V1T: large-scale mouse V1 response prediction using a Vision Transformer

Anonymous authors
Paper under double-blind review

Abstract

Accurate predictive models of the visual cortex neural response to natural visual stimuli remain a challenge in computational neuroscience. In this work, we introduce V1T, a novel Vision Transformer based architecture that learns a shared visual and behavioral representation across animals. We evaluate our model on two large datasets recorded from mouse primary visual cortex and outperform previous convolution-based models by more than 12.7% in prediction performance. Moreover, we show that the self-attention weights learned by the Transformer correlate with the population receptive fields. Our model thus sets a new benchmark for neural response prediction and can be used jointly with behavioral and neural recordings to reveal meaningful characteristic features of the visual cortex.

1 Introduction

Understanding how the visual system processes information is a fundamental challenge in neuroscience. Predictive models of neural responses to naturally occurring stimuli have shown to be a successful approach toward this goal, serving the dual purpose of generating new hypotheses about biological vision (Bashivan et al., 2019; Walker et al., 2019; Ponce et al., 2019) and bridging the gap between biological and computer vision (Li et al., 2019; Sinz et al., 2019; Safarani et al., 2021). This approach relies on the idea that high performing predictive models, which explain a large part of the stimulus-driven variability, have to account for the nonlinear response properties of the neural activity, thus allowing for the identification of the underlying computations of the visual system (Carandini et al., 2005).

An extensive amount of work on the primary visual cortex (V1) has been dedicated to building quantitative models that accurately describe neural responses to visual stimuli, starting from simple linear-nonlinear models (Heeger, 1992; Jones and Palmer, 1987), energy models (Adelson and Bergen, 1985) and multi-layer models (Lehky et al., 1992; Lau et al., 2002; Prenger et al., 2004). These models, based on neurophysiological data, provide a powerful framework to test hypotheses about neural functions and investigate the principles of visual processing. With the increased popularity of deep neural networks (DNNs) in computational neuroscience in recent years (Kietzmann et al., 2018; Richards et al., 2019; Li et al., 2020; 2021), DNNs have set new standards of prediction performance (Antolík et al., 2016; Klindt et al., 2017; Ecker et al., 2018; Zhang et al., 2019), allowing for a more extensive exploration of the underlying computations in sensory processing (Walker et al., 2019; Bashivan et al., 2019; Burg et al., 2021). DNN-based models are characterized by two main approaches. On the one hand, task-driven models rely on pre-trained networks optimized on standard vision tasks, such as object recognition, in combination with a readout mechanism to predict neural responses (Yamins et al., 2014; Cadieu et al., 2014; Cadena et al., 2019a). Task-driven models have proven to be successful for predicting visual responses in primates by obtaining a shared generalized representation of the visual input across animals (Yamins and DiCarlo, 2016; Cadena et al., 2019a). However, these models do not yield the same generalization and prediction results for the mouse visual cortex (Cadena et al., 2019b). On the other hand, data-driven models share a common representation by being trained end-to-end directly on data from thousands of neurons, without any assumption on the functional properties of the network, and they have been shown to be successful as predictive models for the mouse visual cortex (Lurz et al., 2021; Franke et al., 2022).

Data-driven models for prediction of visual responses across multiple animals typically employ the core-readout framework (Klindt et al., 2017; Cadena et al., 2019a; Lurz et al., 2021; Burg et al., 2021; Franke et al., 2022). Namely, a core module which learns a shared latent representation of the visual stimuli across the animals, followed by animal-specific linear readout modules to predict neural responses given the latent features. This architecture enforces the nonlinear computations to be performed by the shared core, which can in principle capture general characteristic features of the visual cortex (Lurz et al., 2021). The readout models then learn the animal-specific mapping from the shared representation of the input to the individual neural responses. With the advent of large-scale neural recordings, datasets that consist of thousands or even hundreds of thousands of neurons are becoming readily available (Stosiek et al., 2003; Steinmetz et al., 2021). This has led to an increase in the number of parameters needed in the readout network to account for the large number of neurons, hence significant effort in neural predictive modeling has been dedicated to develop more efficient readout networks. On the other hand, due to their effectiveness and computation efficiency (Goodfellow et al., 2016), convolutional neural networks (CNNs) are usually chosen as the shared representation model.

Recently, Vision Transformer (ViT) (Dosovitskiy et al., 2021) has achieved excellent results in a broad range of computer vision tasks (Han et al., 2022) and Transformer-based (Vaswani et al., 2017) models have become increasingly popular in computational neuroscience (Tuli et al., 2021; Schneider et al., 2022; Whittington et al., 2022). For instance, Ye and Pandarinath (2021) proposed a Neural Data Transformer to model spike trains, which was extended by Le and Shlizerman (2022) using a Spatial Transformer to achieve state-of-the-art performance in 4 neural datasets. Berrios and Deza (2022) introduced a data augmentation and adversarial training procedure to train a dual-stream Transformer which showed strong performance in predicting monkey V4 responses. In modeling the mouse visual cortex, Conwell et al. (2021) experimented with a wide range of out-of-the-box DNNs, including CNNs and ViTs, to compare their representational similarity when pre-trained versus randomly initialized. Here, we explore the benefits of the ViT convolution-free approach and self-attention mechanism as the core representation learner in a data-driven neural predictive model. Note that, in this text, the term "attention" strictly refers to the self-attention layer in Transformers (Vaswani et al., 2017), which is distinct from the perceptual process of "attention" in the neuroscience literature.

Since neural variability shows a significant correlation with the internal brain state (Pakan et al., 2016; 2018; Stringer et al., 2019), information about behavior can greatly improve visual system models in the prediction of neural responses (Bashiri et al., 2021; Franke et al., 2022). To exploit this relationship, we also investigate a principled mechanism in the model architecture to integrate behavioral states with visual information.

Altogether, we propose V1T, a novel ViT-based architecture that can capture visual and behavioral representations of the mouse visual cortex. This core architecture, in combination with an efficient per-animal readout (Lurz et al., 2021), outperforms the previous state-of-the-art model by 12.7% and 19.1% on two large-scale mouse V1 datasets (Willeke et al., 2022; Franke et al., 2022), which consist of neural recordings of thousands of neurons across over a dozen behaving rodents in response to thousands of natural images. Moreover, we show that the attention weights learned by the core module correlate with behavioral variables, such as pupil direction. This link between the model and the visual cortex activity is useful for pinpointing how behavioral variables affect neural activity.

2 Neural data

We considered two large-scale neural datasets for this work, Dataset S[1] by Willeke et al. (2022) and Dataset F by Franke et al. (2022). These two datasets consist of V1 recordings from behaving rodents in response to thousands of natural images, providing an excellent platform to evaluate our proposed method and compare it against previous visual predictive models.

We first briefly describe the animal experiment in Dataset S. A head-fixed mouse was placed on a cylindrical treadmill with a 25 inch monitor placed 15 cm away from the animal's left eye and more than 7,000 neurons from layer L2/3 in V1 were recorded via two-photon calcium imaging. Note that the position of the monitor was selected such that the stimuli were shown to the center of the recorded population receptive field. Gray-scale images $x_{\text{image}} \in \mathbb{R}^{c=1\times h=144\times w=256}$ from ImageNet (Deng et al., 2009) were presented to the

[1]The Sensorium Challenge held at NeurIPS 2022 Competition Track Program

animal for 500 ms with a blank screen period of 300 to 500 ms between each presentation. Neural activities were accumulated between 50 and 500 ms after each stimulus onset. In other words, for a given neuron i in trial (stimulus presentation) t, the neural response is represented by a single value $r_{i,t}$. In addition, the anatomical coordinates of each neuron as well as five behavioral variables $x_{\text{behaviors}} \in \mathbb{R}^5$ were recorded alongside with the calcium responses. These variables include pupil dilation, the derivative of the pupil dilation, pupil center (2d-coordinates) and running speed of the animal. Each recording session consists of up to 6,000 image presentations (i.e. trials), where 5,000 unique images are combined with 10 repetitions of 100 additional unique images, randomly intermixed. The 1,000 trials with repeated images are used as the test set and the rest are divided into train and validation sets with a split ratio of 90% and 10% respectively. In total, data from 5 rodents[2] (Mouse A to E) were recorded in this dataset.

Dataset F follows largely the same experimental setup with the following distinction: colored images (UV-colored and green-colored, i.e. $x_{\text{image}} \in \mathbb{R}^{c=2\times h\times w}$) from ImageNet were presented on a screen placed 12 cm away from the animal; 4,500 unique colored and 750 monochromatic images were used as the training set and an additional 100 unique colored and 50 monochromatic images were repeated 10 times throughout the recording; in total, 10 rodents (Mouse F to O) were used in the experiment with $1,000$ V1 neurons recorded from each animal. Table A.1 summarizes the experimental information from both datasets.

3 Previous work

A substantial body of work has recently focused on predictive models of cortical activity that learn a shared representation across neurons (Klindt et al., 2017; Cadena et al., 2019a; Lurz et al., 2021; Burg et al., 2021; Franke et al., 2022), which stems from the idea in systems neuroscience that cortical computations share common features across animals (Olshausen and Field, 1996). In DNN models, these generalizing features are learned in a nonlinear core module, then a subsequent neuron-specific readout module linearly combines the relevant features in this representation to predict the neural responses. Recently, Lurz et al. (2021) and Franke et al. (2022) introduced a shared CNN core and animal-specific Gaussian readout combination that achieved excellent performance in mouse V1 neural response prediction, and this is the current state-of-the-art model on large-scale benchmarks including Dataset S and Dataset F. Here, we provide a brief description for each of the modules in their proposed architecture, which our work is built upon.

CNN core. Typically, the core module learns the shared visual representation via a series of convolutional blocks (Cadena et al., 2019a; Lurz et al., 2021; Franke et al., 2022). In Lurz et al. (2021), given an input image $x_{\text{image}} \in \mathbb{R}^{c\times h\times w}$, the CNN core with filter size k outputs a latent representation vector $z \in \mathbb{R}^{d\times h'\times w'}$ where $h' = h - k + 1$, $w' = w - k + 1$ and d is the hidden dimension. Previous works have shown correlation between behaviors and neural variability, and that the behavioral variables can significantly improve neural predictivity (Niell and Stryker, 2010; Reimer et al., 2014; Stringer et al., 2019; Bashiri et al., 2021). To that end, Franke et al. (2022) proposed to integrate the behavioral variables $x_{\text{behaviors}} \in \mathbb{R}^5$ with the visual stimulus by duplicating each variable to a $h \times w$ matrix and concatenating them with x_{image} in the channel dimension, resulting in a vector of $\mathbb{R}^{c+5\times h\times w}$.

