nature portfolio

Peer Review File

A Concentration of Visual Cortex-Like Neurons in Prefrontal Cortex

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

By using floating microelectrode arrays recorded from awake monkeys performed passive fixation task, Rose and Ponce found that some neurons in vlPFC have retinocentric visual receptive fields with image selectivity, which are similar to the properties of receptive fields in the ventral stream (e.g., area V4). They also showed that those neurons in vlPFC may be clustered.

The lateral prefrontal cortex (lPFC) is the converging stage of the dorsal (mainly project to the dorsal portion of lPFC) and ventral (mainly project to the ventral portion of lPFC) pathways. It is not well known how information (visuospatial and object recognition) computed by the two pathways are integrated into the lPFC. This manuscript makes some important advances in understanding the properties of vlPFC. Overall, the experiments are elegant and well performed, I outline my specific comments below.

1. The microelectrode arrays used in this manuscript have electrodes with different lengths of 1.6-2.4mm. Can one array record two layers simultaneously, or two arrays with different electrode lengths record from two layers separately? Either way, since lPFC has layers with different functions in processing information, the authors should consider which layer(s) the electrodes recorded from, and discuss the influence on the results.

2. In Panel e in Figure 1, the image showing 'Neuronal activity' is a little misleading because it does not represent the extracellular neuronal activity.

3. Line 128, 133,135, and 136, I believe the unit 'mm' are typos. Please verify.

4. In Panel b in Figure 2, the two images are not explicit. I think it is better to use a 4X9 grid to present the array geometry, and put the RF heatmaps (e.g., images in panel a) from all sites into the grid.

5. Panels c and d in Figure 2 are not cited in the main text.

6. In Panel c in Figure 2, why there is one V4 RF located in the ipsilateral visual field?

7. Pane d in Figure 2, V4 has a retinotopic map; the microelectrode array should only record neurons at a certain retinotopic location, e.g., about 10 degrees in this figure. But why there are some V4 RFs located at more than 20 degrees eccentricity?

8. Figure 3 is not well organized. There are three 'Monitor-centered coordinates', I am not sure how to associate them with the panels.

9. Panel c in Figure 5, maybe using transparent symbols to make the plot clear?

10. In Panel f in Figure 5, the symbols show the authors did statistical test, but not explicitly shown in the figure or in the legend.

11. The statistical letters, e.g. F and P, are not correctly formatted in the manuscript.

Reviewer #2 (Remarks to the Author):

The paper by Rose and Ponce examines whether visual responses in ventrolateral prefrontal (vlPFC) cortex of Rhesus monkeys have similar characteristics with those in visual areas. To this end, the authors performed extracellular recordings with chronic 32-channel arrays in vlPFC in two monkeys and addressed three main issues: a) spatial tuning during presentation of visual stimuli in a passive fixation task, b) selectivity for the identity of the visual stimulus across a variety of complex natural images and c) whether visual responses of vlPFC neurons can guide the synthesis of optimal visual stimuli through a closed loop adaptive image generator. In one monkey recordings from visual area V4 were also obtained, while in the other monkey recordings from auditory area CPB were obtained.

Although the paper is well written and it is generally relatively easy to follow the results and analyses, I am afraid I do not see a major conceptual advancement. First, previous studies have already provided evidence on how and what visual attributes are encoded in PFC (including vlPFC). In fact, some of these earlier studies examined a much larger sample than the one used in the study by Rose and Ponce, while also covering a more extended area within PFC. Examples of such studies include work from Constantinidis' lab assessing visual selectivity in untrained monkeys during passive fixation (e.g. Riley et al 2017) and several other labs including Suzuki and Azuma 1983, Viswanathan and Nieder 2017 etc (for a comprehensive review see Constantinidis and Qi, Front. Integr. Neurosci, 2018). Unfortunately, references to some of these studies are missing. It is not at all clear to me how the study by Rose and Ponce advances our knowledge relative to these earlier results. I am certain that the authors are aware of the earlier findings because in the discussion they acknowledge that: "While vlPFC neurons are known to show stimulus-position tuning suggestive of classic RFs14 and image selectivity11–13,15,35–38,44,45, including static13,35,37,38,44 and dynamic faces15,37 in the absence of task training, not all of these properties have been previously and thoroughly conducted in the same neuronal sites". I honestly do not see how looking at those properties on the same sites over days tells us something novel in the specific paradigm that the authors use.

Second, and in contrast to their latter statement above, I would argue that the authors have not used optimal methods to address the questions they are trying to answer. They employed 32

channel implanted arrays, which covered a small patch of cortex, recording from the same sites over different sessions. The problem with chronic arrays is that even if the signal changes from day to day it is impossible to be certain that different neurons are encountered on different experiments. Thus, the sample is largely the same over days and in this particular case it was 32 sites at best. This makes it extremely difficult to draw solid conclusions. In fact, several of the analyses carried out by the authors end up reporting proportions of neurons in a sample of less than 10 sites (e.g. lines 193 and 196, 367-368) and in one case data from only one monkey (e.g. Figure 3d where no active vlPFC sites are reported for Monkey C). In fact, I would argue that several of the analyses are inconclusive because of this limitation with the most prominent example that on retinocentric vs allocentric coding and gain fields.

Besides these general comments, I also found that in certain cases the clarity of presentation and the statistics used should be improved. I list some of these below:

1) To assess visual responsiveness the authors compare responses in a 1-30ms post stimulus presentation to a window 50-150/200ms. It is odd that as a baseline they did not use a window during fixation (pre stimulus presentation). This would allow for more robust data comparing roughly similar sizes window around 100-150ms width. Moreover, a paired t-test should be used for these comparisons (unpaired t-test is reported).

2) When assessing polysensory responses, the number of sites with significant responses for each modality should be clearly stated in the text.

3) In the gain field analysis, a one-way ANOVA is mentioned in the methods, but it is not clear which factor was used. How were gaze position and stimulus position handled?

4) In the cluster analysis, what does the conclusion tell us? Is it possible that the array was at the border between two different PFC regions? Did the authors find any evidence for clusters of spatially selective cells interspersed among clusters of non-spatially selective zones?

5) In the Experimental setup paragraph in the methods, it is mentioned that for most experiments animals were performing a passive fixations task. I am probably missing something because I thought in all experiments animals were performing a passive fixation task. Could you clarify that?

6) A reference to figure 1d, 2c and 5c is missing in the text

7) On lines 211-214 it is not clear what is being compared to what.

Reviewer #3 (Remarks to the Author):

Ventrolateral prefrontal cortex (VLPFC) is thought to be at the apex of the visual streams that converge in frontal cortex. This manuscript details a number of really quite interesting experiments that Rose and Ponce conducted that initially take a quite standard approach to characterizing the visual receptive field (RF) properties of VLPFC neurons and comparing these to V4 and central auditory parabelt neurons. Further, they investigate whether these RFs are retinotopically organized before conducting a really interesting analysis using cutting-edge machine learning driven prototype matching. This final piece of the manuscript is really quite exciting, and the results are rather unexpected; namely that image prototypes that maximally drive responding in VLPFC are often quite amorphous consisting of quite simple visual features.

This is a really quite an interesting manuscript, and the results should be of interest to those working in on the neurophysiology of frontal cortex as well as visual processing. The first part of the manuscript mapping RFs in VLPFC is very standard and robust. This is a good thing and is the foundation for the rest of the manuscript. The second half is more interesting as it is looking for what specific features drive VLPFC neuron responses. Some would have predicted that the prototypes that maximally drive VLPFC responses would have been more detailed.

While this is really quite interesting the biggest weakness of the study is the relatively coarse neurophysiology data that the authors have managed to obtain from their floating micro-arrays. On the one hand this could be seen as not so much of a problem as the main message of the paper relates to the visual tuning properties. On the other hand, the retinotopic mapping and ML-driven prototyping analysis is slightly diminished if the effects are being driven by multi-unit activity (at least I think that is what is happening in one of the monkeys) as opposed to single neurons. The point being that the conglomerated activity of many neurons will likely lead to more general/amorphous prototypes/RFs if individual neurons have very specific tuning profiles. I don't think that this necessarily means that the manuscript isn't sufficiently important, I'm just going to need a little convincing about whether the effects can really be seen at the single neuron level. This main concern as well as a number of others fleshed out in more detail below:

1) As highlighted above I'm interested to know more about the distinction between true single neuron coding and coding by multi-unit/hash activity in the data. The first thing that the authors should do is provide a table accounting for how many single neurons and multi-unit activity they have recorded/analyzed. They allude to there being little difference between these measures, but I'd like to know what the numbers for each are, as this helps a reader to understand the robustness of the effects reported. Because the arrays are also fixed in place and thus less likely to move, it makes sense to also highlight if any of the single neurons recorded across days are thought to be the same (or not).

