathbf{y}|\mathbf{x}, \mathbf{z}^{(l)}) \tag{8}$$

The last equality is obtained, since $\mathbf{x}$ and $\mathbf{z}$ are independent (Fig. 1 **D**). The last integral is approximated via Monte-Carlo sampling. We found 5000 samples to be sufficient for convergence (4).

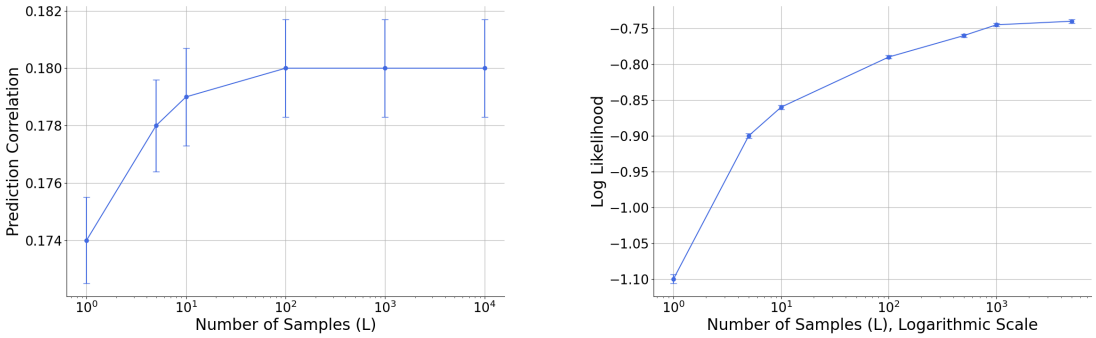

Figure 4: Convergence plots demonstrating the effect of sample size on model prediction correlation and log-likelihood measured in Bits per Neuron and Time

Appendix D. Model Architecture and Hyperparmeters Setting

The video processing part consists of a factorized 3D-CNN block followed by a Gaussian readout (Höfling et al. (2022)). Each convolutional layer consists of a factorized 3D convolution across spatial and temporal dimension followed by a batch normalization layer and an ELU activation function. We use a variational autoencoding approach for the latent representations. A dropout layer is applied to the neuron responses before they are fed into the encoder. This prevents the model from learning correlations between specific neurons thereby encouraging the learning of global latent representations. The encoder applies a linear layer reducing the dimensionality of neuron responses $N \approx 8,000$ (N ranging from 7800 to 8200 depending on the mouse as part of the SENSORIUM dataset (Turishcheva et al., 2024)) to $n << N$. For each time step the same linear layer is applied independently. The linear layer is followed by a layer normalization and an ELU activation function. An individual linear layer was trained for each mouse. The output is processed by a one-layer recurrent neural network with gated recurrent units producing the means $\mu(\mathbf{y}; \phi)$ of the

approximate posterior $q(\mathbf{z}|\mathbf{y};\phi) = \mathcal{N}(\mu(\mathbf{y};\psi), \sigma\mathbf{I})$. σ is a learnable model parameter of the encoder. The resulting latent variables are decoded with another one-layer GRU, denoted as g. The GRUs are shared among the mice. $g(\mathbf{z};\phi)$ is combined with the outputs from the video encoding part computing the parameters for the probability $\log p(\mathbf{y}|\mathbf{z},\mathbf{x})$ as in 2, 3. For a graphical illustration see 1. We searched over various hyperparameters (Table 5) using Optuna (Akiba et al., 2019).