Readout. To compute the neural response of neuron i from mouse m with n_m neurons, the readout module $\mathtt{R}_m : \mathbb{R}^{d\times h'\times w'} \to \mathbb{R}^{n_m}$ by Lurz et al. (2021) computes a linear regression of the core representation z with weights $w_i \in \mathbb{R}^{w'\times h'\times c}$, followed by an ELU activation with an offset of 1 (i.e. $o = \text{ELU}(\mathtt{R}_m(z)) + 1$), which keeps the response positive. The regression is performed by a Gaussian readout, which learns the parameters of a 2d Gaussian distribution whose mean μ_i represents the center of the receptive field of the neuron in the image space and whose variance quantifies the uncertainty of the receptive field position, which decreases over training. The response is thus obtained as a linear combination of the feature vector of the core at a single spatial position, which allows the model to greatly reduce the number of parameters per neuron in the readout. Notably, to learn the position μ_i, the model also exploits the retinotopic organization of V1 by coupling the recorded cortical 2d coordinates of each neuron with the estimated center of the receptive field from the readout. Moreover, a shifter module is introduced to adjust (or shift) the μ_i receptive field center of neuron i to account for the trial-to-trial variability due to eye movement (Franke et al., 2022). The shifter

[2] 2 additional mice were used in the Sensorium challenge (Willeke et al., 2022) and their test sets are not publicly available.

network $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ consists of 3 dense layers with hidden size of 5 and tanh activation; it takes as input the 2d pupil center coordinates and learns the vertical and horizontal adjustments needed to shift μ_i.

4 Methods

The aim of this work is to design a neural predictive model $F(x_{\text{image}}, x_{\text{behaviors}})$ that can effectively incorporate both visual stimuli and behavioral variables to predict responses o that are faithful to real recordings r from mouse V1. With that goal, we first detail the core architectures proposed in this work, followed by the training procedure and evaluation metrics. Code used in this work is attached as supplementary material and will be made publicly available upon publication.

4.1 V1T core

Vision Transformers (Dosovitskiy et al., 2021), or ViTs, have achieved competitive performance in many computer vision tasks, including object detection and semantic segmentation, to name a few (Chen et al., 2020; Carion et al., 2020; Strudel et al., 2021). Here, we propose a data-driven ViT core capable of learning a shared representation of the visual stimuli that is relevant for the prediction of neural responses in the visual cortex. Moreover, we introduce an alternative approach in V1T to encode behavioral variables in a more principled way when compared to previous methods and further improve the neural predictive performance of the overall model.

The original ViT classifier is comprised of 3 main components: (1) a tokenizer first encodes the 3d image (including channel dimension) into 2d patch embeddings, (2) the embeddings are then passed through a series of Transformer (Vaswani et al., 2017) encoder blocks, each consisting of a Multi-Head Attention (`MHA`) and a Multi-Layer Perceptron (`MLP`) module which requires 2d inputs, and finally (3) a classification layer outputs the class prediction. The following sections detail the modifications made to convert the vanilla ViT to a shared visual representation learner for the downstream readout modules. We additionally experiment with a number of recently proposed efficient ViTs that have been emphasized for learning from small to medium size datasets.

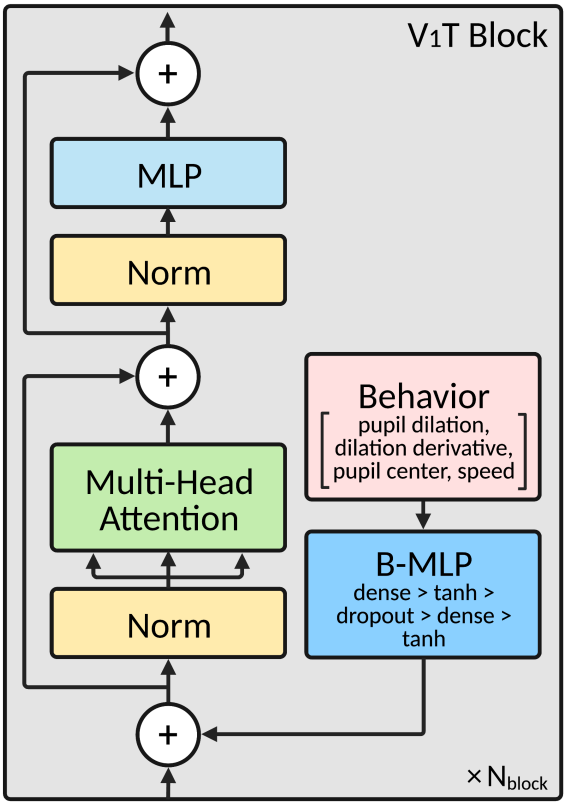


Figure 1: Illustration of the V1T block architecture.

Tokenizer. The tokenizer, or patch encoder, extracts non-overlapping squared patches of size $p \times p$ from the 2d image and projects each patch to embeddings z_0 of size d, i.e. $\mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^{l \times (cp^2)} \rightarrow \mathbb{R}^{l \times d}$, where $l = hw/p^2$ is the number of patches. Dosovitskiy et al. (2021) proposed two tokenization methods in the original ViT, where patches can be extracted either (1) via a $p \times p$ sliding window over the height and width dimensions of the image, followed by a linear layer with d hidden units, or (2) via a 2d convolutional layer with kernel size p and d filters.

Transformer-based models benefit from (or even necessitate) pre-training on large datasets, in the magnitude of millions or even billions of samples, in order to obtain optimal performance (Han et al., 2022). In contrast, typical neural recordings in animal experiments are considerably smaller. To stay consistent with previous work, we instead focus on developing a core architecture that can be effectively trained on limited

amount of data from scratch. To that end, we considered two recently introduced efficient ViT methods that are highly competitive in scarce data settings. Lee et al. (2021) proposed Shifted Patch Tokenization (SPT) to combat the low inductive bias in ViTs and enable better learning from limited data. Conceptually, SPT allows additional (adjacent) pixel values to be included in each patch, thus improving the locality, or receptive field, of the model. Input image $x_{\text{image}} \in \mathbb{R}^{1\times h\times w}$ is shifted spatially by $p/2$ in one of the four diagonal directions (top-left, top-right, bottom-left, bottom-right) with zero padding and the four shifted images (i.e. each shifted in one diagonal direction) are then concatenated with the original image, resulting in a vector $\mathbb{R}^{5\times h\times w}$, which can be processed by the two patch extraction approaches mentioned above. With a similar goal in mind, the Compact Convolutional Transformer (CCT, Hassani et al. 2021) was proposed as a convolutional tokenizer to learn the patch embeddings that can take advantage of the translation equivariance and locality inherent in CNNs. The proposed mini-CNN is fairly simple: it consists of a 2d convolution layer with a $p \times p$ kernel and filter size d, followed by ReLU activation and a max pool layer. In this work, we experimented with and compared all four tokenization methods: sliding window, a single 2d convolutional layer, SPT and CCT.

As ViTs are agnostic to the spatial structure of the data, a positional embedding is added to each patch to encode the relative position of the patches with respect to each other (Dosovitskiy et al., 2021; Han et al., 2022) and this positional embedding can either be learned or sinusoidal. Finally, a learnable BERT (Devlin et al., 2019) `[cls]` token is typically added to the patch embeddings (i.e. $z_0 \in \mathbb{R}^{(l+1)\times d}$) to represent the class of the image.

Transformer encoder. The encoder consists of a series of ViT blocks, where each block comprises two sub-modules: Multi-Head Attention (`MHA`) and Multi-Layer Perceptron (`MLP`). In each `MHA` module, we applied the standard self-attention formulation (Vaswani et al., 2017): $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d})V$, where query Q, key K and value V are linear projections of the input z_b at block b. Conceptually, the self-attention layer assigns a pairwise attention value among all the patches (or tokens). In addition to the standard formulation, we also experimented with the Locality Self Attention (LSA, Lee et al. 2021), where a diagonal mask is applied to QK^T to prevent strong connections in self-tokens (i.e. diagonal values in QK^T), thus improving the locality inductive bias. Each sub-module is preceded by Layer Normalization (`LayerNorm`, Ba et al. 2016), and followed by a residual connection to the next module.

Reshape representation. To make the dimensions compatible with the Gaussian readout module (see Section 3 for an overview), we reshape the 2d core output $z \in \mathbb{R}^{l\times d}$ to $\mathbb{R}^{d\times h'\times w'}$, where $l = h' \times w'$ and $h' \leq w'$. Note that if the number of patches l is not sufficiently large, it is possible for the same position in z to be mapped to multiple neurons, which could lead to adverse effects. For instance, in the extreme case of $l = 1$, all neurons would be mapped to a single $p \times p$ region in the visual stimulus (i.e. they would have the same visual receptive field), which is not biologically plausible given the size of the recorded cortical area (Garrett et al., 2014). We therefore set the stride size of the patch encoder as a hyperparameter and allow for overlapping patches, thus letting the hyperparameter optimization algorithm select the optimal number of patches.

4.1.1 Incorporating behaviors

Previous studies have shown that visual responses can be influenced by behavioral variables and brain states; for example, changes in arousal, which can be monitored by tracking pupil dilation, lead to stronger (or weaker) neural responses (Reimer et al., 2016; Larsen and Waters, 2018). As a consequence, the visual representation learned by the core module should also be adjusted according to the brain state. Here, instead of concatenating the behavioral variables as additional channels in the image (see Section 3), we propose an alternative method to integrate behavioral variables with visual stimuli using a novel ViT architecture – V1T, illustrated in Figure 1.

We introduced a behavior MLP module (`B-MLP` : $\mathbb{R}^5 \to \mathbb{R}^d$) at the beginning of the encoder block which learns to adjust the visual latent vector z based on the observed behavioral states $x_{\text{behaviors}}$. Each `B-MLP` module comprises two fully-connected layers with d hidden units and a dropout layer in between; tanh activation is used so that the adjustments to z can be both positive and negative. Importantly, as layers in DNNs learn different features of the input, usually increasingly abstract and complex with deeper layers (Zeiler and

Fergus, 2014; Raghu et al., 2021), we hypothesize that the influence of the internal brain state should therefore change from layer to layer. To that end, we learned a separate $\texttt{B-MLP}_b$ at each block b in the V1T core, thus allowing level-wise adjustments to the visual latent variable. Formally, $\texttt{B-MLP}_b$ projects $x_{\text{behaviors}}$ to the same dimension of the embeddings z_{b-1}, followed by an element-wise summation between latent behavioral and visual representations, and then the rest of the operations in the encoder block:

$$z_b \leftarrow z_{b-1} + \texttt{B-MLP}_b(x_{\text{behaviors}}) \tag{1}$$

$$z_b \leftarrow \texttt{MHA}_b(\texttt{LayerNorm}(z_b)) + z_b \tag{2}$$

$$z_b \leftarrow \texttt{MLP}_b(\texttt{LayerNorm}(z_b)) + z_b \tag{3}$$

where z_0 denotes the original patch embeddings.

