OMNIMOUSE: SCALING PROPERTIES OF MULTI-MODAL, MULTI-TASK BRAIN MODELS ON 150B NEURAL TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling data and artificial neural networks has transformed AI, driving breakthroughs in language and vision. Whether similar principles apply to modeling brain activity remains unclear. Here we leveraged a dataset of 3.3 million neurons from the visual cortex of 78 mice across 323 sessions, totaling more than 150 billion neural tokens recorded during natural movies, images and parametric stimuli, and behavior. We train multi-modal, multi-task transformer models (1M–300M parameters) that support three regimes flexibly at test time: neural prediction (predicting neuronal responses from sensory input and behavior), behavioral decoding (predicting behavior from neural activity), neural forecasting (predicting future activity from current neural dynamics), or any combination of the three. We find that performance scales reliably with more data, but gains from increasing model size saturate – suggesting that current brain models are limited by data rather than compute. This inverts the standard AI scaling story: in language and computer vision, massive datasets make parameter scaling the primary driver of progress, whereas in brain modeling – even in the mouse visual cortex, a relatively simple and low-resolution system – models remain data-limited despite vast recordings. These findings highlight the need for richer stimuli, tasks, and larger-scale recordings to build brain foundation models. The observation of systematic scaling raises the possibility of phase transitions in neural modeling, where larger and richer datasets might unlock qualitatively new capabilities, paralleling the emergent properties seen in large language models.

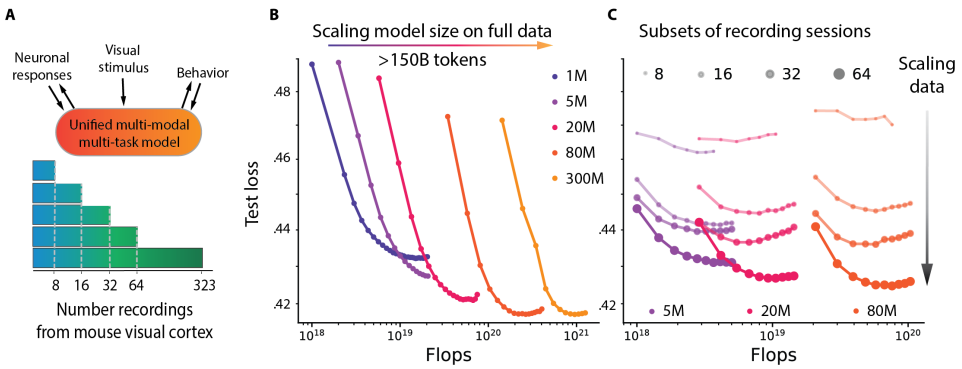

Figure 1: **A.** OmniMouse unifies neural prediction, behavior decoding, and forecasting tasks. **B.** Scaling model size on an 150+ billion neural tokens shows performance saturation, unlike language models. **C.** In contrast, scaling data consistently improves performance across all model sizes, suggesting that neural prediction is currently limited by data.

1 INTRODUCTION

Scaling models and data has driven recent progress in machine learning, with large language, vision, and multi-modal models showing consistent performance gains and enabling foundation models that unify tasks across domains. A natural question is whether models of the brain can also benefit from scaling. In the mouse visual cortex, large datasets (MICrONS Consortium et al., 2021; de Vries et al., 2019; Angelaki et al., 2025) and standardized benchmarks (Willeke et al., 2022; Turishcheva

et al., 2024) exist. Yet, compared with internet-scale corpora, the available datasets are much smaller, more fragmented, and less diverse. The neuroscience community has recently started to work towards foundational models for EEG (Chau et al., 2024; Chen et al., 2024; Cui et al., 2024; Jiang et al., 2024; Kostas et al., 2021; Yang et al., 2023; Thapa et al., 2024; Li et al., 2024), fMRI (Caro et al., 2023; Dong et al., 2024; Kan et al., 2022; Thomas et al., 2022; d'Ascoli et al., 2025), MEG (Csaky et al., 2024), and intracranial signals (Zhang et al., 2023; Wang et al., 2023). But single-neuron resolution, multi-modal foundation models are still missing.

Prior work in this direction focused on isolated modalities (Ye et al., 2023; Azabou et al., 2023), a single predictive task (Wang et al., 2025), lacked scalability across datasets (Ye & Pandarinath, 2021; Mi et al., 2023; Antoniades et al., 2024), or omitted stimulus (Zhang et al., 2025) and behavioral information (Jiang et al., 2025; Mi et al., 2023). These models do not capture the multi-modal and multi-task nature of neural computation. Hence, we cannot systematically study if there are benefits of scaling – a key hallmark of foundational models – in large-scale, single-neuron recordings.

In this work, we introduce OmniMouse, a multi-modal, multi-task architecture for modeling activity in the mouse visual cortex. OmniMouse integrates video stimuli, neuronal responses, and behavioral signals (running speed, eye movements and pupil size) into a single transformer framework. Unlike prior models that are typically restricted to a single modality, task, or dataset, OmniMouse combines single-neuron tokenization, video encoding, and a structured masking framework into a unified architecture. This design enables flexible masking on both the input and output, allowing the model to handle arbitrary combinations of neural forecasting (predicting from past activity), stimulus-conditioned response prediction, sub-population prediction, and behavioral decoding—all within a single model We train OmniMouse on the largest single-neuron dataset to date: 323 recordings from the visual cortex of 78 awake mice viewing naturalistic movies, images, and parametric stimuli, totaling over 150 billion neuronal activity tokens. This unprecedented scale enables a systematic scaling laws analysis, investigating how model and dataset size impact neuronal encoding and behavioral decoding performance.

Our main findings and contributions are:

- **We provide a systematic scaling analysis for neuronal data**: We find that performance improves systematically with more data, but saturates with model size beyond moderate scales. This suggests that **data, not model size, is currently the bottleneck** for predictive accuracy in neural modeling—providing a clear directive for the field that progress requires larger and more diverse neural datasets.
- **We propose a multi-modal multi-task model accounting for a visual stimuli**: OmniMouse handles both single-modality and multi-modal inputs, supporting any combination of forecasting and stimulus-conditioned prediction across neurons, visual stimuli, time, and animals in a single model.
- **OmniMouse achieves state-of-the-art performance**: When compared to strong specialized baselines on the same training data, OmniMouse outperforms prior methods across nearly all tasks (apart from running speed decoding) demonstrating the strength of our approach independent of data scale advantages.

2 Related work

Large-scale deep learning models for single-neuron predictions. Deep learning has advanced predictive modeling in neuroscience, particularly in vision (Cadieu et al., 2014; Batty et al., 2017; Klindt et al., 2017; McIntosh et al., 2016; Cadena et al., 2019; Kindel et al., 2019; Walker et al., 2019; Zhang et al., 2018; Ecker et al., 2018; Sinz et al., 2018; Burg et al., 2021; Cowley & Pillow, 2020). Early CNN-based approaches introduced shared feature cores with per-neuron readouts (Antolík et al., 2016; Klindt et al., 2017; McIntosh et al., 2016), later extended with temporal dynamics (Sinz et al., 2018) and more efficient readouts (Lurz et al., 2021). Building on these advances, Wang et al. (2025) trained a 13-mice CNN model and showed that "digital twins" can capture biological phenomena beyond their training data. With the shift to transformers, new variants have explored ViT cores (Li et al., 2023), hybrid convolution-attention designs (Lin et al., 2024; Pierzchlewicz et al., 2023), and spatial-transformer readouts (Saha et al., 2024), though most still omit video input.

Transformers have also been applied to response-to-response modeling. The Neural Data Transformer (NDT) (Ye & Pandarinath, 2021) predicted spikes from spikes and behavior, later extended to multiple animals (Ye et al., 2023) and neuronal masking strategies (Zhang et al., 2024). While NDT projects all neurons together via linear layers, Quantformer (Calcagno et al., 2024), also a transformer-based forecaster, introduced neuron-specific tokens to handle any number of neurons. POYO (Azabou et al., 2023), a behavior-decoding model, added spike timing to similar tokens, removing the need for time-window binning, and its extension POYO+ (Azabou et al., 2025) also handled discrete classification tasks such as stimulus orientation. POCO (Duan et al., 2025) combined POYO and NDT tokenization to predict neuronal activity from history and other neurons, while STDNT (Le & Shlizerman, 2022) explicitly modeled correlations but did not consistently outperform NDT. Representing the most significant scaling of NDT-based framework, NEDS (Zhang et al., 2025) modeled approximately 30,000 neurons across 74 sessions using a multitask loss to predict neuronal activity and behavior, also using both of them as input. However, the aforementioned models ignore visual stimuli. To study the combined effect of both the 'brain state' and 'visual stimuli' on neuronal activity, Bashiri et al. (2021) used a CNN branch for processing static input stimuli and an additional flow-branch to model trial-to-trial correlations between neurons. For dynamic video stimuli, Schmidt et al. (2025) modeled a latent brain state probabilistically, using NDT-style response tokenization. Similarly, Neuroformer (Antoniades et al., 2024) used past activity and visual input but is limited to single sessions and cannot flexibly condition on subsets of neurons or response history. CEBRA (Schneider et al., 2023), a contrastive encoder, also mapped activity to behavior or stimuli, accounting for inter-neuron correlations. The closest work to ours, outside of single-cell studies, is d'Ascoli et al. (2025), which constructed a multi-modal fMRI predictor using concatenated video, text, and audio embeddings.

General scaling laws in deep learning. Large-scale models in language and vision exhibit predictable improvements with scale, described by empirical "scaling laws". Kaplan et al. (2020) first showed that performance follows power-law trends in model size, dataset size, and compute. Hoffmann et al. (2022) refined this with "Chinchilla scaling", prescribing proportional growth of model and data size for optimal efficiency. Aghajanyan et al. (2023) adjusted scaling laws for models with large per-modality pre-trained tokenizers but newer lightweight tokenization ("early-fusion") approaches (Chameleon, 2024; Piergiovanni et al., 2024; Shukor et al., 2025) achieved stronger performance with fewer parameters. Hence, no universal framework for multi-modal scaling exists: Shukor et al. (2025) estimated power-law coefficients for early-fusion models but did not analyze cross-modal interactions. This gap is especially evident in scientific domains, where data are multimodal, complex, noisy, and limited. Examples such as AlphaFold3 (Abramson et al., 2024) suggest that systematic scaling of both models and datasets can drive major advances in AI for science.