2) Following on from the above, the most interesting part of the manuscript for me are the experiments focusing on determining prototypes for neurons in VLPFC. However, I found it hard to know how consistent the results were for single neurons and multi-unit activity. The authors do, to

their credit, highlight where there are differences between single units/hash/multi-units, but a little clarity/detail here would be good as it potentially alters the interpretation of the findings (see point above). One way to address this would be to include examples of prototypes from single units vs multiunits and if possible compute the degree of visual complexity for each and compare etc to see if there are differences. Again if there are differences it could indicate that single neurons have more visual like receptive fields whereas if there are not then I think it supports the authors current conclusions about VLPFC visual prototypes.

3) I'm also interested to know more about the methods for characterizing prototypes for VLPFC. The authors state that they went through 10 block iterations of images for each site/neuron/activity using the XDream approach. I wondered how this was arrived at and whether this parameter was based on determining prototypes in visual cortex where activity is more stereotyped/consistent to visual stimuli. As an aside the 100ms image presentation seemed quite short given that it often takes ~100ms for PFC neurons to respond to visual images. My thought here is that prototypes in VLPFC may be more variable based on frontal cortex in general being more multi-sensory/exhibiting mixed selectivity than visual areas. I think that this is something that the authors should explore to highlight the differences between VLPFC and V4 neurons/sites. So, as the experiments cannot be performed again, how stochastic are the responses in VLPFC or V4 across the 10 block/iterations of the XDream procedure. I realize that the responses are changing as the images are updated, but I think that there could be interesting differences in how spiking activity converges on a stable response to the prototypes across the two areas across the blocks and that this should be reported here. Of course, if the prototypes converge at the same rate in V4/VLPFC then that would be an interesting finding in itself.

4) The point that the authors make in the discussion that VLPFC neurons can drive adaptive image generators like XDream should not be understated. It also made me wonder if the authors might want to connect to the literature that has emphasized the attentional functions of VLPFC (see Rushworth et al., J Neuroscience, 2005 and related articles).

Minor

The references to figures in the text are a little scarce throughout the manuscript. For instance, in the section describing the positional tuning of RF's in VLPFC, there were no call outs to figures that illustrated the effects being described. Consider including these.

Could the authors provide a little more information about the exact location of the arrays in VLPFC? There are a number of cytoarchitectonic areas that have been identified in this location and it would be good to provide detail on which the authors think the arrays are in.

We are pleased to continue this dialogue with our Reviewers. We have spent a considerable amount of time addressing their concerns, and we thank them for their insight and attention to detail. They made us think deeply about the project in general, and they each led to changes that we hope will satisfy our readers and provide an accurate impression of our contribution in visual neuroscience.

In this document, the original reviewer comments will be in Verdana, dark red font. Our responses will be in Arial, black. All changes to the manuscript and figures will be listed here for the convenience of the reviewers, and changed regions in the main manuscript will be in Arial, blue font.

REVIEWER COMMENTS

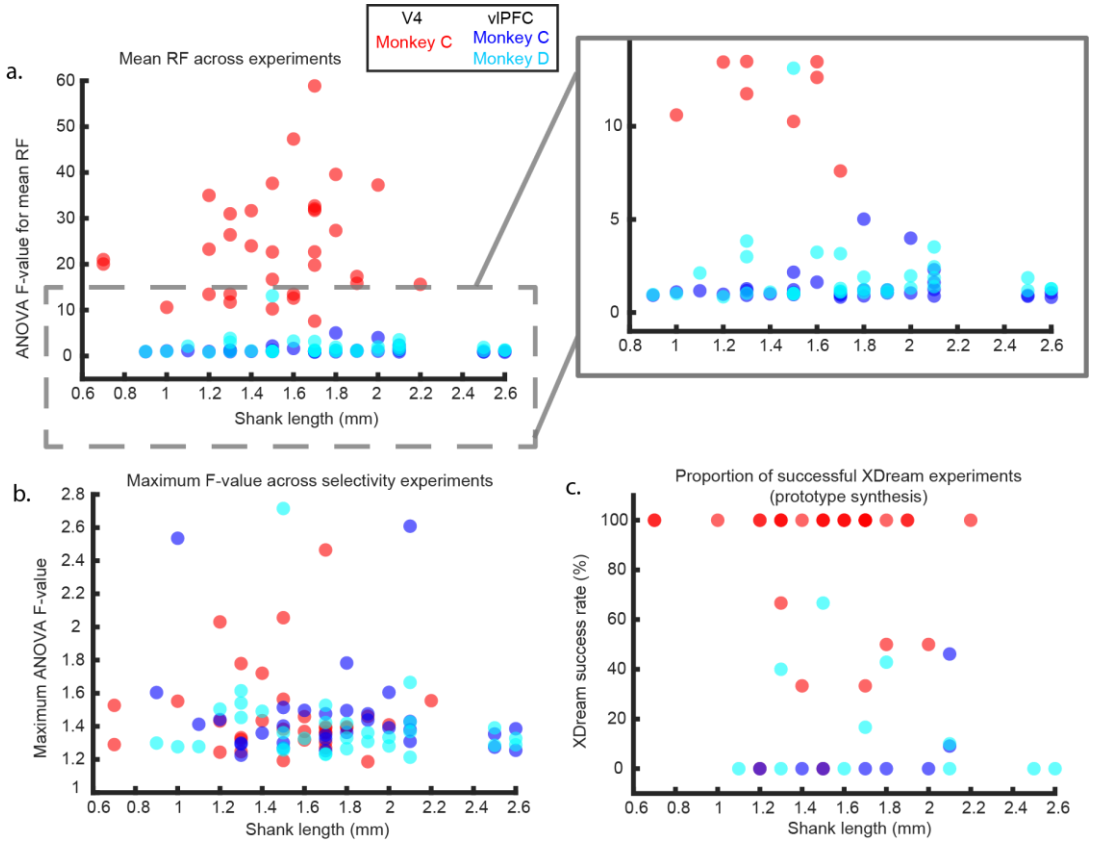
Reviewer #1 (Remarks to the Author):

By using floating microelectrode arrays recorded from awake monkeys performed passive fixation task, Rose and Ponce found that some neurons in vlPFC have retinocentric visual receptive fields with image selectivity, which are similar to the properties of receptive fields in the ventral stream (e.g., area V4). They also showed that those neurons in vlPFC may be clustered.
The lateral prefrontal cortex (lPFC) is the converging stage of the dorsal (mainly project to the dorsal portion of lPFC) and ventral (mainly project to the ventral portion of lPFC) pathways. It is not well known how information (visuospatial and object recognition) computed by the two pathways are integrated into the lPFC. This manuscript makes some important advances in understanding the properties of vlPFC. Overall, the experiments are elegant and well performed, I outline my specific comments below.

1. The microelectrode arrays used in this manuscript have electrodes with different lengths of 1.6-2.4mm. Can one array record two layers simultaneously, or two arrays with different electrode lengths record from two layers separately? Either way, since lPFC has layers with different functions in processing information, the authors should consider which layer(s) the electrodes recorded from, and discuss the influence on the results.

Thank you for raising this question. We agree that in principle, this could reveal fundamental differences between the neuronal computations in the different cortical layers of vlPFC. Laminar anatomical organization likely serves a functional purpose, but despite some efforts to link cortical layers to specific computations (e.g., Chen et al., 2007; Ziemba et al., 2019), the relationship remains largely unexplored.

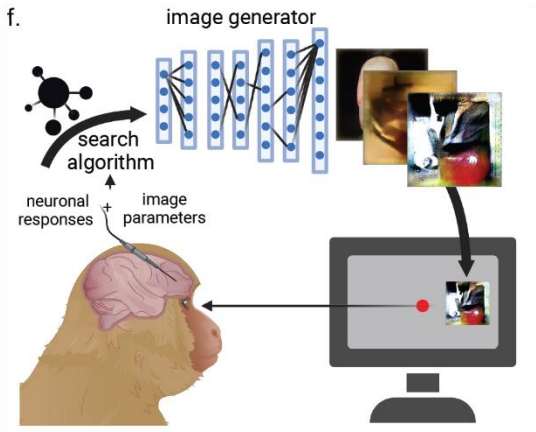
Generally, we design our arrays to have wires with multiple lengths as a precautionary measure— if the arrays do not settle well in cortex, or if alternatively, they dimple the cortical surface, then either the long or short electrodes have the best chance to land in the cortical band. In the past, we have not seen systematic relationships between electrode length and functional features, but inspired by this suggestion, we carried out new analyses to investigate whether electrode depth, serving as a proxy for cortical depth, was correlated with any of our major findings. Due to the “floating” design of the arrays, the electrode shanks have a vertical degree of freedom within the cortical tissue; as such, any given electrode could sample different cortical layers across days. With this caveat in mind, however, we plotted the length of each electrode with respect to 1) *sensitivity to stimulus position* (the ANOVA *F*-value shown by each given site when calculating its RF, 2) the *sensitivity to image identity* (the ANOVA *F*-value shown by each given site when computing its image selectivity, and 3) the *likelihood of successful closed-loop image synthesis experiments* (or *evolutions)*. These relationships are plotted in **Response Figure 1.** While there were some intriguing trends by visual examination, such as stronger RF reliability in V4 as a function of electrode length, statistical testing was not robust pertaining to these trends. We fit linear regression models to these trends and found that for V4, the slope was 6.3 ± 6.29 (p = 0.32, t-Stat = 1.0, N = 32 observations), for vlPFC, Monkey C, 0.14 ± 0.37 (p = 0.71, t-Stat = 0.37, N = 32), and for vlPFC Monkey D, -0.41 ± 0.89, t-Stat = -0.46, N = 32). This was true for our other analyses. Overall, our dataset does not proffer a strong relationship between electrode lengths as a proxy for cortical depth and/or a relationship between depth and our major findings, unfortunately. However, we have recently obtained many high-density recording probes (Neuropixels), and in principle, these should allow us to ask these questions more accurately.



