SPARSE YET ADAPTIVE: AN INTERPRETABLE TRANSFORMER FOR NEURAL POPULATION READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

How neural systems selectively read out reliable information from high-dimensional and redundant population activity remains a central problem in systems neuroscience. Conventional decoders can establish what information is present in neural activity, but provide limited insight into how that information may be flexibly read out. Here, we introduce a transformer-based framework that uses self-attention as an interpretable probe of population readout. Applied to large-scale macaque V1 recordings, the model revealed internal readout dynamics with clear biological correspondence: highly weighted neurons were tuned near the current stimulus, and removal of dominant readout neurons triggered recruitment of functionally similar substitutes, consistent with redundant and resilient population coding. We further applied the framework to sensory perturbations. Despite reduced contrast or moderate external noise, orientation information remained accurately recoverable, paralleling perceptual robustness. Attention analysis further suggested a possible mechanism for this stability: as sensory evidence degraded, accurate readout recruited additional neurons and weaker population interactions. Together, our framework moves neural decoding beyond black-box prediction toward an interpretable tool for uncovering computational strategies and generating mechanistic insights. Code is available at https://anonymous.4open.science/r/Transformer-model-for-accurate-orientation-coding-0751/.

1 INTRODUCTION

Neural populations represent sensory information through high-dimensional and often redundant patterns of activity. Such redundancy can support robust representations because related information is distributed across many neurons (Cunningham & Yu, 2014; Kriegeskorte & Wei, 2021; Stringer et al., 2019; Stringer & Pachitariu, 2024). However, representation alone does not specify how information is read out. Downstream computations must selectively extract task-relevant signals from populations whose neurons differ in tuning, response strength, variability, and shared fluctuations. A central question is therefore not only what information is present in a neural population, but how reliable readout can be constructed from many partially redundant signals.

Neural decoding provides a powerful framework for quantifying information available in population activity. Linear decoders recover stimulus variables through weighted pooling, whereas nonlinear models such as multilayer perceptrons can capture more complex mappings between neural activity and stimulus or behavioral variables (Hornik et al., 1989; Hornik, 1991; Glaser et al., 2020; Livezey & Glaser, 2021; Mathis et al., 2024). These approaches have substantially advanced the characterization of neural representations, but decoding accuracy alone provides limited insight into the internal organization of population readout. In particular, similar predictive performance may arise from different computational strategies, and conventional feedforward models do not provide an explicit condition-specific interaction structure showing how neuronal contributions vary across stimuli. This distinction is especially important in redundant populations, where many neurons may carry related information but need not contribute equally to a given readout (Averbeck et al., 2006; Butts & Goldman, 2006; Kang et al., 2004; Seriès et al., 2004).

Self-attention provides a natural computational framework for examining such condition-dependent readout. Rather than assigning each input a single fixed contribution, self-attention dynamically

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

weights neuronal representations according to the current population activity (Vaswani et al., 2017), providing an explicit stimulus-dependent interaction structure that can be inspected during decoding. Importantly, attention scores are not assumed to correspond directly to biological synaptic connectivity. Instead, we use self-attention as an interpretable computational probe of population readout. The critical question is therefore not simply whether attention improves decoding, but whether the internal readout structure revealed by the model corresponds to meaningful neuronal properties and reorganizes systematically under perturbation.

Primary visual cortex provides a well-characterized system for studying how redundant population representations are selectively read out. V1 encodes fundamental visual features such as orientation through large populations of neurons with overlapping tuning preferences and heterogeneous responses (De Valois et al., 1982; Hubel & Wiesel, 1959; 1962; Ringach et al., 2002; Schiller et al., 1976). Such distributed coding can support robust sensory representations, but also creates a readout-efficiency problem: recovering relevant information need not require downstream circuits to use every active neuron equally. V1 orientation coding therefore provides a useful testbed for examining how neural systems balance redundancy with selective readout, including which neurons dominate a given readout and how readout strategies reorganize when dominant neural signals are disrupted or sensory evidence is degraded.

Here, we develop a transformer-based framework to reconstruct oriented visual stimuli from large-scale two-photon recordings of macaque V1 populations. The model combines neuron-specific embeddings that preserve neuronal identity with self-attention that captures stimulus-dependent interactions among neuronal representations, enabling joint modeling of neuron-specific properties and population-level readout dynamics. The resulting attention patterns provide an interpretable window into the computational organization of population readout. Using this framework, we identify a sparse and stimulus-dependent readout aligned with neuronal tuning and show that this readout flexibly reorganizes under both neuronal and sensory perturbations. More broadly, the framework uses self-attention not only as a decoding mechanism, but as a computational probe for generating testable hypotheses about neural population readout.

Our main contributions are as follows:

1. We develop an interpretable transformer-based framework that jointly models neuron-specific properties and stimulus-dependent population interactions.
2. We show that self-attention reveals a sparse, biologically aligned population readout that cannot be inferred from decoding accuracy alone.
3. We demonstrate adaptive reorganization of this readout under both neuronal and sensory perturbations, providing computational insight into redundancy and robustness in neural population coding.

2 Dataset

2.1 Visual stimuli

Two sets of visual stimuli were used in this study. For the main Gabor dataset, visual stimuli consisted of Gabor patches spanning 12 orientations from 0° to 165° in 15° steps, four spatial frequencies (1, 2, 4, and 8 cpd), and three stimulus sizes (sigma of the Gaussian envelop of the Gabor function), yielding 144 stimulus conditions in total. Stimulus contrast was fixed at 90%.

For the sensory-perturbation dataset, drifting Gabor stimuli were systematically manipulated in signal contrast and external noise. Each Gabor had a spatial frequency of 4 cpd, a diameter of 1.5 cycles, and drifted at 4 Hz. Zero-mean Gaussian white noise was superimposed on the Gabor and updated dynamically across frames at the same temporal rate. Stimulus conditions differed slightly between the two macaques. For Monkey A (MA), stimuli comprised 12 orientations in 15° steps, five signal contrasts (3.125%, 6.25%, 12.5%, 25%, and 50%), and five RMS noise contrasts (0, 0.05, 0.09, 0.16, and 0.29); the 3.125% contrast condition was excluded from subsequent analyses. For Monkey B (MB), stimuli comprised 10 orientations in 18° steps, four signal contrasts (6.25%, 12.5%, 25%, and 50%), and six RMS noise contrasts (0, 0.03, 0.05, 0.09, 0.16, and 0.29). Representative stimulus images across contrast and noise conditions are shown in Appendix Fig. A.1.

