Neural Attention Maps Alignment in Vision Transformers and Mammalian Visual Cortex

Hamd Jalil[*1], **Ahmed Qazi**[*1], **Asim Iqbal**[1†]

[1]Tibbling Technologies
asim@tibbtech.com

Abstract

Image parsing with Vision Transformers has achieved state-of-the-art results, but how these models process visual information compared to biological vision systems is an open question. In this study, we present an extensive benchmarking between the attention mechanisms in the Vision Transformer-based models, such as Segment Anything, and its several variants that capture long-range dependencies in understanding the generalized features in natural images, with the neural responses captured from the mouse visual cortex for the same visual inputs. We found a significant correspondence between self-attention and convolutional maps in these models and cortical neural activity in the mouse visual cortex. This trend is observed to be consistent across similar model architectures with varying numbers of parameter units and provides an explainable trade-off between the accuracy and efficiency on real-world object segmentation datasets. This relationship is observed to be generalized across the subregions and neuronal genotypes, capturing diverse functional units in the mouse visual cortex. Our work proposes a pioneering effort in identifying important parallels between hierarchical representational learning in vision-based transformers and the biological visual cortex. To advance the development of neuro-AI models, these neural correlates suggest that aspects of cortical computation, captured by the state-of-the-art vision models, can potentially contribute to their effectiveness for image understanding tasks as well as guiding the advancement of novel model architecture design. We anticipate that this practice will also lead to future interpretability work to better understand the encoding and decoding principles of computation in the mammalian visual cortex.

Introduction

The rise of AI models in computer vision applications such as image classification (He et al. 2016) and segmentation (Long, Shelhamer, and Darrell 2015) has sparked increasing interest from neuroscientists in investigating biological similarities with brain computations (Kriegeskorte 2015). Deep learning models exhibit striking similarities to neural representations when processing natural visual stimuli (Cadieu et al. 2014), with recent focus shifting to Vision Transformers (ViTs) (Dosovitskiy et al. 2021). ViTs employ self-attention mechanisms for computing spatial dependencies, enhancing explainability in semantic feature processing—paralleling neuroscience findings where visual neurons respond to semantically aligned features (Blumberg and Kreiman 2010; Kriegeskorte et al. 2008).

Although transformers have achieved state-of-the-art results in computer vision tasks (Thisanke et al. 2023), systematic investigations of cortical-ViT alignments remain limited. This gap is particularly important given that the hierarchical ventral visual stream shares intriguing similarities with multiscale ViT encoders - both process visual information across increasing receptive fields through feedforward and feedback pathways (Berezovskii, Nassi, and Born 2011; Ghiasi et al. 2022). The mouse cortex provides a tractable model for probing these relationships via in vivo calcium imaging of neural populations.

This study presents extensive benchmarking between state-of-the-art ViTs and neural recordings from the mouse visual cortex using matching natural stimuli from the Allen Brain Observatory. We systematically compare attention maps with neural responses across 15 transformer architectures, revealing strong alignments between artificial and biological kernels. Our experimental framework (**Figure 1**) establishes a foundational approach for quantifying neural plausibility across diverse ViT scales and architectures, providing concrete insights for developing biologically-inspired AI models.

Related work

Functional Similarities - Models and Visual Cortex

Research into the primate visual system has been instrumental in designing AI models (Kubilius et al. 2019, 2018), with CNNs showing high similarity to biological visual processing mechanisms (Lindsay 2021). However, biological plausibility is not always proportional to model complexity (Kubilius et al. 2019, 2018). Models with recurrent shallow architectures, such as CORnet-S (Kubilius et al. 2019, 2018), demonstrate stronger brain alignment, emphasizing the importance of recurrent structures in the ventral stream's object recognition processes (Kar et al. 2019).

Recent studies have explored functional similarities with the mammalian visual system, particularly mouse. For in-

[*]These authors contributed equally.
[†]Corresponding author.

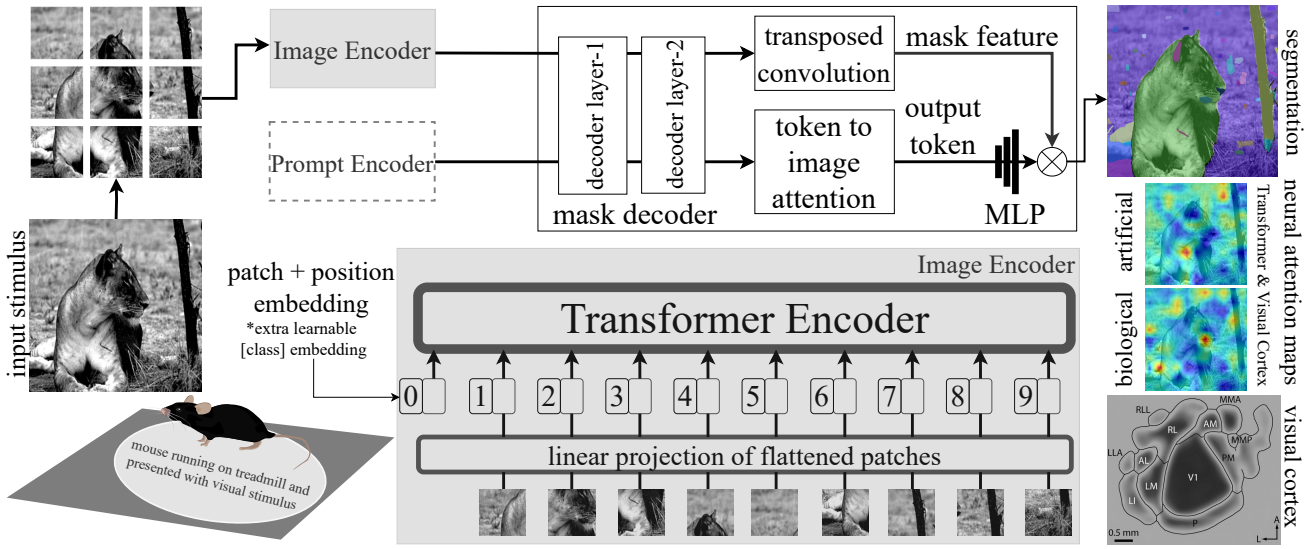


Figure 1: Block diagram flowchart of the experimental setup to perform a systematic comparison between artificial and neural kernels is shown. On the left, a sketch of a mouse running on a treadmill is drawn, which is presented with a sample visual stimulus of a lion image. The same image is passed to the model framework, where it first passes through the image encoder, followed by the mask decoder module. The model framework of SAM is presented here, where the zoomed-in version of the image encoder is shown at the bottom, where the patch and positional embeddings are forwarded to the transformer encoder. The mask decoder consists of multiple decoders and convolutional stages with a Multilayer Perceptron at the end, where the output (semantic) segmentation mask is shown overlaid on the input image. On the right side, a qualitative sample comparison of an artificial kernel (attention map) with a biological kernel is shown, overlaid on top of the input image. The anatomy of the mouse visual cortex is shown at the bottom right with several highlighted sub-regions that are involved in visual information processing of the input stimulus.

stance, the MouseNet model (Shi et al. 2022) provides insights into mouse visual cortex (MVC) computations, though DNNs effective for primates show diminished efficacy in mouse systems (Cadena et al. 2019). This has driven research into rodent visual cortex neural taskonomies (Conwell et al. 2021) and robustness in neuro-inspired CNN architectures (Li et al. 2019).

Transformer-based models have recently gained prominence in neuroscience; models such as V1T (Li et al. 2023a) and Brain Network Transformer (Kan et al. 2022) provide insights into neural information processing. (Kozachkov, Kastanenka, and Krotov 2023) explored biological foundations of transformers, suggesting that neuron-astrocyte networks can implement transformer computations, while (Toneva and Wehbe 2019) demonstrated brain alignment approaches to improve model performance (Iqbal et al. 2025). Exploring ViT-mammalian brain similarities has become increasingly active (Li et al. 2023a; Kan et al. 2022; Johnson et al. 2023; Robinson and Drenkow 2022).

Evolution of Vision Model Architectures

CNN-based models have dominated computer vision applications through hierarchical feature extraction, with architectures like LeNet (Lecun et al. 1998), ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2015), YOLO (Redmon et al. 2016) excelling in classification and detection, while U-Net (Ronneberger, Fischer, and Brox 2015), DeepLab (Chen et al. 2016), and Mask-RCNN (He et al. 2017b) pioneered semantic and instance segmentation. Transformers, originally designed for Natural Language Processing (NLP) (Vaswani et al. 2017), evolved into ViTs (Dosovitskiy et al. 2021), extending self-attention to images with impressive results in classification, detection (Carion et al. 2020), and segmentation (Li et al. 2022).

ViTs have become the backbone architectures for domain-agnostic models (Kirillov et al. 2023a; Ke et al. 2023), with variants like ViT-H/B/L, TinyViT (Wu et al. 2022), and Swin Transformers (Liu et al. 2021) forming foundations for SAM (Kirillov et al. 2023a), HQ-SAM (Ke et al. 2023), Semantic-SAM (Li et al. 2023b), Mask2Former (Cheng et al. 2022), and Mask DINO (Li et al. 2022). While these architectures show promising results across vision tasks, limited work explores their neural plausibility. This paper introduces a novel comparative approach to assess the biological plausibility of state-of-the-art architectures with the mammalian visual cortex, enabling novel ideas in building neuro-inspired AI models.