4.2 Training and evaluation

In order to isolate the change in prediction performance that is solely due to the proposed core architectures, we employed the same readout architectures by Lurz et al. (2021), as well as a similar data preprocessing and model training procedure. We used the same train, validation and test split provided by the two datasets (see Section 2). Natural images, recorded responses, and behavioral variables (i.e. pupil dilation, dilation derivative, pupil center, running speed) were standardized using the mean and standard deviation measured from the training set and the images were then resized to 36×64 pixels from 144×256 pixels. The shared core and per-animal readout modules were trained jointly using the AdamW optimizer (Loshchilov and Hutter, 2019) to minimize the Poisson loss

$$\mathcal{L}_m^{\text{Poisson}}(r, o) = \sum_{t=1}^{n_t} \sum_{i=1}^{n_m} \Big(o_{i,t} - r_{i,t} \log(o_{i,t}) \Big) \tag{4}$$

between the recorded responses r and predicted responses o, where n_t is the number of trials in one batch and n_m the number of neurons for mouse m. A small value $\varepsilon = 1e-8$ was added to both r and o prior to the loss calculation to improve numeric stability. Gradients from each mouse were accumulated before a single gradient update to all modules. We tried to separate the gradient update for each animal, i.e. one gradient update per core-readout combination, but this led to a significant drop in performance. We suspect this is because the core module failed to learn a generalized representation among all animals when each update step only accounted for gradient signals from one animal. We used a learning rate scheduler in conjunction with early stopping: if the validation loss did not improve over 10 consecutive epochs, we reduced the learning rate by a factor of 0.3; if the model still had not improved after 2 learning rate reductions, we then terminated the training process. Dropout (Srivastava et al., 2014), stochastic depths (Huang et al., 2016), and L1 weight regularization were added to prevent overfitting. The weights in dense layers were initialized by sampling from a truncated normal distribution ($\mu = 0.0, \sigma = 0.02$), where the bias values were set to 0.0; whereas the weight and bias in `LayerNorm` were set to 1.0 and 0.0. Each model was trained on a single Nvidia RTX 2080Ti GPU and all models converged within 200 epochs. Finally, we employed Hyperband Bayesian optimization (Li et al., 2017) to find the hyperparameters that achieved the best validation loss. This included finding the optimal tokenization method and self-attention mechanism. The initial search space and final hyperparameter settings are detailed in Table A.2. While learning rate warm-up and pre-training on large datasets are considered the standard approach to train Transformers (Xiong et al., 2020; Han et al., 2022), in order to stay consistent with previous work and to isolate the performance gain solely due to the architectural change, all models presented in this work are trained from scratch and follow the same procedure stated above.

The prediction performance of our models was measured by the single trial correlation metric, used by Willeke et al. (2022) and Franke et al. (2022), which can also account for the trial-to-trial variability in the test set where the same visual stimuli were shown multiple times. We computed the correlation between recorded r and predicted o responses:

$$\text{trial corr.}(r, o) = \frac{\sum_{i,j}(r_{i,j} - \bar{r})(o_{i,j} - \bar{o})}{\sqrt{\sum_{i,j}(r_{i,j} - \bar{r})^2 \sum_{i,j}(o_{i,j} - \bar{o})^2}} \tag{5}$$

where $\bar{r}$ and $\bar{o}$ are the average recorded and predicted responses across all trials in the test set.

5 Results

Here, we first discuss the final core architecture chosen after the Bayesian hyperparameter optimization, followed by a comparison of our proposed core against baseline models on the two large-scale mouse V1 datasets. Moreover, we analyze the trained core module and present the insights that can be gained from it. We present the cross-animal and cross-dataset generalization, as well as sample efficiency performance of the model, in Appendix A.2 and A.3.

Table 1: The single trial correlation (TRIAL CORR.) between predicted and recorded responses across all animals in DATASET S (Willeke et al., 2022) and DATASET F (Franke et al., 2022) test set. Δ% shows the relative improvement against the CNN model (Lurz et al., 2021) with behavior variables; we additionally fitted a CNN and ViT core with stimulus-response pairs (BEHAVIOR: ✕) to evaluate the prediction performance without behavioral information. Detailed per-animal results are available in Appendix A.1.

		DATASET S		DATASET F	
METHOD	BEHAVIOR	TRIAL CORR.	Δ%	TRIAL CORR.	Δ%
LN	✓	0.275	-27.2%	0.223	-27.8%
CNN	✕	0.300	-20.6%		
CNN	✓	0.378	0.0%	0.309	0.0%
ViT	✕	0.319	-15.6%		
ViT	✓	0.414	+9.5%	0.344	+11.3%
V1T	✓	**0.426**	+12.7%	**0.368**	+19.1%
ENSEMBLE OF 5 MODELS					
CNN	✓	0.404	+6.9%	0.340	+10.0%
V1T	✓	**0.439**	+16.1%	**0.378**	+22.3%

V1T benefits from smaller and overlapping patches. We first looked at how hyperparameters of ViT and V1T affect model performance. We observed the predictive performance to be quite sensitive towards number of patches, patch size and patch stride. The most performant models used a patch size of 8 and a stride size of 1, thus extracting the maximum number of patches. We note that this allows the readout to learn a mapping from the shared core representation of the stimulus to the cortical position of each neuron that spans across the whole image, and not just a part of the image. Since the visual receptive fields of neurons are distributed across a large area of the monitor given the size of the recorded cortical area, this leads to more accurate response predictions from the model. Furthermore, we found that the two efficient tokenizers, SPT and CCT, whose aim is to reduce the number of patches, both failed to improve the model performance, reiterating that a finer tiling of the image is crucial for accurate predictions of cortical activity. Moreover, we found that the LSA attention mechanism, which encourages the model to learn from inter-tokens by masking out the diagonal self-token, led to worse performance, suggesting information from adjacent patches in this task is not as influential as it is in image classification. Lastly, we found that V1T with layer-wise `B-MLP` modules yields the best results, indicating that the modulation introduced by behavioral information varies as the core learns different visual representations at deep layers. Further analysis and discussion on the `B-MLP` module are presented in Appendix A.5.

V1T outperforms CNN. Next, we compared the tuned ViT and V1T cores against a baseline linear-nonlinear (LN) model and the previous state-of-the-art CNN model (Lurz et al., 2021) on the two large scale mouse V1 datasets (see Section 2). We also trained a CNN and ViT core on response-stimuli pairs only on `Dataset S`, to evaluate the importance of behavioral information in response predictions. Table 1 summarizes the test performance on the two datasets, results of per-animal performance and an alternative metric are available in Appendix A.1. By simply replacing the CNN core module with the tuned ViT architecture, we observed a considerable improvement in response predictions across all animals, with an average increase of 9.5% and 11.3% in single trial correlation over the CNN model in DATASET S and DATASET F respectively. Thus far, the core module encoded the brain state of the animals by concatenating behavioral variables as additional channels in the natural image. With that said, our proposed V1T core, which encodes the brain

state via the `B-MLP` nonlinear transformations, further improved the average prediction performance by 2.9% and 7.0% in the two datasets, or 12.7% and 19.1% over the CNN model.

As demonstrated in the Sensorium Challenge (Sensorium Workshop, 2022) and Franke et al. (2022), ensemble learning is a common approach to improve neural predictive models. Following the procedure in Franke et al. (2022), we trained 10 models with different random seeds and selected the 5 best models based on their validation performance. The average of the selected models constituted the output of the ensemble model. The CNN ensemble model achieved an average improvement of 6.9% in Dataset S as compared to its non-ensemble variant. Nevertheless, the individual V1T model still outperformed the CNN ensemble by 5.4%. A V1T ensemble trained with the same procedure achieved an average single trial correlation of 0.439, which corresponds to an 8.7% improvement over the CNN ensemble model. Altogether, our proposed core architecture set a new benchmark in both gray-scale and colored visual response prediction.

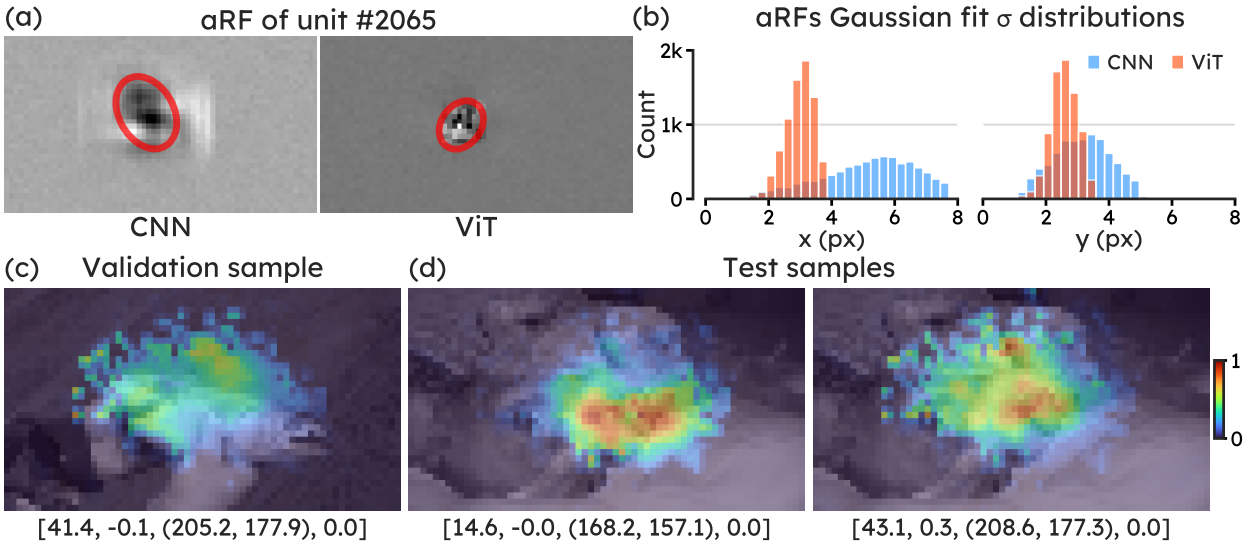


Figure 2: (a) Estimated artificial receptive field (aRF) and 2d Gaussian fit (red circle shows 1 standard deviation ellipse) of the same artificial unit from the CNN and ViT models trained without behaviors. Visually, the ViT learns narrower aRFs, more examples in Appendix A.4. To quantify the size of the aRFs, we compared the fitted Gaussian over all units from `Mouse A` and (b) the distributions of the standard deviations showed that the ViT learns notably narrower aRFs. V1T attention visualization on `Mouse A` (c) validation and (d) test samples. Each attention map was normalized to $[0, 1]$, and the behavioral variables of the corresponding trial are shown below the image in the format of [pupil dilation, dilation derivative, pupil center (x, y), speed]. More examples in Appendix A.6.