2.2 NEURAL RECORDINGS

We analyzed two-photon calcium recordings from two datasets. The main Gabor dataset contained seven V1 fields of view (FOVs), each recorded from a different awake, fixating macaque. Data from four animals (Monkeys D–G) were previously reported (Ju et al., 2021; 2022), whereas data from the remaining three animals (Monkeys A–C) were previously unpublished. The sensory-perturbation dataset contained three additional V1 FOVs, all previously unpublished. The first FOV was recorded from one macaque (Monkey H), whereas the remaining two FOVs were recorded from a second macaque (Monkey I).

Recording and preprocessing procedures were otherwise identical across the two datasets. Each FOV covered an area of 850×850 $\mu\mathrm{m}^2$ at a cortical depth of approximately 150 $\mu\mathrm{m}$. Representative recording sites and two-photon images are shown in Appendix Fig. B.1A–B. Imaging data were aligned using a normalized cross-correlation-based translation algorithm, and neuronal regions of interest (ROIs) were identified from differential fluorescence images across stimulus conditions using band-pass filtering, intensity thresholding, and iterative overlap-based refinement. Only approximately circular ROIs with roundness greater than 0.9 were retained to reduce neuropil contamination. The resulting ROIs ranged from 3.3 to 19.3 μm in diameter, with a mean of 8.9 ± 1.3 μm. Representative orientation functional maps are shown in Appendix Fig. B.1D.

For each neuron and trial, baseline fluorescence F_0 was estimated from the four frames preceding stimulus onset (-500 to 0 ms), and evoked fluorescence F_n from the response window 500–1,000 ms after stimulus onset. The baseline was averaged across the 12 repeats of each stimulus condition to reduce measurement noise, and the response of trial n was quantified as $\Delta F_n/F_0 = (F_n - F_0)/F_0$. The mean calcium-response time course is shown in Appendix Fig. B.1C. Condition-level responses used for decoding were obtained by averaging across the 12 repeated trials. Population distributions of OSI, preferred orientation, and SNR are shown in Appendix Fig. B.1E.

3 METHODS

3.1 TASK

We formulate neural decoding as stimulus reconstruction rather than single-feature prediction. Visual stimuli vary jointly along multiple dimensions, including orientation, spatial frequency, size, and contrast, and V1 population responses reflect these combined stimulus properties. Single-feature decoding therefore captures only a partial projection of the information contained in the population. We instead reconstruct the complete stimulus image from the population response vector, requiring the model to preserve joint stimulus information rather than merely discriminate individual feature labels. Formally, given a population response vector $\mathbf{r} \in \mathbb{R}^N$ from N simultaneously recorded neurons, the decoder learns a mapping $\mathcal{F} : \mathbf{r} \to \hat{\mathbf{I}}$, where $\hat{\mathbf{I}}$ is the reconstructed stimulus image.

3.2 ARCHITECTURE

Given the responses of N simultaneously recorded neurons, the model takes a population response vector $\mathbf{r} = [r_1, \ldots, r_N]^\top \in \mathbb{R}^N$ as input and reconstructs the corresponding stimulus image $\hat{\mathbf{I}} \in \mathbb{R}^{H \times W}$. The model consists of neuron-specific embedding, positional encoding, single-layer self-attention, unembedding, and a linear reconstruction layer. The overall architecture is illustrated in Fig. 1.

Neuron-specific embedding. Each scalar neuronal response r_i is mapped to a d-dimensional latent representation using a neuron-specific learnable embedding vector $\mathbf{w}_i^{\mathrm{emb}}$:

$$\mathbf{e}_i = r_i \mathbf{w}_i^{\mathrm{emb}}. \tag{1}$$

The embedded representations are stacked to form $\mathbf{E} \in \mathbb{R}^{N \times d}$. Using neuron-specific embeddings preserves the identity of individual neurons while placing their responses in a common latent space.

Positional encoding. Neurons are ordered according to their preferred orientations before a fixed sinusoidal positional encoding is added. Let $\mathbf{P} \in \mathbb{R}^{N \times d}$ denote the positional encoding matrix. The

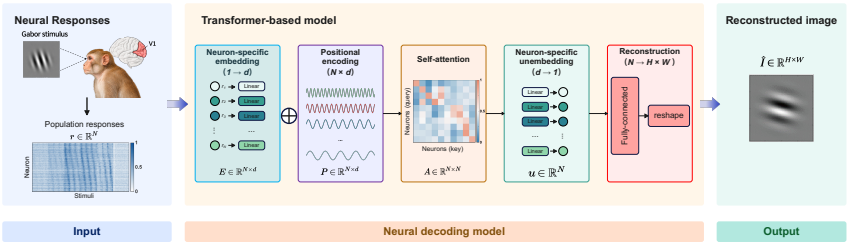

Figure 1: **Architecture of the transformer-based model.** The input is a population response vector from simultaneously recorded V1 neurons. The model consists of neuron-specific embedding, positional encoding, self-attention, neuron-specific unembedding, and a fully-connected reconstruction layer. The output is the reconstructed visual stimulus image.

input to self-attention is

$$\mathbf{X} = \mathbf{E} + \mathbf{P}. \tag{2}$$

This ordering incorporates the functional organization of orientation preference into the population representation.

Self-attention. The embedded population activity is transformed into query, key, and value representations,

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \qquad \mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \qquad \mathbf{V} = \mathbf{X}\mathbf{W}^{V}, \tag{3}$$

and the self-attention matrix is computed as

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right). \tag{4}$$

The attention-weighted neuronal representations are then

$$\mathbf{Z} = \mathbf{A}\mathbf{V}. \tag{5}$$

For each stimulus, $\mathbf{A} \in \mathbb{R}^{N \times N}$ provides a stimulus-dependent weighting among neuronal representations and is used in subsequent analyses of population readout.