Scaling neuroscience models. There is no consensus on whether classic machine learning scaling laws apply to single-neuron data. Jiang et al. (2025) questioned their applicability, analyzing the NDT-based model of Zhang et al. (2024). Jiang et al. (2025) argued that cross-session variability – and thus implicit data heterogeneity – is crucial for scaling benefits, though it remains unclear if these results generalize to different mouse tasks or model architectures. Again using an NDT-based model but on motor cortex microelectrode data from monkeys and humans, Ye et al. (2025) reported that scaling is constrained by data variability, which pretraining alone cannot fully overcome. Consistent with this view, POCO (Duan et al., 2025) used calcium imaging to show that longer recordings improve predictive performance, aligning with earlier results of Lurz et al. (2021). However, POCO included fewer than 90,000 neurons, mostly from zebrafish ($\sim$77,000). Neural saturation has also been observed: Gokce & Schrimpf (2024) found that behavioral alignment improves with model size, but neural alignment plateaus, with gains concentrated in higher-level visual areas. In contrast, Antonello et al. (2023) reported no such saturation when predicting language and audio fMRI responses, suggesting that scaling limits may depend on the modality and data regime. The largest single-cell response-to-behavior prediction model is POYO+ Azabou et al. (2025) with $\sim$100,000 neurons, which did not analyze scaling. Together, these findings highlight the need for large, multimodal, single-neuron datasets to test how scaling laws manifest in systems neuroscience.

3 Large-scale single-neuron dataset

Neuronal responses. We used a dataset of over 3 million single-unit neuronal recordings (Fig. 2) – an order of magnitude larger than the recently published Brain-Wide Map dataset (BWD, 621,733

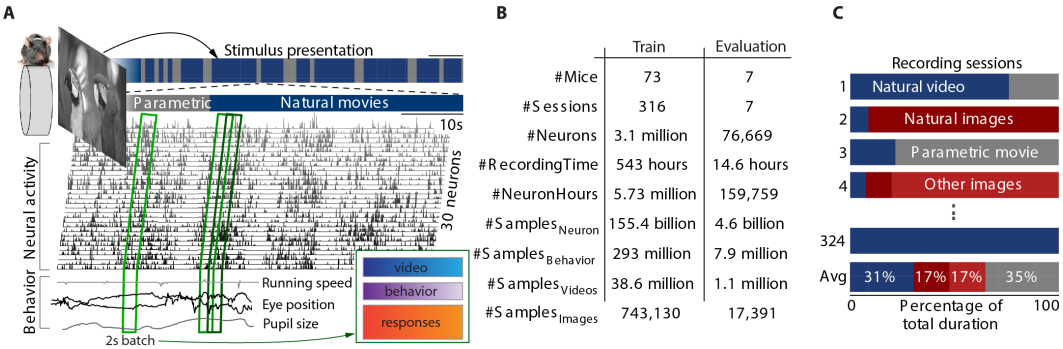

Figure 2: **Data**. **A**. Data were collected from head-fixed mice running on a wheel while viewing videos. Neuronal responses were recorded via calcium imaging, 4210 to 11284 neurons per session. Behavior variables include pupil center x and y positions, pupil dilation and its derivative and running speed. **B**. Dataset statistics. The total number of unique mice in our dataset is 78, since some mice had sessions in both train and evaluation sets. **C**. Different visual stimuli were presented across sessions, with stimulus types varying by session. The bottom row shows their overall distribution.

neurons, $\leq 1 \times 10^6$ neuron-hours) (Angelaki et al., 2025). The dataset contains excitatory neurons' responses in visual cortex recorded via wide-field two-photon calcium imaging at 6–14 Hz in awake, head-fixed, behaving mice (Sofroniew et al., 2016), with spiking activity extracted by CAIMAN (Giovannucci et al., 2019).

Visual stimuli. The mice were presented with naturalistic images sampled from ImageNet (Russakovsky et al., 2015) and videos sampled from cinematic movies and the Sports-1M dataset (Karpathy et al., 2014). In addition, mice were shown parametric stimuli such as static and drifting Gabors (Petkov & Subramanian, 2007), directional pink noise, flashing Gaussian dots, random dot kinematograms (Morrone et al., 2000), and model-generated stimuli (similar to Walker et al., 2019). All stimuli were presented at 30–60 Hz, with images presented for 500 ms and preceded by a 300–500 ms blank screen.

Behavior variables. Our dataset contains five behavior variables: running speed, recorded at 50 -100 Hz, and four pupil variables: pupil center x and y positions, pupil dilation and its derivative, all recorded at 20 Hz.

Data utilization. Similar to Azabou et al. (2023), we sample 2-second windows from any point in the experiment, including inter-trial intervals and blank screens. Critically, we reconstruct the visual stimulus presented throughout the entire recording, enabling continuous representation of the full experimental timeline including blank periods across all diverse visual paradigms. For model training, we downsample all behaviors to 20 Hz, visual stimuli to 30 Hz, and linearly upsample all neuronal responses to 30Hz to be comparable to the SENSORIUM 2023 benchmark.

4 OmniMouse architecture

We sample 2-second chunks of multi-modal data: video frames $\mathbf{V} \in \mathbb{R}^{h \times w \times ch \times time}$ ($\mathbb{R}^{36 \times 64 \times 1 \times 60}$), neural calcium traces $\mathbf{X} \in \mathbb{R}^{P \times time}$ ($\mathbb{R}^{P \times 60}$) for population P, and behavioral traces $\mathbf{B} \in \mathbb{R}^{ch \times time}$

Table 1: **Scaling variants of OmniMouse.** L: multi-modal transformer layers; d_m: model dimension; h: number of attention heads; d_e: dimensions of all embeddings; p_L: multi-modal transformer layer parameters; p_M: model parameters (excluding neuronal embeddings); p_N: all neuronal, session, and animal parameters; p_T: total parameters; S: sequence length.

Model	L	d_m	h	d_e	p_L	p_M	p_N	p_T	S
OmniMouse-1M	2	256	4	256	1.7M	6M	779M	885M	4096
OmniMouse-5M	6	256	8	256	5.1M	10.4M	779M	891M	4096
OmniMouse-20M	6	512	8	256	19.1M	29.1M	779M	810M	4096
OmniMouse-80M	12	768	12	256	88M	115M	779M	894M	4096
OmniMouse-300M	24	1024	16	256	308M	348M	779M	1.1B	4096

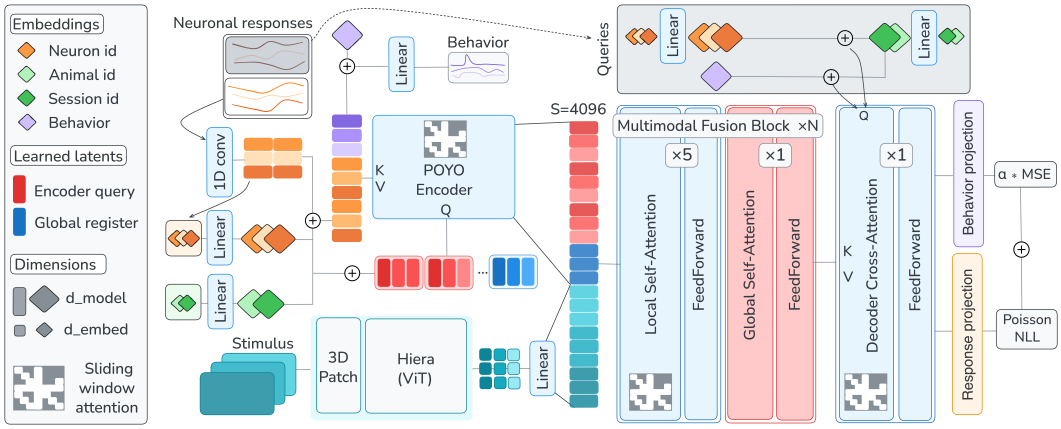

Figure 3: **Model architecture.** OmniMouse introduces a unified framework that handles arbitrary combinations of neural forecasting, sub-population prediction, stimulus encoding, and behavioral decoding through flexible masking. We adopt single-neuron, single-time-chunk tokenization and a cross-attention encoder (following POYO+ (Azabou et al., 2025)), along with analogous queries to the multi-modal cross-attention decoder, enabling per-neuron, per-chunk masking by simply removing tokens from the input and adding corresponding queries to the decoder. A lightweight hierarchical vision transformer tokenizes video at frame-level granularity, allowing temporal masking of visual context. These video features fuse with encoded neural and behavioral embeddings through our transformer stack, creating a unified multi-modal representation from which masked neural activity or behavior can be decoded. Training across 119 App. D.4.1, diverse masking configurations—spanning both core tasks, as well as partial combinations with varying context from each modality—drives strong multi-task performance and enables seamless task switching purely through mask configuration at test time

($\mathbb{R}^{5\times40}$) (running speed, pupil xy-position / size / size derivative). Alongside the chunk, we sample a masking configuration for each modality.

For *video*, the sampled mask defines a starting frame v_0 and the length of visible frames v_c such that $v_0 + v_c \leq 60$, $v_c \in [10, 20, 30, 40, 50, 60]$. The resulting sequence $\mathbf{V}_{v_0:v_0+v_c}$ is encoded through a lightweight, randomly-initialized Hiera vision transformer (Ryali et al., 2023), followed by a linear projection to our model dimension, d_M, producing spatiotemporal embeddings $\widetilde{\mathbf{V}} \in \mathbb{R}^{h'*w'*v'_c \times d_M}$, where h', w', and v'_c result from the stride of the Hiera module.

For *neural responses*, during training we randomly sample $S = 4096$ neurons from population P. From these we select $P_{target} = 3072$ neurons whose final second of activity serves as our prediction target. From the remaining data, we collect activity sequences of each neuron's *unmasked* samples. For OmniMouse, we developed a novel and a flexible neural activity masking scheme that allows for any combination of input masks, down to single-neuron single-sample precision (Fig. 5). The scheme defines a *population prefix* — activity from the population before the last 30 samples — and a *population context* — activity from neurons not being predicted, possibly overlapping in time with the prediction targets. To avoid inflated scores from upsampling artifacts, a gap of at least 0.17 seconds (5 samples) was enforced between the *prefix* and the prediction target. To tokenize the unmasked activity, we apply a strided 1D-convolution to each neuron's sequence and concatenate the outputs, creating a unified sequence of activity embeddings, $\widetilde{\mathbf{X}} \in \mathbb{R}^{S*T\times d_M}$, where T is the number of strides per neuron sequence. Following POYO (Azabou et al., 2023), we add learned identity embeddings for each neuron, session, and animal to the activity features. We use a smaller dimension, d_e, for these embeddings and up-project to d_M in order to reduce the number of parameters learned per-neuron. For *behavior*, we either fully mask or fully unmask the input. When unmasked, we use a shared linear layer to project the traces along the temporal dimension and add learned channel-specific embeddings (as well as the session/animal embeddings), yielding

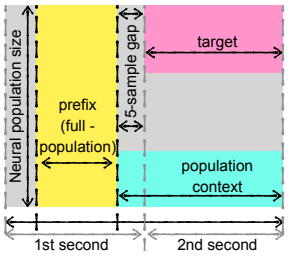

Figure 4: **Neuronal response masking.** We introduce a flexible scheme that supports arbitrary input masks, down to single-neuron, single-sample, and single-frame precision.