Response Figure 1. Relationship of receptive-field (a), image selectivity (b), and evolutionary (c) experiments to electrode length.

2. In Panel e in Figure 1, the image showing 'Neuronal activity' is a little misleading because it does not represent the extracellular neuronal activity.

Yes, we agree that the layout of this panel was suboptimal. Figure 1e had initially intended to depict a schematic of extracellular spikes (e.g., rasters) to signal the neuronal activity, rather than a peristimulus time histogram (PSTH), but we agree that the sample trace used was ambiguous. We have replaced this panel in Figure 1 with the goal of improving clarity: we replaced the electrophysiology schematic with the words "neuronal responses," as well as updated the corresponding icons used. Hopefully, this shows that the search algorithm uses electrophysiology-based responses and image parameters towards the overall optimization goal.

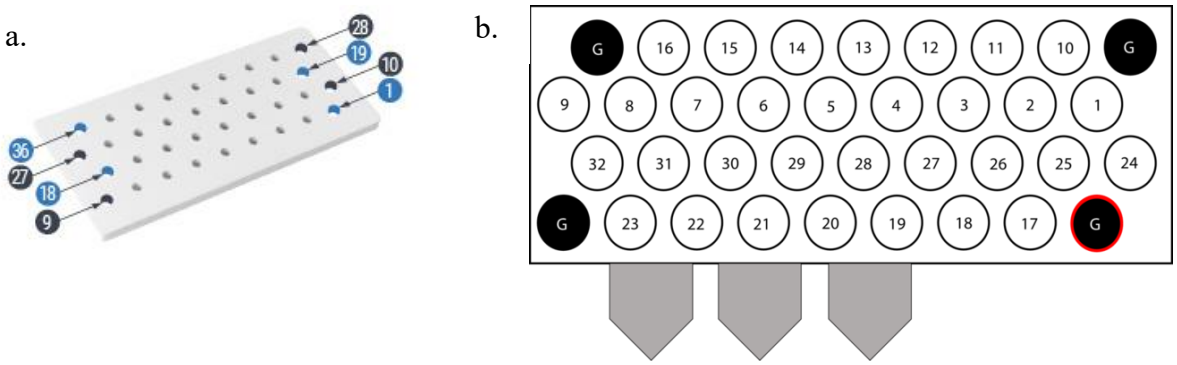


3. Line 128, 133,135, and 136, I believe the unit 'mm' are typos. Please verify.

Thank you for catching this— we discovered that that Greek letters automatically convert to the Latin alphabet in Microsoft Word when the font style is changed. We have updated these lines to now state "μm" and, from this point on, will specifically check for this in our proofreading.

4. In Panel b in Figure 2, the two images are not explicit. I think it is better to use a 4X9 grid to present the array geometry, and put the RF heatmaps (e.g., images in panel a) from all sites into the grid.

Thank you for pointing this out — we see this as an opportunity to clarify the geometric layout of the Microprobes floating microelectrode arrays used for data acquisition. The shanks in these FMAs are arranged in offset rows—more honeycomb-like, rather than a perfectly rectangular grid (**Response Figure 2a**). These FMAs have 36 wires total: 32 recording channels and four ground/reference channels (one per corner in the FMA; **Response Figure 2b**). The locations of the ground channels thus give the recording channels an asymmetric hexagonal appearance, but one in which all channels are spaced equidistantly. To accurately depict the relative distances between each cortical site and the functional properties, we preserved the original geometric configuration of the array.



Response Figure 2. *(a)* Schematic of Microprobes 36-channel chronically implanted Floating Microelectrode Array (FMA), reproduced with permission from Microprobes for Life Science. *(b)* Schematic demonstrating the locations of the four ground electrodes (black circles marked with "G"; red outline on bottom right ground electrode denotes this channel functions as both ground and reference) with respect to the 32 recording electrodes (numbers correspond to the *channel* number for each of the 32 recording channels; as a result of mapping Microprobes hardware to Plexon software, final channel numbers do not maintain their original sequential ordering in the array).

To clarify the arrangement of the RF heatmaps, we added the following line to the legend of Figure 2:

The arrangement of the RF heatmaps reflects the electrode geometry of the Microprobes Inc.'s floating microelectrode array (FMA).

5. Panels c and d in Figure 2 are not cited in the main text.

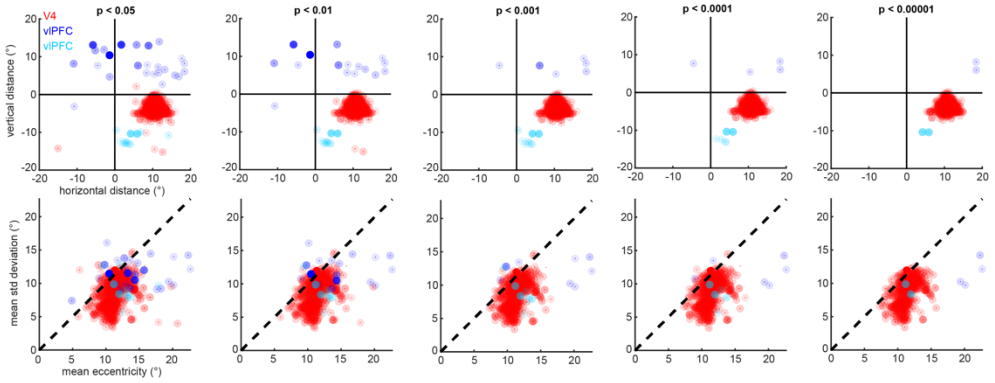
Thank you, we have now ensured those panels are referenced in the main text:

These RFs mostly represented the contralateral visual hemifield to the implanted hemisphere, though a small fraction spanned the midline between contra- and ipsilateral hemifields (**Figure 2c**). Compared to V4 RFs, vlPFC RFs tended to cover more of the visual field (i.e., were larger in estimated width; see **Figure 2d**), though in our samples, vlPFC RFs were also more eccentric than the V4 RFs (**Supplemental Table 1**).

6. In Panel c in Figure 2, why there is one V4 RF located in the ipsilateral visual field?
7. Pane d in Figure 2, V4 has a retinotopic map; the microelectrode array should only record neurons at a certain retinotopic location, e.g., about 10 degrees in this figure. But why there are some V4 RFs located at more than 20 degrees eccentricity?

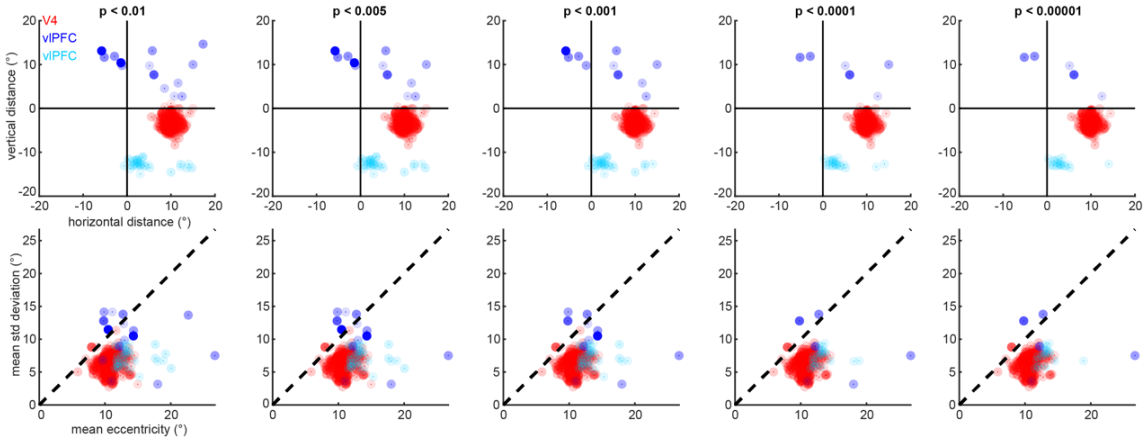
Thank you for these observations. We determined that our initial submission of Figure 2 included an early version of panels *c* and *d*, before we implemented the inclusion criteria described in the **Methods**. In the version originally provided, we included estimated RFs from experiments using all tested stimulus sizes, ranging from 1° to 10° in width. These were early experiments, where too-small or too-large image sizes

presented at large eccentricities can result in false negatives or false positives. We found that the most robust and specific stimulus size was 5°, so we updated our criteria to include only experiments using 5°-width stimuli. We thought this would be a good opportunity to illustrate how the inclusion threshold changes the figure itself, and how some false positives can survive the false-discovery rate correction. When probed further, the seemingly ipsilateral V4 RF failed to reach significance at a threshold of $p < 0.01$; likewise, the V4 RFs with apparently 20° eccentricity failed to reach significance at a threshold of $p < 0.001$ (**Response Figure 3**).