Unembedding and image reconstruction. Each attention-modulated representation $\mathbf{z}_i$ is projected back to a scalar using a neuron-specific unembedding vector $\mathbf{w}_i^{\mathrm{unemb}}$:

$$u_i = \left(\mathbf{w}_i^{\mathrm{unemb}}\right)^{\top} \mathbf{z}_i. \tag{6}$$

The resulting vector $\mathbf{u} = [u_1, \ldots, u_N]^{\top}$ is linearly mapped to the image space,

$$\hat{\mathbf{y}} = \mathbf{W}_{\mathrm{out}}\mathbf{u} + \mathbf{b}_{\mathrm{out}}, \tag{7}$$

and reshaped to obtain the reconstructed stimulus,

$$\hat{\mathbf{I}} = \mathrm{reshape}\left(\hat{\mathbf{y}}, H, W\right). \tag{8}$$

3.3 Training objective

The model was trained to reconstruct the stimulus image from the corresponding population response. We randomly held out 10% of the stimulus conditions as a test set, which was not used for parameter optimization. Model performance was evaluated on the held-out test set only after training was completed. Given M training conditions, the reconstruction objective was the mean squared error (MSE) between the reconstructed and target images:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{M}\sum_{j=1}^{M} \mathrm{MSE}\left(\hat{\mathbf{I}}_j, \mathbf{I}_j\right), \tag{9}$$

where $\mathbf{I}_j$ and $\hat{\mathbf{I}}_j$ denote the target and reconstructed stimulus images for condition j, respectively.

Model parameters were optimized using RMSprop with a learning rate of 5×10^{-5} and a smoothing coefficient of 0.9.

3.4 BASELINES

We compared the transformer with three simpler decoding models. First, a linear decoder directly mapped the population response vector to a two-dimensional circular representation of stimulus orientation. Second, a nonlinear multilayer perceptron (MLP) with one hidden layer provided a control for generic nonlinear decoding capacity. Both models predicted orientation using the representation $\mathbf{y} = [\sin(2\theta), \cos(2\theta)]$. We also implemented a linear image-reconstruction decoder that mapped the same population responses directly to the Gabor image using the same reconstruction targets as the transformer. This model therefore provided a matched control for the image-reconstruction objective without self-attention. Implementation details and training procedures for all baseline models are provided in Appendix A.2.

4 RESULTS

4.1 HIGH-PRECISION RECONSTRUCTION

The transformer-based model achieved a mean structural similarity index (SSIM) of 0.983 between reconstructed and target Gabor images in testset. Orientations extracted from the reconstructions closely matched the true stimulus orientations, yielding a median absolute error of only 0.12° (see Appendix A.3 for orientation extraction). For subsequent analyses, we focused on the 12 orientations presented at the spatial frequency and stimulus size preferred by the majority of neurons in each FOV.

We next examined whether this high precision depended on the self-attention operation. We therefore removed the self-attention module while keeping the remaining architecture, reconstruction target, and training objective unchanged. This ablation reduced reconstruction quality and increased orientation error to 2.95°, indicating that self-attention contributed substantially to the high-precision reconstruction achieved by the full model.

We next compared the orientation information recovered by the transformer against conventional orientation decoders. A linear decoder and a one-hidden-layer multilayer perceptron (MLP) were trained to predict stimulus orientation directly from the population responses, without reconstructing the stimulus image. These models yielded median absolute orientation errors of 2.83° and 2.92°, respectively, compared with 0.12° for the orientation extracted from the transformer reconstructions. Thus, even when optimized specifically for orientation prediction, the direct decoders recovered orientation less precisely than the reconstruction-based transformer.

4.2 SELF-ATTENTION REVEALS SPARSE, STIMULUS-DEPENDENT READOUT STRUCTURE

We used the self-attention module to examine how the model dynamically selects population signals for accurate stimulus reconstruction. For each stimulus condition, self-attention generated an $N \times N$ attention matrix with values ranging from 0 to 1, where each entry quantified the model-assigned influence of one neuronal representation on another during reconstruction. These attention scores therefore provide a stimulus-specific measure of directed functional interactions within the model and allow us to identify neurons that receive particularly strong population-level weighting.

For a 0° stimulus in Monkey A, the full attention matrix and the corresponding mean incoming attention score for each neuron are shown in Fig. 2A. The latter was obtained by averaging the attention received by each neuron from the rest of the population. The attention matrix was highly sparse, with most neuron-to-neuron interactions carrying near-zero weights, whereas only a small subset of neurons received substantial attention from others. Consistently, mean incoming attention scores were strongly concentrated in a few neurons. To further visualize this sparse, non-uniform readout structure, we represented the attention-defined interactions as a network topology (top) and mapped the same interactions onto the recorded neurons at their physical cortical locations (bottom) (Fig. 2B).

Across all stimuli, attention scores consistently followed a strongly right-skewed distribution, with most interactions carrying very small weights and a small fraction extending into a long high-weight tail (Fig. 2C). The overall distributional form was preserved across stimulus orientations, despite condition-specific differences in which neurons received the strongest weights. This organi-

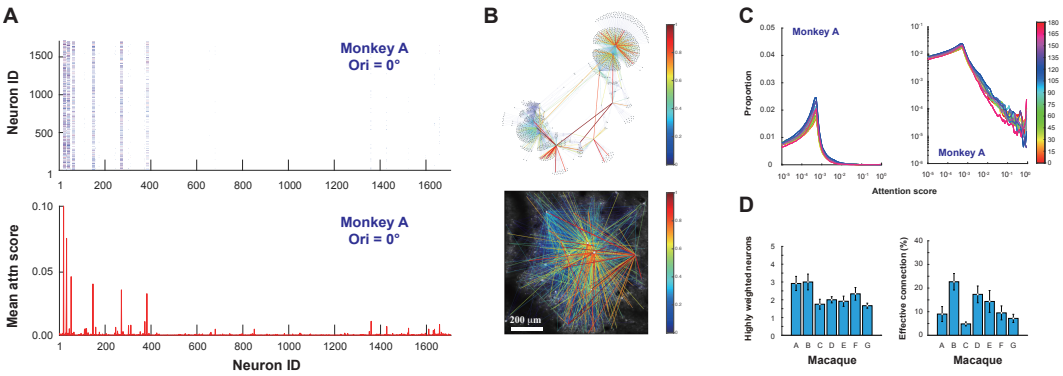

Figure 2: **Attention maps and Highly weighted neurons. A.** Attention matrix for a 0° stimulus in Monkey A (top) and the corresponding mean incoming attention score for each neuron (bottom). **B.** Network visualization of attention-defined interactions (top) and the same interactions mapped onto the recorded cortical locations (bottom). Edge color indicates attention score. **C.** Exemplar distribution of attention scores across stimulus orientations in Monkey A, shown on linear (left) and logarithmic (right) scales. **D.** Number of highly weighted neurons (left) and percentage of effective connections per highly weighted neuron (right) required to reach the 5° orientation-error criterion across macaques.

zation qualitatively resembles mixed lognormal and power-law characteristics and is consistent with strongly skewed connectivity distributions reported in the brain (Buzsáki & Mizuseki, 2014).