Methodology

Allen Brain Observatory dataset

The Allen Brain Observatory offers an invaluable dataset that facilitates a quantitative exploration of the functional properties underpinning the coding of sensory stimuli through the visual pathway. This dataset is character-

ized by its in-depth examination of visually evoked cellular responses, achieved using in vivo calcium imaging from GCaMP6-expressing neurons. These neurons are meticulously selected from specific regions of the MVC, various cortical layers, and Cre lines that label a diverse range of neuron types.

A standout feature of this dataset is its emphasis on the neural responses to natural visual stimuli. To this end, a library of 118 natural scenes, sourced from three distinct databases (Berkeley Segmentation Dataset, van Hateren Natural Image Dataset and McGill Calibrated Colour Image Database) (Martin et al. 2001; van Hateren and van der Schaaf 1998; Olmos and Kingdom 2004), was employed. Each scene image from this collection was briefly displayed for 250 ms and subsequently replaced by the subsequent scene. This process was repeated 50 times for each image, presented in a randomized sequence, interspersed with blank intervals. Such a detailed presentation of natural scenes, including images of various animals like owls, cheetahs, and ducks, provides a comprehensive perspective on both individual cellular and collective cell population responses to intricate visual stimuli in the MVC. This dataset is collected from 256 different mice, ensuring a diverse and representative sample for understanding neural responses. The distribution of neurons across brain regions and their response characteristics are detailed in Supplementary **Figure S1**. The use of multiple mice allows for a more generalized understanding of the MVC's response to natural visual stimuli, accounting for potential individual variations and offering a broader insight into the mammalian visual processing mechanisms.

In our endeavor to draw meaningful connections between ViT-based model architectures and the MVC, we leveraged this publicly available dataset from the Allen Brain Observatory (Allen Institute for Brain Science 2023). Specifically, our analysis focused on the peak DF/F and time-to-peak attributes of neurons, offering insights into the neural dynamics in response to visual stimuli.

MVC Neural Response dataset

Given a collection of natural scene stimuli and the associated neural activity for each stimulus display, we derived the neural representations (N) for the neurons. The neural information is structured into a matrix, with designated columns for the experiment ID, neuron ID, Trial Traces, and stimulus shown. Each neuron possesses a set of Trial Traces (T) that correspond to its neural activity during the stimulus display. These traces are amalgamated with the flattened stimuli (FS) to create the neural representations. Initially, the stimuli are resized and flattened as:

$$FS = s_1, s_2, \ldots, s_n \tag{1}$$

where s_i represents the FS for the i^{th} natural scene. Subsequently, the T for every neuron is represented as:

$$T = t_1, t_2, \ldots, t_n \tag{2}$$

where t_i denotes the set of T for the i^{th} stimulus display. Ultimately, N is deduced as:

$$N = (T^T \cdot FS) \tag{3}$$

Following the generation of neural representations, we embarked on a meticulous selection procedure to curate a balanced subset of these representations. To achieve this, we employed UMAP embeddings, applying them distinctly to both excitatory and inhibitory neural responses. The UMAP settings we utilized were: number of neighbors set to 15, minimum distance of 0.1, and a spread of 1. These parameters were chosen to ensure that the local and global structure of the data was preserved, allowing for a faithful representation of the high-dimensional data in a reduced space.

Subsequent to the UMAP embeddings, we employed K-means clustering, with K set at 10, for each neural response category. From each of the 10 clusters, we selected 50 neural representations centered around the centroids, leading to a total of 500 representations for both excitatory and inhibitory categories. This culminated in a combined dataset comprising 1000 neural representations. The genotypes analyzed span both excitatory and inhibitory neurons, with complete genotype naming conventions and Cre-labelling details provided in Supplementary Table S1. The rationale behind using K=10 in the K-means clustering was to ensure that our sample was both balanced and comprehensive, capturing the diversity and intricacies of the larger dataset. The diversity of neuronal genotypes in our dataset is shown in Supplementary **Figure S2**. This selection of parameters ensures computational tractability while maintaining representational diversity

Computing Neural Kernels

In the MVC response dataset, which comprises 1000 neural representations, each representation can be visualized as a 2D spatial pattern. These 2D patterns, derived from the neural activity in response to specific stimuli, capture the essence of how the visual cortex processes different visual inputs. Representative examples of these neural responses are shown in Supplementary **Figure S3**. For the purpose of our study and to simplify terminology, we will henceforth refer to these 2D neural representations as "neural kernels". Just as model kernels in deep learning architectures (like attention and convolutional layers) are pivotal in processing and understanding input data, these neural kernels play a crucial role in deciphering the visual stimuli presented to the MVC. The term "neural kernel" underscores the parallel between the computational units in AI models and the functional units in the biological neural systems.

Shortlisted model architectures

We shortlisted a set of 15 state-of-the-art models that have shown promising results on segmentation and object detection tasks. All models contain transformer encoders and decoder layers. SAM, MobileSAM, and HQ-SAM use pretrained ViT variants, including ViT-H, ViT-L, ViT-B, and ViT-Tiny, as image encoders, while Semantic-SAM uses Swin-L/T transformers as encoders. Unlike ViT, which computes self-attention globally, Swin Transformer merges input image patches to build hierarchical feature maps using shifted window partitioning with modified self-attention (SW-MSA) and regular window partitioning based multi-head attention (W-MSA) to establish connections across

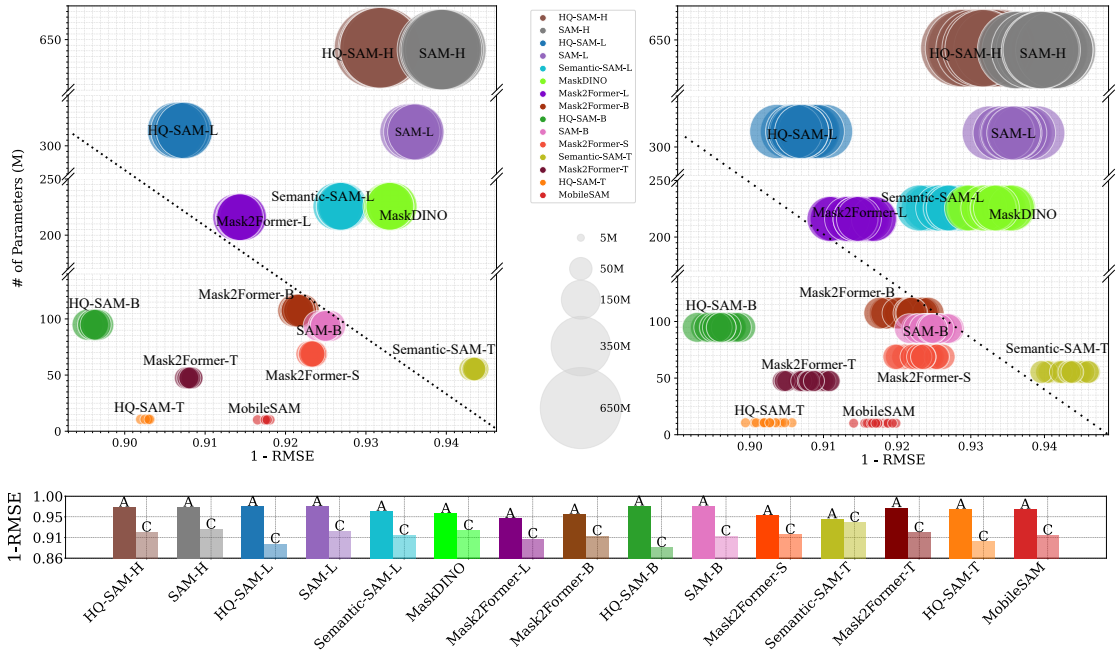


Figure 2: Similarity across attention and neural maps between model architectures and mouse visual cortex. **Top left** plot shows the average similarity (1-RMSE) of the population of convolutional+attention kernels from models and neural kernels drawn from different functional sub-regions in the mouse visual cortex. Different colors indicate different model architectures, where the number of bubbles within the same color cluster represents different brain regions where neural kernels are drawn. **Top right** plot shows the same trend, but each cluster of bubbles with the same color represents different genotypes, covering a diverse population of neurons from all the sub-regions in the mouse visual cortex. The size of each bubble represents the total number of parameters (in millions) present in the model architecture. **Bottom** plot shows the average similarity (1-RMSE) of the convolutional (C) and attention (A) kernels from all the model architectures with the neural kernels from the mouse visual cortex.

windows. Hence, Swin computes self-attention only within each local window. This results in a linear computational complexity as compared to ViT, which computes at a quadratic complexity. Mask2Former and Mask DINO variants also use Swin backbones and transformer decoder layers with varying attention mechanisms. The set of shortlisted models covers a large variety within and across transformer encoder blocks.