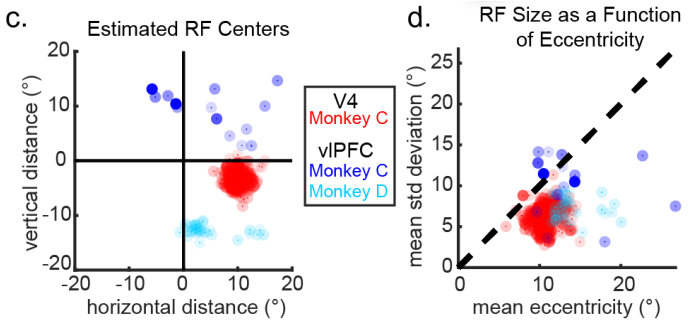
Response Figure 3. Receptive field per different threshold values. This shows the location of putative RFs as a function of *p*-value thresholds (obtained via a one-way ANOVA, with position as the sole factor). Stimulus probe was 10° in width.

We do not have much confidence in the reliability of those outlier data points, though it is essential to accurately present all data that pass significance-threshold testing. In contrast, below we have plotted the locations of estimated RF centers and estimated RF width as a function of eccentricity, for varying cutoffs of statistical significance, although in this analysis restricting to only experiments that mapped RFs with a 5°-wide image (**Response Fig. 4**).



Response Figure 4. Receptive field per different threshold values. This shows the location of putative RFs as a function of *p*-value thresholds (obtained via a one-way ANOVA, with position as the sole factor). Stimulus probe was 5° in width.

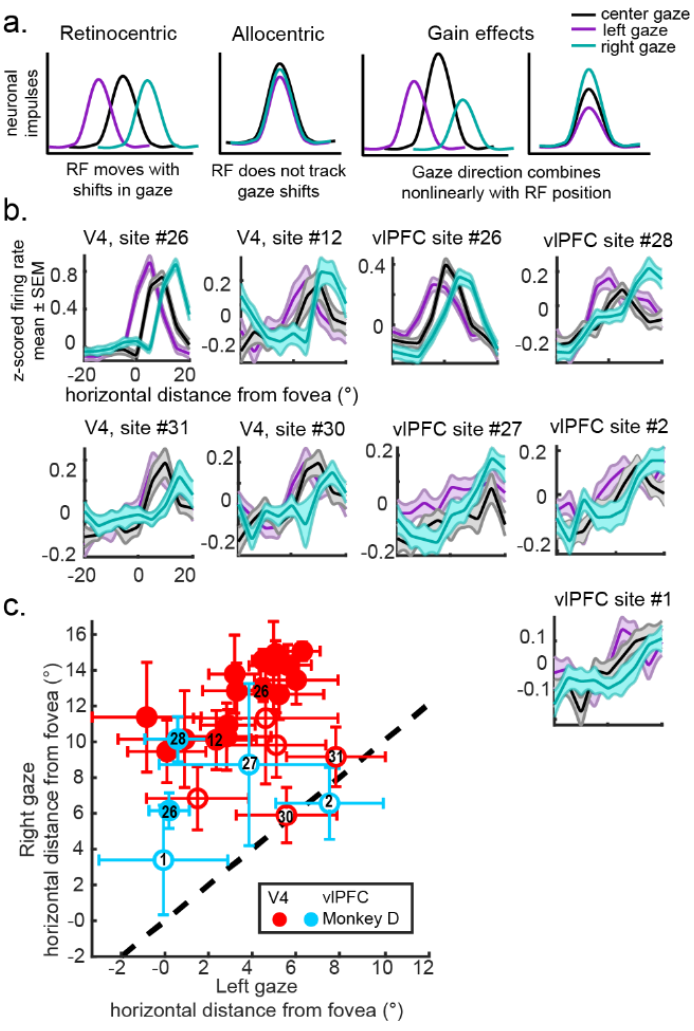
In response to this observation, we have now corrected the main manuscript figure to results from experiments using 5°-wide stimuli ($p < 0.01$ after FDR-correction); additionally, we confirmed that all population statistics and the information in Figure 2a-b we initially provided correctly only reported data from RF experiments using a 5° stimulus. We have also edited **Supplemental Figure 1** to show these two figures with varying significance thresholds (ranging from $p < 0.01$ to $p < 0.00001$ after FDR-correction).



8. Figure 3 is not well organized. There are three 'Monitor-centered coordinates,' I am not sure how to associate them with the panels.

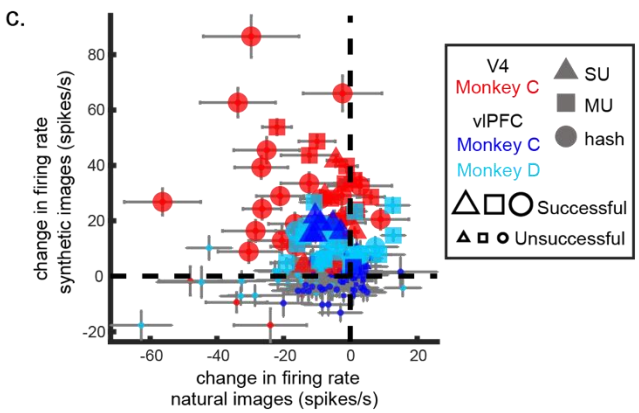
Yes, we struggled trying to find the best way to structure Figure 3. We appreciate the chance to iterate it further. As shown below, we removed the repeated phrase "monitor-centered coordinates" in all three subpanels, then added more precise descriptions of the putative RF schemes in Figure 3a. For example, a "retinocentric" RF should seemingly move horizontally across the monitor with lateral shifts in gaze, as a retinocentric RF encodes the *relative* position on the retina; by contrast, a putative "allocentric" RF would encode the same *absolute* position on the monitor, regardless of whether that real-world location stimulates three different positions on the retina as the monkey shifts his eyes. In either possibility, preferred gaze directions could also produce a gain effect, in which RFs yielded by certain gaze conditions are modulated in amplitude. We hope these modifications have improved the clarity of this result.

Updated Figure 3 is shown below:



9. Panel c in Figure 5, maybe using transparent symbols to make the plot clear?

Thank you very much for this suggestion— the updated figure looks excellent to us. We have adjusted Figure 5c to now use transparent symbols and improved the overall readability of the panel. We also appreciate this prompt, as we also found a minor indexing issue that we corrected.



10. In Panel f in Figure 5, the symbols show the authors did statistical test, but not explicitly shown in the figure or in the legend.

We have now added this to the legend:

Triple asterisks denote statistical significance at $p < 0.001$ by way of a randomization test.

11. The statistical letters, e.g. F and P, are not correctly formatted in the manuscript.

Thank you for highlighting these oversights. We have confirmed APA guidelines for reporting statistics and have now standardized the format throughout the manuscript.

Reviewer #2 (Remarks to the Author):

The paper by Rose and Ponce examines whether visual responses in ventrolateral prefrontal (vlPFC) cortex of Rhesus monkeys have similar characteristics with those in visual areas. To this end, the authors performed extracellular recordings with chronic 32-channel arrays in vlPFC in two monkeys and addressed three main issues: a) spatial tuning during presentation of visual stimuli in a passive fixation task, b) selectivity for the identity of the visual stimulus across a variety of complex natural images and c) whether visual responses of vlPFC neurons can guide the synthesis of optimal visual stimuli through a closed loop adaptive image generator. In one monkey recordings from visual area V4 were also obtained, while in the other monkey recordings from auditory area CPB were obtained.