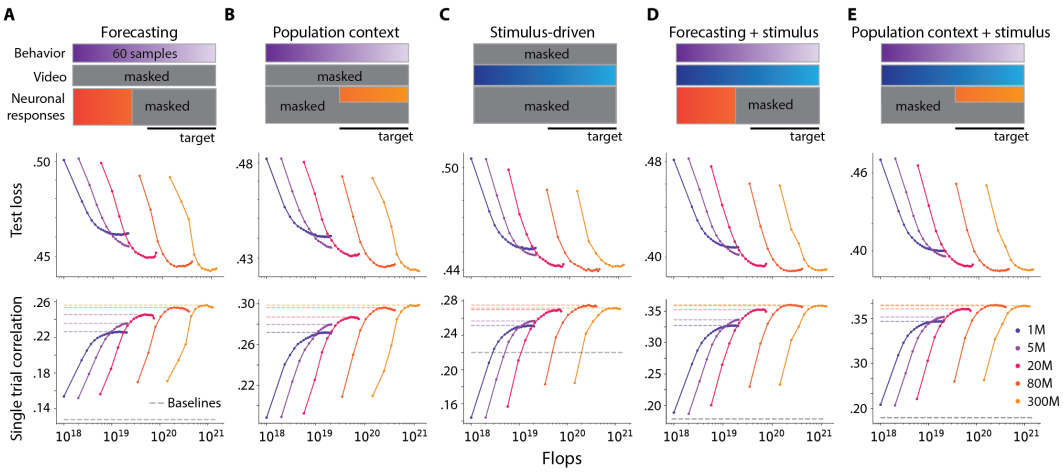

Figure 5: **Task-specific performance gains with model scaling.** Top row: masking schema. Middle row: Test loss. Bottom row: single-trial correlation. Loss and correlation metrics are computed on the held-out test sets of the seven evaluation mice. **A**. Forecasting: predicting one second of future neuronal activity, conditioned only on past neuronal activity (prefix = 25 samples). **B**. Population context; predicting one second of neuronal activity of a sub-population, conditioned only on n = 256 neurons. **C**. Stimulus-driven: Neuronal encoding conditioned on the visual stimulus. **D**. Stimulus-conditioned forecasting: same as forecasting, but also conditioned oprefix = 25 samples. **E**. Stimulus-conditioned population context, context = 256 neurons.

$\widetilde{\mathbf{B}} \in \mathbb{R}^{5 \times d_M}$, for 5 behavior channels.Each token also maintains its timestamp for positional encoding in the input sequence.

Model architecture. After tokenization, we concatenate activity and behavior, $[\widetilde{\mathbf{X}}, \widetilde{\mathbf{B}}]$, and encode using cross-attention with a repeated set of learned latents (Azabou et al., 2023), $\mathbf{Z} \in \mathbb{R}^{M*N \times d_M}$, ($M$ unique latents and N repeats, each repeat with a unique timestamp evenly spaced across the context window), generally reducing the number of input tokens by ~ 10. Within the cross-attention block, we implement *local sliding-window attention*, where latent features only attend to response / behavior features within a fixed temporal window. We also append $g = 256$ "global registers" (Darcet et al., 2023) , $\mathbf{G} \in \mathbb{R}^{g \times d_M}$ which always attend to the entire sequence.

Then we concatenate the cross-attention output and video features, $[\widetilde{\mathbf{Z}}, \widetilde{\mathbf{V}}]$, and pass the sequence to a series of L multi-modal transformer layers (Tab. 1). We interleave local attention (with a sliding-window mask), and global attention blocks at a ratio of $5:1$ (Fig. 3).

To decode neuronal activity and behavior, we use a cross-attention followed by a shared feedforward network, with fused multi-modal features as keys K and values V (Fig. 3). Query construction mirrors input construction: for the response prediction targets, we create a temporal sequence of embeddings using the same learned neuron, animal, and session identity embeddings. Each query also maintains a timestamp indicating the position of the neuronal response and we again employ local causal sliding-window attention. For behavior decoding, we re-use the learned behavioral channel embeddings as queries, with added animal and session embedding. Finally, similar to POYO+ (Azabou et al., 2025), the outputs of the decoder cross-attention block for each modality are routed to modality-specific linear readouts, projecting from d_M back to the original dimensionality. All attentions use RoPE (Su et al., 2024) to encode relative timing between features, both within and across modalities, as well as recent best practices including: RMSNorm pre-normalization layers, query-key normalization, and gated SiLU feed-forward networks (Shazeer, 2020; OLMo et al., 2024; Yang et al., 2025; Biderman et al., 2023).

Training. We trained our model to predict both neuronal responses and behavioral traces, using Poisson loss (averaged across neurons) for neural encoding and mean squared error (MSE) loss for behavior decoding. We used 119 masking configurations (App. D.4.1) during training, varying which modalities were fully or partially masked as well as the amount and duration of neuronal context. To balance the two objectives, the behavioral loss is down-weighted by a factor of 0.1 so that its scale

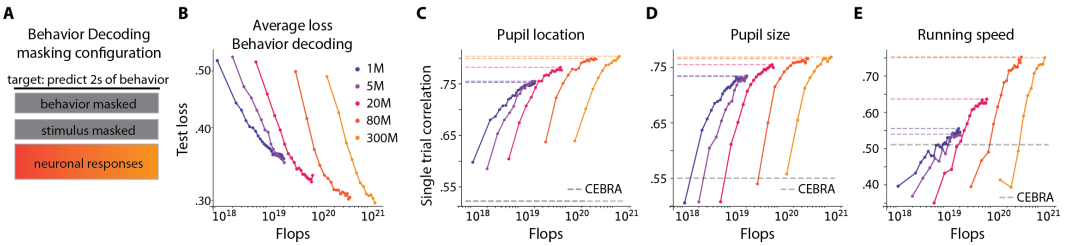

Figure 6: **Behavior decoding scales with model size.** **A**. Masking for behavior decoding. **B**. Decoding loss averaged over all behavioral variables. **C**. Pupil center: correlations computed separately for x and y, then averaged. **D**. Pupil size and its derivative: correlations trace, then averaged. **E**. Running speed: correlation with ground truth.

matches the magnitude of the Poisson loss. For our scaling experiments, we trained models on either the complete dataset of 323 sessions or constructed collections (8, 16, 32, 64 sessions) to study data scaling effects. These nested collections were designed so that larger collections always contained all sessions from smaller ones, ensuring consistent evaluation (see below for evaluation details). We followed Hu et al. (2024); Wen et al. (2024); Hägele et al. (2024) and trained our model with a warmup followed by a constant learning rate for at least 250k steps ($\sim$ 500B tokens), saving checkpoints every 20k steps. After initial training, we continue from each checkpoint for 10k steps using an inverse-square-root learning rate decay, where each decayed checkpoint provides a final evaluation point in Fig. 7.

5 Unified evaluation framework

All scaling experiments use a standardized evaluation protocol on the same mice to ensure fair comparison across models, baselines, and conditions. We chose seven mice (*evaluation mice*) comprised of five publicly available datasets from SENSORIUM 2023 and two test mice from SENSORIUM 2022. For all analyses, we use the held-out set provided by these datasets. We evaluate five regimes of response prediction (Fig. 5) as well as behavior decoding (Fig. 6):

Forecasting conditions predictions on the past activity of the entire population and 40 samples of behavior. We always predict the last second (30 response samples) within each two-second batch, using the first 25 samples of the batch as context. Since NDT-based models (Ye & Pandarinath, 2021) dominate in the forecasting literature, we use IBL (Zhang et al., 2024), a variant of NDT trained with multiple masking strategies similar to ours, as a baseline.

Table 2: **Baseline comparisons.** Results displayed in **bold** indicate the highest score per task in either the data-matched condition (8 sessions; top) or when using the full dataset (323 sessions; bottom). Evaluation conditions in this table were chosen to allow for a fair comparison with all baselines. Baselines were evaluated for all conditions that they support, with ✗ denoting an unsupported condition. Conditions: Forecasting (*Fcst*), forecasting + stimulus (*Fcst+S*), population context (*Pop*) with $n = 256$ visible neurons ($n = 1024$ shown in parentheses), population context + stimulus with $n = 256$ visible neurons(*Pop+S*). Behavioral decoding: Average score across all behaviors (*Avg*), Pupil location (*pupil-loc*), pupil size (*pupil-size*), running speed (*Running*).

	Neuronal Activity Prediction				Behavior Decoding			
Model	Fcst	Fcst+S	Pop	Pop+S	Avg	Pupil-loc	Pupil-size	Running
MtM (Zhang et al., 2024)	0.12	✗	0.07 (0.21)	✗	✗	✗	✗	✗
Latent Model (Schmidt et al., 2025)	✗	0.18	✗	0.16	✗	✗	✗	✗
CEBRA (Schneider et al., 2023)	✗	✗	✗	✗	0.53	0.52	0.55	**0.51**
POYO+ (Azabou et al., 2025)	✗	✗	✗	✗	0.55	0.56	0.63	0.47
OmniMouse-5M (data-matched)	**0.18**	**0.30**	**0.25** (**0.34**)	**0.27**	**0.59**	**0.68**	**0.66**	0.44
OmniMouse-1M (full data)	0.18	0.33	0.27 (0.36)	0.35	0.68	0.75	0.73	0.55
OmniMouse-5M (full data)	0.22	0.34	0.28 (0.37)	0.35	0.69	0.76	0.74	0.57
OmniMouse-20M (full data)	0.23	0.35	0.29 (0.38)	**0.37**	0.75	0.78	0.75	0.73
OmniMouse-80M (full data)	**0.25**	**0.36**	0.29 (**0.39**)	**0.37**	**0.77**	**0.80**	**0.76**	**0.75**
OmniMouse-300M (full data)	**0.25**	**0.36**	**0.30** (**0.39**)	**0.37**	0.76	**0.80**	**0.76**	0.73

Population context conditions predictions on $N = 256$ other simultaneously recorded neurons and 40 samples of behavior. As in the forecasting regime, we predict the last second of each batch and evaluate performance on this interval. This setting assesses how much of the trial-to-trial variability can be explained by simultaneously recorded neurons.
Stimulus-driven conditions predictions on two seconds of video and predicts activity for all neurons in the batch. We provide two seconds of input and evaluate predictions on the final second of neural activity. SENSORIUM 2023 (Turishcheva et al., 2024) establishes a strong baseline for this setting.
Stimulus-conditioned forecasting is identical to forecasting, except that the full 2 seconds of video are also provided as input. We used Schmidt et al. (2025) as a baseline model, which also conditions on neurons, video and behavior.
Stimulus-conditioned population context is identical to population context, except that the full 2 seconds of video are also provided as input. Again, Schmidt et al. (2025) was used as a baseline.
Behavior prediction conditions on the activity of all neurons (without video) and simultaneously predicts all behavioral traces (i. e. pupil size, pupil location and running speed). CEBRA (Schneider et al., 2023) is used as a baseline for this regime.

