A Sensing Whole Brain Zebrafish Foundation Model for Neuron Dynamics and Behavior

Sam Fatehmanesh Vegas
California Institute of Technology
sfatehma@caltech.edu

Matt Thomson
California Institute of Technology
mthomson@caltech.edu

James Gornet
California Institute of Technology
jgornet@caltech.edu

David Prober
California Institute of Technology
dprober@caltech.edu

Abstract

Neural dynamics underlie behaviors from memory to sleep, yet identifying mechanisms for higher-order phenomena (e.g., social interaction) is experimentally challenging. Existing whole-brain models often fail to scale to single-neuron resolution, omit behavioral readouts, or rely on PCA/conv pipelines that miss long-range, non-linear interactions. We introduce a sparse-attention whole-brain foundation model (SBM) for larval zebrafish that forecasts neuron spike probabilities *conditioned on sensory stimuli* and links brain state to behavior. SBM factorizes attention across neurons and along time, enabling whole-brain scale and interpretability. On a held-out subject, it achieves mean absolute error < 0.02 with calibrated predictions and stable autoregressive rollouts. Coupled to a permutation-invariant behavior head, SBM enables gradient-based synthesis of neural patterns that elicit target behaviors. This framework supports rapid, behavior-grounded exploration of complex neural phenomena.

1 Introduction

Predicting large-scale neural dynamics at single-neuron resolution is essential for linking brain activity to behavior and for running rapid in silico experiments that can guide in vivo work (Ahrens et al., 2013; Naumann et al., 2016). Whole-brain light-sheet imaging and modern calcium indicators now make it possible to record brain-wide activity at cellular resolution in larval zebrafish during visually guided behavior (Ahrens et al., 2013; Vladimirov et al., 2014; Dana et al., 2019). Yet current modeling approaches struggle to meet five goals at once: accurate next-step prediction, fidelity to the distribution of brain states, scalability to whole brains, behavioral coverage, and interpretability. PCA pipelines compress activity into low-dimensional embeddings that discard neuron-level structure, limiting connectomic or functional interpretation and often failing at whole-brain scale (Jolliffe and Cadima, 2016). Convolutional video architectures such as U-Nets and their 3D variants emphasize local receptive fields and demand substantial compute when extended to high-resolution spatiotemporal volumes, and popular video diffusion systems continue to rely on convolutional backbones, underscoring the computational burden for long-range interactions (Ronneberger et al., 2015; Cicek et al., 2016; Ho et al., 2022b,a). Recent forecasting benchmarks highlight the opportunity for foundation-model style approaches in neural video but do not yet provide single-neuron interpretability together with behavior-level readouts (Immer et al., 2025; Duan et al., 2025).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

To address this gap we introduce the Sparse Brain Model (SBM), which factorizes attention along space and time with two modules. A dynamic connectome layer applies time-independent self-attention across neuron tokens to expose neuron-neuron influences. A temporal neuron layer applies causal self-attention within each neuron's history, enabling long-range temporal dependence without inter-neuron confounds. We derive neuron-level spike probabilities from DF/F traces using a causal spike-inference pipeline so that the model learns on spiking statistics rather than raw fluorescence (Rupprecht et al., 2021). The attention mechanism provides global, distance-independent interactions and efficient parallelism, and rotary position embeddings provide a compact way to encode relative position for temporal attention (Vaswani et al., 2017; Su et al., 2021). To connect brain state to action, a Peripheral Neural Model (PNM) reads either ground truth or predicted brain states and maps them to behavior, enabling both behavioral prediction and optimization of neural inputs that elicit target actions (Naumann et al., 2016). We find that sparse attention at neuron resolution yields accurate next-step predictions with strong calibration, lower error with longer context, and sublinear error growth during autoregressive rollout. Predicted and true brain states occupy similar low-dimensional manifolds in PCA and UMAP spaces, supporting distributional fidelity rather than only pointwise accuracy (Jolliffe and Cadima, 2016; McInnes et al., 2018). Coupled with the PNM, the system predicts fish behaviors from short histories of brain state and permits gradient-based synthesis of neural patterns that expand the reachable behavioral space beyond random stimulation, providing testable hypotheses for mechanisms that link brain-wide dynamics to action (Naumann et al., 2016). We train and validate all models on the publicly released larval zebrafish whole-brain calcium-imaging dataset of Chen et al. (2018).