Although the paper is well written and it is generally relatively easy to follow the results and analyses, I am afraid I do not see a major conceptual advancement. First, previous studies have already provided evidence on how and what visual attributes are encoded in PFC (including vlPFC). In fact, some of these earlier studies examined a much larger sample than the one used in the study by Rose and Ponce, while also covering a more extended area within PFC. Examples of such studies include work from Constantinidis' lab assessing visual selectivity in untrained monkeys during passive fixation (e.g. Riley et al 2017) and several other labs including Suzuki and Azuma 1983, Viswanathan and Nieder 2017 etc (for a comprehensive review see Constantinidis and Qi, Front. Integr. Neurosci, 2018). Unfortunately, references to some of these studies are missing. It is not at all clear to me how the study by Rose and Ponce advances our knowledge relative to these earlier results. I am certain that the authors are aware of the earlier findings because in the discussion they acknowledge that: "While vlPFC neurons are known to show stimulus-position tuning suggestive of classic RFs14 and image selectivity11–13,15,35–38,44,45, including static13,35,37,38,44 and dynamic faces15,37 in the absence of task training, not all of these properties have been previously and thoroughly conducted in the same neuronal sites". I honestly do not see how looking at those properties on the same sites over days tells us something novel in the specific paradigm that the authors use.

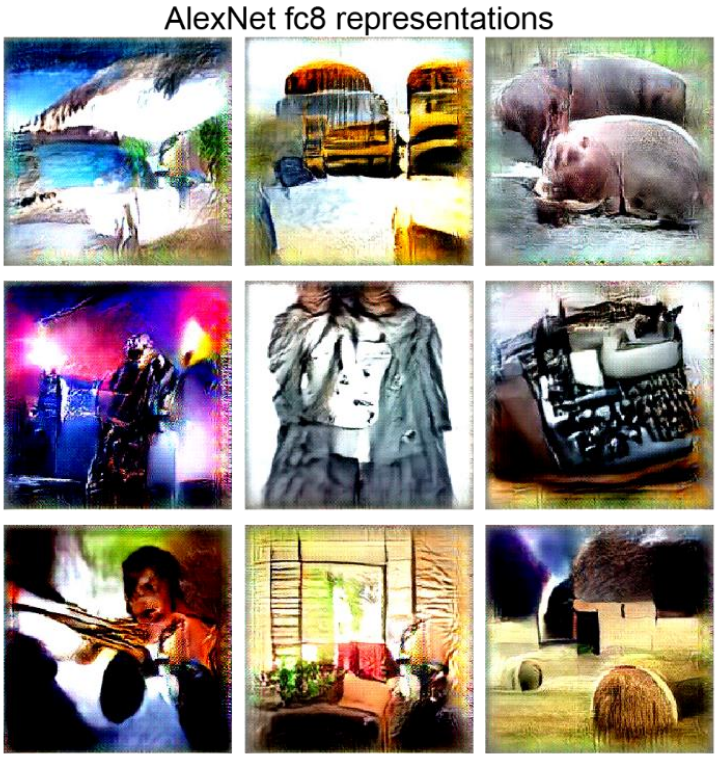
We appreciate these views, and the opportunity to clarify how our manuscript advances an understanding of vision beyond visual cortex. We will address the major conceptual advances in two parts.

Conceptual advances, part I. How similar is the primate brain to modern artificial neural networks?

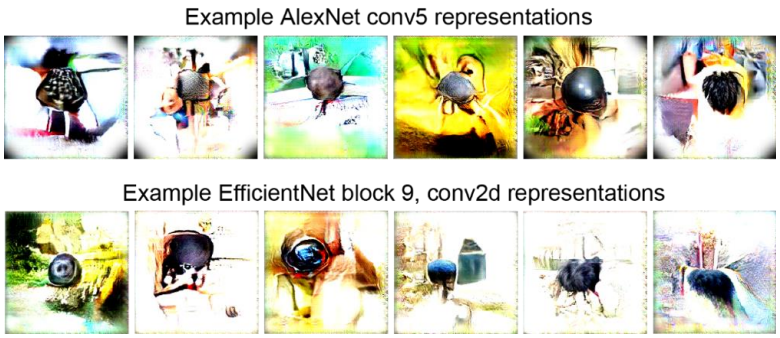
Our motivation for studying vlPFC was related to a specific and significant prediction about the neural computations subserving visual object recognition. For several decades, a majority of investigators have believed that inferotemporal cortex (ITC) mostly contains neurons that encode objects categories, such as "faces" or "bodies." Recently, studies in ITC are raising the view that this is not entirely true. For example, recent work from our laboratory[1,2] and the teams of investigators such as Pinglei Bao (SfN 2023, Washington DC), Talia Konkle[3], Gabriel Kreiman[4] and Margaret Livingstone[5], there is an emerging view that ITC neurons behave more like convolutional filters[6] selective for primitive generic features; these features occur across many types of object categories, but share common visual attributes that may transcend semantic description. In this view, object representation is a *distributed property* across neuronal populations, not one necessarily present at the level of an individual neuron. Even electrophysiology studies of the human brain show that, despite interesting reports of "Jennifer Aniston" neurons, most human temporal-lobe neurons appear to encode specific lower-level visual features, rather than invariant concepts or categories[7]. As this is how convolutional filters in artificial neural networks (ANN) work, these results strengthen the idea that the visual system is comparable to such kinds of artificial vision models[8]. But how robust is this analogy?

If we grant that ANNs are good models of the visual brain, the first two questions should be *which kind of ANN* and *which parts of it*? Modern ANNs are trained explicitly to represent semantic categories (using categories and images from WordNet and ImageNet). If one embraces the hypothesis that supervised-training ANNs are models of the primate visual system, then it follows that the visual system might terminate in the same kinds of representations. They should act as fully connected hidden units lacking proper receptive fields which allows them to assemble local primitive features (from the convolutional layer activations) into incredible scenes like this (**Response Fig. 5**):

Response Figure 5. Visual encoding of fully connected hidden units in a neural network. The output units are trained to classify images according to semantic categories, in these examples, the categories of *seashore*, *school bus*, *hippopotamus*, *stage*, *trench coat*, *typewriter keyboard*, *violin*, *window shade*, and *hay*.



This type of image is a prediction. If conventional understanding is correct — if PFC neurons encode objects like the output unit of a CNN, then in our experiments, visual neurons in PFC should lead to similar representations. Previous studies would predict this: O´Scalaidhe, Wilson, and Goldman-Rakic (1997) suggested that neurons were tuned to faces and not to face sub-features. While we really liked that study, we were not satisfied by this interpretation, as their method of object scrambling can destroy key internal features or crowd them out. In fact, most studies testing for selectivity did not thoroughly test for the possibility that lower-level features could be the central unit of encoding in vlPFC neurons. Our study tests for this in an innovative way, and our results show that intrinsically visual vlPFC neurons function more like V4 and ITC neurons, encoding lower-level "critical features"[9] not encoding conjunctions of lower-level features as in fully connected units of supervised networks. We have now examined thousands of filters in different convolutional neural networks, and found that the type of PFC neurons' encoded prototypes appear systematically in the *convolutional* layers of multiple neural networks (**Response Fig. 6**):



Response Figure 6. Visual encoding of convolutional hidden units (AlexNet layer conv5, EfficientNet block 9, conv2), reminiscent of the discovered PFC neuronal evolutions. Intermediate convolutional units learn simpler patterns commonly present in their training set (their "visual diet"), which can be used to represent multiple object categories. When patterns are useful, they are developed multiple times in a given layer (akin to a cortical region).

Our study suggests that monkey visual neurons are more *convolutional* than *fully connected,* and thus less *categorical* or *semantic* than previously assumed. Perhaps this is true everywhere in the brain, and categories or semantics are a much more distributed cognitive phenomenon. We find this exciting, and if one adheres to the view that vlPFC neurons encodes objects, our results should be surprising, perhaps contentious, but not a simple rehash of established understanding. In short, one key advance from this project is that (1) many visual vlPFC neurons are more like V4 and ITC neurons, (2) that they are more comparable to convolutional vs. fully connected units, and that (3) object representation may be a distributed process, not one that reduces to individual neuronal activity.

In response to our Reviewer's observations, we have decided to redraft the manuscript to emphasize the overall significance of our results as they apply not just to vlPFC, but to this view of vision models in general. The relevant changes are as follows:

(1) Introduction. The revised introduction contextualizes this larger debate about vision and deep networks:

> Visual processing in the primate cortex is classically understood through the dual stream model[10], which introduced the ventral and dorsal streams for shape recognition and visuospatial computations. The streams begin in primary visual cortex (V1) and diverge after visual area V2. The ventral stream comprises V2, V4, and inferotemporal cortex (IT), and it contains a high proportion of neurons that respond strongly to specific shapes (*selectivity*) at preferred retinal positions (*receptive field*, RF). Overall, ventral visual neurons appear to function as filters that fire maximally when their encoded pattern occurs within an input image[6], so they are frequently modeled as kernels in artificial neural networks (ANNs), such as convolutional networks[8,11,12]. While there is excitement in using ANNs as models of the visual system, how far can we carry this analogy? One way to answer

this question is to investigate the state of visual information beyond largely sensory areas. IT projects to ventral portions of lateral prefrontal cortex (vlPFC)[13,14]. Unlike V1, V4, and IT, vlPFC contains more functionally heterogeneous populations, some showing visual tuning[15,16], but with most representing behavioral task variables testing cognition and action processes such as decision-making. The most common ANNs are trained for classification, where "decisions" are made by an output layer, usually a fully connected or global-pooling stage that combines the activity of units with spatially limited RFs and produces activations that undergo a softmax operation[17]. These units have access to all regions of the visual input, and they encode complex conjunctions of lower-level features resembling *objects* and *scenes*. This can be demonstrated using modern feature visualization techniques, which synthesize images from noise and are generally better tests of tuning because they avoid image pre-selection[18] (**Figure 1**). So, is this decision stage in ANNs comparable to what happens in primate cortex — e.g., do visual vlPFC neurons encode full objects and scenes, like fully connected units in an ANN, or do they encode simpler features like convolutional units?