We train all state-of-the-art baselines on the collection of eight mice, used in our smallest data-scaling experiment (Fig. 7) to reduce computational cost. Implementation details and hyperparameters for each baseline are provided in App. D. Consistent with SENSORIUM 2022/2023 competitions, we use single-trial correlation as an evaluation metric. Additionally we evaluate our model on the SENSORIUM 2023 competition test set, which allows direct comparison against the state of the art model of predicting mouse visual cortex responses from video stimuli. We use OmniMouse-80M, freeze the entire model, and train only the neuron and animal embeddings using the released training data of five mice provided by the competition.

6 Results: The benefits of scaling

Current neuronal-predictive models are not compute- or parameter-limited. Because collecting neuronal data is costly, we first asked if existing models are already limited by compute or parameters, or if more data would still improve performance. To answer this question, we trained models on all 323 sessions while scaling width and depth as in Tab. 1. We evaluated five neuronal response masking strategies (Fig. 5, top row): two based on response dynamics (forecasting and population context), two analogous variants that additionally condition on video (video-conditioned forecasting and video-conditioned population context), and one stimulus-driven strategy (video & behavior). For each strategy, models ranged from 1M to 300M parameters, and we tracked both test loss and single-trial correlation as a function of total compute (model FLOPs, excluding FLOPS of neuron-specific parameters). Performance improved across all neuronal prediction tasks as model size increased up to 80M parameters (Fig. 5). Beyond this point, gains were minimal, as loss curves saturated or overfit, indicating that current models are data-limited rather than compute- or parameter-limited.

Table 3: **Sensorium 2023 benchmark results.** Models with Σ suffix denote ensemble predictions. We use n=5 models in the OmniMouse ensemble, from different random seeds, which determines model initialization and ordering of training batches. $\uparrow$ indicates higher is better. We either run the full multi-modal training, or only train the model with a single masking condition(*Unimodal*) – predicting neuronal responses conditioned on behavior and visual stimulus – comparable to all other models of the competition.

Model	Training	Main track ↑	OOD track ↑
DwiseNeuro-Σ (Turishcheva et al., 2024)	end-to-end	0.291	0.221
OmniMouse-5M-Unimodal	end-to-end	0.288 ± .003	0.256 ± .002
OmniMouse-5M-Unimodal-Σ	end-to-end	0.332	0.296
OmniMouse-5M	end-to-end	0.295 ± .005	0.263 ± .003
OmniMouse-5M-Σ	end-to-end	0.327	0.293
OmniMouse-80M	frozen	0.313 ± .001	0.274 ± .001
OmniMouse-80M-Σ	frozen	0.327	0.288

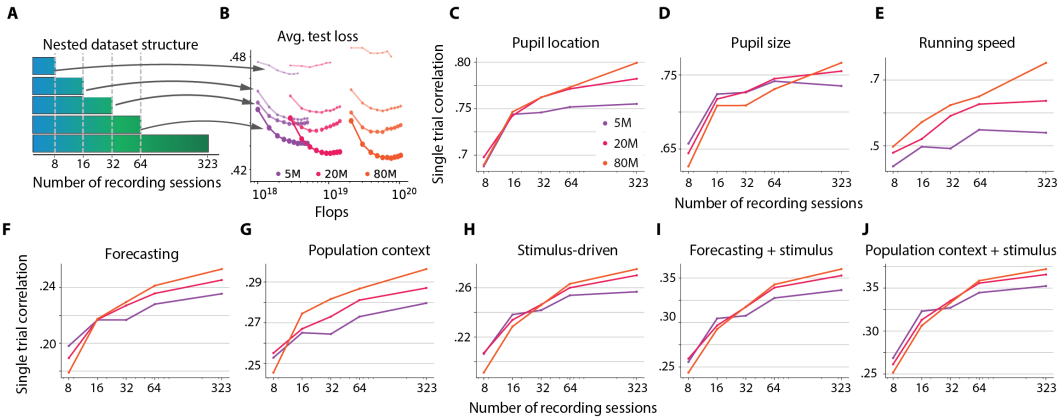

Figure 7: **Scaling data improves model performance.** **A**. Nested datasets structure. **B**. Test loss for different model and data sizes, averaged across all response prediction tasks. **C-K**. Performance improvements when scaling dataset from 8 to 323 sessions: **C**. Pupil center location. **D**. Pupil size and rate of pupil change. **E**. Running speed. **F**. Forecasting, prefix = 25 samples. **J**. Population context, context = 256 neurons. **H**. Stimulus-driven. **I**. Stimulus-conditioned forecasting, prefix = 25 samples. **K**. Stimulus-conditioned population context, context = 256 neurons.

OmniMouse achieves state-of-the-art performance. Our large-scale model outperforms all baselines across six evaluation regimes for both response prediction and behavior decoding (Tab. 2). Crucially, these gains are not simply due to training on more data: in data-matched comparisons, where OmniMouse and baselines are trained and evaluated on identical datasets, our model still outperforms strong specialized methods across nearly all tasks. This demonstrates that the architectural and masking design of OmniMouse provides advantages independent of data scale. We set a new state of the art on the Sensorium 2023 competition (Tab. 3), surpassing the winning entry on both the main and out-of-distribution (OOD) tracks in two evaluation settings: (1) with a frozen pretrained OmniMouse-80M backbone and only neuron-specific parameters trained, and (2) with full end-to-end training on the same 10-mouse competition dataset. In both cases, OmniMouse outperforms prior methods even without the ensembling strategies employed by competition entries. The improvements are particularly pronounced on the OOD track, which evaluates generalization to novel stimuli. We note that while our data-matched setting uses same mice sessions, our framework additionally enables training across video boundaries — an information not available for the previous models.

Behavior prediction shows the most promising scaling dynamics on the available data. To characterize the scaling of behavior prediction, we used the same models and evaluated their ability to predict pupil location, pupil size, and running speed from neuronal activity only (Fig. 6). Across all three settings, performance improved smoothly with compute budget, reminiscent of classic scaling-law behavior. Larger models consistently achieved higher single-trial correlations, albeit with an indication of saturation at the largest scale tested. Note, though, that training was stopped to avoid overfitting for the response prediction task. The models had not yet fully converged for the behavior prediction task and longer training could have improved performance further even on the largest model. OmniMouse not only matches, but surpasses the performance of all strong baselines such as CEBRA, particularly for running speed prediction, where correlation improves by over 0.15% relative to the baseline. These results show that behavioral prediction continues to improve with model scaling and may benefit from further increases in capacity.

Scaling dataset size improves performance. To study how dataset size affects performance, we trained three model sizes – 5M, 20M, and 80M – on nested collections of 8, 16, 32, 64, and 323 sessions such that the larger collections are supersets of the smaller ones (Fig. 7A). For evaluation, we test the model on the same held-out test set of the same seven mice that were contained in all collections (Fig. 7C–J). In all cases, performance improved with the number of sessions, exhibiting predictable data-scaling trends. Larger models consistently benefited more from additional data. The larger models required a minimum size of the training set to outperform the smaller models and the performance gap widened as the dataset increased in size. Behavior decoding benefited the most from data scaling (Fig. 7C–E), showing no saturation and large performance differences between 5M and 80M models. For responses, the strongest gains were observed for tasks that included video

input (Fig. 7C–E), where the 80M models continued to improve even beyond 100 sessions, suggesting that they remained data-limited rather than capacity-limited. The *forecasting* and *population context* showed bigger benefits from scaling of both data and model sizes. The gaps between 20M and 80M models (Fig. 7A, B) increased faster compared to the tasks with video input, which could indicate a lack of diversity of the visual stimuli in our dataset. Overall, these results highlight that scaling both model size and data quantity is synergistic and necessary to approach peak predictive performance.

OmniMouse enables systematic evaluation of how neuronal context shapes predictive performance. Lastly, we assessed the model's generalization by testing on masking conditions not seen during training, varying neuronal history duration (10–25 samples) and population context size (16–2048 neurons). Performance scaled smoothly with additional context demonstrating that OmniMouse learns generalizable representations that enables systematic analyses of contextual contributions to neural variability (see Fig. S2, Fig. S3, and App. B).

7 Discussion

In this work we introduce OmniMouse, a multi-modal, multi-task model of mouse visual cortex that integrates neural activity, video, and behavior across animals, making one step towards a foundation model of mouse vision. A single model achieves state-of-the-art performance on diverse tasks – predicting neural responses from visual stimuli, forecasting activity and decoding behavior. Trained on the largest neural dataset to date (3.3M neurons, 78 mice, 323 sessions), OmniMouse enables systematic study of scaling in brain models.

Our motivation for studying scaling laws is practical: if brain models are to become foundation models for neuroscience, it is essential to ask whether current data can sustain scaling. Despite using naturalistic movies and images, we find that performance saturates with model size, suggesting data – not compute – as the limiting factor. Even in the relatively simple mouse visual system, richer tasks, more varied stimuli, and larger-scale recordings are needed to support continued scaling. At the same time, relatively sparse sampling already yields strong models: with 60,000 neurons from just eight mice, predictive accuracy is high, likely due to redundancy in neural codes. Additional gains from larger datasets appear modest, paralleling language and vision models – yet in those domains, such small improvements have triggered phase transitions to qualitatively new abilities. By analogy, richer neuroscience data may similarly unlock new capabilities in brain models, revealing deeper principles of neural computation.

Limitations. Our work has several limitations. First, OmniMouse parameters scale linearly with the number of neurons, as it learns per-neuron embeddings. This makes training computationally prohibitively expensive may limit scaling to even larger datasets. Second, large-scale transformers remain difficult to interpret, and like deep learning models, they are prone to optimization issues and overparameterization, which constrain the biological insights that can be drawn. Furthermore, the behavioral data present in our data is limited to spontaneous activity and it is thus unclear if this approach can transfer to more complex behaviors.

Future work. Future work could extend to stimulus decoding (Benchetrit et al., 2023; Bauer et al., 2024; Zhu et al., 2025) and more precise study of training dynamics of modality interactions and multi-task learning to improve the masking recipe. Beyond calcium imaging in mouse visual cortex, models could integrate other data types such as electrophysiological recordings, diverse animal species, and more multi-modal stimuli such as audio. Alternatively, one could test generalization of the existing model across new tasks, stimuli, and species via (semi) closed-loop in-silico experiments (Ustyuzhaninov et al., 2022; Li et al., 2025), potentially finding biological insights about neuronal functional properties as in Walker et al. (2019); Li et al. (2025). Finally, jointly modeling visual input, neuronal responses, and behavior enables analysis of spontaneous and evoked activity (Stringer et al., 2019), revealing how brain state shapes sensory processing and core principles of computation.