Spatial tuning difference. As expected, models trained without behavioral information led to worse results (see behavior: ✕ in Table 1). Nevertheless, we observed an average 6.3% improvement in stimuli-response prediction with the tuned ViT core over the CNN model in `Dataset S`. To further our understanding of why the ViT might be performing better in visual response prediction, we evaluated the discrepancies in spatial tuning of the two models by comparing their artificial receptive fields (aRFs). Appendix A.4 details the procedure. Briefly, we presented the models with thousands of white noise images and then summed the images weighted by the response prediction to estimate the aRF of each artificial unit. Figure 2a shows the aRF of the same artificial unit from the CNN and ViT model. Visually, the aRFs of the ViT model appear to be narrower and qualitatively different from the aRFs of the CNN. In order to quantify the aRF sizes, we fitted a 2d Gaussian to each aRF and observed a significant difference in the standard deviation distributions, shown in Figure 2b. Overall, the aRFs of the ViT model have a much narrower spread, with a mean standard deviation of 3.0 ± 0.5 and 2.6 ± 0.4 in the horizontal and vertical directions over all artificial units, considerably lower than the 5.1 ± 1.5 and 3.1 ± 0.9 of the CNN. These results show that the artificial units in the CNN and ViT learn notably different aRFs. Given that we did not constrain the aRF size, our

results suggest that the narrower fields allow ViT to learn location-dependent features that are beneficial for visual response prediction.

Self-attention visualization. In addition to the performance gain in the proposed core architecture, the self-attention mechanism inherent in Transformers can be used to visualize areas in the input image that the model learns to focus on. In our case, it allows us to detect the regions in the visual stimulus that drive the neural responses. To that end, we extracted the per-stimulus attention map learned by the V1T core module via Attention Rollout (Abnar and Zuidema, 2020; Dosovitskiy et al., 2021). Briefly, we aggregated the attention weights (i.e. $\text{Softmax}(QK^T/\sqrt{d})$) across all heads in `MHA`, and then multiplied the weights over all layers (blocks), recursively. Figure 2 shows the normalized average attention weights superimposed to the input images from Mouse A, with more examples available in Appendix A.6. Given that the position of the computer monitor was chosen in order to center the population receptive field, V1 responses from the recorded region should be mostly influenced by the center of the image Willeke et al. (2022). Here, we can see a clear trend where the core module is focusing on the central regions of the images to predict the neural response, which aligns with our expectation from the experiment conditions. Interestingly, when the core module inputs the same image but with varying behaviors (i.e. Figure 2d), we noticed variations in the attention patterns. This suggests that the V1T core is able to take behavioral variables into consideration and adjust its attention solely based on the brain state.

These attention maps can inform us of the area of the image (ir)relevant for triggering visual neuronal responses which, in turn, allow us to build more sophisticated predictive models. For instance, the core module consistently assigned higher weights to patches in the center of the image, suggesting information at the edges of the image are less (or not at all) relevant for the recorded group of neurons. As a practical example, we eliminated irrelevant information in the stimuli by center cropping the image to $\alpha 144 \times \alpha 256$ pixels where $0 < \alpha \leq 1$, prior to downsampling the input to 36×64 pixels. We found that a crop factor of $\alpha = 0.8$ (i.e. removing 36% of the total number of pixels) further improved the single trial correlation to 0.430, or 13.8% better than the CNN. Note that we also obtained similar improvement with the CNN model.

Self-attention correlates with pupil center. To further explore the relationship between the attention weights learned by the core module and the behavioral information, we measured the absolute correlation between the center of mass of the attention maps and the pupil centers in the vertical and horizontal axes. The correlation coefficient of each animal in Dataset S is summarized in Table 2. Overall, we found a moderate correlation between the attention maps and the pupil center of the animal, with an average correlation of 0.525 ± 0.079 and 0.409 ± 0.105 in the horizontal and vertical directions across animals. This relationship demonstrates that the attention maps can reveal the impact of behavioral variables on the neural responses. Therefore, this framework can be particularly useful for studies investigating the coding of visual information across visual cortical areas (V1 and higher visual areas), as the model could determine what part(s) of the visual stimulus is processed along the "hierarchy" of visual cortical areas. Since higher visual areas are known to have larger receptive fields (Wang and Burkhalter, 2007; Glickfeld et al., 2014), we would expect a larger part of the image to be relevant for the core module. Further investigation of the attention map could also be used to determine which part of a visual scene was relevant when performing more specific tasks, such as object recognition, decision-making, or spatial navigation.

Table 2: Correlations between the center of mass of the attention maps and pupil centers in the (x-axis) horizontal and (y-axis) vertical direction in Dataset S test set, all with a p-value $\ll 0.0001$.

Mouse	x-axis	y-axis
A	0.682 (****)	0.568 (****)
B	0.489 (****)	0.493 (****)
C	0.505 (****)	0.370 (****)
D	0.484 (****)	0.310 (****)
E	0.464 (****)	0.302 (****)

6 Discussion

In this work, we presented a novel core architecture V1T to model the visual and behavioral representations of mouse V1 activities in response to natural visual stimuli. The model outperformed the previous state-of-the-art CNN (Lurz et al., 2021) model on two large-scale mouse V1 datasets by a considerable margin (12.7% and 19.1%). In contrast to the winning submissions at the Sensorium Challenge (Sensorium Workshop, 2022), which focused on data augmentation and building large ensembles based on the CNN model, we instead introduced a new architecture as the shared core module. Our best model achieved a single trial correlation of 0.428 and 0.444 (correlation to average: 0.634 and 0.650) in the two held-out test sets, which would place us 2nd place in the leaderboard, and the best method across all models not taking the neuronal response trends over time into account. To the best of our knowledge, our approach is also the first ViT-based model to outperform CNNs in mouse V1 response prediction.

With a strong neural predictive performance, this model also provides a framework to investigate *in silico* the computations in the visual system. For instance, the center of the self-attention maps learned by our model correlates with the pupil center of the animals, highlighting how features of this architecture do reflect properties of cortical neurons' receptive fields, in this case, the retinotopy. This does not imply that our model can directly map onto the anatomy or biophysics of the brain. However, a good predictive models can be a valuable guide for understanding functional properties of the system and provide a compact description of the transformation of stimuli into the response, thus establishing a link between neural activity and perception (Bashivan et al., 2019; Fu et al., 2023).

Nevertheless, the V1T model has a number of limitations. Firstly, only one-dimensional behavioral information can be incorporated since the model integrates scalars into the latent embedding via the `B-MLP` module. Additional architecture engineering is needed if the behavioral variables have varying (and higher) dimensions such as, for instance, 3D poses (Mathis et al., 2018). Secondly, in the case of very limited data (e.g. 500 samples, see Appendix A.3), CNN-based models are likely to outperform ViTs, which typically require considerable amount of data to be performant (Han et al., 2022).

In future work, we plan to further investigate the relationship between behavioral variables and neural responses. The attention visualization technique, for instance, enables ablation studies on the effect of each behavioral variable, such as pupil dilation or running speed, on the neural activity. Moreover, we plan to extend the method to recordings of the visual cortex in response to natural videos, to track how this relationship may evolve over time, as well as experiments in naturalistic settings, to know which part of a visual scene is relevant for certain behaviors.

References

Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299.

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

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bashiri, M., Walker, E., Lurz, K.-K., Jagadish, A., Muhammad, T., Ding, Z., Ding, Z., Tolias, A., and Sinz, F. (2021). A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34:15801–15815.

Bashivan, P., Kar, K., and DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436.

Berrios, W. and Deza, A. (2022). Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art brain-score for area v4. In *SVRHM 2022 Workshop @ NeurIPS*.

Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., and Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. (2019a). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897.

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., and Ecker, A. S. (2019b). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons & Hidden Units: Future directions at the intersection of neuroscience and artificial intelligence @ NeurIPS 2019*.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. (2020). Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

Conwell, C., Mayo, D., Barbu, A., Buice, M., Alvarez, G., and Katz, B. (2021). Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex. *Advances in Neural Information Processing Systems*, 34.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Ecker, A. S., Sinz, F. H., Froudarakis, E., Fahey, P. G., Cadena, S. A., Walker, E. Y., Cobos, E., Reimer, J., Tolias, A. S., and Bethge, M. (2018). A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv preprint arXiv:1809.10504*.

Franke, K., Willeke, K. F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F. H., et al. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930):128–134.

Fu, J., Shrinivasan, S., Ponder, K., Muhammad, T., Ding, Z., Wang, E., Ding, Z., Tran, D. T., Fahey, P. G., Papadopoulos, S., et al. (2023). Pattern completion and disruption characterize contextual modulation in mouse visual cortex. *bioRxiv*, pages 2023–03.

Garrett, M. E., Nauhaus, I., Marshel, J. H., and Callaway, E. M. (2014). Topography and areal organization of mouse visual cortex. *Journal of Neuroscience*, 34(37):12587–12600.

Glickfeld, L. L., Reid, R. C., and Andermann, M. L. (2014). A mouse model of higher visual cortical function. *Current opinion in neurobiology*, 24:28–33.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*.

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. (2021). Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.

Heeger, D. J. (1992). Half-squaring in responses of cat striate cells. *Visual neuroscience*, 9(5):427–443.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer.

Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, page 133504.

Klindt, D., Ecker, A. S., Euler, T., and Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". *Advances in Neural Information Processing Systems*, 30.

Larsen, R. S. and Waters, J. (2018). Neuromodulatory correlates of pupil dilation. *Frontiers in neural circuits*, 12:21.

Lau, B., Stanley, G. B., and Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences*, 99(13):8974–8979.

