Multimodal Deep Learning Model Unveils Behavioral Dynamics of V1 Activity in Freely Moving Mice

Aiwen Xu
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93117
aiwenxu@ucsb.edu

Yuchen Hou
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93117
yuchenhou@ucsb.edu

Cristopher M. Niell
Department of Biology, Institute of Neuroscience
University of Oregon
Eugene, OR 97403
cniell@uoregon.edu

Michael Beyeler
Department of Computer Science,
Department of Psychological and Brain Sciences
University of California, Santa Barbara
Santa Barbara, CA 93117
mbeyeler@ucsb.edu

Abstract

Despite their immense success as a model of macaque visual cortex, deep convolutional neural networks (CNNs) have struggled to predict activity in visual cortex of the mouse, which is thought to be strongly dependent on the animal's behavioral state. Furthermore, most computational models focus on predicting neural responses to static images presented under head fixation, which are dramatically different from the dynamic, continuous visual stimuli that arise during movement in the real world. Consequently, it is still unknown how natural visual input and different behavioral variables may integrate over time to generate responses in primary visual cortex (V1). To address this, we introduce a multimodal recurrent neural network that integrates gaze-contingent visual input with behavioral and temporal dynamics to explain V1 activity in freely moving mice. We show that the model achieves state-of-the-art predictions of V1 activity during free exploration and demonstrate the importance of each component in an extensive ablation study. Analyzing our model using maximally activating stimuli and saliency maps, we reveal new insights into cortical function, including the prevalence of mixed selectivity for behavioral variables in mouse V1. In summary, our model offers a comprehensive deep-learning framework for exploring the computational principles underlying V1 neurons in freely-moving animals engaged in natural behavior.

1 Introduction

Computational models have been crucial in providing insight into the underlying mechanisms by which neurons in the visual cortex respond to external stimuli. Deep convolutional neural networks (CNNs) have had immense success as predictive models of the primate ventral stream, in cases where

Preprint. Under review.

the animal was passively viewing stimuli or simply maintaining fixation [1–5]. Despite their success, these CNNs are poor predictors of neural responses in mouse visual cortex [6], which is thought to be shallower and more parallel than that of primates [7, 8]. According to the best models in the literature [9–14], the mouse visual system is more broadly tuned and operates on relatively low-resolution inputs to support a variety of behaviors [15]. However, these models were limited to predicting neural responses to controlled (and potentially ethologically irrelevant) stimuli that were passively viewed by head-fixed animals.

Movement is a critical element of natural behavior. In the visual system, eye and head movements during locomotion and orienting transform the visual scene in potentially both beneficial (e.g., by providing additional visual cues) and detrimental ways (e.g., by introducing confounds due to self-movement) [16, 17]. Movement-related activity is widespread in mouse cortex [18, 19] and prevalent in primary visual cortex (V1) [20, 21]. For instance, V1 neurons of freely moving mice show robust responses to head and eye position [22, 23], which may contribute a multiplicative gain to the visual response [24] that cannot be replicated under head fixation. V1 activity may be further modulated by variables that depend on the state of the animal and its behavioral goals [19, 21, 25, 26]. However, how these behavioral variables may integrate to modulate visual responses in V1 is unknown. Furthermore, a comprehensive predictive model of V1 activity in freely moving animals is still lacking.

To address these challenges, we make the following contributions:

- We introduce a multimodal recurrent neural network that integrates gaze-contingent visual input with behavioral and temporal dynamics to explain V1 activity during natural vision in freely moving mice.
- We show that the model achieves state-of-the-art predictions of V1 activity during free exploration based on visual input and behavior, demonstrating the ability to accurately model neural responses in the dynamic regime of movement through the visual scene.
- We uncover new insights into cortical neural coding by analyzing our model with maximally activating stimuli and saliency maps, and demonstrate that mixed selectivity of visual and behavioral variables is prevalent in mouse V1.

2 Related Work

Despite their success in predicting neural activity in the macaque visual cortex, deep CNNs trained on ImageNet have had limited success in predicting mouse visual cortical activity [6]. This is perhaps not surprising, as most ImageNet stimuli belong to static images of human-relevant semantic categories and may thus be of low ethological relevance for rodents. More importantly, these deep CNNs may not be the ideal architecture to model mouse visual cortex, which is known to be shallower and more parallel than primate visual cortex [27, 28]. In addition, mice are known to have lower visual acuity than that of primates [7, 8], and much of their visual processing may be devoted to active, movement-based behavior rather than passive analysis of the visual scene [21, 29, 30]. Although the majority of V1 neurons is believed to encode low-level visual features [31], their activity is often strongly modulated by behavioral variables related to eye and head position [22–24], locomotion [17, 20, 21], arousal [26, 32], and the recent history of the animal [25]. Furthermore, mouse V1 is highly interconnected with both cortical and subcortical brain areas, which contrasts with feedforward, hierarchical models of visual processing [21].

A common architectural approach that has proved quite successful is to split the network into different components (first introduced by [11]):

- a “core” network, which typically consist of a CNN used to extract convolutional features from the visual stimulus [11, 12, 24, 33], sometimes in combination with a recurrent network [11];
- a “shifter” network, which mimics gaze shifts by learning a (typically affine) transformation from head- to eye-centered coordinates, either applied to the pixel input [11, 24] or a CNN layer [12];
- a “readout” network, which learns a mapping from artificial to biological neurons [11, 12, 33].

Owing to the difficulty of developing a predictive model of mouse cortex, Willeke *et al.* [14] recently invited submissions to the Sensorium competition held at NeurIPS ’22. The competition introduced a benchmark dataset of V1 neural activity recorded from head-fixed mice on a treadmill viewing static images, with simultaneous measurements of running speed, pupil size, and eye position. A baseline model was provided as well, which consisted of a 4-layer CNN core in combination with a shifter and

readout network [12]. Even though 26 teams submitted 194 different models, the overall improvement to the baseline performance was modest, raising the single trial correlation from .287 to .325 in the Sensorium and from .384 to .453 in the Sensorium+ competition. Architectural innovations (e.g., Transformers, Normalizing Flows, YOLO, and knowledge distillation), were unable to make an impact, as most improvements were gained from ensemble methods. A promising direction was taken by the winning model, which attempted to learn a latent representation of the "brain state" from the various behavioral variables, inspired by [19]. However, the model utilized the timestamps of the test set to estimate recent neuronal activities, which the other competitors did not have access to.