An interactive web demo for region- and neuron-level simulated optogenetic experiments utilizing the whole brain foundation model is available at `https://virtbrain.samfv.systems`. Model code, data loading, and training scripts available at `https://anonymous.4open.science/r/GBM-8E94/`.

2 Model architecture and training

SBM. The Sparse Brain Model (SBM) predicts per-neuron spike probabilities conditioned on sensory input, operating directly at single-cell resolution. At each time step we represent N neurons plus a stimulus token; neuron tokens include DF/F-derived spike probabilities and 3D soma locations.

Factorized attention. Each block applies (*i*) a *dynamic connectome* layer (time-independent self-attention across neurons and the stimulus token) followed by (*ii*) a *temporal neuron* layer (causal self-attention along the history of each neuron, independent of other neurons). Routed, variable-length attention reduces spatial cost from $\mathcal{O}(N^2)$ to $\mathcal{O}(k\,w^2)$ with $k \approx \lceil N/w \rceil$ (effectively linear in N for fixed cluster width w), while causal temporal attention captures long-range dependencies per cell. Geometric information is injected via directional rotary encodings.

Learning target and rollout. We infer *causal* spike rates from calcium traces using CASCADE and convert them to spike probabilities for training with binary cross-entropy. Inference uses fixed-window autoregression (default 4 s; 12 steps). Full equations, masking, and complexity details appear in Appendix A.

Behavior head. A permutation-invariant *Peripheral Neural Model* (PNM) pools neuron features with positions and maps short brain-state histories to behavior; it also enables differentiable input synthesis for targeted behavior generation. See Appendix A.3.

Data, splits, and optimization. We train on a public larval zebrafish whole-brain calcium-imaging dataset with a subject-level split (one held-out fish for validation). Training uses bf16 mixed precision, FlashAttention, TF32 matmuls, `torch.compile`, and a Muon+AdamW optimizer scheme with warmup-cosine scheduling and gradient clipping. Full hyperparameters and engineering practices are in Appendix A.

3 Results

Setup. We trained and evaluated the Sparse Brain Model (SBM) on brain-wide light-sheet recordings of behaving larval zebrafish, using CASCADE to infer causal spike rates that we convert to spike probabilities for learning (Chen et al., 2018; Ahrens et al., 2013; Vladimirov et al., 2014; Rupprecht