Computing Artificial Kernels

Selected models have been trained on a range of datasets. Mask DINO, Mask2former were trained on the COCO dataset (Lin et al. 2014) while SAM, Semantic-SAM, and Mobile SAM were trained on the SA-1B (Kirillov et al. 2023a) dataset. HQ-SAM, which was trained on HQSeg-44K, consists of six image datasets (Ke et al. 2023) covering a wide range of semantic classes. The weights of these model architectures comprise learnable and non-learnable parameters from attention and convolutional layers in the encoder, decoder, and backbones of the models. For our analysis, the attention layers in these models that are inherently structured in 2D were directly incorporated into the analysis without any alterations as "artificial kernels". This 2D structure of the attention layers naturally aligns with our objective, allowing us to directly use them as "artificial kernels".

On the other hand, convolutional layers required a more nuanced approach. We specifically excluded layers with 1×1 and 2×2 size filters. The rationale behind this exclusion was twofold: individual 1×1 and 2×2 kernels were too small to capture global information, and their inclusion would not allow for a uniform comparison between the small 1×1 and 2×2 kernels and the sizeable attention layers. The artificial kernels from the convolutional layers were computed through the following approach: Given the convolutional layer L characterized by dimensions $[F, C, H, W]$, where F denotes the number of filters, C the number of channels, H the height, and W the width, we considered the channels of all filters as separate entities. This resulted in a reshaped structure of $[F \times C, H, W]$, essentially treating each channel of every filter as a distinct 2D kernel. To further refine our convolutional layer selection, UMAP embeddings were employed, ensuring that both local and global relationships within the data remained intact. This was followed by K-means clustering with $K = 10$, aiming to extract a representative subset of these 2D kernels. This methodology, combining UMAP and $K = 10$, was consistent with our approach for neural representations and kernels. K=10 was selected to balance computational efficiency with adequate

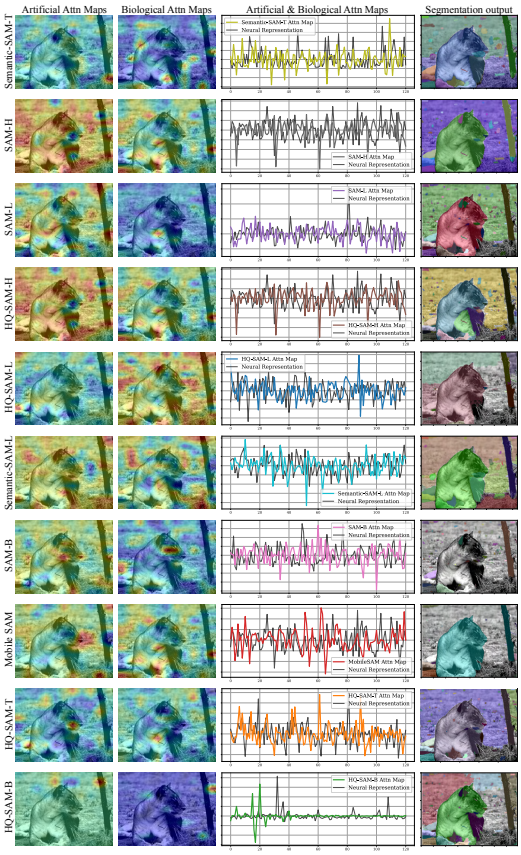


Figure 3: Qualitative comparison of randomly selected artificial and neural kernels. Each row depicts the overlaid artificial and neural kernels on top of an input stimulus image of a lion (same as in Figure 1), followed by their 1D representation, plotted as curves in the plots. Here, the black curve shows the neural kernel representation, and the colored ones are randomly picked from the model (here, 1D representation of attention maps). The last column shows the qualitative output of the models with the segmentations.

sampling diversity, ensuring each cluster contained sufficient samples for robust statistical comparison while maintaining distinct cluster identities.

In conclusion, from both the attention and convolutional layers, we derived 2D "artificial kernels" that serve as the foundation for our subsequent analyses with the "neural kernels" drawn from the mouse visual cortex.

Artificial & Neural kernel similarity analysis

Before diving into the specifics of our analysis technique, it is crucial to understand the foundation of our comparison. We have two primary sets of kernels for our study: the neural kernels and the artificial kernels. The neural kernels are derived from natural visual stimuli, capturing the MVC's response to these stimuli. On the other hand, the artificial kernels, originating from the models, are pre-trained on the COCO, SA-1B, and HQSeg-44K datasets. These datasets are rich in natural stimuli, ensuring a consistent basis for our comparison. The choice of dataset for model training is pivotal in similarity analysis. As highlighted by (Conwell et al. 2021), the dataset on which a model is trained plays a significant role in determining the nature and structure of the features it learns. Given that both our neural and artificial kernels are grounded in natural stimuli, we have a consistent and robust foundation for our subsequent analyses.

While our models underwent different pretraining procedures, we normalized all kernels using z-score normalization to ensure consistent comparison scales. This normalization removes scale differences from varied pretraining, allowing fair structural comparison. Additionally, all models were evaluated on the same natural image stimuli present in both training datasets and the Allen Brain Observatory, providing a common reference framework that mitigates potential biases from varied training protocols.

With this groundwork laid, we proceed to the comparison between the two kernel types. To avoid any dimensional mismatch, each 2D artificial kernel was resized to a fixed size, matching the size of the neural kernels. To ensure a consistent scale and distribution, the kernels underwent normalization. Given a kernel array k, the normalized image k_{norm} is computes as following, where μ is the mean of k and σ is its standard deviation:

$$k_{norm} = \frac{k - \mu}{\sigma} \tag{4}$$

After normalization, we employed the Root Mean Square Error (RMSE) to conduct our analysis. We employed RMSE as it provides a pixel-wise measure of spatial correspondence between normalized 2D patterns, making it suitable for comparing the spatial structure of attention maps with neural activity patterns. Mathematically, the RMSE between two kernels k_1 and k_2 is defined as following where N represents the total number of pixels:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(k_{1i} - k_{2i})^2} \tag{5}$$

A lower RMSE value indicates that the artificial and neural kernels are more similar, implying a better match. Conversely, a higher RMSE value suggests a greater disparity between the two kernels. To provide a more intuitive interpretation, especially when comparing multiple models' artificial kernels with neural kernels, we used $1 - \text{RMSE}$. This metric transforms the RMSE such that a higher value is better. Specifically, a $1 - \text{RMSE}$ value close to 1 indicates a near-perfect match between the artificial and neural kernels, while a value close to 0 indicates a poor match. By using both RMSE and $1 - \text{RMSE}$, we offer a comprehensive view of the similarities and differences between the artificial and neural kernels under examination.

Results & Discussion

From our experimental setup, we ran systematic comparisons of artificial (convolutional and attention) kernels with the neural kernels from all the brain regions and unique neural genotypes. These results are summarised in **Figure 2**,

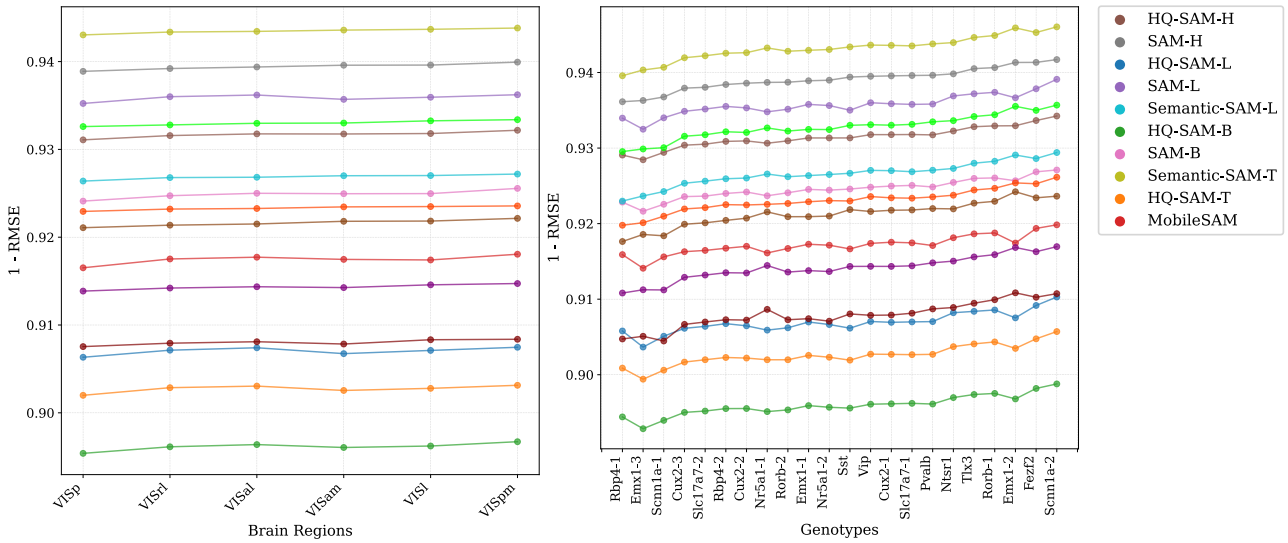


Figure 4: Left - Trend of 1-RMSE scores for neurons from various brain regions across different models, highlighting region-specific neural similarities. Right - Trend of 1-RMSE scores for different neuronal genotypes across the models, illustrating genotype-specific neural correspondences.