Taken together, we identified three main limitations of previous work that this study aims to address:

- **Head-fixed preparations.** Most previous models operated on data from animals in head-fixed conditions with static stimuli, which do not mirror natural behavior and thus provide limited insight into visual processing in real-world environments. In contrast, the present work is applied to state-of-the-art neurophysiological recordings of V1 activity in freely moving mice. This represents a dramatic shift in the "parameter space" of visual input, from static images to dynamic, real-world visual input. One could imagine that this will make the modeling process more difficult, because the stimulus set is more complex, or easier, because it is more matched to the computational challenge the brain evolved for.
- **Limited influence of behavioral state.** Previous models often limited the influence of behavioral state to eye measurements and treadmill running speed, which were either concatenated with the visual features [14, 32], utilized in the shifter network to determine the gaze-contingent retinal input [11, 14], or used to predict a multiplicative gain factor [11].
- **Missing temporal dynamics.** Most previous modeling works ignored the temporal factors that might influence V1 activity and overlooked the dynamic nature of visual processing (but see [11]). We overcome this limitation by utilizing approximately 1-hour-long recordings of three mice freely exploring an arena, and our model is capable of handling continuous data streams of any length.

3 Methods

Head-mounted recording system We had access to data from three adult mice who were freely exploring $48\,\mathrm{cm}$ long $\times$ $37\,\mathrm{cm}$ wide $\times$ $30\,\mathrm{cm}$ high arena (Fig. 1A), collected with a state-of-the-art recording system [24] that combined high-density silicon probes with miniature head-mounted cameras (Fig. 1B). One camera was aimed outwards to capture the visual scene from the mouse's perspective ("worldcam") at $16\,\mathrm{ms}$ per frame (downsampled to 60×80 pixels). A second camera, aimed at the eye, was used to extract eye position (θ, ϕ) and pupil radius (σ) at $30\,\mathrm{Hz}$ using DeepLabCut [34], which allowed for the worldcam video to be corrected for eye movements (see [24] for details). Pitch (ρ) and roll (ω) of the mouse's head position were extracted at $30\,\mathrm{kHz}$ from the inertial measurement unit (IMU). Locomotion speed (s) was estimated from the top-down camera feed using DeepLabCut [34]. Electrophysiology data was acquired at $30\,\mathrm{kHz}$ using a $11\,\mu\mathrm{m} \times 15\,\mu\mathrm{m}$ multi-shank linear silicon probe (128 channels) implanted in the center of the left monocular V1, then bandpass-filtered between $0.01\,\mathrm{Hz}$ and $7.5\,\mathrm{kHz}$, and spike-sorted with Kilosort 2.5 [35]. Single units were selected using Phy2 (see [36]) and inactive units (mean firing rate $< 3\,\mathrm{Hz}$) were removed. This yielded 68, 32, and 49 active units for Mouse 1–3, respectively. To prepare the data for machine learning, all data streams were deinterlaced and resampled at $20.83\,\mathrm{Hz}$ ($48\,\mathrm{ms}$ per frame; Fig. 1C).

Model architecture We used a 3-layer CNN (kernel size 7, $128 \times 64 \times 32$ channels) to encode the visual stimulus. Each convolutional layer was followed by a BatchNorm layer, a ReLU, and a Dropout layer (0.5 rate). A fully-connected layer transformed the learned visual features into a visual feature vector, $\boldsymbol{v}$ (Fig. 2, *top-right*). In a purely visual version of the model, $\boldsymbol{v}$ was fed into a fully-connected layer, followed by a softplus layer, to yield a neuronal response prediction.

To encode behavioral state, we constructed an input vector from different sets of behavioral variables:

- $\mathcal{S}$: all behavioral variables used in the Sensorium+ competition [14], consisting of running speed (s), pupil size (σ), and its temporal derivative ($\dot{\sigma}$);
- $\mathcal{B}$: all behavioral variables used in [24], consisting of eye position (θ, ϕ), head position (ρ, ω), pupil size (σ), and running speed (s);
- $\mathcal{D}$: the first-order derivatives of the variables in $\mathcal{B}$, namely $\dot{\theta}$, $\dot{\phi}$, $\dot{\omega}$, $\dot{\rho}$, $\dot{\sigma}$, and $\dot{s}$.

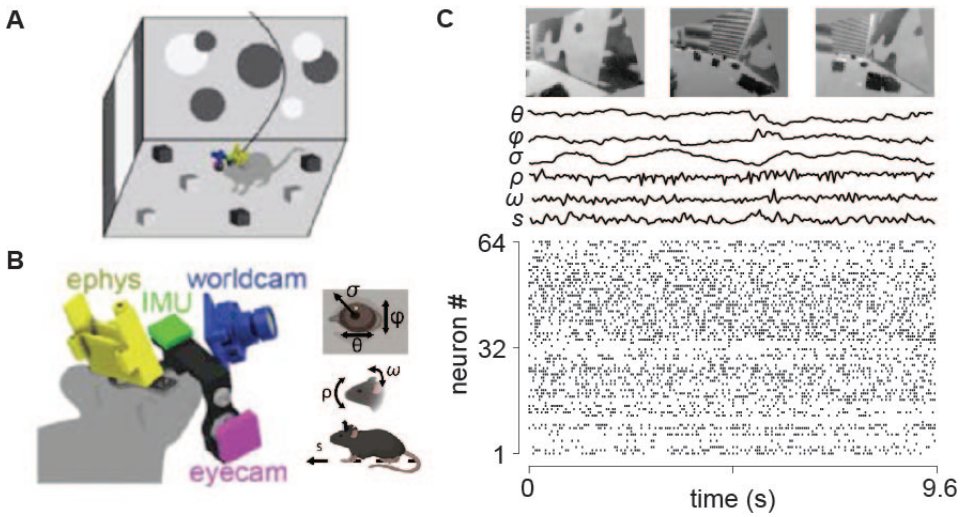

Figure 1: Schematic of the head-mounted recording system for freely moving mice (adapted from [24]). A) Three mice freely explored a $48\,\mathrm{cm}$ long $\times$ $37\,\mathrm{cm}$ wide $\times$ $30\,\mathrm{cm}$ high arena. B) Preparation included a silicon probe for electrophysiological recording in V1 (yellow), miniature cameras for recording the mouse's eye position and pupil size (θ, ϕ, and σ; magenta), and visual scene (blue), and inertial measurement unit for measuring head orientation (ρ and ω; green). C) Sample data from a $9.6\,\mathrm{s}$ period during free movement showing (from top) visual scene, horizontal and vertical eye position, pupil size, head pitch and roll, locomotor speed, and a raster plot of 64 units.

To test for interactions between behavioral variables, these sets could also include the pairwise multiplication of their elements; e.g., $\mathcal{B}_\times = \{b_i b_j \;\forall\; (b_i, b_j) \in \mathcal{B}\}$. The input vector was then passed through a batch normalization layer and a fully connected layer (subjected to a strong L1 norm for feature selection) to produce a behavioral vector, $\boldsymbol{b}$.