We then asked how many highly weighted neurons were actually required to support decoding. For each stimulus orientation, neurons were ranked by their mean incoming attention scores and progressively retained from highest to lowest weight, while the remaining neurons were nullified. Reconstruction accuracy improved as more neurons were included, with orientation error progressively decreasing (see Appendix Fig. B.2 for three examples of the procedure). Using an orientation error of $< 5^\circ$ as the criterion for accurate reconstruction, only approximately 2–3 highly weighted neurons were required (Fig. 2D, left). Moreover, retaining only the top $12.12 \pm 1.31\%$ of attention-defined interactions received by these neurons was sufficient to reach the same criterion (Fig. 2D, right), corresponding to only $0.021 \pm 0.003\%$ of all possible pairwise interactions in the population. Thus, despite the high dimensionality of the recorded population, the transformer's effective readout was extremely sparse.

4.3 Model-derived readout aligns with biologically meaningful tuning

The sparse readout identified above reflects the internal behavior of the decoder and does not by itself establish biological relevance. The biological relevance of this sparse readout was assessed by determining whether highly weighted neurons showed response properties systematically aligned with the current stimulus. Across stimulus orientations, highly weighted neurons were preferentially tuned near the presented orientation (Fig. 3A). Despite the broad distribution of orientation preferences in the recorded population, highly weighted neurons consistently showed preferred orientations close to the presented stimulus across all monkeys, with median orientation differences of approximately 5–10° (Fig. 3B).

This stimulus-aligned tuning was robust across model initializations. Different random seeds did not necessarily select the same individual neurons, but the resulting highly weighted neurons consistently preferred orientations close to the presented stimulus (Appendix Fig. B.3). Thus, the model-derived readout was more stable at the level of functional neuronal properties than at the level of individual neuron identity. This functional stability is consistent with sparse and variable visual representations (Yoshida & Ohki, 2020) and with causal evidence that small functionally defined neuronal ensembles can influence perception (Afraz et al., 2006; Carrillo-Reid et al., 2019; Marshel et al., 2019).

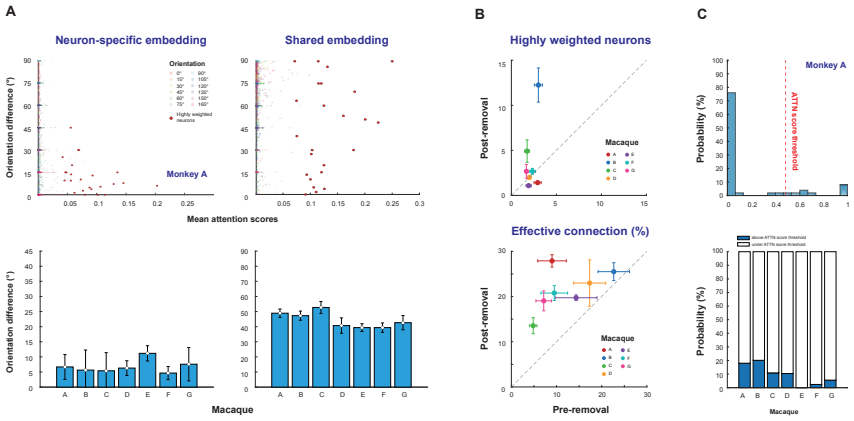

Figure 3: **Biological alignment and reorganization of highly weighted neurons. A.** Relationship between mean attention score and orientation difference under neuron-specific (left) and shared (right) embeddings in an example FOV from Monkey A (top). Colored points indicate neurons across different stimulus-orientation conditions, and red outlined points indicate the identified highly weighted neurons. Bottom, mean orientation differences of highly weighted neurons across macaques. **B.** Number of highly weighted neurons (top) and percentage of effective connections per highly weighted neuron (bottom) before and after removal of the original highly weighted neurons. Dashed lines indicate equality between pre- and post-removal values. **C.** Example distribution of attention scores from substitute to original highly weighted neurons in Monkey A relative to the effective attention-score threshold (top), and proportions of connections above and below the threshold across macaques (bottom).

We next examined whether this biologically aligned readout depended on preserving neuron-specific functional heterogeneity in the embedding space. In the original model, each neuron was assigned an independent learnable embedding, allowing the latent representation to retain neuron-specific identity and response characteristics. We therefore replaced these embeddings with a single shared embedding across the population, effectively removing neuron-specific identity from the latent representation. Reconstruction accuracy remained high under the shared embedding, with a median orientation error of only $0.09°$. However, the orientation differences of highly weighted neurons increased markedly, and their preferred orientations were no longer systematically aligned with the presented stimulus (Fig. 3A). Thus, comparable decoding accuracy could be achieved with substantially different internal readout structures, indicating that preserving neuron-specific functional heterogeneity promotes closer alignment between the model-derived readout and biologically meaningful properties of neural population coding.

4.4 Readout reorganizes after removal of highly weighted neurons

To test whether the highly weighted neurons were indispensable for accurate reconstruction, we removed them from the population and retrained the model using the remaining neurons. Reconstruction performance recovered to a mean SSIM of 0.989, accompanied by the recruitment of a new set of substitute highly weighted neurons. After exclusion of one outlier FOV, the number of substitute neurons did not differ significantly from that of the original highly weighted neurons ($F(1, 66) = 1.644$, $p = 0.204$; Fig. 3B, top). In contrast, the percentage of effective connections per highly weighted neuron increased significantly after removal ($F(1, 77) = 37.224$, $p < 0.001$; Fig. 3B, bottom), indicating that the substitute neurons required broader population input to support the same decoding criterion.