(2) Figure 1. The added text is reflected in an updated Figure 1, which now sets up our central motivating question in new subpanel (a).

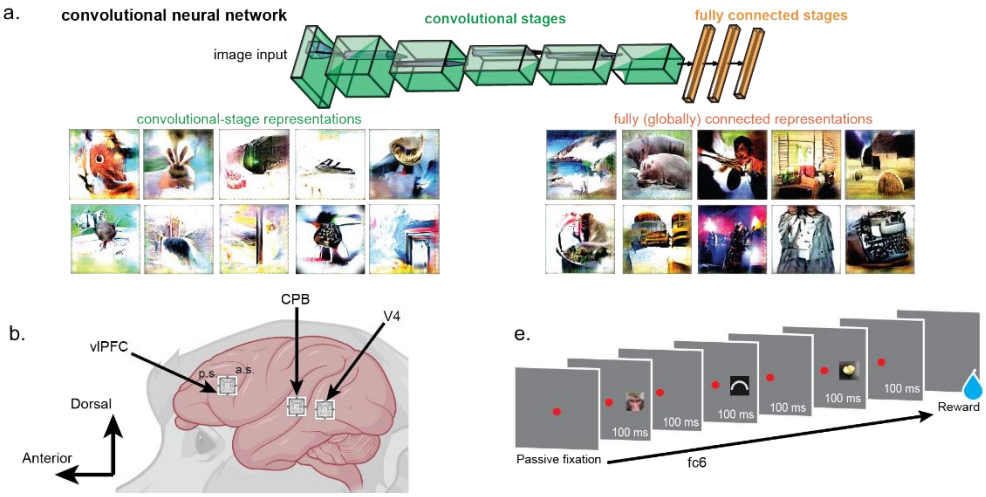


Figure 1. Experimental designs. *(a)* Typical convolutional neural network architecture with convolutional layers and fully connected layers, with the last used for classification. Feature visualization shows low-level pattern encoding in convolutional layers and more object- and scene-like encoding in connected layers.

<u>Conceptual advances, part II. Does PFC have "real" RFs?</u>

Next, we agree with our Reviewer that there were key references that should be prominently featured in our manuscript and including them gives us the opportunity to explain how we build on them. The Constantinidis lab has convincingly demonstrated that vlPFC neurons have visual tuning properties to simple objects, such as squares and triangles (Riley *et al.,* 2017), theirs and other labs have also shown that RFs can be found in PFC (Mikami, Ito and Kubota, 1982, Suzuki and Azuma, 1983; Viswanathan and Nieder, 2017a and 2017b). Our work builds on these findings because we had to establish why and how the image-synthesis approach could work in vlPFC (again, this approach tells us not only that a neuron is selective and has an RF, but also whether it encodes for objects or more primitive sub-features). None of the previous studies analyzed the same neuronal populations for <u>all of the tests</u> we performed: for example, (i) the Riley *et al.* paper used simple geometric shapes, but rough and sparsely sampled positions for RF mapping, and they did not try to find optimal representations (as written, "we did not exhaustively probe neurons for their preferred stimulus object"), (ii) Suzuki and Azuma did more comprehensive spatial RF sampling, but only with dots, (iii) none performed controls for eye position and auditory responses at the same time. So, while we agree with our Reviewer that our work has some overlap with those studies (and we will be careful not to claim innovation in showing that vlPFC neurons have RFs and image selectivity), our work builds on these findings by thoroughly profiling these functions in the same populations while testing the representational content of these neurons.

One more personal motivation for the battery of tests above is that, in early scientific discussions of the visual tuning in vlPFC (*e.g.,* in talks and posters), we were surprised to find unexpected resistance from vision neuroscientists that spent their years studying neurons in V1, V2, or MT. They expressed skepticism that vlPFC neurons had "real receptive fields" (phrasing is *ad verbatim*), similar to those in the visual cortex

electrophysiology literature. We realized that while early studies had answered this question to our satisfaction (and perhaps our Reviewer's), there remained too much doubt from the general vision community that we truly knew what attributes are encoded by PFC neurons or what mechanisms they possessed. This is the reason we had to go back and perform additional experiments, checking for object selectivity *and* testing for RFs *and* multimodal tuning, and then, after the critiques above, we added additional experiments controlling for motor and attentional tuning (this is the reason array signal degradation is mentioned in the Methods, see *Coordinate-transformation experiments*). Our study added many concurrent controls and also made the first application of a closed-loop paradigm in vlPFC known to work well with typically visual neurons. With our results, we believe that we can finally ensure that a wider acceptance and conceptual advance of visual attributes encoded in vlPFC can finally be achieved, at least in the minds of the general vision community.

In response to this comment, we made further modifications to the Introduction, explicitly mentioning the previous work above, and explaining how we worked to build on it.

Introduction. The following text was added:

> So, is this decision stage in ANNs comparable to what happens in primate cortex — *e.g.,* can visual vlPFC neurons lack RFs while still encoding for objects and scenes?
>
> This question has not been answered yet. In many studies, vlPFC neurons were reported to show visual tuning to objects in behavioral tasks, but the majority of these studies required months-long training, so many of these visual tuning properties could have been induced experimentally. Other studies have tested PFC neurons during passive viewing (without cognitive task training[15,16,19]): some showed that many neurons have RFs[20–23], others that neurons could be selective for some image categories over others (such as simple geometric shapes[16] or faces[15]). However, these results do not prove conclusively that vlPFC neurons encode objects or scenes. Visual neurons can appear object-selective when they respond to lower-level features contained within the image[4,9], and discovering these features takes a systematic deconstruction of the image (via feature removal[9] or feature visualization[24]). Previous experiments worked to show that in PFC, intact faces are elemental units of representation[15], but these relied on coarse image-scrambling techniques which can eradicate or crowd out internal critical features. To explore if PFC neurons work like ANN fully connected units, we set out to determine if vlPFC neurons could drive the synthesis of images using deep networks. This is a new technique that combines electrophysiology, search algorithms[25] and generative adversarial networks, which are pretrained to generate images based on natural statistics[26]. It is an excellent test because the natural statistics encoded by the DeepSim[27] generative network correspond to low-level features (*e.g.,* short continuous lines, color patches), not pre-defined objects or scenes as in recent networks[28,29]. To contextualize the significance of this potential outcome, we also characterized the basic visual properties of vlPFC neurons and determine if visual vlPFC neurons perform similar computations to ventral stream neurons.

Second, and in contrast to their latter statement above, I would argue that the authors have not used optimal methods to address the questions they are trying to answer. They employed 32 channel implanted arrays, which covered a small patch of cortex, recording from the same sites over different sessions. The problem with chronic arrays is that even if the signal changes from day to day it is impossible to be certain that different neurons are encountered on different experiments. Thus, the sample is largely the same over days and in this particular case it was 32 sites at best. This makes it extremely difficult to draw solid conclusions.

It is a fair observation. No approach is perfect, but we can explain why we had to carry out our experiments this way. For context, we know that the conventional approach to study PFC is to use single-electrode techniques, where a chronic chamber is implanted over PFC, allowing investigators to sample a region thoroughly, across months, focusing on well-isolated single neurons. This provides great recordings, but there are tradeoffs to this approach. One is that signal isolation of single units is harder to maintain over hours, because of monkey-initiated movements, electrode drift, and electrical noise. The closed-loop image synthesis approach requires excellent signal stability because every synthetic stimulus (of which there are thousands of instances) is presented just one time. Previously, we carried out closed-loop image synthesis with acute chamber recordings and with chronic microelectrode arrays (Ponce, Xiao *et al.,* 2019), and the signal stability and noise reduction was much better with the chronic arrays. For image synthesis experiments, using chronic arrays is, in fact, an optimal method.

We also cared about pursuing the question of stability. When we use arrays, we can track evolved representations in cortical patches over months or even close to two years. This cannot be done with acute recordings. The Constantinidis lab has shown that using chronic arrays is the proper way to address this type of question, and our work builds on that by applying the image synthesis approach on the same sites over months. Our Reviewer is correct that it is impossible to be certain that different neurons are encountered on different experiments, but in fact, we do find measurable correlation from day to day, and we report this considerable effect size as a property of the PFC cortical site.