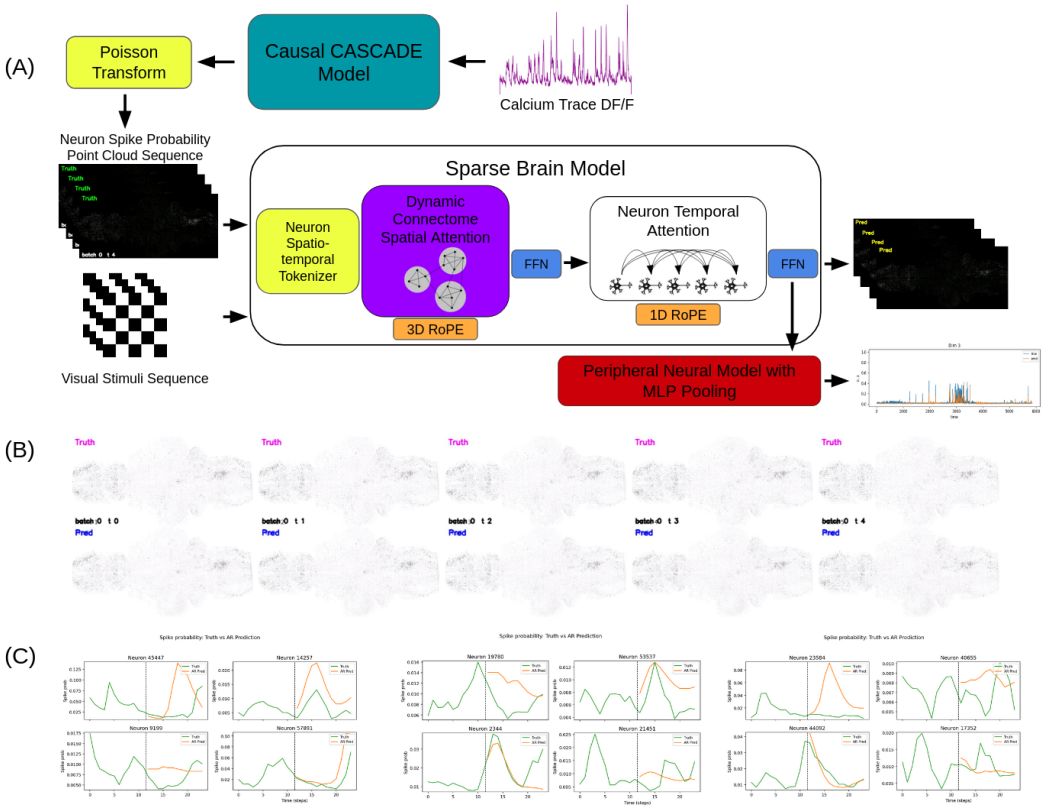

Figure 1: Sparse Attention Architecture enables whole-brain foundation models. (A) This panel illustrates the model architecture and data-processing pipeline. A causal CASCADE model (Ruprecht et al., 2021) is applied to DF/F neuron calcium traces to estimate spike rates, which are then converted to spike probabilities under a Poisson assumption. The Sparse Brain Model (SBM) comprises two key layers: a dynamic connectome layer that infers functional neural clusters and applies time-independent self-attention across neuron tokens, and a temporal neuron layer that applies causal self-attention to each neuron's token sequence independently of other neurons. (B) Comparison of ground-truth and next-step predictions from the SBM, visualized as neuron spike-probability point clouds voxelized to a 512×256 grid and mean-pooled across the z-axis. (C) Comparison of ground-truth spike-probability traces with autoregressive predictions generated using a 4-second sliding window.

et al., 2021). Unless noted otherwise, models use a 4 s (τ=12 steps) context window with teacher forcing at train time and fixed-window autoregression at test time. We want to underscore that even with such a small context window the model is able to auto regressively predict future neural activity with high fidelity and without predictions collapsing into either over or under activation.

3.1 Next-step accuracy, calibration, and rollout stability

The SBM accurately predicts single-neuron next-step activity while preserving cell-level structure (Fig. 1A–C). In qualitative overlays, predicted spike-probability point clouds (voxelized to a 512×256 grid and mean-pooled across z) closely track ground truth across time, and single-neuron traces show both transient and sustained events being followed (Fig. 1B–C). Quantitatively, next-step mean absolute error is < 0.02 under typical contexts and improves further with longer histories (Fig. 2B). Predictions are *well calibrated*: the predicted mean spike probabilities closely match the empirical means, indicating reliable uncertainty at the neuron level (Fig. 2A). During autoregression with a 4 s sliding window, error accumulation is *sublinear* over horizon length and rollouts remain stable (Fig. 2C).