The neurons that later became substitutes contributed only weakly to the original readout before removal. In an example FOV, most attention scores from these neurons to the original highly weighted neurons remained below the effective-connection threshold (Fig. 3C, top). Across macaques, only $9.62 \pm 2.86\%$ of these connections exceeded the threshold (Fig. 3C, bottom). Thus, neurons capable of supporting the same stimulus readout were not necessarily recruited simultaneously: a small

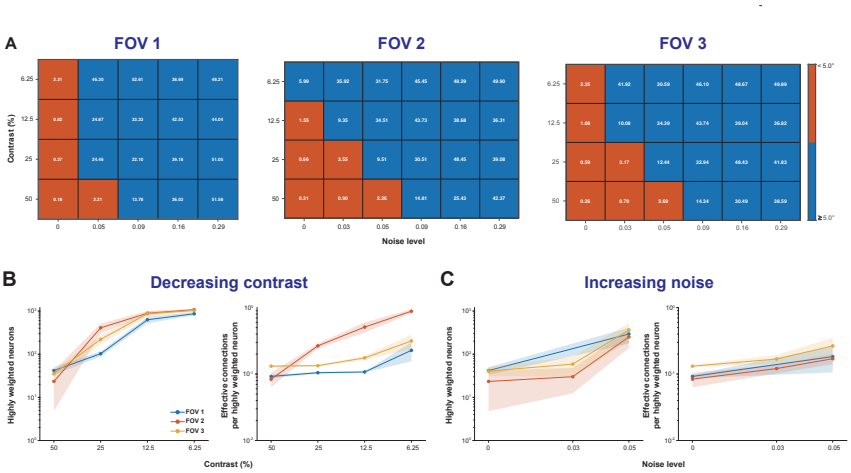

Figure 4: **Population readout under degraded visual stimuli. A.** Mean orientation reconstruction error across stimulus-contrast and noise levels for the three FOVs. Numbers indicate orientation error in degrees; orange and blue cells denote errors below and above the $5°$ criterion, respectively. **B.** Number of highly weighted neurons (left) and percentage of effective connections per highly weighted neuron (right) required to reach the $5°$ criterion as stimulus contrast decreased. **C.** Number of highly weighted neurons (left) and percentage of effective connections per highly weighted neuron (right) required to reach the $5°$ criterion as external noise increased.

subset dominated the original readout, whereas alternative neurons remained weakly engaged and were recruited only after the dominant contributors were removed.

Together, these results reveal a redundant but selectively recruited population readout. Functionally overlapping neurons need not contribute equally at the same time; instead, alternative population signals can be recruited when the dominant readout is disrupted. This result is consistent with recent in vivo evidence that ablation of an individual neuron can leave the function of its associated visual ensemble largely intact, supporting functional redundancy in cortical populations (Wang & He, 2025). The transformer-based framework captures a similar redundancy principle computationally and further shows how adaptive recruitment and reweighting of alternative neurons can preserve population-level readout after perturbation.

4.5 Population readout adapts to degraded sensory evidence

Visual orientation judgments can remain relatively stable across a broad range of stimulus contrasts and moderate levels of external noise, despite substantial degradation of sensory responses in early visual cortex. The transformer showed a similar robustness when reconstructing stimulus orientation from macaque V1 population activity (Fig. 4A).

In the absence of external noise, reducing stimulus contrast produced little loss of orientation accuracy. Reconstruction errors remained below the $5°$ criterion across all tested contrasts in FOVs 1 and 3, and across all but the lowest contrast (6.25%) in FOV 2, where the error increased only slightly to $5.99°$. External noise produced a stronger dependence on sensory quality. At the highest stimulus contrast, orientation errors remained below the $5°$ criterion up to moderate noise levels ($\mathrm{RMS} \leq 0.05$) in all three FOVs, but increased sharply at higher noise levels ($\mathrm{RMS} \geq 0.09$), where the Gabor orientation became largely obscured by the external noise and was difficult to visually discern. Thus, orientation information remained accurately recoverable when the sensory signal was weakened by contrast reduction or moderately corrupted by external noise, before breaking down under more severe degradation (Fig. 4A).

The attention-based readout further showed how this robustness was achieved. As stimulus contrast decreased, progressively more highly weighted neurons were required to maintain the same reconstruction criterion (Fig. 4B, left). These neurons also relied on a larger fraction of effective connections per highly weighted neuron (Fig. 4B, right). A similar pattern emerged with increasing

external noise: maintaining accurate reconstruction at moderate noise levels required recruitment of more highly weighted neurons together with additional effective connections (Fig. 4C).

These results suggest that stable orientation readout under degraded sensory input is supported by adaptive recruitment of additional population signals. When sensory evidence is strong, a small set of strongly weighted neurons is sufficient; as the signal becomes weaker or noisier, the readout incorporates more neurons and additional, predominantly weaker, connections to preserve orientation information.

Furthermore, contrast reduction and external noise induced distinct reorganizations of the learned readout. In the low-dimensional space of self-attention maps, the first three principal components explained 64.4%, 80.4%, and 80.5% of the variance in FOV 1, FOV 2, and FOV 3, respectively. Across orientations and contrast conditions, the resulting intersection angles between the contrast- and noise-related directions were consistently near orthogonal ($68.58°$, $70.59°$, and $65.04°$ across the three FOVs), exceeding the chance-level expectation of $57.3°$. Thus, the two forms of sensory degradation were associated with distinct readout trajectories rather than a single generic compensatory pattern.

5 Discussion

Summary. In this work, we developed a transformer-based framework for studying how information can be selectively read out from high-dimensional and redundant neural population activity. By combining neuron-specific embeddings with self-attention, the model achieved high-precision stimulus reconstruction while providing a stimulus-dependent representation of population readout. Interpretable analysis of self-attention revealed internal computational strategies with clear biological correspondence. Highly weighted neurons consistently showed preferred orientations close to the current stimulus, while removal of dominant readout neurons triggered recruitment of functionally similar substitutes. Under degraded visual input, accurate orientation readout was maintained by recruiting broader neuronal populations together with additional weak interactions. Together, these results suggest a sparse but adaptive population readout that flexibly exploits redundant neural signals to maintain robustness under both neural and sensory perturbations.

Limitations and future directions. First, self-attention is a model-derived quantity and should not be interpreted as anatomical, synaptic, or functional connectivity. The attention-defined interactions instead show systematic correspondence with neuronal tuning and perturbation-induced reorganization, suggesting that they may capture functionally relevant relationships within the population. Establishing whether these relationships correspond to genuine biological functional connectivity will require additional causal and circuit-level validation. Second, the present study focuses on orientation coding in macaque V1 using relatively simple visual stimuli; future work could extend the framework to more complex visual tasks to examine whether similar readout principles generalize beyond basic feature encoding. Third, neuronal perturbation was implemented computationally by removing highly weighted neurons and retraining the model, which does not fully capture biological adaptation to neuronal loss. The sensory-perturbation experiments also lacked simultaneous behavioral measurements, limiting direct conclusions about perception. Future work combining targeted perturbations, behavior-coupled recordings, and downstream-area measurements could test whether model-derived readout structures predict causal changes in neural computation and behavior.