So, while the Reviewer is correct that our methodology was not optimal in some ways, it was for the key questions we wanted to address. In response to the above concern, we have added a short justification that addresses our methodological choices, hopefully allowing the wider readership to see both perspectives:

Under Methods:

> We used microelectrode arrays because they are very stable in terms of electrode drift, both within individual experimental sessions and across weeks and months; stability is important for the closed-loop experiments because any given synthetic stimulus is presented only once. Because the single-unit data yield is lower with chronically implanted arrays, most of our results are relevant to cortical *sites*, not individual neurons, although every major conclusion is replicated with both single units, multi-unit signals, and visual hash.

In fact, several of the analyses carried out by the authors end up reporting proportions of neurons in a sample of less than 10 sites (e.g. lines 193 and 196, 367-368) and in one case data from only one monkey (e.g. Figure 3d where no active vlPFC sites are reported for Monkey C). In fact, I would argue that several of the analyses are inconclusive because of this limitation with the most prominent example that on retinocentric vs allocentric coding and gain fields.

As our Reviewer notes it is true that using chronic arrays means that we are sampling a limited patch of cortex. How well will our findings generalize? One reason that extensive sampling is important is that if a given functional type of PFC cell is found in one animal, especially after extensive training, then the cell type might be unique or idiosyncratic to the animal – the cell could be a quirk, not a principle. That cell type might never be found in a different individual. However, in our study, both animals' brains showed similar results: both had neurons with stable RFs, neurons that had image selectivity, could drive image synthesis algorithms, and comprised populations of mixed sensory selectivity. Put succinctly, we have shown that our results already generalize.

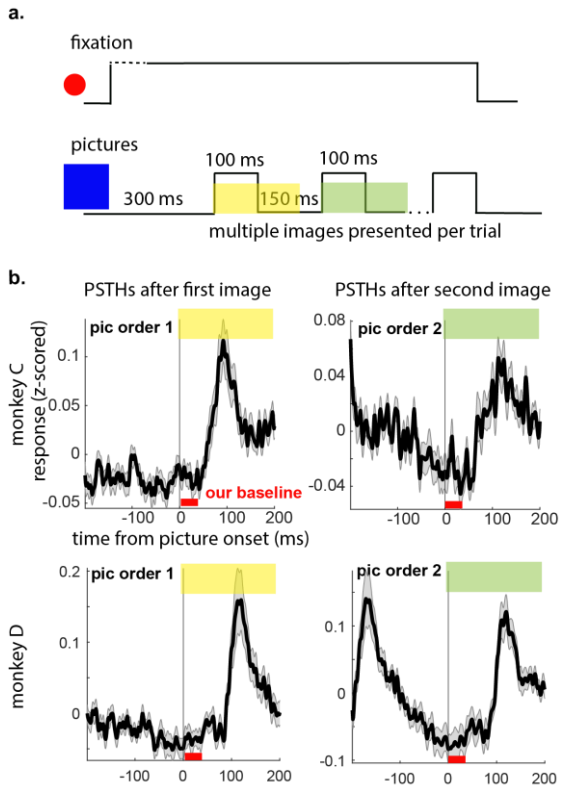
Part of the reason for this is that our methods rely on true random sampling of neurons: we implanted arrays without fMRI pre-mapping, without extensive task training, and we report everything we find. While this was a front-loaded risk, in this case, the outcome was successful. Further, we have checked the literature, and our percentages fit the pattern of foundational studies of the visual and prefrontal cortex. For example, the initial result of finding face-tuning in PFC was based on 11 of 137 stimulus selective neurons (8%, Wilson, O Scalaidhe, and Goldman-Rakic, 1993), and a separate study reported 37/779 (5%, O Scalaidhe, Wilson, Goldman-Rakic, 1997). The Goldman-Rakic lab also presented results for only one monkey trained to perform a fixation task, where 15/156 neurons were face-selective; one of the working-memory trained monkeys had only 9/437 (2%) neurons that were face-selective. Romanski and Diehl (2011) based results on 95 neurons, finding that ~20% showed selectivity for identity. Miller, Erickson, and Desimone (1996) recorded from 264 PFC neurons total, out of which 41 (16%) were selective for visual stimuli. Fuster and Alexander (1971) investigated an average of 22 neurons in a given animal (110 neurons total). This latter trend of fewer neurons per animal were standard in visual neuroscience: Hubel and Wiesel reported the activity of an average of one electrode track per animal (40 cats, 45 electrode tracks), and the seminal study in inferotemporal cortex by Gross, Bender, and Rocha-Miranda (1969) began with 41 visually responsive neurons out of 51. So, small numbers do not mean that results are misleading. More broadly, we think that *every* study is inconclusive, and that only the systematic replication of findings across laboratories provides confidence on a theory. We believe that reporting our results provides an opportunity for replication.

Overall, we certainly understand our Reviewer's concern on this sampling — we only wish to highlight that it is well-within the range of classic studies of visual and prefrontal cortex. We have purchased new sets of Neuropixels probes to increase our sampling numbers in future studies (the Livingstone lab, down the hallway, has been using them chronically with exciting results).

Besides these general comments, I also found that in certain cases the clarity of presentation and the statistics used should be improved. I list some of these below:

1) To assess visual responsiveness the authors compare responses in a 1-30ms post stimulus presentation to a window 50-150/200ms. It is odd that as a baseline they did not use a window during fixation (pre stimulus presentation). This would allow for more robust data comparing roughly similar sizes window around 100-150ms width. Moreover, a paired t-test should be used for these comparisons (unpaired t-test is reported).

This is a good observation. We had a specific motivation for using such different time windows to estimate baseline vs. evoked responses, due to our experimental design. In any given trial, a fixation point appears, and we allow the animal up to 5 seconds to look at it. Once fixation is acquired, images start flashing in short succession — we can have between 2-5 images at a time before the trial ends and the animal receives reward. At the start of the trial, the single-unit/multiunit activity is generally stable for 200-300 ms before the first image is flashed, but afterwards, site activity does not always settle back down to baseline. In other words, for subsequent image presentations, the preceding time window could still have "evoked" activity. To illustrate this, we plotted the mean activity of all vlPFC channels (for each monkey) in response to the first image per trial, and then to the second image (**Response Fig. 7**).



Response Figure 7. Rationale for using a short early response window. a. In our visual responsiveness/selectivity experiments, the fixation point appears and once the monkey looks at it, images are shown in succession, 100-ms on, 150-ms off. **b.** Mean responses from vlPFC sites in monkeys C and D (top and bottom rows), after the first image in the trial is flashed (left, yellow highlights), and after the second image in the trial is flashed (right, green highlights). Red shows our previously proposed time window to assess baseline activity.

The figure illustrates that the Reviewer's idea of using a longer time window to assess baseline activity works well for the first image, but in subsequent image presentations within the trial, this longer window might capture responses to the previous stimulus. This is why we prefer a shorter time window that can work for any image presentation.

However, our Reviewer is correct that this, in turn, creates a less reliable estimate. So, we ran a control analysis measuring channel responsiveness using a short time window OR a longer time window. Specifically, we tested how estimates of the number of responsive sites varied as a function of the duration of the early time window. The early window could range from -150 to 0 ms relative to image onset ("long"), or from 0 to 30 ms relative to image onset ("short"). The late window ranged from 50 to 200 ms after image onset. We then measured the mean rate within each window, comparing the median response in the late window vs. the short early window or vs. the long early window. We used all trials in the *selectivity* experiments but using only the first image presentation within each trial, and compared the statistical robustness of this responsiveness metric via using a paired two-sample test (Wilcoxon signed rank test for zero median; the Student's T test assumes normality in the distribution of residuals, which our data did not show), at a threshold of $p = 0.04$:

	Long window (no. responsive, % total)	**Short window (no. responsive, % total)**
Monkey	Wilcoxon	Wilcoxon
C	465, 41.9%	517 (46.6%)
D	491, 40.6	519 (43%)

Interestingly, the short-baseline window was associated with a 2.4-5% increase in the estimate in visual responsiveness. When focusing on N = 434 channel recordings collected across days, we found the difference in the estimates of responsiveness (short-window estimates minus long-window estimate) ranged from Δ = +3.8% (at p = 0.001) to Δ = -6.1 (at P = 0.08; median Δ = 3.8%). Put succinctly, the longer baseline window showed fewer statistically responsive sites, with a discrepancy of about 8.6% (37.4 out of 434 recordings). We are not sure why, it seems as if there could be some visual responsiveness to the fixation point that lowered the difference between "baseline" and "evoked" activity (see Figure above, Monkey D, pic order 1).

To address this observation and confound, we have made the following changes to the manuscript, but did not elaborate since our main conclusions did not change:

Results (under Many vlPFC sites showed position tuning consistent with V4 RFs.)