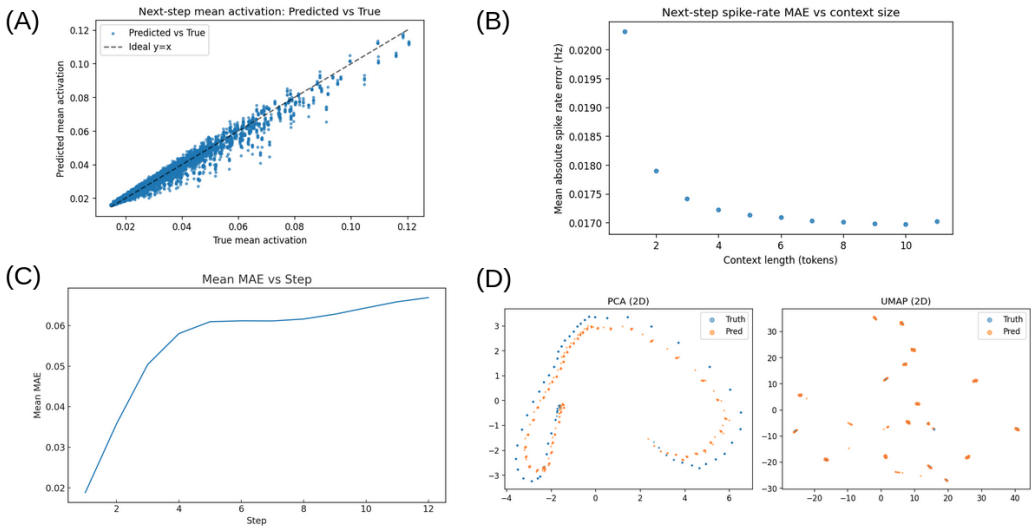

Figure 2: Sparse Brain Models produce more accurate next-step predictions with longer context and preserve low error during autoregression. (A) The SBM is well calibrated, with predicted mean neuron spike probabilities highly correlated with ground-truth means. (B) Prediction error decreases as the history size increases, and during autoregression the error accumulates sublinearly with respect to the number of steps when using a 4-second (12 time-point) sliding window. (C) During next-step prediction, greater temporal context yields significantly lower error.(D) In PCA and UMAP embeddings, the distributions of next-step predictions and ground-truth brain states are highly correlated, indicating that the SBM preserves the brain's original activity distribution and manifold.

3.2 Distributional fidelity of predicted brain states

Beyond pointwise error, predicted and real brain states occupy similar low-dimensional manifolds. In PCA and UMAP embeddings, the distributions of next-step predictions largely overlap with those of ground truth, supporting the claim that SBM learns the *distribution* of brain activity rather than overfitting to mean trends (Jolliffe and Cadima, 2016; McInnes et al., 2018) (Fig. 2D). This property is important for downstream tasks that depend on realistic population-level coordination rather than isolated spikes.

3.3 Linking brain state to behavior

To connect neural dynamics to action, we pair SBM (or ground-truth spikes) with a permutation-invariant behavior head (PNM; Sec. A.3). Using four brain-state time points taken backward in time as input, the PNM predicts behavior on held-out fish with a mean Pearson correlation of **0.42** (Fig. 3A). In a behavioral PCA space, predicted behaviors cover much of the distribution of real behaviors, indicating that the neural representations learned by SBM are behaviorally informative (Naumann et al., 2016; Jolliffe and Cadima, 2016) (Fig. 3B).

3.4 Novel behavior generation via gradient-based neural pattern search

We asked whether brief, structured neural patterns could *generate* new behaviors. A naive baseline—driving the system with *random* neural stimulations—produced only a *small* subset of the behavioral space, collapsing to modes already common in the data (Fig. 3B, "random"). In contrast, optimizing short sequences of neuron-specific inputs by gradient descent (through the differentiable PNM) discovered *diverse* novel behaviors that populate previously sparse regions of the embedding (Fig. 3B, "Novel Behaviors").

Together, these results show that (i) the SBM provides calibrated, accurate forecasts at single-neuron resolution with stable rollouts, (ii) predicted brain states retain distributional fidelity, and, (iii) when

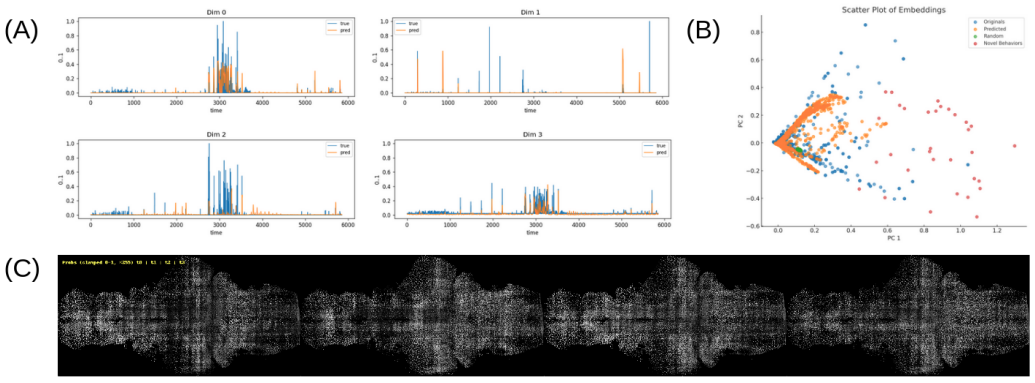

Figure 3: Adding the Peripheral Neural Model (PNM) enables prediction of fish behaviors from brain states—either ground truth or model predictions—and enables the generation of novel behaviors. (A) For a held-out fish, the PNM predicts behaviors from four brain-state time points taken backward in time, achieving a mean Pearson correlation coefficient of 0.42 with ground truth. (B) In PCA space, the PNM replicates a substantial portion of the original behavior distribution. When driven by random neural stimulations, the model generates only a small subset of behaviors, implying that realistic behaviors require highly structured neural patterns; in contrast, optimizing neural inputs via gradient descent to target behaviors yields a much broader distribution of novel behaviors. (C) A visualizes a four-time-point neural state learned by the PNM to elicit a novel behavior.

we coupled the SBM to the PNM, *neural inverse design* enables gradient-learned patterns substantially expand reachable behavior beyond what random stimulation achieves.