Broader impact. More broadly, this work illustrates how modern AI architectures can serve not only as predictive models, but also as hypothesis-generating tools for neuroscience. By exposing stimulus- and perturbation-dependent structure in large-scale neural activity, interpretable decoders can help bridge population-level recordings with mechanistic questions about selective readout, redundancy, and robustness. At the same time, biological constraints provide an important benchmark for evaluating whether learned internal representations are functionally meaningful, rather than merely predictive. This bidirectional interaction between machine learning and neuroscience may therefore enable a new class of models that are both computationally powerful and experimentally grounded, helping connect large-scale neural data to mechanistic theories of brain computation and guide future closed-loop experiments.

AI USE STATEMENT

AI tools were used solely for language editing to improve the clarity, grammar, and readability of the manuscript. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

Seyed-Reza Afraz, Roozbeh Kiani, and Hossein Esteky. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, 442(7103):692–695, 2006. doi: 10.1038/nature04982.

Bruno B. Averbeck, Peter E. Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006. doi: 10.1038/nrn1888.

Daniel A. Butts and Mark S. Goldman. Tuning curves, neuronal variability, and sensory coding. *PLoS Biology*, 4(4):e92, 2006. doi: 10.1371/journal.pbio.0040092.

György Buzsáki and Kenji Mizuseki. The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, 15(4):264–278, 2014. doi: 10.1038/nrn3687.

Luis Carrillo-Reid, Shuting Han, Weijian Yang, Alejandro Akrouh, and Rafael Yuste. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell*, 178(2):447–457.e5, 2019. doi: 10.1016/j.cell.2019.05.045.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. doi: 10.1038/nn.3776.

Russell L. De Valois, E. William Yund, and Norva Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):531–544, 1982. doi: 10.1016/0042-6989(82)90112-2.

Joshua I. Glaser, Ari S. Benjamin, Raeed H. Chowdhury, Matthew G. Perich, Lee E. Miller, and Konrad P. Kording. Machine learning for neural decoding. *eNeuro*, 7(4):ENEURO.0506–19.2020, 2020. doi: 10.1523/ENEURO.0506-19.2020.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4 (2):251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148(3):574–591, 1959. doi: 10.1113/jphysiol.1959.sp006308.

David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962. doi: 10.1113/jphysiol.1962.sp006837.

N. S. Ju, S. C. Guan, L. Tao, S. M. Tang, and C. Yu. Orientation tuning and end-stopping in macaque v1 studied with two-photon calcium imaging. *Cerebral Cortex*, 31:2085–2097, 2021. doi: 10.1093/cercor/bhaa346.

N. S. Ju, S. C. Guan, S. M. Tang, and C. Yu. Macaque v1 responses to 2nd-order contrast-modulated stimuli and the possible subcortical and cortical contributions. *Progress in Neurobiology*, 217: 102315, 2022. doi: 10.1016/j.pneurobio.2022.102315.

Kukjin Kang, Robert M. Shapley, and Haim Sompolinsky. Information tuning of populations of neurons in primary visual cortex. *Journal of Neuroscience*, 24(15):3726–3735, 2004. doi: 10.1523/JNEUROSCI.4272-03.2004.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021. doi: 10.1038/s41583-021-00502-3.

Jesse A. Livezey and Joshua I. Glaser. Deep learning approaches for neural decoding across architectures and recording modalities. *Briefings in Bioinformatics*, 22(2):1577–1591, 2021. doi: 10.1093/bib/bbaa355.

James H. Marshel, Yoon Seok Kim, Timothy A. Machado, Sean Quirin, Brandon Benson, Jonathan Kadmon, Cephra Raja, Adelaida Chibukhchyan, Charu Ramakrishnan, Masatoshi Inoue, Janelle C. Shane, Douglas J. McKnight, Susumu Yoshizawa, Hideaki E. Kato, Surya Ganguli, and Karl Deisseroth. Cortical layer-specific critical dynamics triggering perception. *Science*, 365(6453):eaaw5202, 2019. doi: 10.1126/science.aaw5202.

Mackenzie Weygandt Mathis, Adriana Perez Rotondo, Edward F. Chang, Andreas S. Tolias, and Alexander Mathis. Decoding the brain: From neural representations to mechanistic models. *Cell*, 187(21):5814–5832, 2024. doi: 10.1016/j.cell.2024.08.051.

Dario L. Ringach, Robert M. Shapley, and Michael J. Hawken. Orientation selectivity in macaque v1: diversity and laminar dependence. *Journal of Neuroscience*, 22(13):5639–5651, 2002. doi: 10.1523/JNEUROSCI.22-13-05639.2002.

Peter H. Schiller, Barbara L. Finlay, and Sherwin F. Volman. Quantitative studies of single-cell properties in monkey striate cortex. i. spatiotemporal organization of receptive fields. *Journal of Neurophysiology*, 39(6):1288–1319, 1976. doi: 10.1152/jn.1976.39.6.1288.

Peggy Seriès, Peter E. Latham, and Alexandre Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience*, 7(10):1129–1135, 2004. doi: 10.1038/nn1321.

Carsen Stringer and Marius Pachitariu. Analysis methods for large-scale neuronal recordings. *Science*, 386(6722):eadp7429, 2024. doi: 10.1126/science.adp7429.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019. doi: 10.1038/s41586-019-1346-5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Haipeng Wang and Hao He. Optical precise ablation of targeted individual neurons in vivo. *ACS Chemical Neuroscience*, 16(3):374–383, 2025. doi: 10.1021/acschemneuro.4c00538.

Takashi Yoshida and Kenichi Ohki. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11:872, 2020. doi: 10.1038/s41467-020-14645-x.

A Additional Experimental Details

A.1 Visual stimuli

Signal contrast and RMS noise contrast were systematically varied while the remaining Gabor parameters were held fixed. Decreasing signal contrast reduced the visibility of the Gabor pattern, whereas increasing external noise progressively obscured its orientation structure.