Le, T. and Shlizerman, E. (2022). STNDT: Modeling neural population activity with spatiotemporal transformers. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Lee, S. H., Lee, S., and Song, B. C. (2021). Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*.

Lehky, S. R., Sejnowski, T. J., and Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *Journal of Neuroscience*, 12(9):3568–3581.

Li, B. M., Amvrosiadis, T., Rochefort, N., and Onken, A. (2020). Calciumgan: A generative adversarial network model for synthesising realistic calcium imaging data of neuronal populations. *arXiv preprint arXiv:2009.02707*.

Li, B. M., Amvrosiadis, T., Rochefort, N., and Onken, A. (2021). Neuronal learning analysis using cycle-consistent adversarial networks. *arXiv preprint arXiv:2111.13073*.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816.

Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F., Pitkow, Z., and Tolias, A. (2019). Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lurz, K.-K., Bashiri, M., Willeke, K., Jagadish, A., Wang, E., Walker, E. Y., Cadena, S. A., Muhammad, T., Cobos, E., Tolias, A. S., Ecker, A. S., and Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. In *International Conference on Learning Representations*.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289.

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

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Pakan, J. M., Francioni, V., and Rochefort, N. L. (2018). Action and learning shape the activity of neuronal circuits in the visual cortex. *Current opinion in neurobiology*, 52:88–97.

Pakan, J. M., Lowe, S. C., Dylda, E., Keemink, S. W., Currie, S. P., Coutts, C. A., and Rochefort, N. L. (2016). Behavioral-state modulation of inhibition is context-dependent and cell type specific in mouse visual cortex. *Elife*, 5.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., and Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.

Prenger, R., Wu, M. C.-K., David, S. V., and Gallant, J. L. (2004). Nonlinear v1 responses to natural scenes revealed by neural network analysis. *Neural Networks*, 17(5-6):663–679.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128.

Reimer, J., Froudarakis, E., Cadwell, C. R., Yatsenko, D., Denfield, G. H., and Tolias, A. S. (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *neuron*, 84(2):355–362.

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., and Tolias, A. S. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, 7(1):1–7.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770.

Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., Tolias, A., and Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34:739–751.

Schneider, S., Lee, J. H., and Mathis, M. W. (2022). Learnable latent embeddings for joint behavioral and neural analysis. *arXiv preprint arXiv:2204.00673*.

Sensorium Workshop (2022). Sensorium competition presentation - NeurIPS 2022 workshop. https://youtu.be/lncQPWROFbc.

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., and Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, 103(6):967–979.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539).

Stosiek, C., Garaschuk, O., Holthoff, K., and Konnerth, A. (2003). In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, 100(12):7319–7324.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., and Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272.

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., and Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065.

Wang, Q. and Burkhalter, A. (2007). Area map of mouse visual cortex. *Journal of Comparative Neurology*, 502(3):339–357.

Whittington, J. C. R., Warren, J., and Behrens, T. E. (2022). Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*.

Willeke, K. F., Fahey, P. G., Bashiri, M., Pede, L., Burg, M. F., Blessing, C., Cadena, S. A., Ding, Z., Lurz, K.-K., Ponder, K., et al. (2022). The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666*.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. (2020). On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Ye, J. and Pandarinath, C. (2021). Representation learning for neural population activity with Neural Data Transformers. *Neurons, Behavior, Data analysis, and Theory*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, Y., Lee, T. S., Li, M., Liu, F., and Tang, S. (2019). Convolutional neural network models of v1 responses to complex patterns. *Journal of computational neuroscience*, 46(1):33–54.

A Appendix

Table A.1: Experimental information of Mouse A to E from Dataset S (Willeke et al., 2022) and Mouse F to O from Dataset F (Franke et al., 2022). Each mouse has a unique recording ID (column 2) although we assigned a separate mouse ID (column 1) to use throughout this paper for simplicity.

Mouse	rec. ID	num. neurons	total trials	num. test
A	21067-10-18	8372	5994	998
B	22846-10-16	7344	5997	999
C	23343-5-17	7334	5951	989
D	23656-14-22	8107	5966	993
E	23964-4-22	8098	5983	994
F	25311-10-26	867	7358	1475
G	25340-3-19	922	7478	1497
H	25704-2-12	773	7500	1500
I	25830-10-4	1024	7360	1473
J	26085-6-3	910	7464	1495
K	26142-2-11	1121	7500	1500
L	26426-18-13	1125	7500	1500
M	26470-4-5	1160	7473	1495
N	26644-6-2	824	7500	1500
O	26872-21-6	1109	7466	1495

Table A.2: ViT and V1T cores - Gaussian readout hyperparameter search space and their final settings after a Hyperband Bayesian optimization (Li et al., 2017).

hyperparameter	search space	final value
Core		
num. blocks	uniform, min: 1, max: 8	4
num. heads	uniform, min: 1, max: 12	4
patch size	uniform, min: 2, max: 16	8
patch stride	uniform, min: 1, max: patch size	1
patch method	sliding window, 2d conv, SPT, CCT	sliding window
patch dropout	uniform, min: 0, max: 0.5	0.0229
embedding size	uniform, min: 8, max: 1024, interval: 1	155
MHA dropout	uniform, min: 0, max: 0.5	0.2544
MLP size	uniform, min: 8, max: 1024, interval: 1	488
MLP dropout	uniform, min: 0, max: 0.5	0.2544
stochastic depth dropout	uniform, min: 0, max: 0.5	0.0
L1 weight regularization	uniform, min: 0, max: 1	0.5379
initial learning rate	uniform, min: 0.005, max: 0.0001	0.0016
Readout		
position network num. layers	uniform, min: 1, max: 4, interval: 1	1
position network num. units	uniform, min: 2, max: 128, interval: 2	30
bias initialization	0, mean standardized response	0
L1 weight regularization	uniform, min: 0, max: 1	0.0076

A.1 Prediction results

Table A.3: Single trial correlation between predicted and recorded responses in Dataset S test set. All models were trained with behaviors. To demonstrate that the extracted attention maps can inform us about the (ir)relevant regions in the visual stimulus, we trained an additional V1T core with images center cropped to $\alpha h \times \alpha w$ pixels.

	Mouse					
Method	A	B	C	D	E	avg (gain)
LN	0.262	0.306	0.281	0.263	0.262	0.275 (-27.2%)
CNN	0.350	0.424	0.385	0.371	0.360	0.378 (0.0%)
ViT	0.375	0.455	0.415	0.433	0.392	0.414 (+9.5%)
V1T	0.401	0.464	0.430	0.436	0.401	0.426 (+12.7%)
V1T (center crop $\alpha = 0.8$)	**0.403**	**0.468**	**0.433**	**0.442**	**0.403**	**0.430** (+13.8%)
Ensemble of 5 models						
CNN	0.379	0.443	0.409	0.406	0.385	0.404 (+6.9%)
V1T	**0.414**	**0.475**	**0.443**	**0.452**	**0.413**	**0.439** (+16.1%)

Table A.4: Single trial correlation between predicted and recorded responses in Dataset F test set. All models were trained with behaviors.

	Mouse										
Method	F	G	H	I	J	K	L	M	N	O	avg (gain)
LN	0.194	0.254	0.214	0.279	0.255	0.233	0.148	0.231	0.174	0.243	0.223 (-27.8%)
CNN	0.253	0.371	0.184	0.377	0.329	0.319	0.207	0.331	0.341	0.376	0.309 (0.0%)
ViT	0.310	0.375	0.352	0.379	0.385	0.262	0.294	0.360	0.358	0.368	0.344 (+11.3%)
V1T	**0.326**	**0.386**	**0.387**	**0.394**	**0.398**	**0.373**	**0.298**	**0.377**	**0.363**	**0.379**	**0.368** (+19.1%)

Correlation to Average (Willeke et al., 2022) is another commonly used metric to evaluate neural predictive models. It is the correlation between $r_{i,j}$ recorded and $o_{i,j}$ predicted responses over repeated j trials of stimulus i :

$$\text{avg. corr.}(r, o) = \frac{\sum_i (\bar{r}_i - \bar{r})(o_i - \bar{o})}{\sqrt{\sum_i (\bar{r}_i - \bar{r})^2 \sum_i (o_i - \bar{o})^2}} \tag{6}$$

where $\bar{r}_i = \frac{1}{J}\sum_{j=1}^{J} r_{i,j}$ is the average response across J repeats, and $\bar{r}$ and $\bar{o}$ are the average recorded and predicted responses across all trials.

Table A.5: The Correlation to Average (avg. corr.) between predicted and recorded responses across all animals in Dataset S and Dataset F test sets. Table 1 shows the results in single trial correlation.

		Dataset S		Dataset F	
Method	behavior	avg. corr.	Δ%	avg. corr.	Δ%
Linear	✓	0.387	-33.1%	0.312	-39.7%
CNN	✗	0.551	-4.7%		
CNN	✓	0.578	0.0%	0.516	0.0%
ViT	✗	0.568	-1.7%		
ViT	✓	0.621	7.4%	0.542	4.9%
V1T	✓	**0.629**	8.9%	**0.551**	6.6%
Ensemble of 5 models					
CNN	✓	0.610	5.5%	0.567	9.9%
V1T	✓	**0.644**	11.3%	**0.562**	8.9%

A.2 Cross-animal and cross-dataset generalization

DNN-based neural predictive models are often neuron/animal specific and do not generalize well to unseen neurons/animals. Here, we evaluate the cross-animal and cross-dataset generalization performance of our proposed model against the CNN model.

We first tested the cross-animal performance of the CNN and V1T models by performing cross-validation over animals in Dataset S (Willeke et al., 2022). Specifically, we compare the model fitted on one animal (direct setting) against a model that was pre-trained on $N-1$ animals and whose readout was fine-tuned (with core frozen) on the left-out animal (transfer setting). We repeated this process for all 5 animals, and their results are summarized in Table A.6. On average, the V1T model outperformed the CNN model by 3.3% and 6.7% in the direct and transfer settings, respectively. Moreover, the V1T model experienced a larger level of performance gain in the transfer learning setting, with an average prediction improvement of 5.6% over direct training, whereas the CNN had a 2.2% gain. These results suggest that the V1T core can generalize well to unseen animals, and also benefit from transfer learning to a greater extent.