We then concatenated the vectors $\boldsymbol{v}$, $\boldsymbol{b}$, and their element-wise product $\boldsymbol{v} \odot \boldsymbol{b}$ (all calculated for each individual input frame), fed them through a batch normalization layer, and input them to a 1-layer gated recurrent unit (GRU) (hidden size of 512). To incorporate temporal dynamics, we constructed different versions ($\mathrm{GRU_k}$) of the model that had access to k previous frames. A fully-connected layer and a softplus activation function were applied to yield the neuronal response prediction.

Training and model evaluation To deal with the continuous and dynamic nature of the data, we split the $\sim 1\,\mathrm{h}$-long recording into 10 consecutive segments. The first $70\,\%$ of each segment were then reserved for training (including an 80-20 validation split) and the remaining $30\,\%$ for testing.

Models were separately trained on the data from each mouse. Model parameters were optimized with Adam (batch size: 256, CNN learning rate: .0001, full model: .0002) to minimize the Poisson loss between predicted neuronal response ($\hat{r}$) and ground truth (r) : $\frac{1}{N}\sum_{i=1}^{N}(\hat{r}_i - r_i \log \hat{r}_i)$, where N denotes the number of recorded neurons for each mouse. We used early stopping on the validation set (patience: 5 epochs), which led all models to converge in less than 50 epochs. Due to the large number of hyper-parameters, the specific network and training settings were determined using a combination of grid search and manual exploration on a validation set (see Appendix A).

To evaluate model performance, we calculated the cross-correlation (cc) between a smoothed version ($2\,\mathrm{s}$ boxcar filter) of the predicted and ground-truth response for each recorded neuron [24].

All models were implemented in PyTorch and trained on an NVIDIA RTX 3090 with 24GB memory. All code will be made available on GitHub should this paper get accepted.

Maximally activating stimuli We used gradient ascent [33] to discover the visual stimuli that most strongly activate a particular model neuron in our network. The visual input was initialized with noise sampled in $\mathcal{N}(.5, 2)$. We used the Adam optimizer to repeatedly add the gradient of the target neuron's activity with respect to its inputs. We also applied L2 regularization (weight of .02) and Laplacian regularization (weight of 0.01) [37] on the image. This procedure was repeated 6400 times.

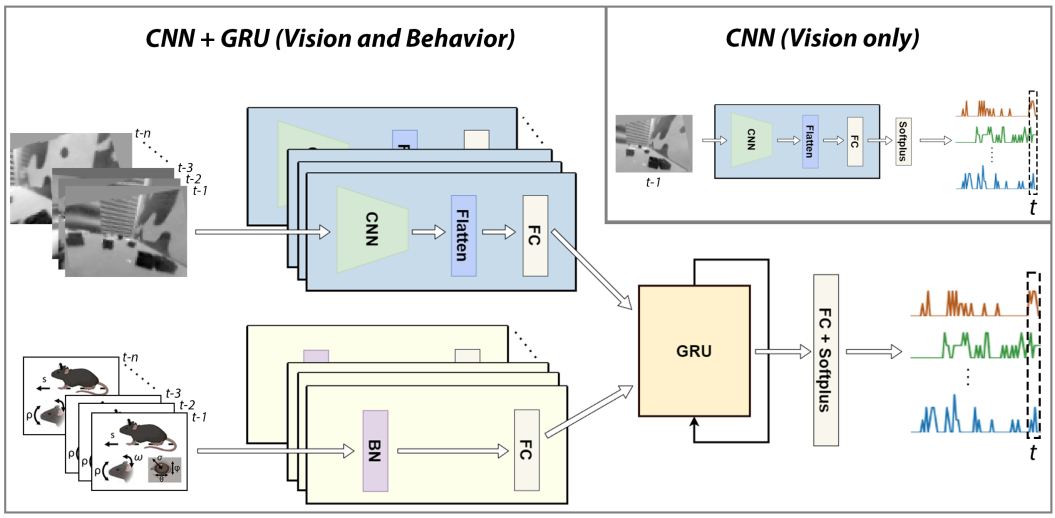

Figure 2: Model architecture diagram. The vision-only network (top-right) was a CNN network, predicting the neural activity at time t given the visual input at time $t-1$ (48 ms bins). The full model combined the CNN with a behavioral encoder and a gated recurrent unit (GRU), predicting the neural activity at time t given the visual and behavioral inputs from time $t-1$ to $t-n$.

The resulting, maximally activating visual stimuli were smoothed with a Butterworth filter (low-pass, .05 cutoff frequency ratio) to reduce the impact of high-frequency noise.

Saliency map We computed a saliency map [38] of the behavioral vector for each neuron to discover which behavioral variables contributed most strongly to each model neuron's activity. We iterated through the test dataset, recorded the gradient of each behavioral input with respect to each neuron's prediction, and then averaged the gradients per neuron to obtain the saliency map.

4 Results

Mouse V1 activity is best predicted with a 3-layer CNN To determine the purely visual contribution to V1 responses, we experimented with a large number of vision architectures (see Appendix A). In the end, a vanilla 3-layer CNN (kernel size 7, $128 \times 64 \times 32$ channels) yielded the best cross-correlation between predicted and ground-truth responses (Table 1), outperforming the best autoencoder architecture (kernel size: 7, encoder: $64 \times 128 \times 256$ channels, decoder: $256 \times 128 \times 64$ channels), ResNet-18 [39] (a 20-layer CNN with the first input channel being replaced by 1), EfficientNet-B0 [40] (a 65-layer CNN with the first input channel being replaced by 1), and the Sensorium baseline [12] (a 4-layer CNN with a readout network). The greatest improvement in cross-correlation was achieved for Mouse 2, whose activity overall proved to be much harder to predict than that of Mouse 1 and 3.

Behavioral variables improve most neuronal predictions Once we identified the 3-layer CNN as the best visual encoder, we added the different sets of behavioral variables to the network. To allow for a fair comparison with the Sensorium+ baseline [14], we first limited ourselves to $\mathcal{S} = \{\sigma, \dot{\sigma}, s\}$, but then gradually added more behavioral variables ($\mathcal{B}$) [24] as well the derivatives of these variables ($\mathcal{D}$) and multiplicative pairs ($\mathcal{B}_\times$ and $\{\mathcal{B} \cup \mathcal{D}\}_\times$).