	MS COCO						
Models	AP_b	AP	AP_l	AP_m	AP_s	Point (max) 1-IoU	Point (oracle) 1-IoU
SAM-H	34.0	48.9	64.5	53.3	**34.4**	-	-
HQ-SAM-H	**34.9**	**49.9**	66.5	**54.0**	**34.2**	-	-
SAM-B	28.2	44.4	57.7	48.7	32.1	52.1	68.2
HQ-SAM-B	31.3	46.7	62.9	50.5	32.0	-	-
SAM-L	33.3	48.5	63.9	53.1	34.1	**55.7**	70.5
HQ-SAM-L	**34.4**	49.5	66.2	53.8	33.9	-	-
Semantic-SAM-L*	-	-	-	-	-	**57.0**	**74.2**
HQ-SAM-T	-	45.0	62.8	48.8	29.2	-	-
MobileSAM	-	44.3	61.8	48.1	28.8	-	-
Semantic-SAM-T*	-	47.4	66.1	50.7	28.3	54.5	**73.8**
Mask2Former-L	-	**50.1**	**72.1**	53.9	29.9	-	-
Mask DINO*	-	**52.1**	**72.5**	**55.4**	32.9	-	-
Mask2Former-B*	-	48.1	71.1	52.0	27.8	-	-
Mask2Former-S*	-	46.3	68.4	50.3	25.3	-	-
Mask2Former-T*	-	45.0	67.4	48.3	24.5	-	-

Table 1: Models' performance on the COCO segmentation dataset. Bold indicates best performance, underlined bold indicates second best, and underlined indicates third best. 1-IoU shows 1 click-IoU. Models marked with * are evaluated on the COCO val2017 dataset.

showcasing the level of biological plausibility (1-RMSE) the experimented models hold with the neural kernels in the MVC. Higher 1-RMSE demonstrates a stronger correlation between artificial kernels and neural kernels. We make our observations keeping in account the biological plausibility of models demonstrated in **Figure 2** and their recorded performance on real-world segmentation tasks using COCO dataset, shown in **Table 1**. Our shortlisted models can be divided into two groups via a diagonal trend-line (shown as dashed curve in the top plots of **Figure 2**). These trends are consistent across different brain regions and neuronal genotypes, as detailed in **Figure 4**, which shows region-specific and genotype-specific neural correspondences. Models on the right of this trend-line tend to have a higher number of parameters (except Semantic-SAM-T), achieve stronger biological plausibility (1-RMSE), and have been reported (Table 1) to hold benchmarking performance on the segmentation task using the COCO dataset. In contrast, models that are lower on the left of this trend-line attain relatively less significance with the biological similarity metric, have relatively fewer parameters, and suffer in real-world performance on the COCO dataset. Following sub-sections explain the observed trends in greater detail. The complete mapping of genotype abbreviations used throughout this analysis to their full Cre-labelling techniques is provided in Supplementary Table S1.

Impact of model size on neural similarity and benchmarking

It can be observed from **Figure 2** (top plots) that models with largest number of parameters (SAM-H and HQ-SAM-H) shows high neural similarity (1-RMSE) and are one of the top performing models on real-world task (Table 1). On the right side of the trend-line, all the models are observed to hold benchmarking (top-3) performance in Table 1. The exception holds for Semantic-SAM-T, which is one of the high-performing models with around 50 million parameters. Interestingly, Semantic-SAM-T and Mask2Former-T, although they have similar number of parameters but they exhibit a strong gap in the neural similarity index, where Semantic-SAM-T outperforms Mask2Former in terms of biological plausibility as well as real-world benchmarking in

Table 1. This performance boost of Semantic-SAM-T with such a low number of parameters as compared to bigger models (e.g., SAM-H and HQ-SAM-H) can be explained by two major differences in its training and underlying model architecture. To highlight the former, Semantic-SAM and its variants (T/L) have also been trained on 1/10th of the SA-1B dataset in addition to the COCO dataset. In terms of architecture, Semantic-SAM variants and Mask DINO (also showing best performance in Table 1 for AP, AP_l, and AP_m while staying above trend-line) have a common query-based transformer decoder architecture that constructs a pixel embedding map using the output from the backbone and the transformer encoder layer. This is in contrast to Mask2Former models, which are outperformed by Mask DINO in real-world benchmarking as well as the neural similarity index. In Mask DINO (and Semantic-SAM), the pixel embedding map and query embedding are combined via dot-product to obtain an output mask.

$$m = d \otimes (S(t(Mb) + g(Me))) \tag{6}$$

where S is the segmentation head, t is a convolutional layer to map the channel dimension to the Transformer hidden dimension, and g is a simple interpolation function to perform 2 x upsampling of Me. Mb and Me are backbone and encoder output feature maps respectively, while d is the query embedding from the transformer decoder. On the other hand, Mask2Former models that lie below the trend-line consist of masked attention in the decoder:

$$X_l = \text{softmax}(M_{l-1} + Q_l K_l^T) V_l + X_{l-1} \tag{7}$$

where M_{l-1} is the attention mask in layer l, $X_l \in \mathbb{R}^{N \times C}$ refers to $N\ C - dimensional$ query features at the l^{th} layer and $Q_l = f_Q(X_{l1}) \in \mathbb{R}^{N \times C}$. $K_l, V_l \in \mathbb{R}^{H_l W_l \times C}$ are the image features under transformation. The attention masks predicted from a previous layer are of high resolution and used as hard-constraints for attention computation. They are neither efficient nor flexible for box prediction. Hence, query-based decoders boost biological plausibility and real-world dataset performance unless the number of parameters is boosted (such as that in Mask2Former-L, where increasing the number of parameters also enhances their biological plausibility and real-world dataset performance).

One consistent finding is a pattern of decreasing biological plausibility amongst ViT-based encoders from H, L, B, and T variants (in order). In **Figure 2**, it can be seen that this is true for SAM and HQ-SAM variants. The trend is consistent for brain regions (top left, **Figure 2**) as well as the genotypes (top right, **Figure 2**). One factor here is the decreasing number of parameters; we observe a depreciation in biological plausibility amongst these variants. Furthermore, when comparing variant groups for SAM and HQ-SAM, we observe a deterioration of plausibility for HQ-SAM variants as compared to SAM variants. In HQ-SAM, an HQ-Output Token and Global-local Feature Fusion is introduced in SAM for high-quality mask prediction. Moreover, the lightweight HQ-Output Token reuses SAM's mask decoder to generate new MLP layers for performing point-wise product with fused HQ-Features. This is also consistent for HQ-SAM-T and MobileSAM, where the number of parameters are approximately the same, but the introduction of HQ tokens and Fusion results in lower biological plausibility between HQ-SAM-T and MobileSAM. Hence, the additional tokens to SAM happen to reduce biological plausibility even though performance remains in the same ballpark.

Neural response activation similarity with convolutional & attention layers in ViTs

From **Figure 2**, the top plots provide an overarching view, showcasing the average biological plausibility across all layers for each model. At a cursory glance, Semantic-SAM-T stands out as the most biologically plausible model. However, a deeper dive into the individual layers, as depicted in the bottom plots of **Figure 2** and detailed layer-by-layer analysis in **Figure 5**, reveals a more nuanced explanation. The attention layers consistently exhibit a higher degree of biological plausibility, i.e. higher $1 - RMSE$ score as compared to their convolutional counterparts. In fact, the attention layers in most models surpass even the overall biological plausibility of Semantic-SAM-T, which is the front-runner in the top plots of **Figure 2**. This stark difference underscores the inherent biological alignment of attention mechanisms. However, it is crucial to note that while attention layers outshine in terms of biological plausibility, the convolutional layers still play a critical role in these architectures, contributing to their overall performance. The averaged biological plausibility, as seen in the top plots (**Figure 2**), is a testament to the combined efficacy of both these layers. Yet, the standout performance of attention layers suggests a clear direction for future research. Given their closer alignment with biological neural kernels and their integral role in transformer-based vision architectures, attention mechanisms seem poised to be the cornerstone of future advancements in both deep learning and our understanding of the neural underpinnings of visual perception.

Qualitative analyses reveal synchronized responses and deformations, indicating the early ViT layers encode lower-level visual features similar to the neural populations. This is depicted in **Figure 3**, with additional qualitative comparisons across diverse stimuli (bird, wolf) shown in Supplementary **Figure S5**, and further explained in Supplementary data. Through rigorous analysis, we quantify which specific cortical visual areas align most strongly to different ViT layers. Applying explainability techniques on a per-head basis reveals increasing receptive field size and complexity through the depth, matching the ventral stream hierarchy. Pre-trained representations transferred to downstream tasks also demonstrate enhanced neural plausibility, suggesting generalizable biological compatibility.