Next, we evaluated the cross-dataset generalization performance. To that end, we fitted a model on the gray-scaled version of Dataset F (Franke et al., 2022) only (by averaging across the color channel of the image). We then froze the core module and trained the readouts on Dataset S and compared the loss in performance in this transfer setting. The results are presented in Table A.7 for the two core architectures. We observed a larger performance drop with the frozen V1T model compared to the model trained directly, with an average deficit of -19.0%, versus the -12.7% drop in the frozen CNN model. Similar to the cross-animal generalization, the CNN model exhibits a higher level of variation in prediction performance over the 5 animals. While the relative performance drop was greater for the V1T core than for the CNN core, V1T achieved better single trial correlations in both the original and the transfer settings, with the readout modules trained on the frozen V1T core achieving an average single trial correlation of 0.345, or about 4.9% better than the frozen CNN counterpart (0.329).

Table A.6: **CNN vs V1T cross-animal generalization in Dataset S**. We compare the test performance between (direct) fitting one model per animal and (transfer) pre-training a model on $N-1$ animals and fine-tuning the readout for the Nth animal. We repeat the same leave-one-out process for all animals. gain shows the relatively prediction performance of the transfer models over the direct models.

	Mouse						
	A	B	C	D	E	avg.	gain
CNN							
direct	0.332	0.422	0.389	0.400	0.335	0.376	
transfer	0.357	0.420	0.386	0.398	0.359	0.384	2.2%
V1T							
direct	0.368	0.417	0.394	0.414	0.347	0.388	
transfer	0.384	0.450	0.414	0.415	0.385	0.410	5.6%

Table A.7: **CNN vs V1T cross-dataset generalization**. We first pre-trained the core module on a gray-scale version of Dataset F, then (transfer) froze the core and fine-tuned the readouts on Dataset S. gain shows the test performance drop in the cross-dataset transfer learning setting as compare (original) a model directly trained on Dataset S .

	Mouse						
	A	B	C	D	E	avg.	gain
CNN							
original	0.350	0.424	0.385	0.371	0.360	0.378	
transfer	0.314	0.353	0.337	0.316	0.327	0.329	-12.7%
V1T							
original	0.401	0.464	0.430	0.436	0.401	0.426	
transfer	0.327	0.382	0.347	0.343	0.328	0.345	-19.0%

A.3 Sample efficiency

Most neural datasets are constrained by their limited size, due to technical and/or ethical limitations, while typical DNNs require a large amount of data to train on, especially Transformer-based models (Han et al., 2022). Here, we evaluate the sample efficiency of the CNN and V1T models. We trained the two models on 500 (11%), 1500 (33%), 2500 (55%), 3500 (77%) and 4500 (100%) samples per animal in Dataset S (Willeke et al., 2022). Figure A.1 shows the single trial correlation in the test set for models trained on different sample sizes, each with 30 different random seeds. Overall, we found that V1T outperforms the CNN model even at 1500 training samples per animal. Moreover, the predictive performance of the CNN model plateaued at around 3500 training samples, while V1T continues to improve, suggesting that the ViT-based model can continue to improve with more data.

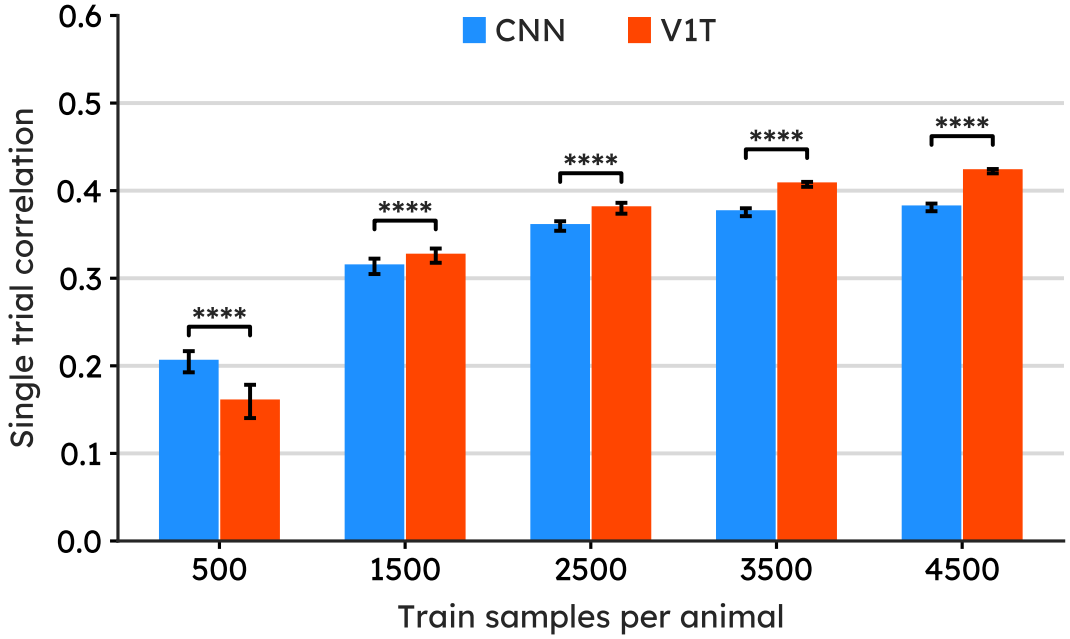


Figure A.1: CNN vs V1T prediction performance when trained with 500, 1500, 2500, 3500 and 4500 (all) samples per animal in Dataset S. The two models were each trained with 30 different random seeds, the error bar shows the standard deviations of the repeated experiments, and the statistical difference (two-sided t-test) between CNN and V1T in each sample group is shown above each bar pairs (****: $p \leq 0.0001$).

A.4 Artificial receptive fields

Here, we outline the procedure to estimate the artificial receptive fields (aRFs) of the CNN and ViT models (not V1T, since there is no behavior involved) and the process to compare their spatial positions and sizes. We first present each trained model with $N = 500,000$ images of white noise drawn from a uniform distribution. The aRF of unit i is then computed as the summation of all noise images, weighted by the respective output:

$$\text{aRF}_i = \sum_{n}^{N} \texttt{M}(x_n)_i * x_n, \quad x_n \sim \mathcal{U}^{1\times 36\times 64} \tag{7}$$

where model `M` can be either the CNN or ViT, and $\texttt{M}(x_n)_i$ denotes the response of unit i given white noise image x_n. Figure A.2 shows the estimated aRFs of 3 randomly selected artificial units (out of 8372 in the readout for Mouse A) from the two models.

To quantify the location and size of the aRFs, we fitted a 2d Gaussian to each aRF and compared the mean and covariance of the fitted parameters. We repeated the same process for all 8372 artificial units. Concretely, we first subtracted the mean from each aRF to center the values on the baseline, then took their absolute values and fitted a 2d Gaussian using SciPy's `curve_fit()` function. Note that not all aRFs have good fit. We thus dropped the bottom 5% of the fitted results. Figure A.2c shows the KDE plot of the fitted Gaussian means from the aRFs of the CNN and ViT. The vast majority of the aRFs are centered with respect to the image, aligning with our expectations from the attention rollout maps (see Section 5). Figure 2b compares the standard deviations in horizontal and vertical direction of the fitted Gaussian.

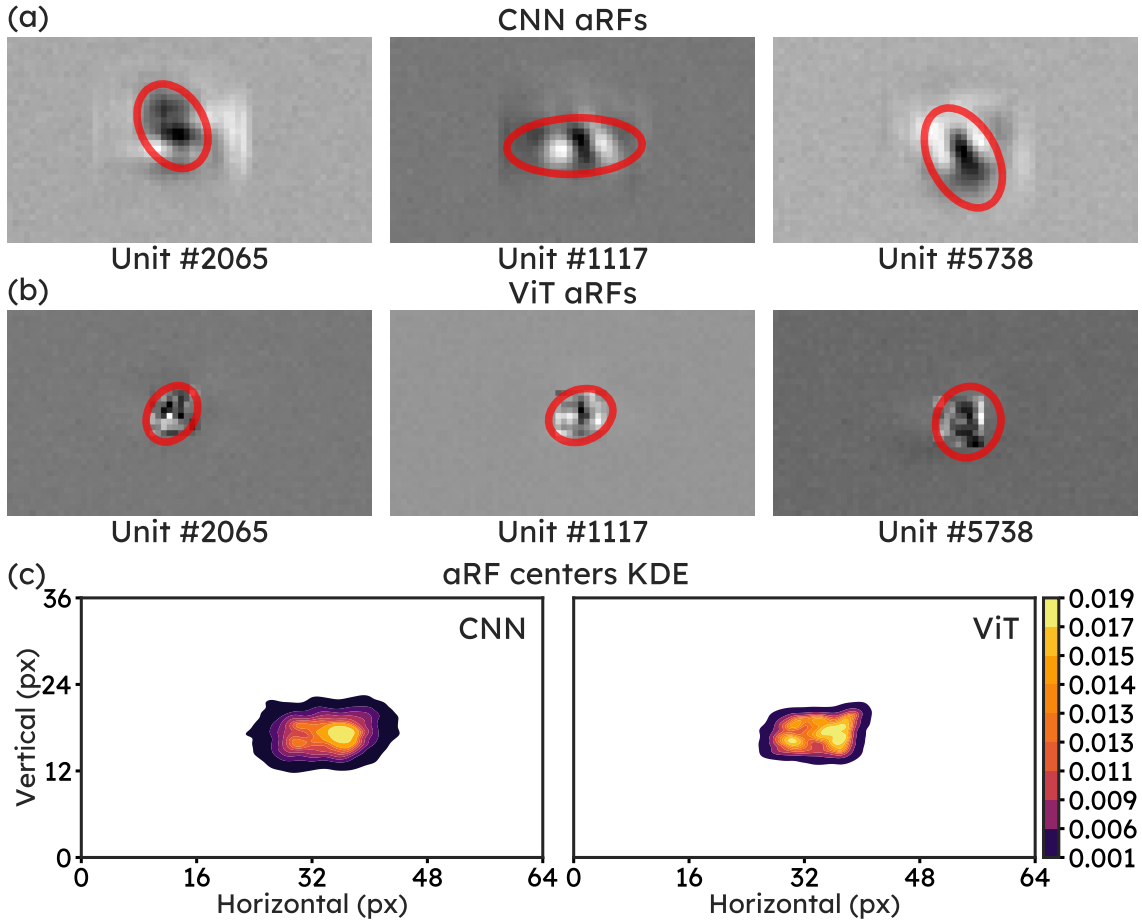


Figure A.2: Estimated artificial receptive fields (aRFs) of (a) CNN and (b) ViT over the same set of randomly selected artificial units from `Mouse A`. The red circles (1 standard deviation ellipse) show the 2d Gaussian fit. (c) KDE of the Gaussian centers of the two models.