The results are shown in Table 2. All models were able to outperform the Sensorium+ baseline, and the addition of behavioral variables and their interactions further improved model performance. Note that although the full model used a GRU to combine visual and behavioral features, the input sequence length was always 1 (i.e., GRU_1). That being said, it is possible that the GRU learned long-term correlations that the Sensorium+ baseline model did not have access to. Nevertheless, the biggest performance improvements were gained through the addition of behavioral variables

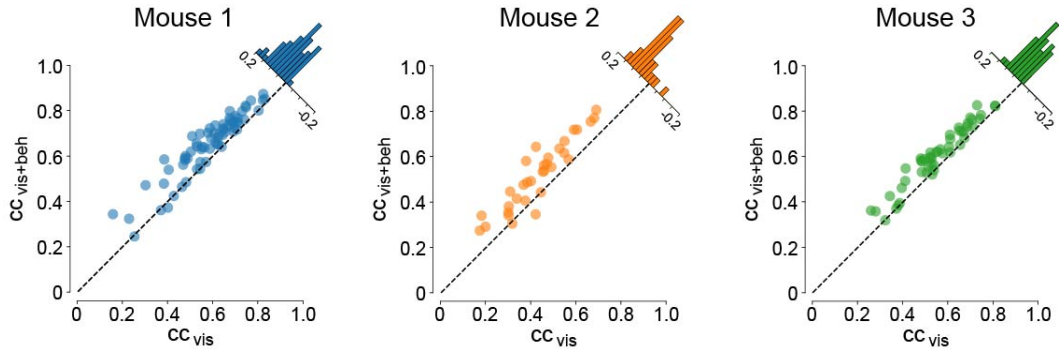

Figure 3: The integration of behavioral variables improved the cross-correlation (cc) for the majority of neurons. Each dot represents a neuron. A dot above the dashed diagonal indicates a higher cc with the inclusion of behavioral variables. Histograms (small insets) illustrate the distribution of the improvement in cc across the neuronal population.

related to head and eye position (which are present in $\mathcal{B}$ but not in $\mathcal{S}$), their derivatives ($\mathcal{D}$), and multiplicative interactions between these variables ($\{\mathcal{B} \cup \mathcal{D}\}_\times$).

We also wondered whether the prediction of only some V1 neurons would benefit from the addition of these behavioral variables. To our surprise, the cross-correlation between predicted and ground-truth responses improved for almost all recorded V1 neurons (Fig. 3).

	Mouse 1		Mouse 2		Mouse 3	
Model	$cc \uparrow$	MSE $\downarrow$	$cc \uparrow$	MSE $\downarrow$	$cc \uparrow$	MSE $\downarrow$
CNN	$\mathbf{.583 \pm .140}$	**.0665**	$\mathbf{.430 \pm .135}$	**.0991**	$\mathbf{.554 \pm .133}$	**.0896**
Autoencoder	$.553 \pm .134$	.0688	$.369 \pm .104$	.110	$.530 \pm .136$	.0992
ResNet-18 [39]	$.540 \pm .145$	.0734	$.373 \pm .144$	.112	$.516 \pm .132$	.0950
EfficientNet-B0 [40]	$.531 \pm .160$	.0687	$.361 \pm .159$	.112	$.497 \pm .138$	.0946
Sensorium [12]	$.551 \pm .131$	.0699	$.303 \pm .117$	.138	$.462 \pm .131$	.108

Table 1: Best-performing vision models, compared to the Sensorium baseline [12] (see Appendix A for more). Best-performing networks are indicated in bold. cc: cross-correlation, mean $\pm$ standard deviation across neurons ($\uparrow$: the higher the better), MSE: mean-squared error ($\downarrow$: the lower the better).

		Mouse 1		Mouse 2		Mouse 3	
	Feature Set	$cc \uparrow$	MSE $\downarrow$	$cc \uparrow$	MSE $\downarrow$	$cc \uparrow$	MSE $\downarrow$
CNN+GRU$_1$	$\{\mathcal{B} \cup \mathcal{D}\}_\times$	$\mathbf{.649 \pm .135}$	**.0539**	$\mathbf{.520 \pm .148}$	**.0841**	$\mathbf{.607 \pm .130}$	**.0793**
	$\mathcal{B} \cup \mathcal{D}$	$.641 \pm .141$	.0549	$.464 \pm .145$	.0956	$.596 \pm .130$	.0808
	$\mathcal{B}_\times$	$.632 \pm .136$	.0588	$.468 \pm .152$	.0919	$.590 \pm .131$	.0835
	$\mathcal{B}$	$.637 \pm .137$	.0549	$.471 \pm .151$	.0911	$.584 \pm .132$	.0835
	$\mathcal{S}$	$.610 \pm .137$	.0647	$.465 \pm .149$	.0976	$.563 \pm .142$	.0907
	Sensorium+	$.551 \pm .142$	.0708	$.465 \pm .157$	.0973	$.520 \pm .148$	.0936

Table 2: CNN+GRU$_1$ model, compared to the Sensorium+ baseline [12], trained on different behavioral feature sets. $\mathcal{S} = \{\sigma, \dot{\sigma}, s\}$: the set of variables used in the Sensorium+ competition [14]. $\mathcal{B} = \{\theta, \phi, \omega, \rho, \sigma, s\}$: the set of variables from [24]. $\mathcal{D} = \{\dot{\theta}, \dot{\phi}, \dot{\omega}, \dot{\rho}, \dot{\sigma}, s\}$: the derivatives of $\mathcal{B}$. $\mathcal{A}_\times = \{a_i a_j \; \forall \; (a_i, a_j) \in \mathcal{A}\}$ denotes the set of all multiplicative pairs. $\cup$ denotes the union operator. Best performing networks are indicated in bold. cc: cross-correlation, mean $\pm$ standard deviation across neurons ($\uparrow$: the higher the better), MSE: mean-squared error ($\downarrow$: the lower the better).

Access to longer series of data in time further improves predictive performance After we identified the full behavioral feature set ($\{\mathcal{B} \cup \mathcal{D}\}_\times$) as the one yielding the best model performance, we extended the GRU's temporal dependence by allowing the input to vary from one frame ($48\,\mathrm{ms}$) to a total of eight frames ($384\,\mathrm{ms}$), and assessed the model's performance.

The results are shown in Table 3. The number of frames needed by the model to reach peak predictive performance varied for different mice (6, 5, and 3, respectively). This indicates that temporal information is important for predicting dynamic neural activity. However, the dependence on temporal information has a limit, and different neurons in V1 might possess different temporal capacities.

Well-defined visual receptive fields emerge To assess whether the CNN+GRU$_1$ model learned meaningful visual receptive fields, we used gradient ascent (see Methods) to find the maximally activating stimulus for each neuron. Receptive fields for the 32 best-predicted neurons are shown in Fig. 4. Interestingly, most of them had well-defined excitatory and inhibitory subregions, often resembling receptive fields of orientation-selective neurons. Most excitatory and inhibitory subregions spanned approximately 30° of visual angle (the full width of the frame, 80 pixels, roughly corresponding to 120° of visual angle), which is roughly on the same order of magnitude compared to receptive field sizes typically observed in mouse V1, varying from 10° to 30° [8, 24, 41].