Reproducibility Statement

To ensure the reproducibility of our results, we provide the complete source code for our multimodal model, including scripts for training, evaluation, fine-tuning, and inference, available at `https://anonymous.4open.science/r/unraveling-70BA/`. Additionally, the data-loading logic is provided at `https://anonymous.4open.science/r/experanto-iclr/`. Regarding the dataset, which consists of large-scale neuronal responses from the visual cortex and naturalistic visual stimulation, we have detailed the data acquisition and processing pipeline in Appendix E. While the full dataset is currently undergoing final preparation due to its unprecedented scale, we are committed to releasing it publicly within six months.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500, 2024.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In International Conference on Machine Learning, pp. 265–279. PMLR, 2023.

Dora Angelaki, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, Kcénia Bougrova, Sebastian A Bruijns, Matteo Carandini, Joana A Catarino, et al. A brain-wide map of neural activity during complex behaviour. Nature, 645(8079):177–191, 2025.

Ján Antolík, Sonja B Hofer, James A Bednar, and Thomas D Mrsic-Flogel. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. PLoS computational biology, 12(6):e1004927, 2016.

Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. Advances in Neural Information Processing Systems, 36:21895–21907, 2023.

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. Advances in Neural Information Processing Systems, 36:44937–44956, 2023.

Mehdi Azabou, Krystal Xuejing Pan, Vinam Arora, Ian Jarratt Knight, Eva L Dyer, and Blake Aaron Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In The Thirteenth International Conference on Learning Representations, 2025.

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. Advances in Neural Information Processing Systems, 34:15801–15815, 2021.

Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, EJ Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. In International Conference on Learning Representations, 2017.

Joel Bauer, Troy W Margrie, and Claudia Clopath. Movie reconstruction from mouse visual cortex activity. bioRxiv, pp. 2024–06, 2024.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. arXiv preprint arXiv:2310.19812, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

David H Brainard and Spatial Vision. The psychophysics toolbox. Spatial vision, 10(4):433–436, 1997.

Max F Burg, Santiago A Cadena, George H Denfield, Edgar Y Walker, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Learning divisive normalization in primary visual cortex. PLOS Computational Biology, 17(6):e1009028, 2021.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. PLOS Computational Biology, 15(4):e1006897, April 2019. doi: 10.1371/journal.pcbi.1006897.

Charles F Cadieu, Ha Hong, Daniel L K Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol., 10(12):e1003963, 2014.

Salvatore Calcagno, Isaak Kavasidis, Simone Palazzo, Marco Brondi, Luca Sità, Giacomo Turri, Daniela Giordano, Vladimir R. Kostic, Tommaso Fellin, Massimiliano Pontil, and Concetto Spampinato. Quantformer: Learning to quantize for neural activity forecasting in mouse visual cortex, 2024. URL https://arxiv.org/abs/2412.07264.

Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. bioRxiv, pp. 2023–09, 2023.

Team Chameleon. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv. org/abs/2405.09818, 9(8), 2024.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. ArXiv, pp. arXiv–2406, 2024.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and interpretable large-scale eeg foundation model. arXiv preprint arXiv:2401.10278, 2024.

BR Cowley and JW Pillow. High-contrast "gaudy" images improve the training of deep neural network models of visual cortex. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), Advances in Neural Information Processing Systems 33, pp. 21591–21603. Curran Associates, Inc., 2020.

Richard Csaky, Mats WJ van Es, Oiwi Parker Jones, and Mark Woolrich. Foundational gpt model for meg. arXiv preprint arXiv:2404.09256, 2024.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. arXiv preprint arXiv:2309.16588, 2023.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Tribe: Trimodal brain encoder for whole-brain fmri response prediction. arXiv preprint arXiv:2507.22229, 2025.

Saskia E. J. de Vries, Jerome A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel K. Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, Kate Roll, Marina Garrett, Tom Keenan, Leonard Kuan, Stefan Mihalas, Shawn Olsen, Carol Thompson, Wayne Wakeman, Jack Waters, Derric Williams, Chris Barber, Nathan Berbesque, Brandon Blanchard, Nicholas Bowles, Shiella D. Caldejon, Linzy Casal, Andrew Cho, Sissy Cross, Chinh Dang, Tim Dolbeare, Melise Edwards, John Galbraith, Nathalie Gaudreault, Terri L. Gilbert, Fiona Griffin, Perry Hargrave, Robert Howard, Lawrence Huang, Sean Jewell, Nika Keller, Ulf Knoblich, Josh D. Larkin, Rachael Larsen, Chris Lau, Eric Lee, Felix Lee, Arielle Leon, Lu Li, Fuhui Long, Jennifer Luviano, Kyla Mace, Thuyanh Nguyen, Jed Perkins, Miranda Robertson, Sam Seid, Eric Shea-Brown, Jianghong Shi, Nathan Sjoquist, Cliff Slaughterbeck, David Sullivan, Ryan Valenza, Casey White, Ali Williford, Daniela M. Witten, Jun Zhuang, Hongkui Zeng,

Colin Farrell, Lydia Ng, Amy Bernard, John W. Phillips, R. Clay Reid, and Christof Koch. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nature Neuroscience, 23(1):138–151, December 2019. doi: 10.1038/s41593-019-0550-9. URL https://doi.org/10.1038/s41593-019-0550-9.

Zhiwei Ding, Dat T. Tran, Kayla Ponder, Zhuokun Ding, Rachel Froebe, Lydia Ntanavara, Paul G. Fahey, Erick Cobos, Luca Baroni, Maria Diamantaki, Eric Y. Wang, Andersen Chang, Stelios Papadopoulos, Jiakun Fu, Taliah Muhammad, Christos Papadopoulos, Santiago A. Cadena, Alexandros Evangelou, Konstantin Willeke, Fabio Anselmi, Sophia Sanborn, Jan Antolik, Emmanouil Froudarakis, Saumil Patel, Edgar Y. Walker, Jacob Reimer, Fabian H. Sinz, Alexander S. Ecker, Katrin Franke, Xaq Pitkow, and Andreas S. Tolias. Bipartite invariance in mouse primary visual cortex. bioRxiv, 2025a. doi: 10.1101/2023.03.15.532836. URL https://www.biorxiv.org/content/early/2025/04/19/2023.03.15.532836.

Zhuokun Ding, Paul G Fahey, Stelios Papadopoulos, Eric Y Wang, Brendan Celii, Christos Papadopoulos, Andersen Chang, Alexander B Kunin, Dat Tran, Jiakun Fu, et al. Functional connectomics reveals general wiring rule in mouse visual cortex. Nature, 640(8058):459–469, 2025b.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. Advances in Neural Information Processing Systems, 37:86048–86073, 2024.

Yu Duan, Hamza Tahir Chaudhry, Misha B Ahrens, Christopher D Harvey, Matthew G Perich, Karl Deisseroth, and Kanaka Rajan. Poco: Scalable neural forecasting through population conditioning. arXiv preprint arXiv:2506.14957, 2025.

Alexander S. Ecker, Fabian H. Sinz, Emmanouil Froudarakis, Paul G. Fahey, Santiago A. Cadena, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Andreas S. Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex. arXiv, 2018.

Paul G Fahey, Taliah Muhammad, Cameron Smith, Emmanouil Froudarakis, Erick Cobos, Jiakun Fu, Edgar Y Walker, Dimitri Yatsenko, Fabian H Sinz, Jacob Reimer, et al. A global map of orientation tuning in mouse visual cortex. BioRXiv, pp. 745323, 2019.

Emmanouil Froudarakis, Philipp Berens, Alexander S Ecker, R James Cotton, Fabian H Sinz, Dimitri Yatsenko, Peter Saggau, Matthias Bethge, and Andreas S Tolias. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. Nat. Neurosci., 17(6):851–857, June 2014.

Marina E Garrett, Ian Nauhaus, James H Marshel, and Edward M Callaway. Topography and areal organization of mouse visual cortex. J. Neurosci., 34(37):12587–12600, September 2014.

Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown, Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou, et al. Caiman an open source tool for scalable calcium imaging data analysis. elife, 8:e38173, 2019.

Abdulkadir Gokce and Martin Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. arXiv preprint arXiv:2411.05712, 2024.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 76232–76264. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8b970e15a89bf5d12542810df8eae8fc-Paper-Conference.pdf.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395, 2024.

Linxing Preston Jiang, Shirui Chen, Emmanuel Tanumihardja, Xiaochuang Han, Weijia Shi, Eric Shea-Brown, and Rajesh PN Rao. Data heterogeneity limits the scaling effect of pretraining neural data transformers. bioRxiv, pp. 2025–05, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765, 2024.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. Advances in Neural Information Processing Systems, 35:25586–25599, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale video classification with convolutional neural networks. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732, June 2014.

William F Kindel, Elijah D Christensen, and Joel Zylberberg. Using deep learning to probe the neural code for images in primary visual cortex. Journal of vision, 19(4):29–29, 2019.

Mario Kleiner, David Brainard, and Denis Pelli. What's new in psychtoolbox-3? 2007.

David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". Advances in neural information processing systems, 30, 2017.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. Frontiers in Human Neuroscience, 15:653659, 2021.

Trung Le and Eli Shlizerman. Stndt: Modeling neural population activity with spatiotemporal transformers. Advances in Neural Information Processing Systems, 35:17926–17939, 2022.

Bryan M Li, Isabel M Cornacchia, Nathalie L Rochefort, and Arno Onken. V1t: large-scale mouse v1 response prediction using a vision transformer. arXiv preprint arXiv:2302.03023, 2023.

Bryan M Li, Wolf De Wulf, Danai Katsanevaki, Arno Onken, and Nathalie LI Rochefort. Movie-trained transformer reveals novel response properties to dynamic stimuli in mouse visual cortex. bioRxiv, 2025. doi: 10.1101/2025.09.16.676524. URL https://www.biorxiv.org/content/early/2025/09/17/2025.09.16.676524.

Yamin Li, Ange Lou, Ziyuan Xu, Shengchao Zhang, Shiyu Wang, Dario Englot, Soheil Kolouri, Daniel Moyer, Roza Bayrak, and Catie Chang. Neurobolt: Resting-state eeg-to-fmri synthesis with multi-dimensional feature mapping. Advances in Neural Information Processing Systems, 37:23378–23405, 2024.

Isaac Lin, Tianye Wang, Shang Gao, Shiming Tang, and Tai Sing Lee. Incremental learning and self-attention mechanisms improve neural system identification. arXiv e-prints, pp. arXiv–2406, 2024.

Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Willeke, Akshay Jagadish, Eric Wang, Edgar Y. Walker, Santiago A Cadena, Taliah Muhammad, Erick Cobos, Andreas S. Tolias, Alexander S Ecker, and Fabian H. Sinz. Generalization in data-driven models of primary visual cortex. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=Tp7kI90Htd.

Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci., 21(9):1281–1289, September 2018.

Lane T McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A Baccus. Deep learning models of the retinal response to natural scenes. Adv. Neural Inf. Process. Syst., 29 (Nips):1369–1377, 2016.