Conclusion

We performed systematic comparisons of artificial and neural kernels in high-performing vision transformer architectures and the mouse visual cortex. Our results provide empirical evidence that intrinsic attention and convolutional maps in transformer architectures are well-suited for capturing core representations in hierarchical biological vision sys-

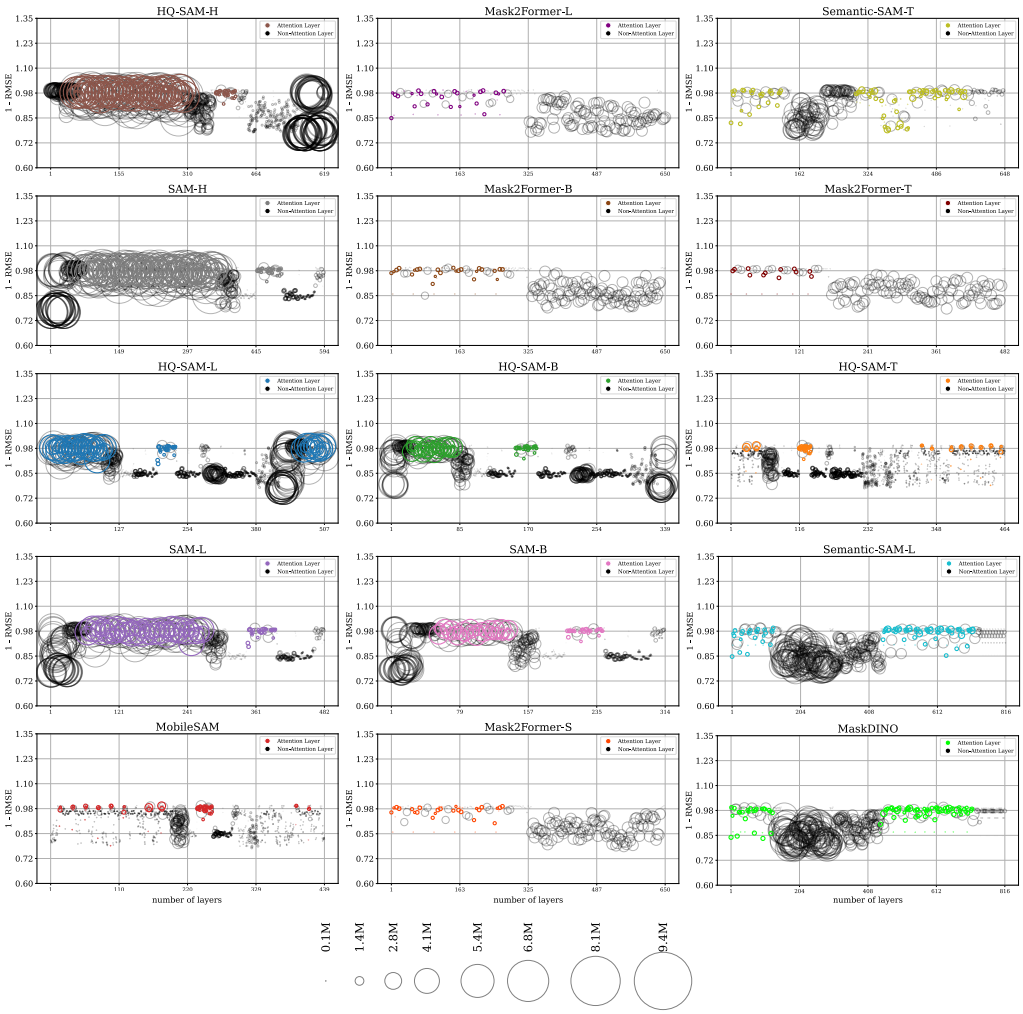


Figure 5: This figure presents a detailed performance evaluation of various neural network models, each depicted in a distinct subplot. The y-axis represents the "1 - RMSE" performance metric, while the x-axis enumerates the layers within each model. The size of each scatter point is indicative of the number of parameters in the respective layer, with larger points signifying layers with a higher parameter count. Attention layers are distinctly marked by their model-specific color with a transparent fill, while non-attention layers are represented in black. This visualization offers a comprehensive view of the relative performance of attention versus non-attention layers across a range of neural network architectures. A similar visualization in Supplementary **Figure S4** compares attention and convolutional layers.

tems. The global receptive fields and long-range dependencies modeled by multi-headed self-attention appear to mimic cortical computations' primitives. These findings provide concrete design principles for developing more efficient, brain-inspired vision architectures. The attention mechanism alignment suggests prioritizing multi-head attention in resource-constrained environments where biological plausibility and computational efficiency must be balanced. Our neural analysis motivates hybrid architectures augmented with convolution that balance local and global processing, while dynamic routing and sparse attention can improve efficiency while retaining long-range interactions.

Several limitations should be acknowledged. Mouse visual cortex, while providing a tractable biological model, is less complex than primate visual systems that more closely resemble human visual processing. This work represents foundational research establishing Neuro-AI correspondences rather than direct performance improvements. Our analysis focuses on kernel-level similarities without exploring dynamic temporal processing differences between biological and artificial systems.

Future research will apply these neural constraints to design biologically plausible, efficient attention mechanisms. Extended primate data analysis and investigation of neural temporal dynamics will further advance our understanding of visual principles in both artificial and biological systems.

References

Allen Institute for Brain Science. 2023. Visual Coding - Natural Scenes. https://observatory.brain-map.org/visualcoding/stimulus/natural_scenes. Accessed: 2023-07-10.

Berezovskii, V. K.; Nassi, J. J.; and Born, R. T. 2011. Segregation of feedforward and feedback projections in mouse visual cortex. *The Journal of Comparative Neurology*, 519(18): 3672–3683.

Blumberg, J.; and Kreiman, G. 2010. How cortical neurons help us see: visual recognition in the human brain. *J Clin Invest*, 120(9): 3054–3063.

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. 2019. How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons Hidden Units @ NeurIPS 2019*.

Cadieu, C. F.; Hong, H.; Yamins, D. L. K.; 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.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *CoRR*, abs/1606.00915.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. *CoRR*, abs/1406.2031.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation.

Cheng, B.; Schwing, A. G.; and Kirillov, A. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. arXiv:2107.06278.

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H. S.; and Hu, S.-M. 2015. Global Contrast Based Salient Region Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3): 569–582.

Conwell, C.; Mayo, D.; Barbu, A.; Buice, M. A.; Alvarez, G. A.; and Katz, B. 2021. Neural Regression, Representational Similarity, Model Zoology Neural Taskonomy at Scale in Rodent Visual Cortex. In *NeurIPS Proceedings*.

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*, 248–255.

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. arXiv:2010.11929.

Ghiasi, A.; Kazemi, H.; Borgnia, E.; Reich, S.; Shu, M.; Goldblum, M.; Wilson, A. G.; and Goldstein, T. 2022. What do Vision Transformers Learn? A Visual Exploration. arXiv:2212.06727.

He, J.; Yang, S.; Yang, S.; Kortylewski, A.; Yuan, X.; Chen, J.; Liu, S.; Yang, C.; and Yuille, A. L. 2021. PartImageNet: A Large, High-Quality Dataset of Parts. *CoRR*, abs/2112.00933.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017a. Mask R-CNN. *CoRR*, abs/1703.06870.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017b. Mask R-CNN. Cite arxiv:1703.06870Comment: open source; appendix on more results.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Iqbal, A.; Mahmood, H.; Stuart, G. J.; Fishell, G.; and Honnuraiah, S. 2025. Biologically grounded neocortex computational primitives implemented on neuromorphic hardware improve vision transformer performance. *Proceedings of the National Academy of Sciences*, 122(41): e2504164122.

Johnson, E. C.; Robinson, B. S.; Vallabha, G. K.; Joyce, J.; Matelsky, J. K.; Norman-Tenazas, R.; Western, I.; Villafañe-Delgado, M.; Cervantes, M.; Robinette, M. S.; Reddy, A. V.; Kitchell, L.; Rivlin, P. K.; Reilly, E. P.; Drenkow, N.; Roos, M. J.; Wang, I.-J.; Wester, B. A.; Gray-Roncal, W. R.; and Hoffmann, J. A. 2023. Exploiting Large Neuroimaging Datasets to Create Connectome-Constrained Approaches for more Robust, Efficient, and Adaptable Artificial Intelligence. arXiv:2305.17300.

Kan, X.; Dai, W.; Cui, H.; Zhang, Z.; Guo, Y.; and Yang, C. 2022. Brain Network Transformer. In *NeurIPS Proceedings*.

Kar, K.; Kubilius, J.; Schmidt, K.; Issa, E. B.; and DiCarlo, J. J. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22: 974–983.

Ke, L.; Ye, M.; Danelljan, M.; Liu, Y.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2023. Segment Anything in High Quality. *arXiv:2306.01567*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023a. Segment Anything. *arXiv:2304.02643*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023b. Segment Anything. arXiv:2304.02643.

Kozachkov, L.; Kastanenka, K. V.; and Krotov, D. 2023. Building transformers from neurons and astrocytes. *Proceedings of the National Academy of Sciences*, 120(34): e2219150120.

Kriegeskorte, N. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing.

Kriegeskorte, N.; Mur, M.; Ruff, D. A.; Kiani, R.; Bodurka, J.; Esteky, H.; Tanaka, K.; and Bandettini, P. A. 2008. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*.

Kubilius, J.; Schrimpf, M.; Kar, K.; Rajalingham, R.; Hong, H.; Majaj, N. J.; Issa, E. B.; Bashivan, P.; Prescott-Roy, J.; Schmidt, K.; et al. 2019. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *NeurIPS Proceedings*.