4 Discussion

Current practice trades off scale, interpretability, and behavioral grounding. POCO-style pipelines are not spike rate native and are not conditioned on sensory information (Duan et al., 2025). Video U-Net approaches demand multiple large GPUs and rely on local convolutions that cannot flexibly attend between distant neurons (Ronneberger et al., 2015; Cicek et al., 2016), and state-of-the-art video diffusion systems still inherit heavy convolutional backbones (Ho et al., 2022b,a). These constraints hinder models that must predict single-neuron activity at whole-brain scale while remaining faithful to the distribution of brain states and usable for behaviorally relevant analysis.

Our results show that sparse attention can meet these requirements by separating spatial from temporal reasoning and by operating directly on spike probabilities inferred from calcium imaging (Rupprecht et al., 2021). The dynamic connectome layer provides interpretable neuron–neuron influences, while the temporal layer captures within-cell dynamics with causal self-attention informed by rotary position embeddings (Vaswani et al., 2017; Su et al., 2021). Fidelity in PCA and UMAP spaces indicates that the model preserves the structure of the brain's activity distribution rather than overfitting to pointwise errors (Jolliffe and Cadima, 2016; McInnes et al., 2018). The Peripheral Neural Model links predicted brain states to behavior and enables gradient-based synthesis of neural patterns that broaden the repertoire of achievable actions, creating an efficient path for rapid in silico experiments that can guide targeted in vivo tests (Naumann et al., 2016).

If successful, this framework can make in-silico experimentation routine: researchers could screen perturbations, prioritize targets, and design closed-loop interventions that elicit desired behaviors before committing to in vivo trials. Interpretable sparse attention may provide a bridge between data-driven prediction and mechanistic theory by exposing testable neuron-to-neuron influences, enabling integration with anatomical connectomes and cell-type maps, and guiding targeted stimulation or pharmacological interventions. Extending the approach to richer sensory contexts, longer timescales, and other species could yield general-purpose whole-brain foundation models that support hypothesis generation for complex states such as sleep and learning, improve the data efficiency of experimental programs, and ultimately enable principled control of neural circuits.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2025/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, 2013. doi: 10.1038/nmeth.2434.

Xiuye Chen, Yu Mu, Yu Hu, Aaron T. Kuan, Maxim Nikitchenko, Owen Randlett, Alex B. Chen, Jeffery P. Gavornik, Haim Sompolinsky, Florian Engert, and Misha B. Ahrens. Brain-wide organization of neuronal activity and convergent sensorimotor transformations in larval zebrafish. *Neuron*, 100(4):876–890.e5, 2018. doi: 10.1016/j.neuron.2018.09.042.

O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D u-net: Learning dense volumetric segmentation from sparse annotation. *arXiv preprint arXiv:1606.06650*, 2016.

H. Dana, Y. Sun, B. Mohar, et al. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods*, 16(7):649–657, 2019. doi: 10.1038/s41592-019-0435-6.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

S. Duan, A. Raja, E. Alba, et al. POCO: Scalable neural forecasting through population conditioning. *arXiv preprint arXiv:2506.14957*, 2025.

J. Ho, W. Chan, C. Saharia, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, volume 35, pages 8633–8646, 2022b.

A. Immer, J.-M. Lueckmann, J. Lin, et al. Forecasting whole-brain neuronal activity from volumetric video. *arXiv preprint arXiv:2503.00073*, 2025.

I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.

Jordan Keller. Muon optimizer (muonwithauxadam). `https://github.com/KellerJordan/Muon`, 2024.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

E. A. Naumann, J. E. Fitzgerald, T. W. Dunn, et al. From whole-brain data to functional circuit models of behavior. *Cell*, 167(4):947–960.e20, 2016. doi: 10.1016/j.cell.2016.10.019.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

A. Roy, M. Saffar, A. Vaswani, and D. Grangier. Efficient content-based sparse attention with routing transformers. *arXiv preprint arXiv:2003.05997*, 2020.