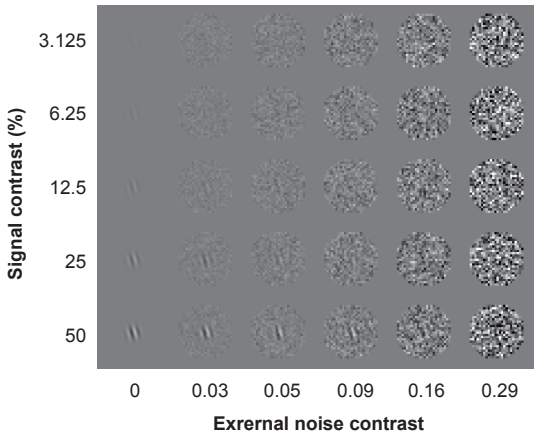

Figure A.1: **Representative stimuli from the sensory-perturbation dataset.** Example Gabor stimuli across signal-contrast and RMS noise levels. Columns indicate increasing signal contrast, and rows indicate increasing RMS noise contrast. Gabor orientation and other stimulus parameters are held fixed across examples.

A.2 Baseline orientation decoders

We implemented two commonly used neural decoding models as controls: a linear decoder and a nonlinear multilayer perceptron (MLP). Because orientation is a circular variable with 180° periodicity, both models predicted orientation using a two-dimensional circular representation rather than directly regressing angular values. Specifically, each orientation θ was encoded as

$$\mathbf{y} = \begin{bmatrix} \sin(2\theta) \\ \cos(2\theta) \end{bmatrix}. \tag{10}$$

The double-angle representation maps orientations separated by 180° to the same point while preserving their relative angular relationships. Each decoder produced two continuous outputs,

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_{\sin} \\ \hat{y}_{\cos} \end{bmatrix}, \tag{11}$$

which were normalized before loss computation:

$$\tilde{\mathbf{y}} = \frac{\hat{\mathbf{y}}}{\|\hat{\mathbf{y}}\|_2}. \tag{12}$$

The predicted orientation was recovered as

$$\hat{\theta} = \frac{1}{2} \operatorname{atan2}\left(\tilde{y}_{\sin}, \tilde{y}_{\cos}\right), \tag{13}$$

and converted to degrees within the range $[0^\circ, 180^\circ)$. Both models were optimized using a cosine-based circular loss,

$$\mathcal{L}_{\text{ori}} = 1 - \tilde{\mathbf{y}}^\top \mathbf{y}, \tag{14}$$

where $\tilde{\mathbf{y}}$ and $\mathbf{y}$ denote the normalized predicted and target orientation vectors, respectively.

Linear decoder. The linear decoder consisted of a single fully connected layer that directly mapped the population response vector $\mathbf{r} \in \mathbb{R}^N$ to the two-dimensional orientation representation:

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{r} + \mathbf{b}. \tag{15}$$

The model contained no hidden layers or nonlinear activations and therefore provided a baseline for orientation information recoverable through fixed linear weighting of population activity.

Multilayer perceptron. The nonlinear decoder consisted of a single hidden layer with 128 units. The hidden representation was computed as

$$\mathbf{h} = \phi\left(\mathrm{BN}\left(\mathbf{W}_1\mathbf{r} + \mathbf{b}_1\right)\right), \tag{16}$$

where BN denotes batch normalization and ϕ denotes the nonlinear activation function. The hidden representation was then mapped to the two-dimensional orientation space:

$$\hat{\mathbf{y}} = \mathbf{W}_2\mathbf{h} + \mathbf{b}_2. \tag{17}$$

Together, these two models allowed us to distinguish fixed linear pooling and generic nonlinear feed-forward computation from the attention-based population readout implemented by the transformer.

Optimization. Both control models were trained using AdamW with an initial learning rate of 10^{-3} and weight decay of 10^{-4}. The learning rate was reduced by a factor of 0.5 when the evaluation loss failed to improve for 100 epochs. Training was performed for a maximum of 5,000 epochs with early stopping after 1,000 consecutive epochs without improvement. The batch size was 16. Both models used the same circular orientation loss and optimization procedure.

A.3 Gabor orientation extraction

To quantify orientation from the reconstructed Gabor images, we estimated the dominant image orientation from spatial gradients. Let $I \in \mathbb{R}^{H\times W}$ denote a reconstructed Gabor image. Horizontal and vertical image gradients were first computed as

$$G_x = \frac{\partial I}{\partial x}, \qquad G_y = \frac{\partial I}{\partial y}. \tag{18}$$

The gradient components were vectorized and combined into a matrix

$$\mathbf{G} = \begin{bmatrix} G_x^{(1)} & G_y^{(1)} \\ G_x^{(2)} & G_y^{(2)} \\ \vdots & \vdots \\ G_x^{(P)} & G_y^{(P)} \end{bmatrix} \in \mathbb{R}^{P\times 2}, \tag{19}$$

where $P = H \times W$ is the total number of image pixels. Principal component analysis (PCA) was then applied to $\mathbf{G}$. Let

$$\mathbf{v}_1 = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \tag{20}$$

denote the first principal component, corresponding to the direction of maximum variance in the image-gradient space. The dominant direction was calculated as

$$\theta = \mathrm{atan2}(v_y, v_x). \tag{21}$$

The resulting angle was converted from radians to degrees,

$$\theta_{\mathrm{deg}} = \theta \frac{180}{\pi}, \tag{22}$$

and wrapped to the orientation range $[0°, 180°)$:

$$\theta_{\mathrm{ori}} = \mathrm{mod}(\theta_{\mathrm{deg}}, 180°). \tag{23}$$

The same procedure was applied to both target and reconstructed images, and orientation reconstruction error was calculated as the minimum circular difference between their extracted orientations.

B Additional Results

B.1 Orientation functional maps and neuronal population properties

Additional characterization of the neural recordings used in the main Gabor dataset is shown in Appendix Fig. B.1. Data were obtained from seven awake, fixating macaques, with one V1 FOV recorded from each animal, including previously unpublished data from Monkeys A–C and previously reported data from Monkeys D–G (Ju et al., 2021; 2022). Each FOV covered an area of $850 \times 850\ \mu\mathrm{m}^2$ at a cortical depth of approximately $150\ \mu\mathrm{m}$.