Kubilius, J.; Schrimpf, M.; Nayebi, A.; Bear, D.; Yamins, D.; and DiCarlo, J. 2018. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*. Preprint.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, B. M.; Cornacchia, I. M.; Rochefort, N. L.; and Onken, A. 2023a. V1T: large-scale mouse V1 response prediction using a Vision Transformer. *arXiv preprint*.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; yue Li, C.; Zhang, L.; and Gao, J. 2023b. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *ArXiv*, abs/2307.04767.

Li, F.; Zhang, H.; xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2022. Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation. arXiv:2206.02777.

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. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Liew, J. H.; Cohen, S.; Price, B.; Mai, L.; and Feng, J. 2021. Deep Interactive Thin Object Selection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 305–314.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Lindsay, G. W. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10): 2017–2031.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, 416–423.

Olmos, A.; and Kingdom, F. A. A. 2004. McGill Calibrated Colour Image Database.

Qin, X.; Dai, H.; Hu, X.; Fan, D.-P.; Shao, L.; and Gool, L. V. 2022. Highly Accurate Dichotomous Image Segmentation. arXiv:2203.03041.

Ramanathan, V.; Kalia, A.; Petrovic, V.; Wen, Y.; Zheng, B.; Guo, B.; Wang, R.; Marquez, A.; Kovvuri, R.; Kadian, A.; Mousavi, A.; Song, Y.-Z.; Dubey, A.; and Mahajan, D. K. 2023. PACO: Parts and Attributes of Common Objects. *ArXiv*, abs/2301.01795.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Robinson, B. S.; and Drenkow, N. 2022. Cortical Transformers: Robustness and Model Compression with Multi-Scale Connectivity Properties of the Neocortex. *SVRHM Poster*. Published: 18 Oct 2022, Last Modified: 05 May 2023.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Cite arxiv:1505.04597Comment: conditionally accepted at MICCAI 2015.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Shi, J.; Tripp, B.; Shea-Brown, E.; Mihalas, S.; and Buice, M. A. 2022. MouseNet: A biologically constrained convolutional neural network model for the mouse visual cortex. *PLOS Computational Biology*.

Shi, J.; Yan, Q.; Xu, L.; and Jia, J. 2016. Hierarchical Image Saliency Detection on Extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4): 717–729.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Thisanke, H.; Deshan, C.; Chamith, K.; Seneviratne, S.; Vidanaarachchi, R.; and Herath, D. 2023. Semantic Segmentation using Vision Transformers: A survey. arXiv:2305.03273.

Toneva, M.; and Wehbe, L. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Adv. Neural. Inf. Process. Syst.*, volume 32, 14954–14964.

van Hateren, J.; and van der Schaaf, A. 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc Biol Sci*, 265(1394): 359–366.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Wei, T.; Li, X.; Chen, Y. P.; Tai, Y.; and Tang, C. 2019. FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation. *CoRR*, abs/1907.12347.

Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; and Yuan, L. 2022. TinyViT: Fast Pretraining Distillation for Small Vision Transformers. In *European conference on computer vision (ECCV)*.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency Detection via Graph-Based Manifold Ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 3166–3173.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022. Unified Contrastive Learning in Image-Text-Label Space. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19141–19151.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. arXiv:2306.14289.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2016. Semantic Understanding of Scenes through the ADE20K Dataset. *CoRR*, abs/1608.05442.

Supplementary Material

Allen Brain Observatory Dataset

The Allen Brain Observatory dataset is a comprehensive dataset containing important observations, and encompassing a vast array of neurons, each uniquely characterized by distinctive brain regions, genotypes, and neural responses. This section delves into the exploratory data analysis, shedding light on the diversity and characteristics of the neurons observed.

Supplementary **Figure S1** (middle) presents a bar chart illustrating the average peak response times of neurons in relation to the stimuli across various brain regions. Notably, the VISam region registers the longest average peak time, approximately 0.34 seconds, hinting at a more gradual or lagged neuronal response in this region compared to others. On the opposite end, the VISpm region records the shortest average peak time, suggesting a more rapid neuronal reaction.

Supplementary **Figure S1** (right) offers a detailed overview of the distribution of top-performing neurons (those with Peak DF/F values greater than 10) across various brain regions. The bar chart distinctly highlights that the VISp region is home to the largest count of these top neurons, with numbers surpassing 20. Other regions, such as VISl, VISal, and VISpm, also feature these top neurons, albeit in lesser quantities.

Supplementary **Figure S1** (left) and Supplementary **Figure S2** provide a comprehensive overview of the distribution of neurons across different brain regions and genotypes. Notably, the VISp region stands out, hosting the most significant number of neurons, while VISam appears less densely populated. This distribution suggests that our selected dataset is representative of the overall population, emphasizing the dataset's richness in terms of diversity.

Supplementary **Figure S3** offers a detailed visualization of DF/F traces for various neurons, capturing their unique activity profiles over time. These traces, which represent neuronal activation across different trials, provide a comprehensive view of the temporal dynamics of neuronal responses. The x-axis represents time in seconds, while the y-axis denotes the neural amplitude. By observing the evolution of these signals, both before and after stimulus presentation, and comparing them to peak DF/F values, we can discern the distinct response patterns of neurons. The variability in these DF/F traces, both in terms of their profiles and magnitudes, underscores the rich diversity of neuronal responses. This diversity is influenced by various factors, including the brain region in which the neuron is located, its genotype, and its functional characteristics. This figure, therefore, serves as a testament to the intricate and varied ways in which neurons respond to stimuli, highlighting the depth and richness of the data collected.

In order to completely understand Supplementary **Figures S2** and **S4**, refer to Supplementary Table S1, which provides the comprehensive mapping of genotype abbreviations to their corresponding full names, including Cre-labelling techniques. The best performing brain region in all models is VISpm, and the best performing genotype is Scnn1a-2.

Shortlisted Model Architectures

As explained in the main section of the manuscript, we deployed 15 model architectures for the comparisons with the neural kernels' data from mice visual cortex. All the models used in our comparisons consist of transformer encoder-decoder architecture as well as Swin (Liu et al. 2021)/ViT (Dosovitskiy et al. 2021) backbones pre-trained on the Imagenet-21k (Deng et al. 2009) dataset. The following is a comprehensive explanation of each model:

SAM (Kirillov et al. 2023b) SAM consists of an Image Encoder from a pre-trained ViT, adapted using Multilayer Autoencoder (MAE) techniques. For prompts, positional encodings and learned embeddings are used for sparse prompts, while convolutions are employed for dense prompts. The Mask Decoder is a modification of a transformer decoder block, which integrates prompt self-attention and cross-attention mechanisms followed by an MLP-driven output token that feeds into a dynamic linear classifier for mask foreground probabilities. Training was performed on SA-1B (Kirillov et al. 2023b), which contains more than 1B masks across 11M images.

Semantic-SAM (Li et al. 2023b) The Image Encoder consists of pre-trained Swin-T/L(Liu et al. 2021). For the decoder, nine decoder layers are used from MaskDINO(Li et al. 2022) for all the segmentation tasks. The pre-trained base model in UniCL (Yang et al. 2022) is used as a language model. Training datasets include SA-1B (Kirillov et al. 2023b), followed by COCO (Lin et al. 2014), ADE20k (Zhou et al. 2016), Objects365 (Shao et al. 2019), part segmentation data of PASCAL part (Chen et al. 2014), PartImageNet (He et al. 2021) and PACO (Ramanathan et al. 2023) for joint training.

MobileSAM (Zhang et al. 2023) MobileSAM architecture is built using decoupled distillation where a lightweight Image Encoder (ViT tiny) (Wu et al. 2022) is distilled directly from the heavyweight Image Encoder (ViT-H) in SAM. It is trained on only 1% of the SA-1B (Kirillov et al. 2023b).

HQ-SAM (Ke et al. 2023) In addition to the encoder-decoder architecture of SAM, HQ-SAM contains an HQ-Output Token and Global-local Feature Fusion for high-quality mask prediction. The HQ-Output Token generates new MLP layers for performing pointwise product with fused HQ-Features. It is trained using HQSeg-44K dataset which comprises of DIS (Qin et al. 2022) (train set), ThinObject-5K (Liew et al. 2021) (train set), FSS-1000 (Wei et al. 2019), DUT-OMRON (Yang et al. 2013), ECSSD (Shi et al. 2016), MSRA10K (Cheng et al. 2015) each one consisting of an average of 7.4K mask labels.