Lu Mi, Trung Le, Tianxing He, Eli Shlizerman, and Uygar Sümbül. Learning time-invariant representations for individual neurons from population dynamics. Advances in Neural Information Processing Systems, 36:46007–46026, 2023.

MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. Nature, 640(8058):435–447, April 2025.

MICrONS Consortium, J Alexander Bae, Mahaly Baptiste, Agnes L Bodor, Derrick Brittain, Joann Buchanan, Daniel J Bumbarger, Manuel A Castro, Brendan Celii, Erick Cobos, Forrest Collman, Nuno Maçarico da Costa, Sven Dorkenwald, Leila Elabbady, Paul G Fahey, Tim Fliss, Emmanouil Froudakis, Jay Gager, Clare Gamlin, Akhilesh Halageri, James Hebditch, Zhen Jia, Chris Jordan, Daniel Kapner, Nico Kemnitz, Sam Kinn, Selden Koolman, Kai Kuehner, Kisuk Lee, Kai Li, Ran Lu, Thomas Macrina, Gayathri Mahalingam, Sarah McReynolds, Elanine Miranda, Eric Mitchell, Shanka Subhra Mondal, Merlin Moore, Shang Mu, Taliah Muhammad, Barak Nehoran, Oluwaseun Ogedengbe, Christos Papadopoulos, Stelios Papadopoulos, Saumil Patel, Xaq Pitkow, Sergiy Popovych, Anthony Ramos, R Clay Reid, Jacob Reimer, Casey M Schneider-Mizell, H Sebastian Seung, Ben Silverman, William Silversmith, Amy Sterling, Fabian H Sinz, Cameron L Smith, Shelby Suckow, Zheng H Tan, Andreas S Tolias, Russel Torres, Nicholas L Turner, Edgar Y Walker, Tianyu Wang, Grace Williams, Sarah Williams, Kyle Willie, Ryan Willie, William Wong, Jingpeng Wu, Chris Xu, Runzhe Yang, Dimitri Yatsenko, Fei Ye, Wenjing Yin, and Szi-Chieh Yu. Functional connectomics spanning multiple areas of mouse visual cortex. bioRxiv, pp. 2021.07.28.454025, July 2021.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. arXiv preprint arXiv:2402.06196, 2024.

M C Morrone, M Tosetti, D Montanaro, A Fiorentini, G Cioni, and D C Burr. A cortical area that responds specifically to optic flow, revealed by fMRI. Nat. Neurosci., 3(12):1322–1328, December 2000.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 16(5):1–72, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. December 2024.

Denis G Pelli. The videotoolbox software for visual psychophysics: transforming numbers into movies. Spatial vision, 10(4):437–442, 1997.

Nicolai Petkov and Easwar Subramanian. Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal gabor filters with surround inhibition. Biol. Cybern., 97(5-6):423–439, December 2007.

AJ Piergiovanni, Isaac Noble, Dahun Kim, Michael S Ryoo, Victor Gomes, and Anelia Angelova. Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26804–26814, 2024.

Paweł A Pierzchlewicz, Konstantin F Willeke, Arne F Nix, Pavithra Elumalai, Kelli Restivo, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Katrin Franke, Andreas S Tolias, and Fabian H Sinz. Energy guided diffusion for generating neurally exciting images. In Advances in Neural Processing Systems (NeurIPS 2023), pp. 2023.05.18.541176, May 2023.

Jacob Reimer, Emmanouil Froudarakis, Cathryn R Cadwell, Dimitri Yatsenko, George H Denfield, and Andreas S Tolias. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. Neuron, 84(2):355–362, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis., 115(3):211–252, December 2015.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In International conference on machine learning, pp. 29441–29454. PMLR, 2023.

Shreya Saha, Ishaan Chadha, et al. Modeling the human visual system: Comparative insights from response-optimized and task-optimized vision models, language models, and different readout mechanisms. arXiv preprint arXiv:2410.14031, 2024.

Finn Schmidt, Polina Turishcheva, Suhas Shrinivasan, and Fabian H. Sinz. Modeling dynamic neural activity by combining naturalistic video stimuli and stimulus-independent latent factors, 2025. URL https://arxiv.org/abs/2410.16136.

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

Noam Shazeer. GLU variants improve transformer. February 2020.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. arXiv preprint arXiv:2504.07951, 2025.

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. Advances in neural information processing systems, 31, 2018.

Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. elife, 5:e14472, 2016.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. Science, 364(6437):eaav7893, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, Gauri Ganjoo, Emmanuel Mignot, and James Zou. Sleepfm: Multi-modal representation learning for sleep across brain activity, ecg and respiratory signals. arXiv preprint arXiv:2405.17766, 2024.

Armin Thomas, Christopher Ré, and Russell Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. Advances in neural information processing systems, 35: 21255–21269, 2022.

Polina Turishcheva, Paul Fahey, Michaela Vystrčilová, Laura Hansel, Rachel Froebe, Kayla Ponder, Yongrong Qiu, Konstantin Willeke, Mohammad Bashiri, Ruslan Baikulov, et al. Retrospective for the dynamic sensorium competition for predicting large-scale mouse primary visual cortex activity from videos. Advances in Neural Information Processing Systems, 37:118907–118929, 2024.

Ivan Ustyuzhaninov, Max F Burg, Santiago A Cadena, Jiakun Fu, Taliah Muhammad, Kayla Ponder, Emmanouil Froudarakis, Zhiwei Ding, Matthias Bethge, Andreas S Tolias, et al. Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. BioRxiv, pp. 2022–02, 2022.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. Nat. Neurosci., 22(12):2060–2065, December 2019.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. arXiv preprint arXiv:2302.14367, 2023.

Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. Nature, 640(8058):470–477, 2025.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. October 2024.

Konstantin F Willeke, Paul G Fahey, Mohammad Bashiri, Laura Pede, Max F Burg, Christoph Blessing, Santiago A Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, et al. The sensorium competition on predicting large-scale mouse primary visual cortex activity. arXiv preprint arXiv:2206.08666, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. May 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. Advances in Neural Information Processing Systems, 36:78240–78260, 2023.

Joel Ye and Chethan Pandarinath. Representation learning for neural population activity with neural data transformers. arXiv preprint arXiv:2108.01210, 2021.

Joel Ye, Jennifer Collinger, Leila Wehbe, and Robert Gaunt. Neural data transformer 2: multi-context pretraining for neural spiking activity. Advances in Neural Information Processing Systems, 36: 80352–80374, 2023.

Joel Ye, Fabio Rizzoglio, Adam Smoulder, Hongwei Mao, Xuan Ma, Patrick Marino, Raeed Chowdhury, Dalton Moore, Gary Blumenthal, William Hockeimer, et al. A generalist intracortical motor decoder. bioRxiv, pp. 2025–02, 2025.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. Advances in Neural Information Processing Systems, 36:26304–26321, 2023.

Yimeng Zhang, T-S Tai Sing Lee, Ming Li, Fang Liu, Shiming Tang, Tai Sing, Lee Ming, Li Fang, Liu Shiming, T-S Tai Sing Lee, Ming Li, Fang Liu, and Shiming Tang. Convolutional neural network models of V1 responses to complex patterns. J. Comput. Neurosci., pp. 1–22, 2018.

Yizi Zhang, Yanchen Wang, Donato Jiménez-Benetó, Zixuan Wang, Mehdi Azabou, Blake Richards, Renee Tung, Olivier Winter, Eva Dyer, Liam Paninski, et al. Towards a" universal translator" for neural dynamics at single-cell, single-spike resolution. Advances in Neural Information Processing Systems, 37:80495–80521, 2024.

Yizi Zhang, Yanchen Wang, Mehdi Azabou, Alexandre Andre, Zixuan Wang, Hanrui Lyu, The International Brain Laboratory, Eva Dyer, Liam Paninski, and Cole Hurwitz. Neural encoding and decoding at scale, 2025. URL https://arxiv.org/abs/2504.08201.

Yu Zhu, Bo Lei, Chunfeng Song, Wanli Ouyang, Shan Yu, and Tiejun Huang. Multi-modal latent variables for cross-individual primary visual cortex modeling and analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 1228–1236, 2025.

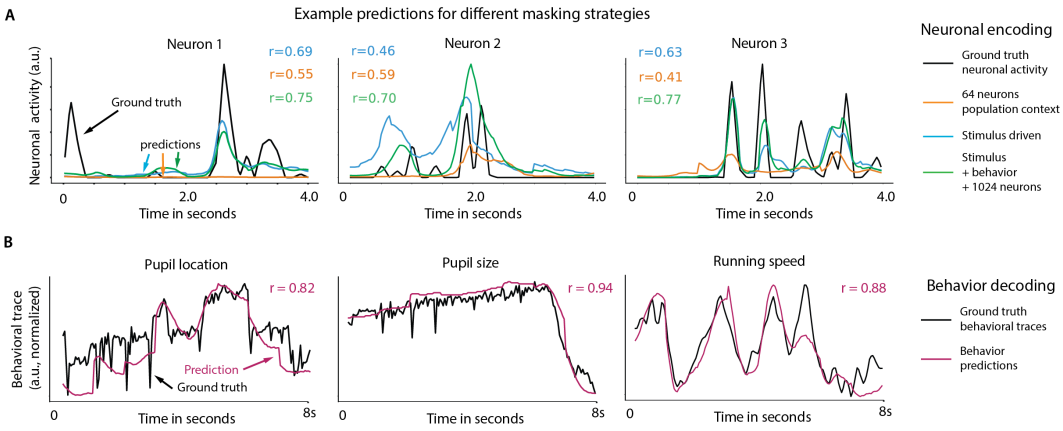

Figure S1: **Example predictions of neuronal activity and behavioral variables**. **A**. Here we show three example neurons and their ground truth neuronal activity for 4 seconds (black). We also show the model prediction of OmniMouse for three evaluation conditions: *population context of 64 neurons* (orange), *stimulus-driven* (blue), *stimulus + behavior + neuron context* (green). The predictive performance, shown as pearson correlation r is increasing with more information provided to the model. Our model is designed to disentangle the relative contributions of sensory input, behavior, and population dynamics to individual neurons' activity. **B**. Ground truth and predictions for behavioral variables.

A Qualitative visualizations

B Supplemental results

OmniMouse enables systematic evaluation of how neuronal context shapes predictive performance. We evaluated OmniMouse on conditions not seen during training, systematically varying neuronal history duration (10-25 samples) and population context size (16-2048 neurons) for population context tasks. Performance scaled smoothly with context availability across all conditions Fig. S2. When video was available, performance plateaued more quickly for forecasting but continued to improve for population context, suggesting that nearby neurons carry complementary information beyond visual input. These systematic evaluations demonstrate that OmniMouse has learned generalizable representations of neural variability, enabling quantitative assessment of how different sources of context—temporal history contribute to explaining variability in neural responses.