A.5 B-MLP activation

We investigated different variations of the B-MLP module. The motivation of the proposed behavior module is to enable the core to learn a shared representation of the visual and behavioral variables across the animals. Moreover, the level-wise connections allow the self-attention module in each V1T block to encode different behavioral features with the latent visual representation. We experimented with a per-animal B-MLP module (while the rest of the core was still shared across animals) which did not perform any better than the shared counterpart, suggesting that the behavior module can indeed learn a shared internal brain state presentation. We also tested having the module in the first block only, as well as using the same module across all blocks (i.e. all $\texttt{B-MLP}_b$ shared the same weights). Both cases, however, led to worse results with a $2-4\%$ reduction in predictive performance on average. To further examine the proposed formulation, we analyzed the activation patterns of the shared behavior module at each level in V1T, shown in Figure A.3. We observed a noticeable distinction in B-MLP outputs in earlier versus deeper layers, with a higher spread in deeper layers, which corroborates our hypothesis that the influence of the behavioral variables differs at each level of the visual representation process.

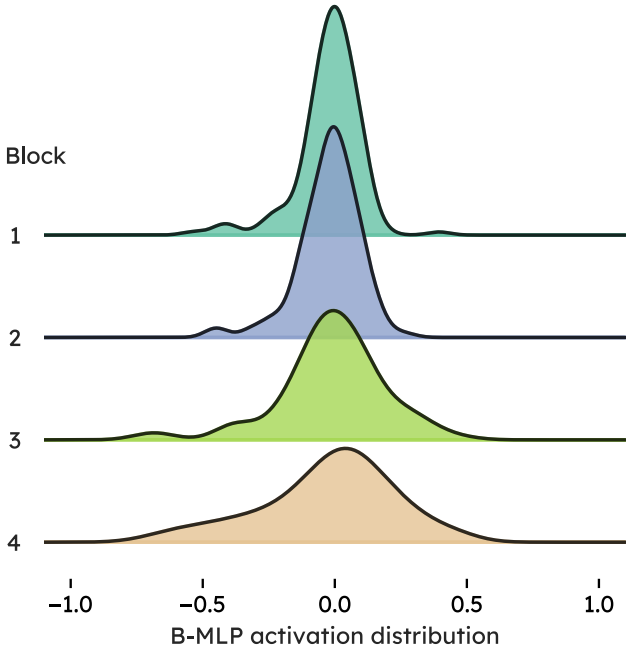
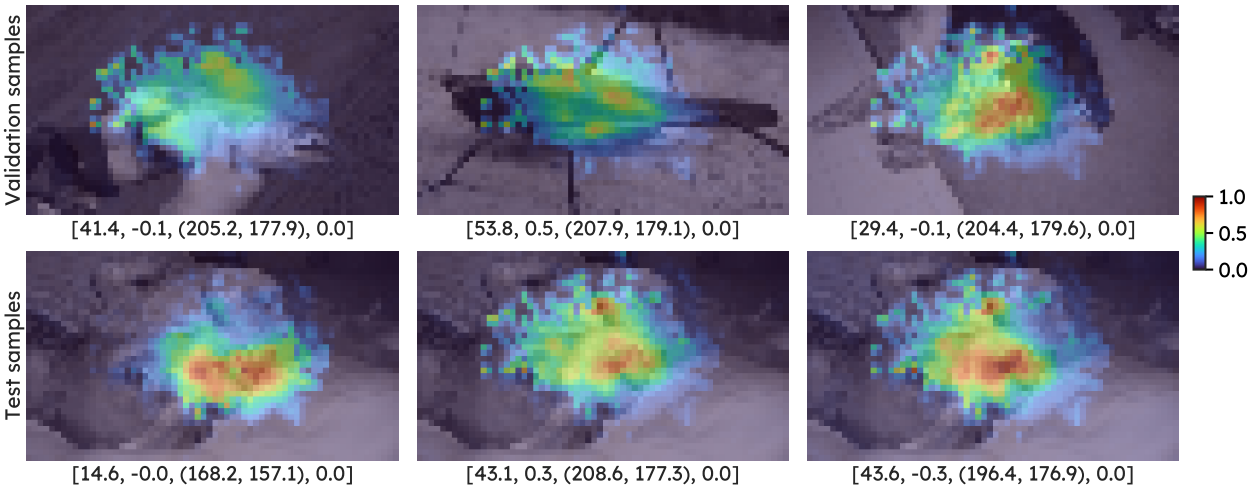
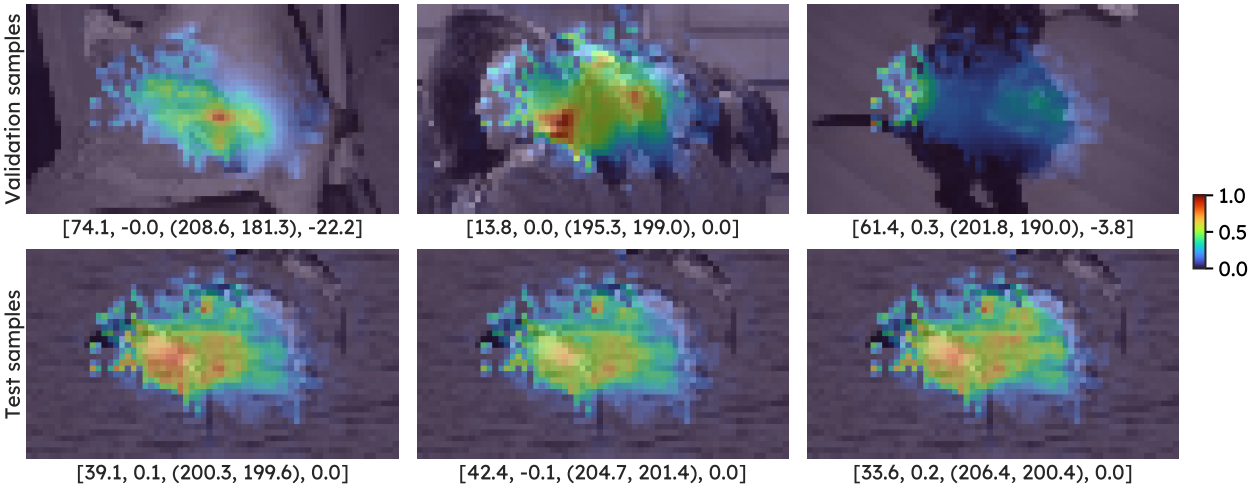


Figure A.3: tanh activation distributions of B-MLP at each level (block) in the V1T core. The spread of activation distributions indicates varying influence of behavioral variables at the block in the core module.