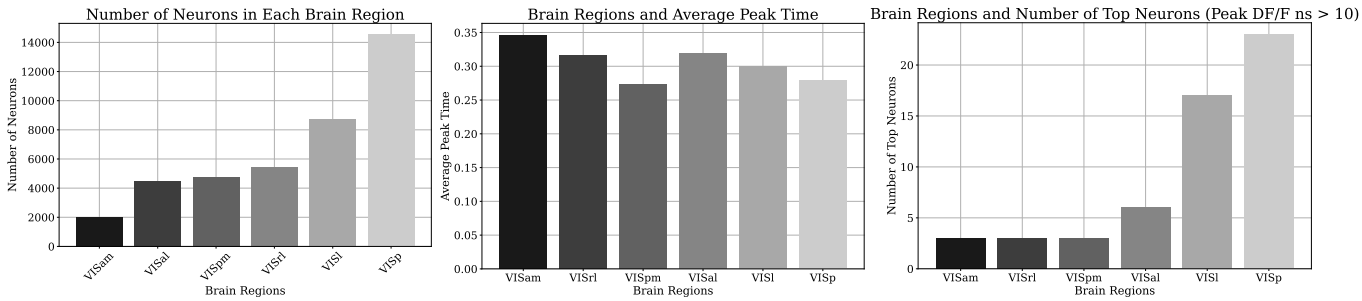
Mask2Former (Cheng et al. 2022) The model's setting is based on the same fundamental architecture as MaskFormer(Cheng, Schwing, and Kirillov 2021) - a Swin backbone (Liu et al. 2021) pre-trained on ImageNet-21K (Deng et al. 2009), pixel decoder, and a Transformer decoder. Mask2Former introduces a transformer decoder with masked attention. It uses Detectron2 (Wu et al. 2019), and

Abbreviation	Full Name
Cux2_1	Cux2-CreERT2/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Rbp4_1	Rbp4-Cre_KL100/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Cux2_2	Cux2-CreERT2/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Slc17a7_1	Slc17a7-IRES2-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Cux2_3	Cux2-CreERT2/Cux2-CreERT2;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Emx1_1	Emx1-IRES-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Rorb_1	Rorb-IRES2-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Fezf2	Fezf2-CreER/wt;Ai148(TIT2L-GC6f-ICL-tTA2)/wt
Scnn1a_1	Scnn1a-Tg3-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Ntsr1	Ntsr1-Cre_GN220/wt;Ai148(TIT2L-GC6f-ICL-tTA2)/wt
Rbp4_2	Rbp4-Cre_KL100/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Emx1_2	Emx1-IRES-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Nr5a1_1	Nr5a1-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Nr5a1_2	Nr5a1-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Rorb_2	Rorb-IRES2-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/wt
Tlx3	Tlx3-Cre_PL56/wt;Ai148(TIT2L-GC6f-ICL-tTA2)/wt
Slc17a7_2	Slc17a7-IRES2-Cre/wt;Camk2a-tTA/wt;Ai94(TITL-GCaMP6s)/wt
Emx1_3	Emx1-IRES-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)-hyg/wt
Scnn1a_2	Scnn1a-Tg3-Cre/wt;Camk2a-tTA/wt;Ai93(TITL-GCaMP6f)/Ai93(TITL-GCaMP6f)
Vip	Vip-IRES-Cre/wt;Ai148(TIT2L-GC6f-ICL-tTA2)/wt
Pvalb	Pvalb-IRES-Cre/wt;Ai162(TIT2L-GC6s-ICL-tTA2)/wt
Sst	Sst-IRES-Cre/wt;Ai148(TIT2L-GC6f-ICL-tTA2)/wt

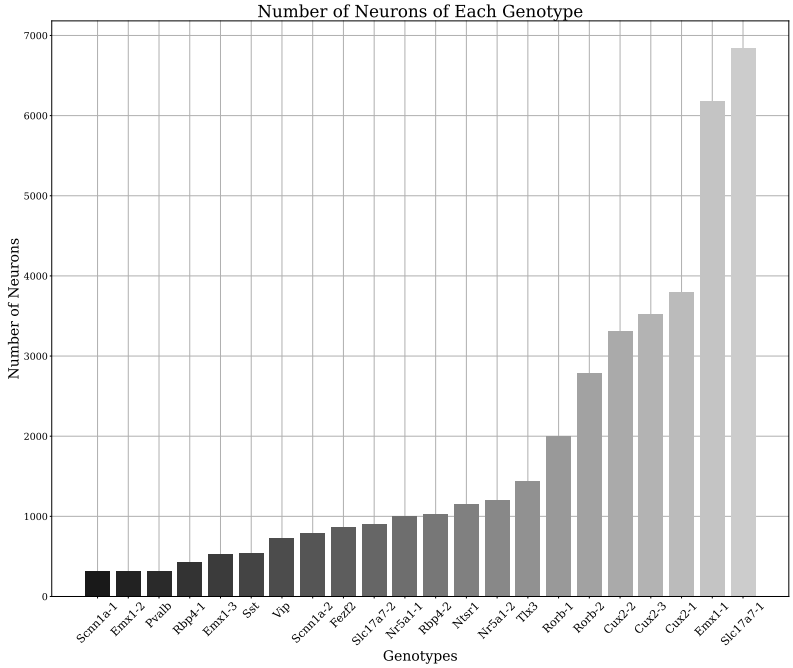
Supplementary Table S1: Mapping of Genotype abbreviations to full names with Cre-labelling techniques.

the updated Mask R-CNN (He et al. 2017a) baseline settings(Wu et al. 2019) for the COCO dataset.

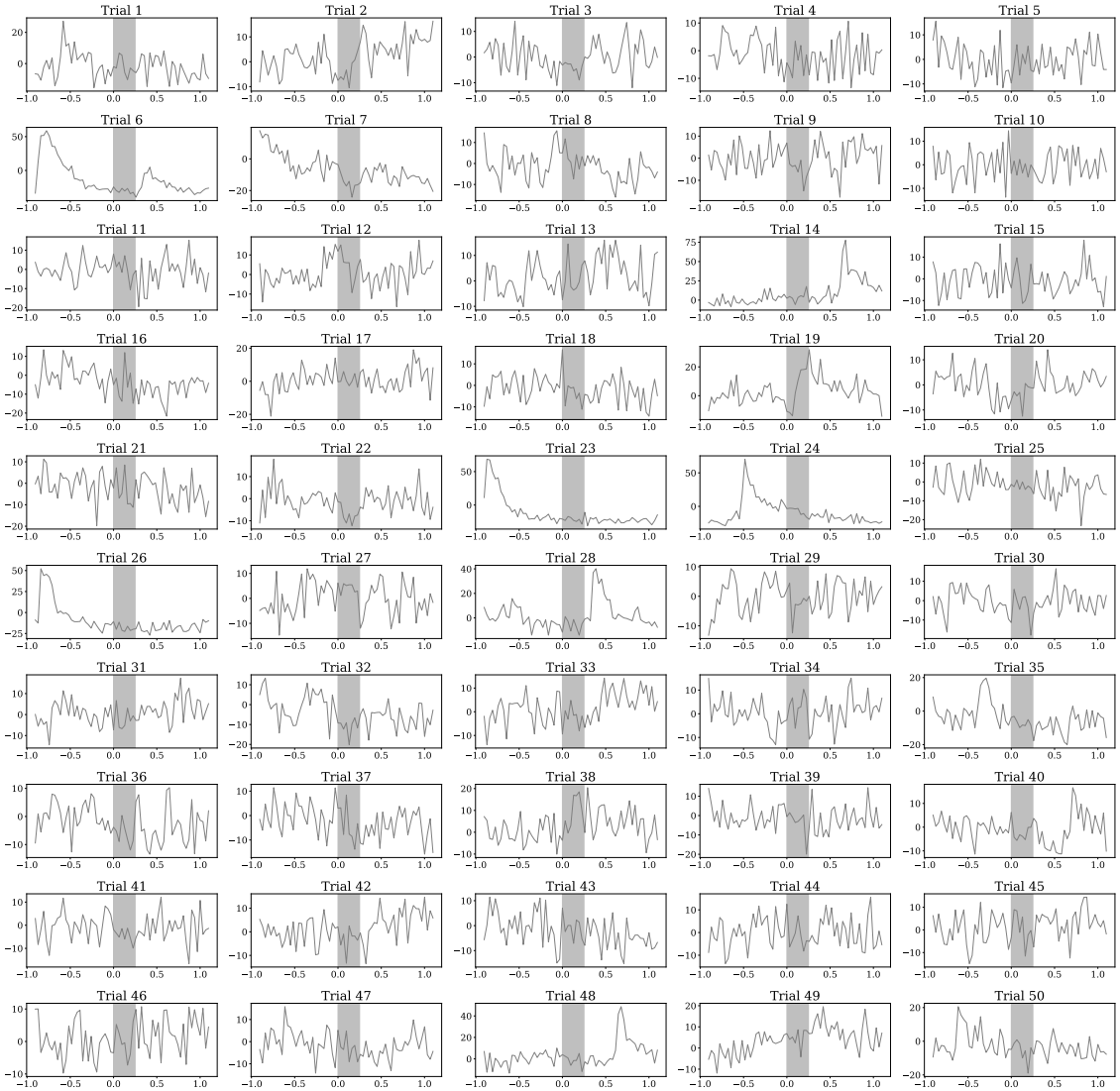
MaskDINO (Li et al. 2022) It builds from the fundamentals of Mask2Former (Cheng et al. 2022) to construct a pixel embedding map using the output from the backbone and Transformer encoder layer. The pixel embedding map and query embedding are combined via dot-product to obtain an output mask. The baseline settings (Wu et al. 2019) for segmentation tasks are the same as Mask2Former (Cheng et al. 2022), as well as the pre-trained ImageNet-21K (Deng et al. 2009) backbone.