Appendix E. Forecasting Model

For the forecasting, we consider the latent representations to be a Markov process developing smoothly over the time (Fig. 5). Thus, we trained a forecasting model, where we adapted variational autoencoder architecture to an autoregressive model by shifting the objective from reconstructing the original data point to predicting the subsequent time point as in Wang et al. (2023b). Hence, we minimize:

$$\sum_t \left(\log p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}, \mathbf{x}, \psi)\rangle_{\mathbf{z}\sim q(\mathbf{z}_{t+1}|\mathbf{y}_t,\phi)} + D_{KL}(q(\mathbf{z}_{t+1}|\mathbf{y}_t,\phi)|p(\mathbf{z}_{t+1})) \right) \tag{9}$$

Again, $p(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}, \mathbf{x}, \psi)$ is the ZIG likelihood of responses at time $t+1$ given a latent $\mathbf{z}_{t+1}$, a video $\mathbf{x}$ and ψ are the parameters of the video-encoding model (Fig. 1B). $q(\mathbf{z_{t+1}}|\mathbf{y_t},\phi) = \mathcal{N}(\mu(\mathbf{y_t},\phi), \sigma\mathbf{I})$ is the approximate posterior of the future latent at time z_{t+1} given the responses y_t.

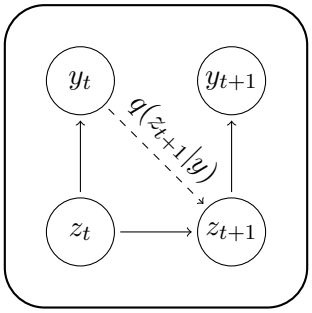

Figure 5: Illustration of the forecasting model. The dashed arrow shows the variational approximation of the Encoder $q(z_{t+1}|y)$. The solid lines show the dependencies of the current and future latent z_t, z_{t+1} and the current and future responses y_t, y_{t+1}

Table 4: Summary of Reviewed Literature: **Column 2** categorizes, if the recorded neural activities are dynamic or static, and its role as input/output of the model. **Column 3** details the type of stimulus components. **Column 4** indicates if the used model has a latent space and if a distribution is computable for the latent (probabilistic). **Column 5** describes the involvement and role of behavioral data. **Column 6** lists the datasets, including subject types, data collection methods, tasks performed during recordings, and neuron sample sizes.

Literature work	Neural activity	Stimulus	Learned latent	Behavior	Datasets
Schneider et al. (2023)	Dynamic Input	No	Yes, probabilistic	Input	Rats, Mice, Monkey 2p and electro-physiology 10-1000 Neurons
Wang et al. (2023a)	Dynamic Output	Video	No	Input	Mice, 2p, passive ∼ 140,000 Neurons
Turishcheva et al. (2023)					∼ 40,000 Neurons
Sinz et al. (2018)					
Kim et al. (2023)	Dynamic In/Output	Audio	Yes, probabilistic	No	Synthetic and Rats, audio decision-making 67 Neurons
Antoniades et al. (2024)	Dynamic Output	Video	Yes, not probabilistic	Output	Mice, 2p, passive 386 Neurons
Azabou et al. (2023)		No			Monkey 27,373 Neurons
Gokcen et al. (2022)	Dynamic Input	No	Yes, probabilistic	No	Macaque V1-3 120 Neurons
Sussillo et al. (2016)					Synthetic 30 Neurons
Zhou and Wei (2020)	Dynamic In/Output	No	Yes, probabilistic	Task Input	Monkey, reaching-task Rat, running 192,120 Neurons
Wang et al. (2023c)				No	Monkey, Rat task 200 x 200 Neurons
Bjerke et al. (2023)				No	Mouse,Rats 26 and 149 Neurons
Jensen et al. (2021)					Macaque, reaching-task 200 Neurons
Geenjaar et al. (2023)	Dynamic In/Output	No	Yes not, probabilistic	No	fMRI
Seeliger et al. (2021)	Dynamic Output	Video	No	No	fMRI
Bashiri et al. (2021)	Static Output	Image	Yes, probabilistic	Output	Mouse V1/LM, 2p,passive ∼ 4,000 Neurons

Table 5: Hyperparamter Configuration

General	
Learning rate	0.005
Core	
Number Layers	3
Temporal Kernel Size first Layer	11
Spatial Kernel Size first Layer	(11,11)
Spatial Kernel Size other Layer	(5,5)
Spatial Kernel Size other Layer	(5,5)
Channels per Layer	(32,64,128)
Encoder	
Dropout probability	50%
Output Dim Linear Layer	42
Output Dim Encoder GRU	12
Decoder	
Output Dim Decoder GRU	12