P. Rupprecht, S. Carta, A. Hoffmann, et al. A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nature Neuroscience*, 24(9):1324–1337, 2021. doi: 10.1038/s41593-021-00895-5.

J. Su, Y. Lu, S. Pan, et al. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. In *NeurIPS*, volume 30, pages 5998–6008, 2017.

N. Vladimirov, Y. Mu, T. Kawashima, et al. Light-sheet functional imaging in fictively behaving zebrafish. *Nature Methods*, 11(9):883–884, 2014. doi: 10.1038/nmeth.3040.

A Model Architecture and Training

A.1 Problem setup

Let $x_{t,n} \in [0,1]$ denote the spike *probability* of neuron $n \in \{1,\dots,N\}$ at time $t \in \{1,\dots,T\}$, inferred *causally* from calcium traces via CASCADE (Rupprecht et al., 2021). Let $p_n \in \mathbb{R}^3$ be the soma location and $s_t \in \mathbb{R}^{d_s}$ encode exogenous input (stimuli/task). The task is next-step forecasting: given a context window $\mathcal{C}_t = \{(x_{t-\tau:t-1,\cdot},\ s_{t-\tau:t-1})\}$ of length τ, predict $x_{t,\cdot}$; during rollout we iterate this autoregressively.

A.2 Sparse Brain Model (SBM)

SBM factorizes spatiotemporal reasoning into two attention modules per block: (*i*) a *dynamic connectome layer* that attends across neurons (plus a stimulus token) within each time step, and (*ii*) a *temporal neuron layer* that applies *causal* attention along the history of each neuron independently. This preserves single-cell interpretability while scaling to whole brains.

Tokenization and embeddings. At time t we form N neuron tokens and one stimulus token. Neuron n is embedded from its scalar activity and 3D position using RMSNorm:

$$h_{t,n}^{(0)} = \mathrm{RMSNorm}\big(W_{\text{neuron}}\,[x_{t,n};\,p_n] + b_{\text{neuron}}\big) \in \mathbb{R}^d, \tag{1}$$

and the stimulus token is

$$h_{t,\text{stim}}^{(0)} = \mathrm{RMSNorm}\big(W_{\text{stim}}\,s_t + b_{\text{stim}}\big) \in \mathbb{R}^d. \tag{2}$$

We stack $H_t^{(0)} = [h_{t,1}^{(0)},\dots,h_{t,N}^{(0)},h_{t,\text{stim}}^{(0)}] \in \mathbb{R}^{(N+1)\times d}$. Variable neuron counts are padded to $N_{\max}$ with mask $m \in \{0,1\}^{N_{\max}}$ respected by all attention ops.

Block structure. Each of the L blocks applies spatial then temporal attention with residuals and normalization:

$$\widetilde{H}_t^{(\ell)} = \mathrm{SpatialAttn}\big(\mathrm{RMSNorm}(H_t^{(\ell-1)}),\,P\big) + H_t^{(\ell-1)}, \quad P = \{p_n\}_{n=1}^N, \tag{3}$$

$$Z_{\cdot,n}^{(\ell)} = \mathrm{TemporalAttn}\big(\mathrm{RMSNorm}(\widetilde{H}_{\cdot,n}^{(\ell)}),\,\text{causal}\big) + \widetilde{H}_{\cdot,n}^{(\ell)}, \tag{4}$$

where $H_t^{(\ell)}$ stacks $\{Z_{t,n}^{(\ell)}\}_{n=1}^N$ and drops the stimulus token after the spatial layer.

Dynamic connectome layer (spatial attention). Within each time slice we attend *across* neurons (and the stimulus token). 3D geometry is injected via directional rotary encodings (spatial RoPE): project p_n onto random unit directions at log-spaced frequencies and add to Q/K streams (Su et al., 2021). To scale to $N \sim 10^5$, we use *routing* to partition tokens into k balanced clusters of target size $w \ll N$; each centroid selects top-w tokens (multi-membership allowed), attention is computed per cluster with variable-length FlashAttention, and outputs are scatter-added to the original order and averaged over duplicates. This reduces cost from $\mathcal{O}(N^2)$ to $\mathcal{O}(k\,w^2)$ with $k \approx \lceil N/w \rceil$, i.e., effectively linear in N for fixed w (Roy et al., 2020; Dao, 2023).