Receptive fields demonstrated noticeable differences across mice. The model trained on Mouse 1 seems to have generated many robust visual receptive fields, whereas the models trained on Mice 2–3 appear weaker (same colorbar across panels). In addition, even some of the best-predicted neurons lack a pronounced or spatially structured receptive field, implying that these neurons could be primarily driven by behavioral variables.

Analysis of behavioral saliency maps reveals different types of neurons Intrigued by the fact that some neurons lacked pronounced visual receptive fields, we aimed to analyze the influence of behavioral state on the predicted neuronal response by performing a saliency map analysis on the behavioral inputs (see Methods). Since different behavioral variables operate on different input ranges, we first standardized the saliency map activities for each behavioral variable across the model neuron population. Saliency map activities further than 1 standard deviation from the mean were then interpreted as "driving" the neuron, allowing us to categorize each neuron as being driven by one or multiple behavioral variables (Fig. 5).

We first asked which neurons in our model were driven by which behavioral variables (Fig. 5, *top*). Consistent with [24], we found a large fraction of model neurons driven by eye and head position, and smaller fractions driven by locomotion speed and pupil size. 20-25% of neurons were not driven by any of these behavioral variables, rendering their responses purely visual.

However, a particular neuron could be driven by multiple behavioral variables. Repeating the above analysis, we found that most model neurons showed mixed selectivity (Fig. 5, *middle*), with only a minority of cells responding exclusively to a single behavioral variable. Adding the interaction terms

		Mouse 1		Mouse 2		Mouse 3	
Model	History	$cc\uparrow$	MSE $\downarrow$	$cc\uparrow$	MSE $\downarrow$	$cc\uparrow$	MSE $\downarrow$
CNN+GRU$_1$	48 ms	$.649 \pm .135$	.0539	$.520 \pm .148$	.0841	$.607 \pm .130$	.0793
CNN+GRU$_2$	96 ms	$.650 \pm .137$	.0541	$.527 \pm .168$	.0849	$.606 \pm .139$	.0793
CNN+GRU$_3$	144 ms	$.648 \pm .142$	.0556	$.504 \pm .161$	.0909	$\mathbf{.616 \pm .132}$	**.0772**
CNN+GRU$_4$	192 ms	$.647 \pm .149$	.0544	$.528 \pm .169$	.0842	$.610 \pm .132$	.0795
CNN+GRU$_5$	240 ms	$.646 \pm .143$	.0546	$\mathbf{.532 \pm .159}$	**.0825**	$.604 \pm .133$	.0792
CNN+GRU$_6$	288 ms	$\mathbf{.655 \pm .139}$	**.0532**	$.519 \pm .161$	.0833	$.597 \pm .138$	.0814
CNN+GRU$_7$	336 ms	$.643 \pm .151$	.0555	$\mathbf{.532 \pm .180}$	**.0858**	$.613 \pm .133$	.0785
CNN+GRU$_8$	384 ms	$\mathbf{.655 \pm .141}$	**.0532**	$.524 \pm .161$	.0841	$.600 \pm .139$	.0803

Table 3: CNN+GRU$_k$ model trained with input from k timesteps on the full feature set ($\{\mathcal{B} \cup \mathcal{D}\}_\times$). Best performing networks are indicated in bold. cc: cross-correlation, mean $\pm$ standard deviation ($\uparrow$: the higher the better), MSE: mean-squared error ($\downarrow$: the lower the better).

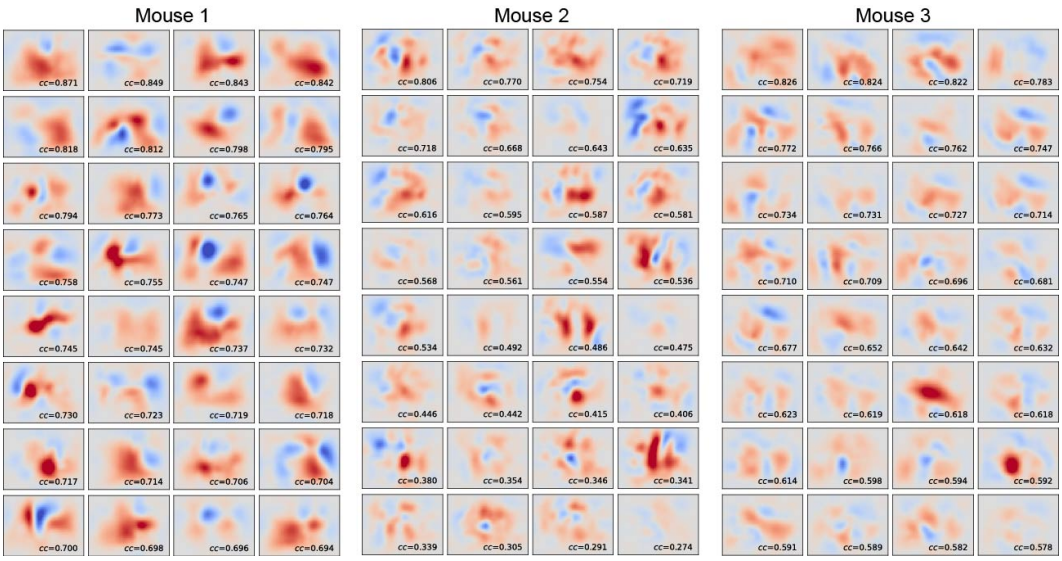

Figure 4: The maximally activating stimuli learned in CNN+GRU$_1$, generated via gradient ascent. The 32 neurons with the highest cross-correlation (cc) from each mouse are shown, sorted by cc.

between behavioral variables (Fig. 5, *bottom*) did not change the fact that most model V1 neurons encoded combinations of multiple behavioral variables, often relating information about the animal's eye position to head position and locomotor speed.

5 Discussion

In this paper, we propose a deep recurrent neural network that achieves state-of-the-art predictions of V1 activity in freely moving mice. We discovered that our model outperforms previous models under these more naturalistic conditions, which could be attributed to the better alignment of this data with the computations performed by the mouse visual system, based on its natural visual environment and behavioral characteristics. Similar to previous models, we found that a simple CNN architecture is sufficient to predict the visual response properties of cells in mouse V1.

In addition, mouse V1 is known to be strongly modulated by signals related to the movement of the animal's eyes, head, and body [20, 21, 24], which are severely restricted in head-fixed preparations. Models trained on head-fixed preparations may thus be limited in their predictive power. In contrast, our model was able to predict V1 activity on a 1-hour continuous data stream, during which the animal freely explored a real-world arena. Our analyses demonstrate the impact of the animal's behavioral state on V1 activity and reveal that most model V1 neurons exhibit mixed selectivity to multiple behavioral variables.