A.6 Attention rollout maps

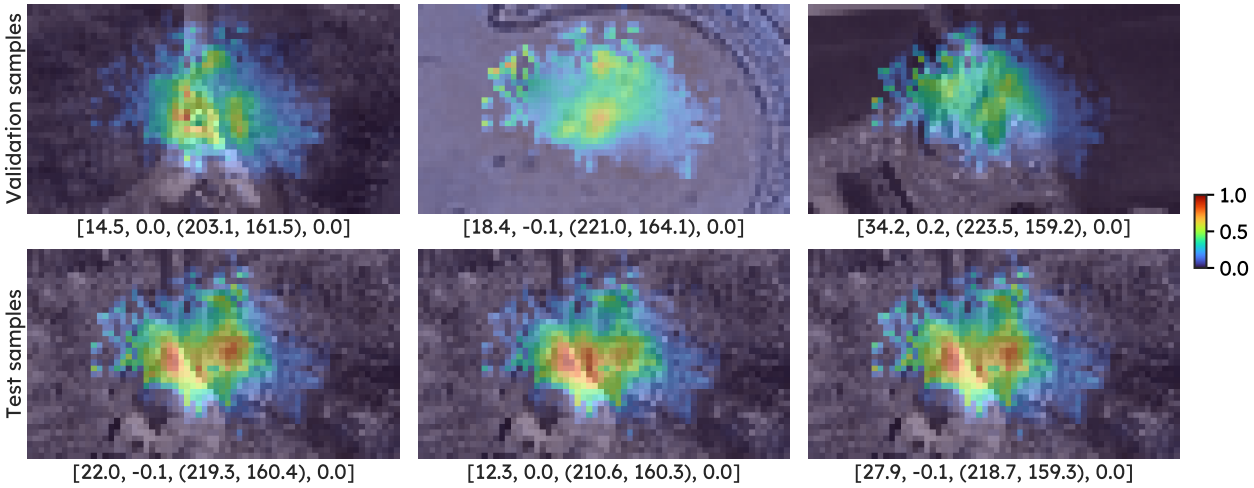


(a) Mouse A

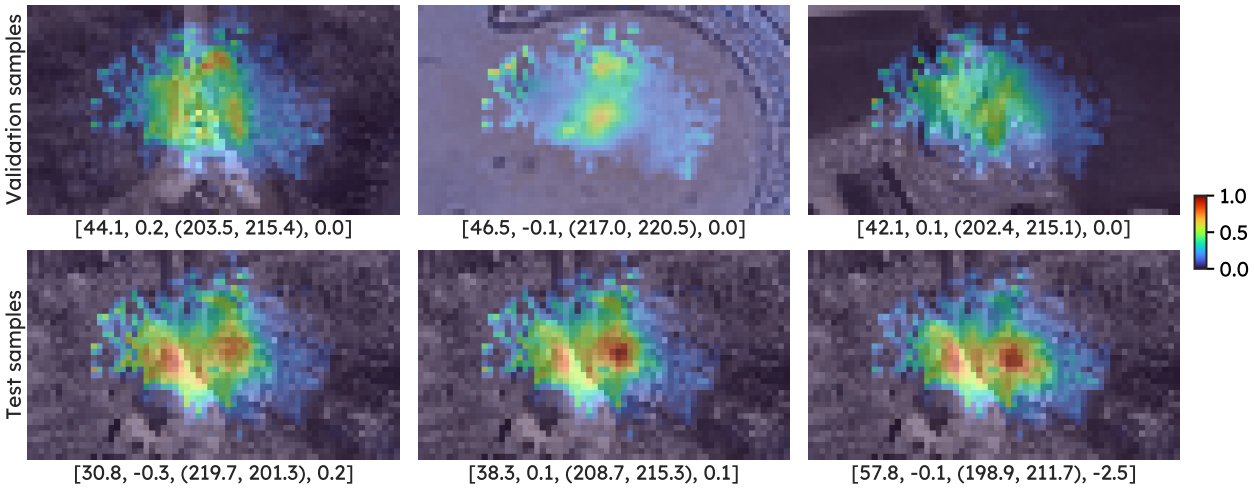


(b) Mouse B

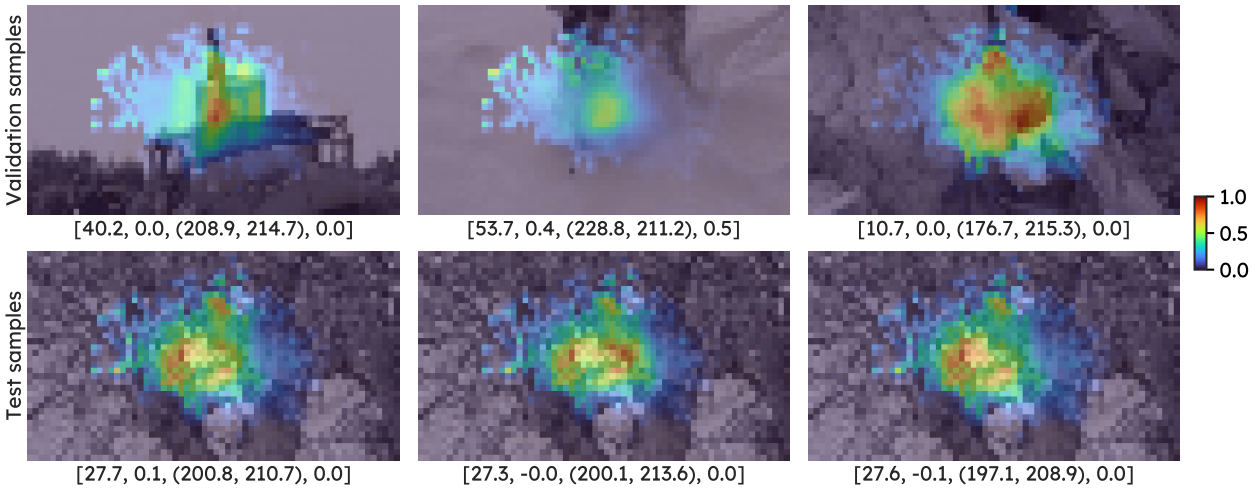
Figure A.4: V1T attention visualization on validation and test samples of Mouse A to E from Dataset S. As the computer monitor was positioned such that the visual stimuli were presented to the center of the receptive field of the recorded neurons (see Dataset S discussion in Section 2), we expected regions in the center of the image to correlate the most with the neural responses, indicating that the core module learned to assign higher attention weights toward those regions. Note that the core module is shared among all mice. For this reason, we also expected similar patterns across animals. We observed small variations in the attention maps in the test set, where the image is the same and behavioral variables vary, suggesting the core module learned to adjust its attention based on the internal brain state. To quantify this result, we further showed that there are moderate correlations between the center of mass of the attention maps and the pupil center, see discussion in Section 5. Each attention map was normalized to $[0, 1]$, and the behavioral variables of the corresponding trial are shown below the image in the format of [pupil dilation, dilation derivative, pupil center (x, y), speed]. The Figure continues to the next page.



(c) Mouse C



(d) Mouse D



(e) Mouse E

A.7 Behaviors and predictive performance

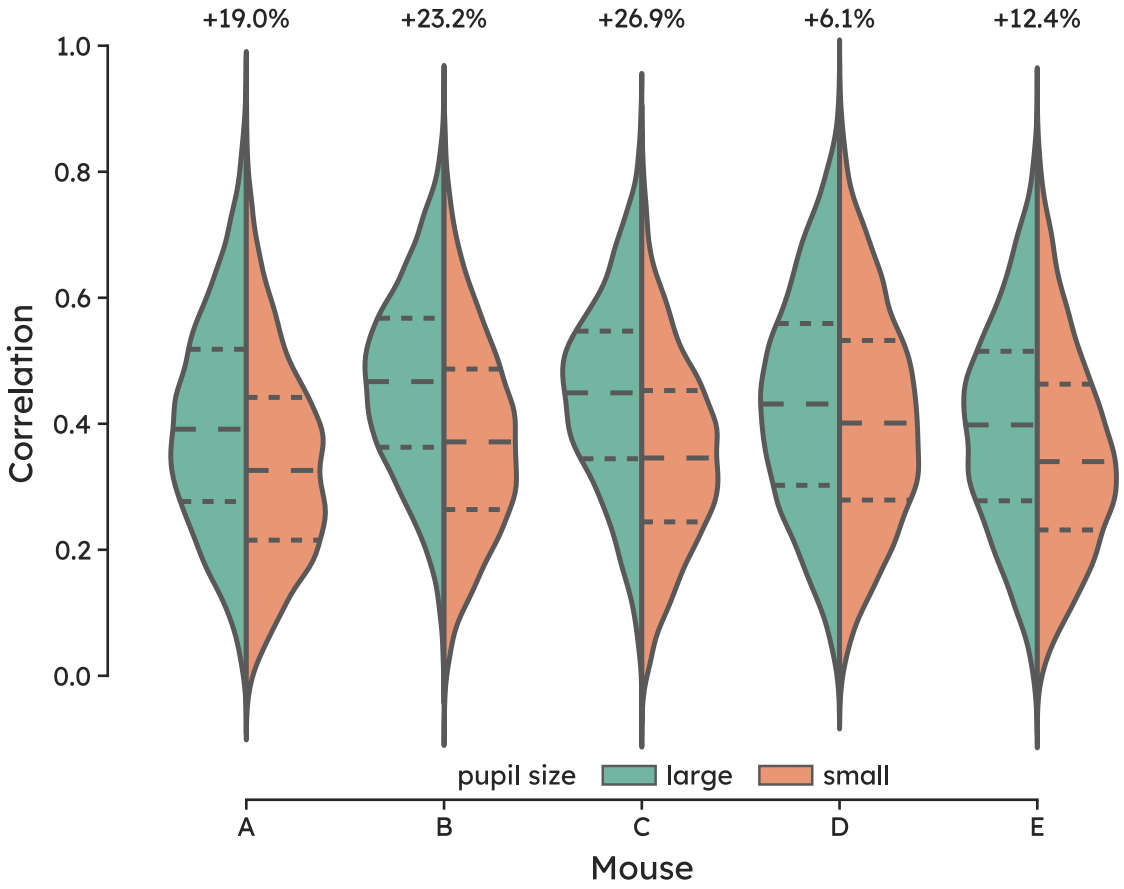


Figure A.5: Predictive performance w.r.t. pupil dilation in Dataset S. Previous work has shown that pupil dilation is an indication of arousal, i.e. stronger (or weaker) neural responses with respect to the visual stimulus (Reimer et al., 2016; Larsen and Waters, 2018). We thus expected a similar tendency could also be observed with our model. Here, we divided the test set into 3 subsets based on pupil dilation. We then compared the predictive performance of the model in the (large) larger third subset against the (small) smaller third subset. We observed that trials with larger pupil sizes are better predicted, with an average difference of +17.5% across animals. The dashed lines indicate the quartiles of the distributions and the percentage above each violin plot shows the relative prediction improvement of the larger set against the smaller set.

A.8 Readout position and retinotopy

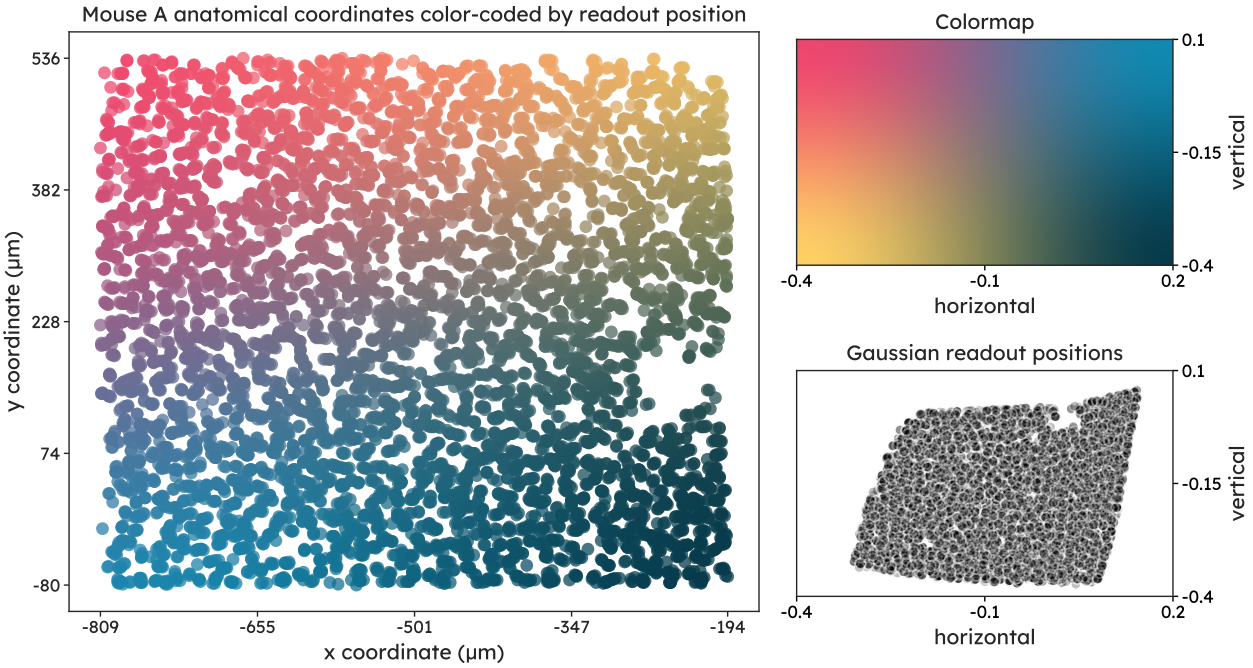


Figure A.6: The learned readout position with respect to neuron anatomical coordinates in Mouse A. The position network in the Gaussian readout (see Section 3) learns the mapping between the latent visual representation (i.e. output of the core, bottom right panel) and the 2d anatomical location of each neuron (left panel). Lurz et al. (2021) and Willeke et al. (2022) demonstrated that a smooth mapping can be obtained when color-coding each neuron by its corresponding readout position unit. This aligned with our expectation that neurons that are close in space should have a similar receptive field (Garrett et al., 2014). Here, we showed that, despite the substantial architectural change, a similar mapping can also be obtained with the V1T core. The code to generate this plot was written by Willeke et al. (2022) and is available at github.com/sinzlab/sensorium.