Appendix Fig. B.1A–C illustrates the recording site, average two-photon image, and mean calcium-response time course for Monkey A. Appendix Fig. B.1D shows exemplar orientation functional maps for Monkeys A–C. Appendix Fig. B.1E quantifies properties of the recorded populations, including distributions of orientation selectivity index (OSI), preferred orientation, and signal-to-noise ratio (SNR) across the seven FOVs. The OSI distribution was centered around moderate values (median = 0.284), preferred orientations were largely uniformly distributed across 0–180°, and SNRs were skewed toward higher values (median = 1.677).

B.2 Progressive reconstruction with attention-ranked neuronal signals

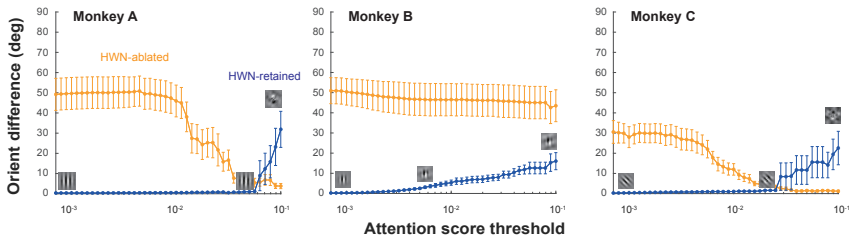

Figure B.2: **Mean differences between the reconstructed orientation and the true stimulus orientation in the attention-retained and attention-ablated conditions.** Orientation difference across attention-score thresholds for Monkeys A–C. Blue curves indicate reconstructions retaining highly weighted neurons above each threshold, whereas orange curves indicate reconstructions after ablating the same neurons. Example reconstructed stimuli at selected thresholds illustrate the corresponding changes in reconstruction quality.

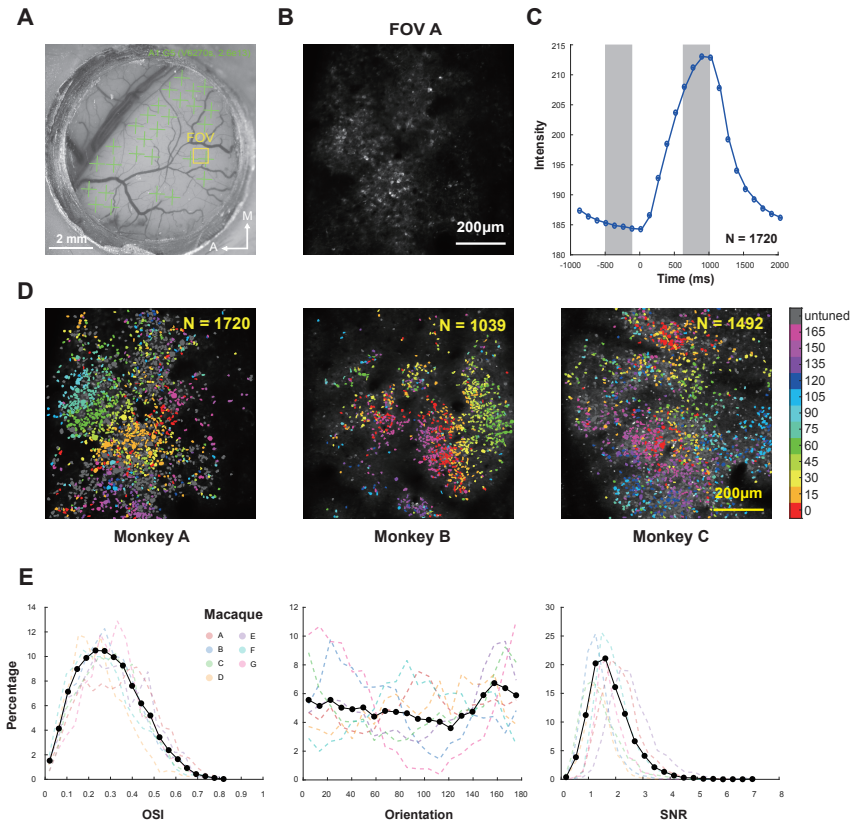

Figure B.1: **Two-photon calcium imaging and orientation functional maps. A.** Exemplar vascular map of Monkey A. Green "+" signs indicate the viral injection regions, and the yellow box indicates the 850×850 μm^2 FOV selected for imaging. A, anterior; M, medial; io, inferior occipital sulcus; lu, lunate sulcus. **B.** Average two-photon image of Monkey A over a recording session. **C.** Time course of calcium responses in Monkey A, averaged across neurons. Each point indicates the mean response intensity from one frame. Error bars represent ± 1 SEM. Shaded regions before and after stimulus onset indicate the frames used to calculate baseline (F_0) and evoked response (F), respectively. **D.** Cellular orientation functional maps imaged at a cortical depth of 150 μm from Monkeys A, B, and C. Each dot represents one neuron, and color indicates its preferred orientation. Grey indicates neurons without significant orientation tuning ($p > 0.05$). **E.** Distributions of neuronal OSI (left), preferred orientation (middle), and SNR (right). Each dashed line represents one FOV, and the black solid line indicates the pooled distribution across all FOVs.

B.3 FUNCTIONAL STABILITY OF HIGHLY WEIGHTED NEURONS ACROSS MODEL INITIALIZATIONS

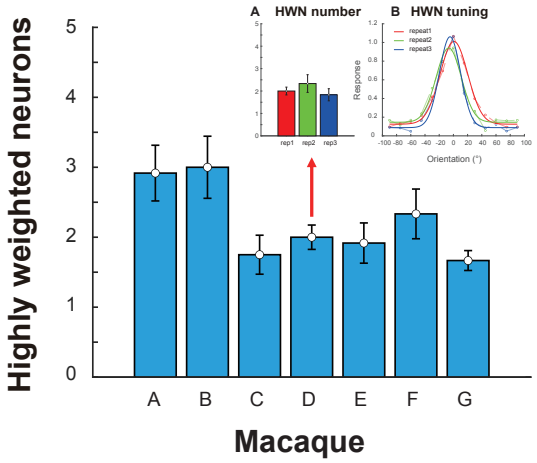

Figure B.3: **Stability of highly weighted neurons across model initializations.** The main panel shows the number of highly weighted neurons across macaques. **A.** Number of highly weighted neurons identified across three independent random-seed initializations in an example FOV from Monkey D. **B.** Orientation tuning curves of the corresponding highly weighted neurons across the three initializations. Although the exact neuron identities varied across runs, the number of highly weighted neurons and their stimulus-aligned tuning remained similar.