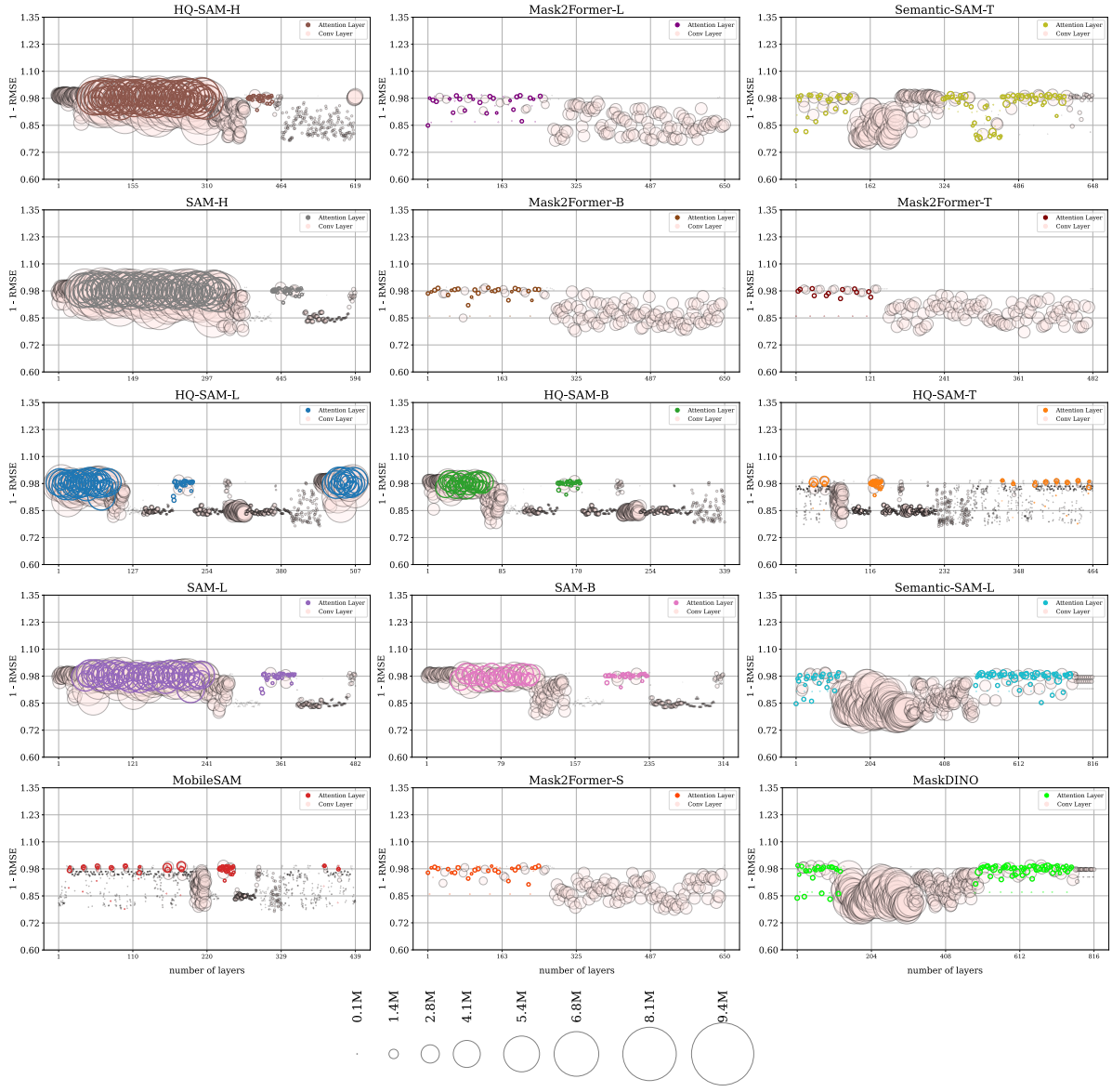
Supplementary Figure S1: Data statistics from Allen Brain Observatory. The left plot shows the unique number of brain regions in the mouse visual cortex from which the neural response dataset is collected. The middle plots show the average peak time of neural response in the individual visual cortex sub-regions. The right plots show the highest active neurons from the same sub-regions.



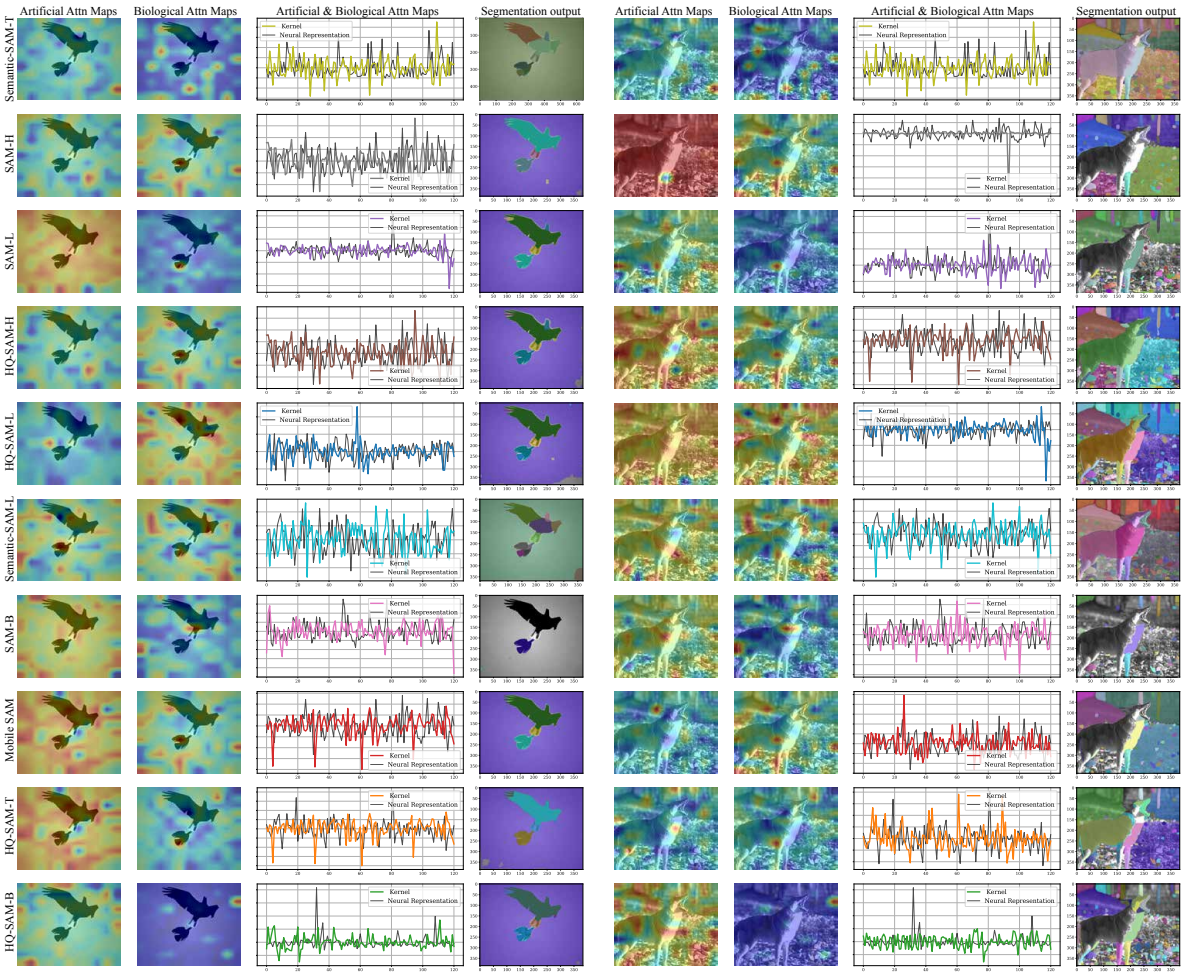
Supplementary Figure S2: Labelled neurons from each genotype. The plot represents the number of neurons that are captured by each genotype on the x-axis. Most of these genotypes capture the excitatory neurons, whereas genotypes like Sst, Vip, and Pvalb capture the subsets of inhibitory neurons in the brain.



Supplementary Figure S3: A range of DF/F traces showcase the activation patterns of distinct neurons across multiple trials, highlighting their individual activity profiles. The x-axis represents time in seconds, while the y-axis denotes neural amplitude.



Supplementary Figure S4: The figure showcases the performance of various models, with each model represented in a separate subplot. The y-axis indicates the performance metric ”1 - RMSE”, while the x-axis enumerates the layers within each model. The size of each scatter point corresponds to the magnitude of the layer’s parameters, with larger points indicating layers with more parameters. Two distinct color schemes are used to differentiate between layer types: attention layers are represented by their respective model’s unique color with a transparent fill, while the ”mistyrose” color is reserved for convolutional layers. This visualization provides insights into the distribution and performance of attention and convolutional layers across different neural network architectures.



Supplementary Figure S5: Extended qualitative comparison of artificial and neural kernels using multiple stimuli. Each row showcases overlaid artificial and neural kernels on top of two distinct input stimulus images: a bird and a wolf. These are followed by their 1D representations, with the black curve illustrating the neural kernel and the colored curves representing randomly selected artificial kernels (e.g., 1D representations of attention maps) from the model. The final column provides a qualitative display of the model outputs, highlighting the segmentations for each stimulus.