Furthermore, we hypothesized that harder tasks might benefit more from scaling, as shown for large language models (Minaee et al., 2024; Naveed et al., 2025). To test this, we varied the neuronal history duration (*full-population prefix* $\in [10, 15, 20, 25]$) for forecasting tasks and context size (*context* $\in [16, 32, ..., 1024, 2048]$) for population context tasks, where shorter contexts represent harder tasks. We also compared performance with and without 2 seconds of video input. Fig. S2 confirms our hypothesis: performance improves consistently as context grows, hence, bigger context indicates easier task. Non-video conditioned regimes scale more steeply, likely due to lower baselines. For forecasting, they never match video-conditioned models, since video provides temporal information unavailable at prediction. For population context, however, sufficient neural responses recover enough information to match video performance. However, contrary to LLMs, in our case scaling does not preferentially benefit harder tasks: across all tasks, curves for different model sizes remain parallel. If harder tasks gained more, larger models (20–80M) would show bigger advantages over smaller ones (1–5M) at minimal context.

C Relation to other neuroscience scaling.

This is the first study to systematically scale both model and data size using only neuro-data, yet our findings align with prior neuro-scaling work. Consistent with Gokce & Schrimpf (2024), behavior prediction improves with larger models, and the greater gains from joint model–data scaling

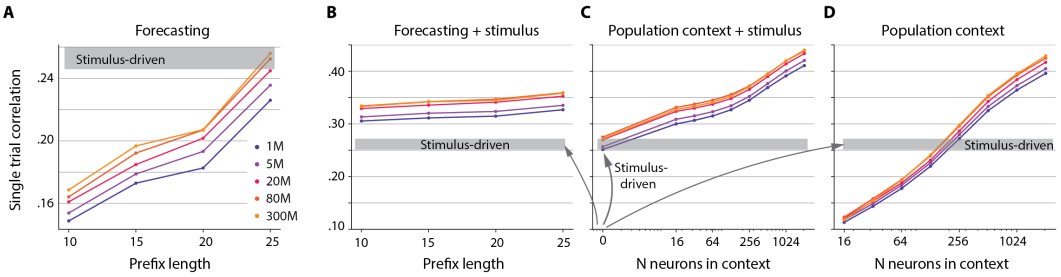

Figure S2: **Using the models capabilities to investigate context lengths for forecasting and population context tasks. A**. Forecasting with a change of prefix length, i.e. how many samples of the full population are unmasked. A prefix length of 10 corresponds to one third of a second of neuronal activity.. **B**. Same change of forecasting context as **A**, but with video. **C**. Performance improvements in addition to video with population context. # neurons in context = 0 means that all neurons are masked, and the model conditions its prediction purely on the visual stimulus. In all panels, the stimulus-driven performance is denoted as the gray box for ease of comparison. Remarkably, as seen in panel A, forecasting a whole second of neuronal activity given the past second (i.e. prefix length = 25) yields to the same performance as showing the entire video. Context increases from 0-16-...-2048. **D**. Population context only, 16 - 2048 neurons

on non-video tasks (Fig. 7A,B) support claims from Jiang et al. (2025); Ye et al. (2025) that data heterogeneity limits scaling: our visual stimuli include many repeats, while neural responses vary with latent brain state and noise even when the visual stimuli is same.

D BASELINES

To establish baseline comparisons while managing computational costs, we train state-of-the-art baseline models on the smallest nested dataset containing eight mice (the seven evaluation mice plus one additional training mouse). This approach ensures that all methods are compared under identical conditions while keeping baseline training tractable. We train all baselines on 8 recordings from 8 unique mice – 5 fully released mice from the sensorium 2023 competition (keeping the original train-validation-test splits), 2 mice from the sensorium 2022 competition that were used for the test split. session from the MiCRONs collection. The same 8 mice were used in the smallest scaling experiment.

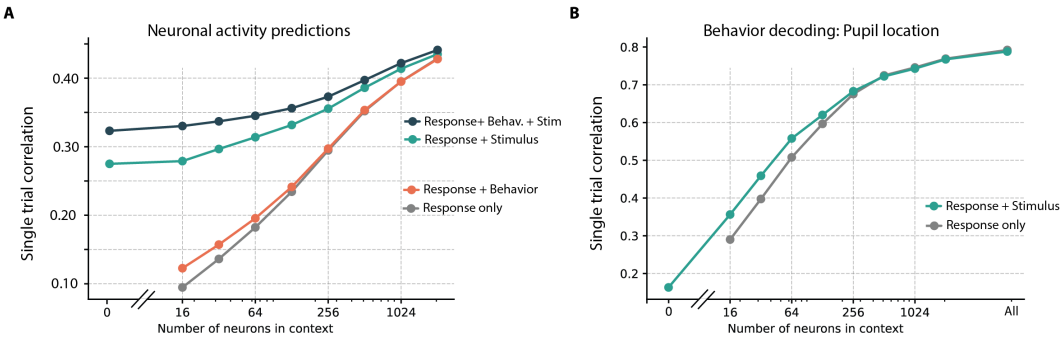

Figure S3: **Systematic evaluation across mask configurations**. We evaluate the neuronal prediction and behavior decoding performance of OmniMouse-80M by systematically varying the model inputs via masking. Only masks using N = [64, 256, 1024] have been seen during training. OmniMouse generalizes to unseen conditions, and allows to systematically study the contribution of visual stimulus, behavioral variables, and neuronal (sub)-population activity. **A**. Neuronal activity predictions given different amounts of visible neurons in context. **B**. Behavioral decoding.

D.1 CEBRA

CEBRA explanation: CEBRA performs dimensionality reduction on neural activity using InfoNCE contrastive learning, where positive and negative pairs are defined by auxiliary variables such as time or behavior. When the auxiliary variable is discrete, for example a left or right wheel turn, it selects positives uniformly from all samples with the same label. When the variable is continuous, such as running speed or pupil direction, it chooses a random point within a time window around the sample and then find the closest match in the dataset using either Euclidean or cosine distance; this sample becomes the positive pair, which adds diversity and prevents repeatedly selecting the same example. Negative pairs are sampled randomly. For decoding, CEBRA encode neural responses, find the nearest latent vectors for responses in the training set, and returns their associated behavioral variables as predictions.

Model hyperparameters: We trained a joint model for 8 mice, using a batch size of 512 and learning rate of $3 \cdot 10^{-4}$. The network contained 256 hidden units and produced 128-dimensional outputs (both doubled relative to the Allen example `https://cebra.ai/docs/demo_notebooks/Demo_Allen.html`). Training ran for up to 50,000 iterations with cosine distance as the loss metric. The model used a temperature of 1, time-delta conditioning to enable behavior mode, and time offsets of 5. As CEBRA requires same frequencies between responses and behavior, both were resamples to 20 Hz, in order to compute correlation on the same predictions as for the OmniMouse. Please note that downsampling from 30 Hz responses is not reducing any information as responses were upsampled from 6-16 Hz to 30 Hz and the upsampling is done with nearest-neighbor interpolation.

D.2 Universal Spike Transalator

Universal Spike Translator explaination: The Universal Spike Translator Zhang et al. (2024) performs a self-supervised modeling approach called multi-task-masking (MtM). The model alternates between masking out and reconstructing neural activity across different time steps and neurons. It uses a learnable token that provides the model with context about the specific masking scheme that is being applied during training, allowing for "mode switching" at test time for different downstream tasks. During training, the masking schemes are sampled randomly which are: **(1) Neuron masking:** Randomly masks individual neurons and reconstructs their activity using the unmasked neurons as context. **(2) Causal masking:** Masks future time steps and predicts them using the past steps as context.

Model hyperparameters: We used the default hyperparameters from "ndt1_stitching_prompting" and "ssl_session_trainer" configs from `https://github.com/colehurwitz/IBL_MtM_model`. Please note that compared to our forecasting settings, IBL does not take behavior as model input.

D.3 Latent dynamic model

Latent dynamic model explanation: This is a probabilistic model that predicts the joint distribution of neuronal responses from naturalistic video stimuli and stimulus-independent latent factors. Specifically, the model predicts time-varying neuronal response using a Zero-Inflated-Gamma (ZIG) distribution to model the distribution of neuronal responses conditioned on the stimulus and the latent factor. This is a modification of the deterministic factorized 3D convolutional core and a Gaussian readout, where we have an additional encoder that takes a subset of neurons as input to derive a latent variable. This latent variable is then combined with the transformed visual input to predict the activity of other neurons in the session. The model is trained by maximizing the Evidence Lower Bound (ELBO) of $p_{ZIG}(y|x)$ via variational inference.

Model hyperparameters: For both SENSORIUM 2023 baseline and Schmidt et al. (2025) baseline we used the default hyperparameters from Schmidt et al. (2025): 3 layer core with both spatial and temporal kernel = 11 in the first layer and 5 on the layer two and three. For more details see App.C from Schmidt et al. (2025). All data modalities were upsampled to 30 Hz as both SENSORIUM 2023 baseline and Schmidt et al. (2025) latent model require all modalities to have the same frequencies. Both SENSORIUM 2023 baseline and Schmidt et al. (2025) latent model predict 42 samples from a 60-frame video input, we always used only last 30 frames for evaluation, to make it consistent with OmniMouse, who was trained to predict 30 samples. Please note that OmniMouse suppost flexible

size of predictions, while SENSORIUM 2023 baseline and Schmidt et al. (2025) latent model cannot do it.

D.4 Implementation details

D.4.1 Masking strategies used during training

Mask	Behavior	Video (last frames visible)	Visible Neurons	Context (from → to)	Prefix (from → to)	Predicted Behavior
1–3	✓	0	[64, 256, 1024]	0 → 60	—	
4	✗	0	4096	0 → 60	—	✓
5–7	✗	0	[64, 256, 1024]	0 → 60	—	✓
8–19	✓	0	[64, 256, 1024, 4096]	—	[0, 10, 15] → 25	
20–28	✓	0	[64, 256, 1024]	25 → 60	[0, 10, 15] → 25	
29–37	✗	0	[64, 256, 1024]	25 → 60	[0, 10, 15] → 25	✓
38–40	✓	10	[64, 256, 1024]	10 → 60	—	
41–52	✓	10	[64, 256, 1024, 4096]	—	[0, 10, 15] → 25	
53–58	✓	10	[64, 256, 1024, 4096]	25 → 60	[10, 15] → 25	
59–61	✓	20	[64, 256, 1024]	20 → 60	—	
62–73	✓	20	[64, 256, 1024, 4096]	—	[0, 10, 15] → 25	
74–79	✓	20	[64, 256, 1024]	25 → 50	[10, 15] → 25	
80–82	✓	20	[64, 256, 1024]	30 → 60	—	
83–94	✓	30	[64, 256, 1024, 4096]	—	[0, 10, 15] → 25	
95–100	✓	30	[64, 256, 1024]	25 → 40	[10, 15] → 25	
101–103	✓	40	[64, 256, 1024]	30 → 50	—	
104–111	✓	40	[64, 256, 1024, 4096]	—	[10, 15] → 25	
112–114	✓	50	[64, 256, 1024]	30 → 40	—	
115–118	✓	50	[64, 256, 1024, 4096]	—	10 → 20	
119	✓	60	—	—	—	

Table 4: **Summary of training mask configurations**. In each batch all behavior traces for the whole 2 seconds were either given as input or predicted. For each batch 4096 neurons were randomly sampled from N neurons per mouse and last second (30 responses) for 3072 neurons of these 4096 the activity was predicted.