Accurate predictions of mouse V1 activity under natural conditions Our brains did not evolve to view stationary stimuli on a computer screen. However, most research on neural coding in vision has been conducted under head-fixed conditions, which do not mirror natural behavior and thus provide limited insight into visual processing in real-world environments. Some visual functions mediated by the ventral stream, such as identifying faces and objects, resemble this condition, but the real visual environment is constantly shifting due to self-motion, leading to dynamic activities such as navigation or object reaching, typically mediated by the dorsal stream. To truly understand visual perception in natural environments, we need to capture the computational principles when the subjects are in motion.

In this research, we take the initial steps towards this by modeling a novel data type encompassing neural activity coupled with a visual scene captured from a freely moving animal's perspective. This represents a dramatic (but, in our opinion, crucial) shift in the "parameter space" of visual input, from static images projected on a screen to dynamic, real-world visual input.

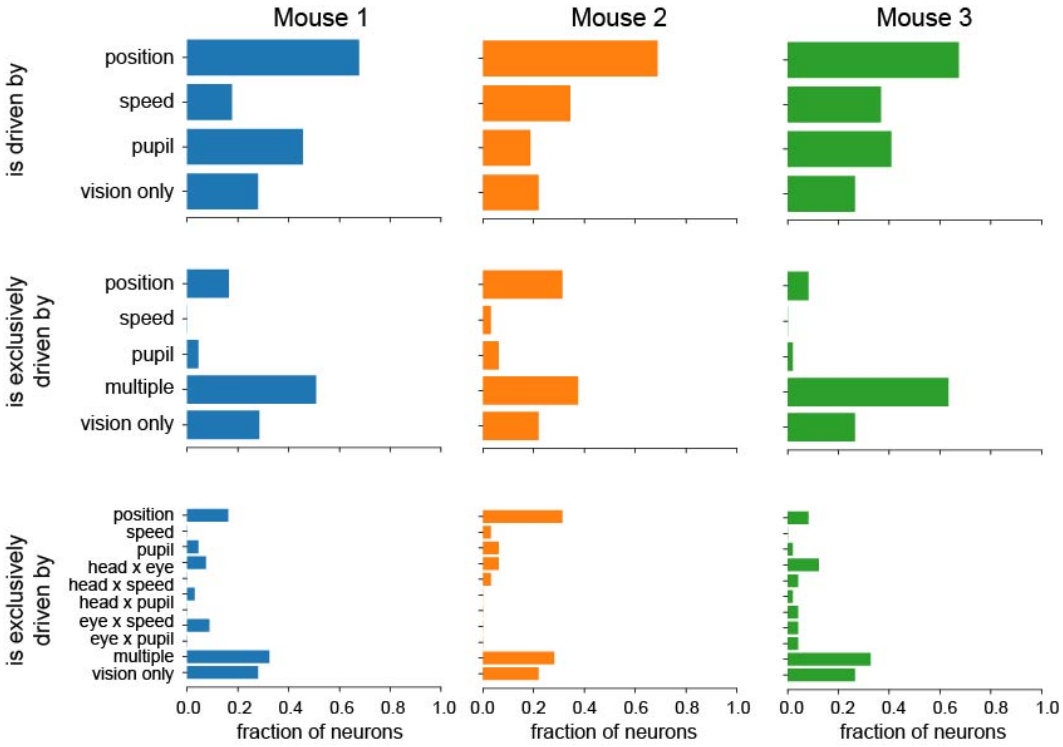

Figure 5: Effect of behavioral variables on model neuron activity, inferred by the saliency analysis. A) Fraction of neurons that are "driven by" (i.e., their saliency map activation is further than 1 standard deviation from the mean) different behavioral variables (similar to [24]). A neuron that responds to (e.g.) both position and speed may be counted twice. Neurons without a strong behavioral drive are categorized as "vision only". B) Fraction of neurons that are uniquely driven by a specific behavioral variable. Still, a large fraction of neurons are driven by multiple behavioral variables. C) Same as B), but split with interaction terms.

Mixed selectivity of behavioral variables Our experiments demonstrated that the models incorporating behavioral variables and their interactions performed substantially better than the models relying exclusively on visual inputs. Moreover, our saliency map analysis showed that only around 25% of model neurons could be considered purely visual, with the majority of model neurons driven by multiple behavioral variables.

This widespread mixed selectivity is consistent with previous literature suggesting that V1 neurons may be modulated by a high-dimensional latent representation of several behavioral variables related to the animal's movement, recent experiences, and behavioral goals [19]. It is also consistent with the idea of a basis function representation [42, 43], which allows a population of neurons to conjunctively represent multiple behaviorally relevant variables. Such representations are often employed by higher-order visual areas in primate cortex to implement sensorimotor transformations [44]. It is intriguing to find computational evidence for such a representation as early as V1 in the mouse. Future computational studies should therefore aim to study the mechanisms by which V1 neurons might construct a nonlinear combination of behavioral signals.

Limitations and future work. While our study opens a new perspective on modeling neural activity during natural conditions, there are a few limitations that need to be acknowledged. First, our data was relatively limited (around 50 neurons per animal, for 3 animals). The development of a Sensorium-style standardized dataset [14] for freely-moving mice would significantly benefit future research in this area, enabling more robust comparisons between different modeling approaches. Second, it would be beneficial to integrate other modalities that are known to be encoded in mouse V1 into the model. One such example is reward signals [45], which could provide additional information about the animal's decision-making processes and motivations during exploration.

Acknowledgments

This work was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number R01-NS121919. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. Publisher: Proceedings of the National Academy of Sciences.

[2] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915, November 2014. Publisher: Public Library of Science.

[3] Umut Güçlü and Marcel A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. Publisher: Society for Neuroscience Section: Articles.

[4] 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. Publisher: Public Library of Science.

[5] 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, April 2019.

[6] Santiago A. Cadena, Fabian H. Sinz, Taliah Muhammad, Emmanouil Froudarakis, Erick Cobos, Edgar Y. Walker, Jake Reimer, Matthias Bethge, Andreas Tolias, and Alexander S. Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? October 2019.

[7] Glen T Prusky, Paul W. R West, and Robert M Douglas. Behavioral assessment of visual acuity in mice and rats. *Vision Research*, 40(16):2201–2209, July 2000.

[8] Cristopher M. Niell and Michael P. Stryker. Highly Selective Receptive Fields in Mouse Visual Cortex. *The Journal of Neuroscience*, 28(30):7520–7536, July 2008.

[9] David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating " what" and " where". In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] 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. December 2018.

[11] Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[12] 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. January 2021.

[13] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 25164–25178. Curran Associates, Inc., 2021.

[14] 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, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The Sensorium competition on predicting large-scale mouse primary visual cortex activity, June 2022. arXiv:2206.08666 [cs, q-bio].

[15] Andrew D. Huberman and Cristopher M. Niell. What can mice tell us about how vision works? *Trends in Neurosciences*, 34(9):464–473, September 2011.

