Appendix

A Data Acquisition

The full procedure of data collection and data preprocessing is described in detail in Ref. [25], but are briefly described below for the reader's convenience.

Visual scenes from the mouse's perspective, neural activity in V1, and various behavioral variables were simultaneously recorded from three adult mice who were freely exploring an arena with a state-of-the-art head-mounted recording system [25]. The recording system consisted of high-density silicon probes, two miniature cameras and an inertial measurement unit (IMU). 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. One wide-angled camera (around $120°$) was aimed outwards to capture the visual scene available to the right eye of the mouse at $16\,\mathrm{ms}$ per frame ("worldcam"). A second camera was aimed at the right eye (illuminated by an infrared-LED) to record a video feed of the right eye at $30\,\mathrm{Hz}$. The IMU acquired three-axis gyroscope and accelerometer data at $30\,\mathrm{kHz}$. In addition, a top-down camera recorded the mouse in the arena at $60\,\mathrm{Hz}$.

During experiments, mice were placed in an arena where they could move around freely for about 1 hour. The arena was approximately $48\,\mathrm{cm}$ long $\times$ $37\,\mathrm{cm}$ wide $\times$ $30\,\mathrm{cm}$ high. The gray floor was covered with black-and-white Legos to provide visual contrast. One wall of the arena was a monitor displaying a moving black-and-white spots stimulus, and the other three walls were covered with wallpaper with static stimuli including white noise, black-and-white high-spatial-frequency gratings, and black-and-white low-spatial-frequency gratings. In order to encourage foraging behavior during the recording sessions, small fragments of tortilla chips were sparsely distributed across the arena.

The worldcam video was downsampled to 60×80 pixels. DeepLabCut [43] was used to extract extract eye position (θ, ϕ) and pupil radius (σ). Pitch (ρ) and roll (ω) of the mouse's head position were extracted from the IMU. Locomotion speed (s) was estimated from the top-down camera feed using DeepLabCut [43]. Electrophysiology data bandpass-filtered between $0.01\,\mathrm{Hz}$ and $7.5\,\mathrm{kHz}$, and spike-sorted with Kilosort 2.5 [44]. Single units were selected using Phy2 (see [45]) 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).

B Vision-Only Models

B.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. Kernel size 7 performed best.

We repeated the grid search for CNNs with different number of convolutional layers. The resulting 3-layer CNN with 0.5 dropout rate 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.

B.2 Autoencoder

We initially hypothesized that an autoencoder could provide regularization benefits over a "vanilla" CNN, because the reconstruction loss might encourage the model to learn visual features that are useful for decoding. We used an encoder ϕ to map 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 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 at 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 them.

	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	$\mathbf{.390 \pm .118}$	.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	$\mathbf{.553 \pm .134}$	**.0688**	$.369 \pm .104$	.111	$\mathbf{.530 \pm .136}$	.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).

C Visual Receptive Fields

To test whether the recovered visual receptive fields are sensitive to the choice of initial values for the behavioral variables, we repeated our experiments by initializing the behavioral variables with noise sampled from the uniform distribution $[-1, 1)$. The values remained the same throughout the process of gradient ascent.

We found that different values of behavioral variables resulted in similar visual receptive fields. A representative example is shown in Fig. C.1. Although some of the absolute values changed, receptive field structure stayed qualitatively the same, with identical excitatory and inhibitory subregions to the ones reported in Fig. 4.

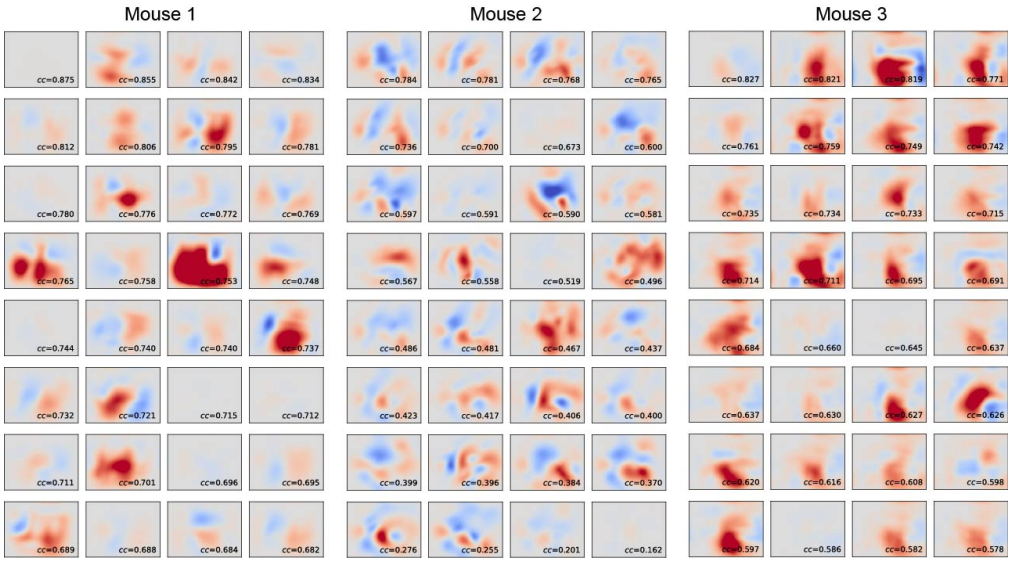

Figure C.1: The maximally activating stimuli learned in CNN+GRU$_1$, generated via gradient ascent with a fixed and randomly initialized behavioral variable. The 32 neurons with the highest cross-correlation (cc) from each mouse are shown, sorted by cc.