D.4.2 Nested scaling dataset construction

The nested dataset was constructed such that for the 7 mice we conducted evaluation on - 3 mice we had repeated sessions, such that the number of repeats grew proportionally to the dataset growth, and 4 other mice had a single session. As session-per-mice distribution is highly skewed, the other sessions were samples randomly.

D.5 Distribution of sessions per mouse

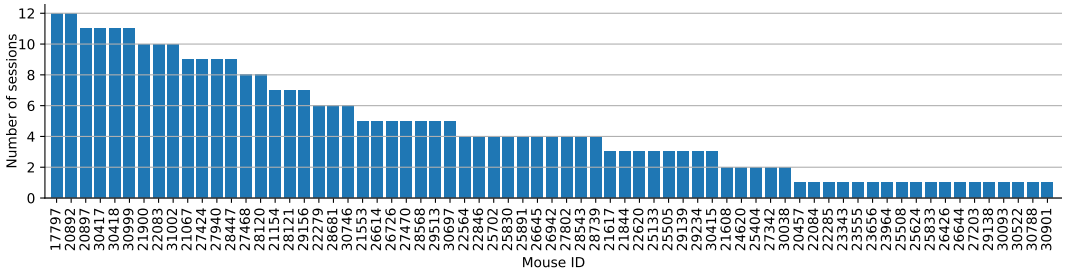

Figure S4: **Distribution of 316 sessions across 69 mice.** More than 100 sessions come from first 10 mice.

E NEUROPHYSIOLOGICAL EXPERIMENTS

Model evaluation was performed on neurophysiological data from Sensorium 2022 ((Willeke et al., 2022), Mouse 1 and 2, evaluation animals for Sensorium and Sensorium Plus tracks) and Sensorium 2023 ((Turishcheva et al., 2024), all animals). Model training was performed on historical data, including data from MICrONS Consortium (2025), Wang et al. (2025), Ding et al. (2025b), Ding et al. (2025a), Fahey et al. (2019), Willeke et al. (2022),Turishcheva et al. (2024), but also included data not previously published.

All procedures were approved by the Institutional Animal Care and Use Committee of Baylor College of Medicine. Seventy-eight mice (Mus musculus, 32 females, 46 males, P50–155 on day of first scan) expressing GCaMP6s in excitatory neurons via Slc17a7-Cre and Ai162 transgenic lines (recommended and generously shared by Hongkui Zeng at Allen Institute for Brain Science; Jackson Labs stock 023527 and 031562, respectively) were anesthetized and a 4 mm craniotomy was made over the visual cortex of the right hemisphere as described previously (Reimer et al., 2014; Froudarakis et al., 2014). In two of the seventy-six animals, GCaMP6s was additionally expressed in inhibitory neurons via DLX5-CreER (Jackson Labs stock 010705), following treatment with tamoxifen (orogastric gavage of tamoxifen (Sigma Aldritch T5648) dissolved in corn oil (Sigma Aldritch C8267) at 15 mg/mL, 200 mg/kg body weight, two doses two days apart, second dose >= 13 days before the first included scan).

Mice were head-mounted above a cylindrical treadmill and calcium imaging was performed using Chameleon Ti-Sapphire laser (Coherent) tuned to 920 nm and a large field of view mesoscope (Sofroniew et al., 2016) equipped with a custom objective (excitation NA 0.6, collection NA 1.0, 21 mm focal length). Laser power after the objective was increased exponentially as a function of depth from the surface according to:

$$P = P_0 \times e^{(z/L_z)} \tag{1}$$

Here P is the laser power used at target depth z, P0 is the power used at the surface (typically not exceeding 25 mW), and L_z is the depth constant (160-220 μm). The greatest laser output of ca. 112 mW was used at approximately 400-500 μm from the surface.

The craniotomy window was leveled with regards to the objective with six degrees of freedom. Pixel-wise responses from an ROI spanning the cortical window (1.7-4 mm diameter FOV, >0.2 px/μm, superficial cortex, >2.47 Hz) to drifting bar stimuli were used to generate a sign map for delineating visual areas (Garrett et al., 2014). In some but not all cases where the imaging field of view spanned multiple areas, area boundaries on the sign map were manually annotated. Imaging FOV of varying dimensions were targeted to lie within the boundaries of visual cortex, and may span between primary visual cortex and surrounding higher visual areas depending on the scan design.

Scan dimensions typically fell into one of three categories. Local field of view scans contained multiple imaging planes at different depths (10-13 planes, most commonly with 5 μm z spacing but ranging between 3 and 45 μm z spacing), with each plane spanning 600-630 × 600-630 μm (240-252 × 240-252 pixels, 0.4 px/μm resolution), acquired most commonly at 7.98 Hz (range 4.34-8.31 Hz). Large field of view scans contained single imaging planes at a single depth, with each plane scanning 1.5 - 3 mm diameter (0.33 - 0.4 px/um resolution), acquired at between 6.5 - 12.4 Hz. In between are scans containing multiple imaging planes at different depths (2-5 planes, with variable interplane spacing between 5 and 150 μm), with each plane spanning approximately 0.8-1.2 mm diameter (0.4-0.6 px/μm resolution), acquired at between 6.3 and 9.6 Hz. Scans with multiple planes, especially at high sampling densities (ex. 5 μm z spacing), have a high likelihood of multiple segmented traces emerging from multiple planes intersecting with the soma of a single neuron in a single scan. Multiple scans were also often collected from the same animal, and as a result single biological neurons may be recorded across multiple scans.

Movie of the animal's eye and face was captured throughout the experiment. A hot mirror (Thorlabs FM02) positioned between the animal's left eye and the stimulus monitor was used to reflect an IR image onto a camera (Genie Nano C1920M, Teledyne Dalsa) without obscuring the visual stimulus. The position of the mirror and camera were manually calibrated per session and focused on the pupil. Field of view was manually cropped for each session to contain the left eye in its entirety, although across different experiments the field of view may have additionally contained more or less of the

face, centered or not centered on the eye, or characterized the pupil at different resolutions. Video was captured at ca. 20 Hz. Frame times were time stamped in the behavioral clock for alignment to the stimulus and scan frame times. Video was compressed using Labview's MJPEG codec with quality constant of 600 and stored in an AVI file.

Light diffusing from the laser during scanning through the pupil was used to capture pupil diameter and eye movements. A DeepLabCut model (Mathis et al., 2018) was trained as previously described (Turishcheva et al., 2024) on 17 manually labeled samples from 11 animals to label each frame of the compressed eye video (intraframe only H.264 compression, CRF:17) with 8 eyelid points and 8 pupil points at cardinal and intercardinal positions. Pupil points with likelihood >0.9 were fit with the smallest enclosing circle, and the radius and center of this circle was extracted. Frames with < 3 pupil points with likelihood >0.9, or producing a circle fit with outlier > 5.5 standard deviations from the mean in any of the three parameters (center x, center y, radius) were discarded. Gaps in behavior were replaced by linear interpolations over the whole session, if there were more than 2 frames with gaps, then the video is removed.

The mouse was head-restrained during imaging but could walk on a treadmill. Rostro-caudal treadmill movement was measured using a rotary optical encoder (Accu-Coder 15T-01SF-2000NV1ROC-F03-S1) with a resolution of 8000 pulses per revolution, and was recorded at approx. 50-100 Hz in order to extract locomotion velocity.

Visual stimuli were presented with Psychtoolbox 3 in MATLAB (Brainard & Vision, 1997; Kleiner et al., 2007; Pelli, 1997) to the left eye with a 31.8 $\times$ 56.5 cm (height $\times$ width) monitor (ASUS PB258Q) with a resolution of 1080$\times$1920 pixels positioned 15 cm away from the eye. When the monitor is centered on and perpendicular to the surface of the eye at the closest point, this corresponds to a visual angle of 3.8 °/cm at the nearest point and 0.7 °/cm at the most remote corner of the monitor. As the craniotomy coverslip placement during surgery and the resulting mouse positioning relative to the objective is optimized for imaging quality and stability, uncontrolled variance in animal skull position relative to the washer used for head-mounting was compensated with tailored monitor positioning on a six dimensional monitor arm. The pitch of the monitor was kept in the vertical position for all animals, while the roll was visually matched to the roll of the animal's head beneath the headbar by the experimenter. In order to optimize the translational monitor position for centered visual cortex stimulation with respect to the imaging field of view, we used a dot stimulus with a bright background (maximum pixel intensity) and a single dark square dot (minimum pixel intensity). Dot locations were randomly ordered from a grid tiling a portion of the screen, either a 10 $\times$ 10 grid tiling a central square (approx. 90° width and height, 10 repeats per location, 200-300 ms presentation at each location), or a 5 $\times$ 8 grid tiling the majority of the monitor (approx. 93° height and 119° width, 20 repeats per location, 200 ms presentation at each location). The final monitor position for each animal was chosen in order to center the population receptive field of the scan field ROI on the monitor, with the yaw of the monitor visually matched to be perpendicular to and 15 cm from the nearest surface of the eye at that position.

A photodiode (TAOS TSL253) was sealed to the top left corner of the monitor, and the voltage was recorded at 10 kHz and timestamped on the behavior clock (MasterClock PCIe-OSC-HSO-2 card). Simultaneous measurement with a luminance meter (LS-100 Konica Minolta) perpendicular to and targeting the center of the monitor was used to generate a lookup table for linear interpolation between photodiode voltage and monitor luminance in cd/m² for 16 equidistant values from 0-255, and one baseline value with the monitor unpowered.

At the beginning of each experimental session, we collected photodiode voltage for 52 full-screen pixel values from 0 to 255 for one second trials. The mean photodiode voltage for each trial V_{pd} was fit as a function of the pixel intensity V_{in}:

$$V_{pd} = B + A \times V_{in}^{\gamma} \tag{2}$$

in order to estimate the γ value of the monitor ($\approx 1.50 - 1.76$). All stimuli were shown with no γ correction.

During the stimulus presentation, sequence information was encoded in a 3 level signal according to the binary encoding of the flip number assigned in-order. This signal underwent a sine convolution, allowing for local peak detection to recover the binary signal. A linear fit was applied to the trial

timestamps in the behavioral and stimulus clocks, and the offset of that fit was applied to the data to align the two clocks, allowing linear interpolation between them. The mean photodiode voltage of the sequence encoding signal at pixel values 0 and 255 was used to estimate the luminance range of the monitor during the stimulus, with typical maximum values of approx. 10-12 cd/m².