[16] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, November 2014. Google-Books-ID: 8BSLBQAAQBAJ.

[17] Philip R. L. Parker, Morgan A. Brown, Matthew C. Smear, and Cristopher M. Niell. Movement-Related Signals in Sensory Areas: Roles in Natural Behavior. *Trends in Neurosciences*, 43(8):581–595, August 2020.

[18] Laura Busse, Jessica A. Cardin, M. Eugenia Chiappe, Michael M. Halassa, Matthew J. McGinley, Takayuki Yamashita, and Aman B. Saleem. Sensation during Active Behaviors. *Journal of Neuroscience*, 37(45):10826–10834, November 2017. Publisher: Society for Neuroscience Section: Symposium and Mini-Symposium.

[19] 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, April 2019. Publisher: American Association for the Advancement of Science.

[20] Cristopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, February 2010.

[21] Emmanouil Froudarakis, Paul G. Fahey, Jacob Reimer, Stelios M. Smirnakis, Edward J. Tehovnik, and Andreas S. Tolias. The Visual Cortex in Context. *Annual Review of Vision Science*, 5(1):317–339, 2019. _eprint: https://doi.org/10.1146/annurev-vision-091517-034407.

[22] Arne F. Meyer, Jasper Poort, John O'Keefe, Maneesh Sahani, and Jennifer F. Linden. A Head-Mounted Camera System Integrates Detailed Behavioral Monitoring with Multichannel Electrophysiology in Freely Moving Mice. *Neuron*, 100(1):46–60.e7, October 2018.

[23] Grigori Guitchounts, Javier Masís, Steffen B. E. Wolff, and David Cox. Encoding of 3D Head Orienting Movements in the Primary Visual Cortex. *Neuron*, 108(3):512–525.e4, November 2020.

[24] Philip R. L. Parker, Elliott T. T. Abe, Emmalyn S. P. Leonard, Dylan M. Martins, and Cristopher M. Niell. Joint coding of visual input and eye/head position in V1 of freely moving mice. *Neuron*, September 2022.

[25] Laura Busse, Asli Ayaz, Neel T. Dhruv, Steffen Katzner, Aman B. Saleem, Marieke L. Schölvinck, Andrew D. Zaharia, and Matteo Carandini. The Detection of Visual Contrast in the Behaving Mouse. *Journal of Neuroscience*, 31(31):11351–11361, August 2011. Publisher: Society for Neuroscience Section: Articles.

[26] Sylvia Schröder, Nicholas A. Steinmetz, Michael Krumin, Marius Pachitariu, Matteo Rizzi, Leon Lagnado, Kenneth D. Harris, and Matteo Carandini. Arousal Modulates Retinal Output. *Neuron*, 107(3):487–495.e9, August 2020.

[27] Julie A. Harris, Stefan Mihalas, Karla E. Hirokawa, Jennifer D. Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, Aaron Feiner, David Feng, Nathalie Gaudreault, Charles R. Gerfen, Nile Graddis, Peter A. Groblewski, Alex M. Henry, Anh Ho, Robert Howard, Joseph E. Knox, Leonard Kuan, Xiuli Kuang, Jerome Lecoq, Phil Lesnar, Yaoyao Li, Jennifer Luviano, Stephen McConoughey, Marty T. Mortrud, Maitham Naeemi, Lydia Ng, Seung Wook Oh, Benjamin Ouellette, Elise Shen, Staci A. Sorensen, Wayne Wakeman, Quanxin Wang, Yun Wang, Ali Williford, John W. Phillips, Allan R. Jones, Christof Koch, and Hongkui Zeng. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202, November 2019. Number: 7781 Publisher: Nature Publishing Group.

[28] Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K. Ramirez, Hannah Choi, Jennifer A. Luviano, Peter A. Groblewski, Ruweida Ahmed, Anton Arkhipov, Amy Bernard, Yazan N. Billeh, Dillan Brown, Michael A. Buice, Nicolas Cain, Shiella Caldejon, Linzy Casal, Andrew Cho, Maggie Chvilicek, Timothy C. Cox, Kael Dai, Daniel J. Denman, Saskia E. J. de Vries, Roald Dietzman, Luke Esposito, Colin Farrell, David Feng, John Galbraith, Marina Garrett, Emily C. Gelfand, Nicole Hancock, Julie A. Harris, Robert Howard, Brian Hu, Ross Hytnen, Ramakrishnan Iyer, Erika Jessett, Katelyn Johnson, India Kato, Justin Kiggins, Sophie Lambert, Jerome Lecoq, Peter Ledochowitsch, Jung Hoon Lee, Arielle Leon, Yang Li, Elizabeth Liang, Fuhui Long, Kyla Mace, Jose Melchior, Daniel Millman, Tyler Mollenkopf, Chelsea Nayan, Lydia Ng, Kiet Ngo, Thuyahn Nguyen, Philip R. Nicovich, Kat North, Gabriel Koch Ocker, Doug Ollerenshaw, Michael Oliver, Marius Pachitariu, Jed Perkins, Melissa Reding, David Reid, Miranda Robertson, Kara Ronellenfitch, Sam Seid, Cliff Slaughterbeck, Michelle Stoecklin, David Sullivan, Ben Sutton, Jackie Swapp, Carol Thompson, Kristen Turner, Wayne Wakeman, Jennifer D. Whitesell, Derric Williams, Ali Williford, Rob Young, Hongkui Zeng, Sarah Naylor, John W. Phillips, R. Clay Reid, Stefan Mihalas, Shawn R. Olsen, and Christof Koch. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, April 2021. Number: 7852 Publisher: Nature Publishing Group.

[29] Angie M Michaiel, Elliott TT Abe, and Cristopher M Niell. Dynamics of gaze control during prey capture in freely moving mice. *eLife*, 9:e57458, July 2020. Publisher: eLife Sciences Publications, Ltd.

[30] Aman B Saleem. Two stream hypothesis of visual processing for navigation in mouse. *Current Opinion in Neurobiology*, 64:70–78, October 2020.

[31] Cristopher M. Niell. Cell Types, Circuits, and Receptive Fields in the Mouse Visual Cortex. *Annual Review of Neuroscience*, 38(1):413–431, 2015. _eprint: https://doi.org/10.1146/annurev-neuro-071714-033807.

[32] Katrin Franke, Konstantin F. Willeke, Kayla Ponder, Mario Galdamez, Na Zhou, Taliah Muhammad, Saumil Patel, Emmanouil Froudarakis, Jacob Reimer, Fabian H. Sinz, and Andreas S. Tolias. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134, October 2022. Number: 7930 Publisher: Nature Publishing Group.

[33] 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. *Nature Neuroscience*, 22(12):2060–2065, December 2019. Number: 12 Publisher: Nature Publishing Group.