Temporal neuron layer (causal attention). For each neuron n, we apply causal multi-head attention along time with standard RoPE over indices:

$$\mathrm{TemporalAttn}(Q,K,V) = \mathrm{MHA}\big(\mathrm{RoPE}_t(Q),\ \mathrm{RoPE}_t(K),\ V;\ \text{causal}\big), \tag{5}$$

batching neurons as independent sequences of length τ (masked rows skipped). This captures long-range temporal dependence per cell without quadratic cross-neuron time interactions.

Decoder and training loss. After L blocks we output a logit per neuron and time,

$$z_{t,n} = w_{\text{dec}}^{\top}\mathrm{RMSNorm}\big(Z_{t,n}^{(L)}\big) + b_{\text{dec}}, \qquad \hat{x}_{t,n} = \sigma(z_{t,n}), \tag{6}$$

and train with binary cross-entropy over unpadded tokens:

$$\mathcal{L}_{\text{BCE}} \;=\; -\frac{1}{|\mathcal{I}|}\sum_{(t,n)\in\mathcal{I}}\Big[x_{t,n}\log\sigma(z_{t,n}) + (1-x_{t,n})\log(1-\sigma(z_{t,n}))\Big]. \tag{7}$$

Autoregressive rollout. Given $(X_{1:\tau}, S_{1:\tau})$, we iterate a fixed window (e.g., $\tau = 12$ steps ≈ 4 s):

$$\hat{x}_{\tau+1,\cdot} = f_\theta(X_{1:\tau}, S_{1:\tau}),\ \hat{x}_{\tau+2,\cdot} = f_\theta([X_{2:\tau}, \hat{x}_{\tau+1,\cdot}], [S_{2:\tau}, S_{\tau+1}]), \ldots \tag{8}$$

A.3 Behavior-from-activity head (PNM)

To map brain states to behavior, we use a permutation-invariant pooling model. Let $X_t \in [0,1]^N$ and $P \in \mathbb{R}^{N\times 3}$. We center/scale P, add Fourier features $\psi(p_n)$, and z-score spikes per time to remove global gain. Each $[\tilde{x}_{t,n}; \psi(p_n)]$ is encoded by a small MLP to $h_{t,n}$; masked mean pooling yields $\bar{h}_t$, a temporal MLP aggregates $\{\bar{h}_{t-\tau+1}, \ldots, \bar{h}_t\}$, and we predict behavior $y \in \mathbb{R}^{d_{\text{beh}}}$. *Training.* We train one model per behavior dimension with L1 loss (ℓ_1) on targets in $[0,1]$, AdamW optimizer, linear warm-up (10% of steps), gradient-clipping at 1.0, and balanced mini-batches that match rare high-magnitude frames to a subset of typical frames for stability.

A.4 Implementation details and training practice

Masks/geometry/kernels. All operations honor neuron masks. Spatial attention uses directional RoPE on positions; temporal attention uses standard RoPE over time (Su et al., 2021). Spatial attention runs with balanced routing and variable-length FlashAttention; temporal attention batches valid neuron rows into a single causal FlashAttention call (Roy et al., 2020; Dao, 2023). In practice w is a few hundred, making compute effectively linear in N.

Numerics. The model runs end-to-end in **bf16**; logits and losses are computed in fp32. We enable TF32 matmuls on CUDA and `torch.compile` with dynamic shapes. A CUDA prefetcher overlaps H2D copies and casts spikes/stim to bf16.

Optimization. We use **MuonWithAuxAdam** (Keller, 2024) for matrix/tensor weights ($\geq$ 2D) in the attention *body* and AdamW for gains/biases plus *embed/head*. Typical hyperparameters: Muon learning rate $2 \cdot 10^{-2}$, AdamW learning rate $5 \cdot 10^{-4}$ (with betas of 0.9 and 0.95), weight decay 10^{-4}, gradient-clipping at 1.0. A warmup-cosine schedule (per batch) is applied; we validate several times per epoch and keep the best checkpoint. Random seeds are fixed, and mixed precision plus caching/prefetching keep throughput high.

Data and splits. We use larval zebrafish whole-brain calcium imaging (Chen et al., 2018; Ahrens et al., 2013; Vladimirov et al., 2014) and adopt a subject-level split: models are trained on the training subject fish and evaluated on a single held-out fish reserved for validation. No frames from the held-out subject are used for training or hyperparameter tuning. Unless otherwise noted, all quantitative results and visualizations are computed on the held-out validation fish.