[34] 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. *Nature Neuroscience*, 21(9):1281–1289, September 2018. Number: 9 Publisher: Nature Publishing Group.

[35] Nicholas A. Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, Susu Chen, Jennifer Colonell, Richard J. Gardner, Bill Karsh, Fabian Kloosterman, Dimitar Kostadinov, Carolina Mora-Lopez, John O'Callaghan, Junchol Park, Jan Putzeys, Britton Sauerbrei, Rik J. J. van Daal, Abraham Z. Vollan, Shiwei Wang, Marleen Welkenhuysen, Zhiwen Ye, Joshua T. Dudman, Barundeb Dutta, Adam W. Hantman, Kenneth D. Harris, Albert K. Lee, Edvard I. Moser, John O'Keefe, Alfonso Renart, Karel Svoboda, Michael Häusser, Sebastian Haesler, Matteo Carandini, and Timothy D. Harris. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, April 2021. Publisher: American Association for the Advancement of Science.

[36] Philip R. L. Parker, Dylan M. Martins, Emmalyn S. P. Leonard, Nathan M. Casey, Shelby L. Sharp, Elliott T. T. Abe, Matthew C. Smear, Jacob L. Yates, Jude F. Mitchell, and Cristopher M. Niell. A dynamic sequence of visual processing initiated by gaze shifts, August 2022. Pages: 2022.08.23.504847 Section: New Results.

[37] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019. Publisher: American Association for the Advancement of Science.

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. arXiv:1312.6034 [cs].

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs].

[40] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. arXiv:1905.11946 [cs, stat] version: 5.

[41] Gert Van den Bergh, Bin Zhang, Lutgarde Arckens, and Yuzo M. Chino. Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys. *Journal of Comparative Neurology*, 518(11):2051–2070, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.22321.

[42] Michael Beyeler, Emily L. Rounds, Kristofor D. Carlson, Nikil Dutt, and Jeffrey L. Krichmar. Neural correlates of sparse coding and dimensionality reduction. *PLOS Computational Biology*, 15(6):e1006908, June 2019.

[43] S. Fusi, E. K. Miller, and M. Rigotti. Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol*, 37:66–74, April 2016.

[44] A. Pouget and L. H. Snyder. Computational approaches to sensorimotor transformations. *Nat Neurosci*, 3 Suppl:1192–8, November 2000.

[45] Marshall G. Shuler and Mark F. Bear. Reward Timing in the Primary Visual Cortex. *Science*, 311(5767):1606–1609, March 2006. Publisher: American Association for the Advancement of Science.

Appendix

A Vision-Only Models

A.1 Hyperparameter tuning

We performed a grid search to find the optimal CNN kernel size (3, 5, 7, 9), number of channels (32, 64, 128, 256, 512; in various combinations), and dropout rate (0, 0.25, 0.5). While other models often rely on kernel size 3 for their CNN, we found these small kernels to lead to worse performance, perhaps due to the mouse's low-resolution vision, and that size 7 performed better. We repeated the grid search for CNNs with different number of convolutional layers. The resulting 3-layer CNN outperformed many differently sized networks, such as a 1-layer CNN with 1024 channels (i.e., a shallow but wide network), a 2-layer CNN, or a 4-layer CNN. Choice of learning rates and optimizers had no notable effect on the final performance of the networks.

A.2 Autoencoder

We hypothesized that an autoencoder could provide regularization benefits, because the reconstruction loss might encourage the model to learn visual features that are useful for decoding. Specifically, an encoder ϕ mapped the original frame $\mathcal{F}$ to a vector $\mathcal{V}$ in the latent space, which was present at the bottleneck, while the decoder ψ then mapped the vector $\mathcal{V}$ from the latent space to the output.

$$\phi: \ \mathcal{F} \rightarrow \mathcal{V} \tag{1}$$

$$\psi: \ \mathcal{V} \rightarrow \mathcal{F} \tag{2}$$

$$\phi, \psi = \text{argmin}_{\phi,\psi} ||\mathcal{F} - (\psi \cdot \phi)\mathcal{F}||^2 \tag{3}$$

After the hyperparameter search, we settled on size 256 for the latent space vector, and the weight of the reconstruction loss relative to the Poisson loss was fixed to 0.5. Both the encoder and the decoder were 3-layer CNNs, and their numbers of channels were symmetric. However, after testing a number of autoencoders with different configurations (Table 4), we found that a simple 3-layer CNN outperformed any and all of the tested autoencoders.

	Mouse 1		Mouse 2		Mouse 3	
Kernel size, encoder #channels	$cc\uparrow$	MSE $\downarrow$	$cc\uparrow$	MSE $\downarrow$	$cc\uparrow$	MSE $\downarrow$
3, 16 × 32 × 64	$.539 \pm .149$	.0728	$.389 \pm .128$	**.107**	$.502 \pm .129$	.0996
5, 16 × 32 × 64	$.550 \pm .147$	.0728	$.363 \pm .116$	.109	$.508 \pm .135$	.0983
7, 16 × 32 × 64	$.525 \pm .152$	.0732	$.353 \pm .121$	.117	$.509 \pm .131$	**.0980**
9, 16 × 32 × 64	$.518 \pm .147$	.0752	$.315 \pm .101$	.119	$.492 \pm .135$	.0997
3, 32 × 64 × 128	$.543 \pm .144$	.0737	$.367 \pm .128$	.109	$.503 \pm .131$	.100
5, 32 × 64 × 128	$.551 \pm .149$	.0723	$.361 \pm .109$	.119	$.514 \pm .132$	.0984
7, 32 × 64 × 128	$.539 \pm .145$	.0739	**$.390 \pm .118$**	.109	$.492 \pm .129$	.100
9, 32 × 64 × 128	$.510 \pm .155$	.0758	$.331 \pm .119$	.112	$.500 \pm .134$	.101
3, 64 × 128 × 256	$.541 \pm .146$	.0758	$.374 \pm .123$	.110	$.514 \pm .127$	.0990
5, 64 × 128 × 256	$.552 \pm .145$	.0777	$.362 \pm .119$	.110	$.508 \pm .134$	.104
7, 64 × 128 × 256	**$.553 \pm .134$**	**.0688**	$.369 \pm .104$	.111	**$.530 \pm .136$**	.0992
9, 64 × 128 × 256	$.537 \pm .146$	.0811	$.355 \pm .109$	.119	$.500 \pm .128$	.105

Table 4: Performance of different autoencoders. The numbers of channels in the decoder were symmetric with those of the encoder. Best performing networks are indicated in bold. cc: cross-correlation, mean $\pm$ standard deviation across neurons ($\uparrow$: the higher the better), MSE: mean-squared error ($\downarrow$: the lower the better).