> While a change in response could signal either excitation or inhibition in response to the stimulus, all statistically reliable changes in firing activity were excitatory. As an additional test, we also measured responsiveness of PFC sites in separate experiments, where PFC sites were stimulated using 5-10°-wide images at the locations near the population RF. We focused on the change in activity at -150 to 0 ms before image onset vs. 50 to 200 ms after. We found that out of 434 recordings across days, 41.9% and 40.6% of sites showed responsiveness (using a shorter baseline window resulted in estimates of 46.6% and 43%). We conclude that many vlPFC neuronal sites signaled the onset of a picture on the screen in the absence of task training, similarly to V4 sites.

2) When assessing polysensory responses, the number of sites with significant responses for each modality should be clearly stated in the text.

We took the opportunity to reanalyze the data and streamline this information. This was good, because we realized that our false discovery correction should have used the Benjamini-Hochberg procedure since the sample size was 32. Our main results held, however: they continue to show that V4 sites were responsive to visual information, parabelt sites responded to sounds, and PFC showed a mixed distribution, although the PFC evolving channels were more comparable to the V4 distribution. We edited the section as follows:

> We found that our choice of images and sounds were effective at driving activity in sites in either V4, CPB, or vlPFC: in V4, 75% of 32 sites responded to images with a median response was 17.3 ± 2.3 events/s (p < 0.02 per Wilcoxon sign rank paired test, after false discovery rate correction using the Benjamini and Hochberg procedure). While none of the V4 sites showed increased activity to sounds, as a population they were *suppressed* weakly by sounds (median response -1.3 ± 0.2, see **Supplemental Figure 2a-b**). No individual V4 site was responsive to both modalities. In CPB, 62.5% of 32 sites showed a median increase in response of 5.0 ± 1.6 events/s and 0% of 32 sites showed reliable changes in median firing rate to images (2.0 ± 0.4 events/s); as in V4, no site was responsive to both modalities. Thus, while visual cortex sites were driven primarily by images and auditory cortex sites by sounds, intermingled sites in both cortices also showed a bit of responsiveness to the other stimulus type. This type of polysensory crosstalk has been described in sensory cortex before[36], although investigations into the topic have been few. Compared to sites from visual and auditory cortex, vlPFC sites were more likely to respond to both images and sounds. vlPFC sites reliably showed image-related modulation of at least a few events per second (Monkey C, 71.9% of 32 sites were responsive to images at a P-threshold of 0.02, showing a median response of 5.0 ± 1.0 events/s; Monkey D, 18.8%, median response 2.2 ± 2.1 events/s). Many of these sites also showed sound-related activity of at least a few events per second (Monkey C, 75.0% of 32 sites, median response 3.5 ± 0.8 events/s ; Monkey D, 6.2% of 32 sites, though the median response was -0.5 ± 0.4, so many sites were also suppressed). In both animals, there were sites that were statistically responsive to both images and sounds (18/32 channels in Monkey C, 2/32 channels in Monkey D).

3) In the gain field analysis, a one-way ANOVA is mentioned in the methods, but it is not clear which factor was used. How were gaze position and stimulus position handled?

Thanks for noticing this. We mentioned one-way ANOVAs in two instances, first, in this statement: "To measure potential gain effects of gaze direction for each cortical site, we performed one-way ANOVAs on the peak firing rates across trials for each of the two gaze directions, then corrected resultant p-values by FDR."

We have now clarified our manuscript as follows

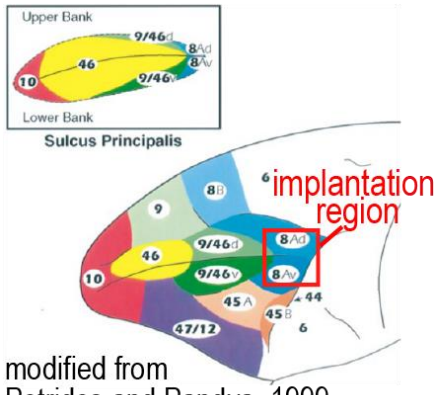
> To measure potential gain effects of gaze direction for each cortical site, we set out to investigate if a given channel's RF changed in retinotopic position or in peak magnitude as a function of gaze condition. To do this, we compared each channel's RF at one gaze condition (-5,0) vs. another gaze condition (+5,0). However, this comparison would only be valid if the channel had an RF in the first place. To determine if a given channel had an RF, we used a third set of trials where the gaze condition was at (0,0). For each gaze condition, we performed a one-way ANOVA test (with stimulus position as the sole factor), correcting the observed *p*-values using false discovery rate tests (*mafdr.m*). We limited our first pass of analysis to channels that showed a significant RF ($p \leq 0.05$ after FDR correction) during the center gaze condition (24 V4 sites in monkey C, four vlPFC sites in Monkey C, and 13 vlPFC sites in Monkey D, for a total of 17 active vlPFC sites), and then analyzed independent trials from the two remaining gaze conditions.

The second statement was "For some sub-analyses, we estimated putative RF centers using trials presented during center gaze direction. For sites exhibiting significant position tuning during center gaze (one-way ANOVA, $p < 0.05$ after correction for false discovery rate)…" This was the same test as above, so we clarify it in the text as well.

> For sites exhibiting significant position tuning during center gaze (one-way ANOVA, stimulus position as sole factor, $p < 0.05$ after correction for false discovery rate)…

4) In the cluster analysis, what does the conclusion tell us? Is it possible that the array was at the border between two different PFC regions? Did the authors find any evidence for clusters of spatially selective cells interspersed among clusters of non-spatially selective zones?

Great questions, but unfortunately we only have partial answers. There is much work that parcellates different regions of the prefrontal cortex. We decided to target the region anterior to the arcuate sulcus and slightly inferior to the principal sulcus, because this area was described as receiving significant IT projections. However, we did not try to implant in any specific region beyond that level of description. In the literature, the gyrus anterior to the arcuate sulcus and inferior to the principal sulcus is generally referred to as ventrolateral prefrontal cortex— it can be parsed into cytoarchitectonic areas 45, 8A, 9/46, and 47/12. The estimated borders of these cytoarchitectonic areas fluctuates between subjects, and precise nomenclature and region borders vary across papers. However, we can refer to the cytoarchitectonic boundaries reported in Petrides and Pandya, 2001 as a start. Per our intraoperative photographs, we targeted the region of vlPFC slightly below the principal sulcus, so that would put us around 8A (**Response Fig. 8**).

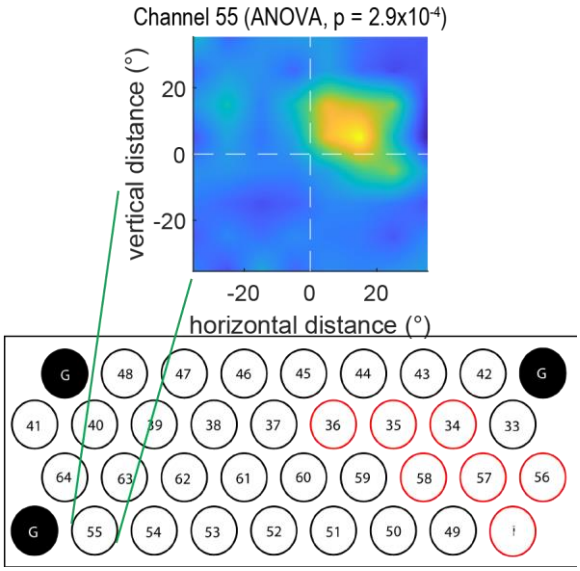


Response Figure 8. Parcellation of lateral prefrontal cortex into areas.

One related question from colleagues was whether we could be at the border of the frontal eye field. This is generally thought to lie along the wall of the arcuate sulcus and on the dorsal aspect of the prearcuate gyrus[30–32], not on the ventral gyrus. However, we still pursued this question in the lab. First, we decided to try and microstimulate the vlPFC array in monkey D, which still had channels with detectable receptive fields. We purchased a PlexStim electrical stimulator (Plexon Inc.) and obtained stimulation parameters recommended by our array manufacturers and compatible with the work by Schwedhelm, Blaudauf, and Treue (*Electrical stimulation of macaque lateral prefrontal cortex modulates oculomotor behavior indicative of a disruption of top-down attention*, **Scientific Reports**). We replicated the task by Schwedhelm *et al.* of a visually guided center-out saccade task, where the monkey acquired fixation, and after 500 ms, a target image appeared 7° to the left or to the right. If the monkey made a saccade to the target, a reward would be dispensed. In half the trials, we randomly applied microstimulation. Because this monkey had vlPFC channels along the vertical meridian, we expected that in stimulation trials, the saccadic endpoints might be biased towards or away from the horizontal meridian, or more directly, that we could evoke saccades due to the stimulation. However, we found no changes in the saccadic distribution at all. We changed stimulation parameters to increase the intensity beyond what other teams have used, but again, this did not cause a detectable change. This was a negative result, so

we did not include it in the manuscript, but it gave us some internal reassurance that we had not missed this possibility.

Next, to address the question if we found spatially selective cells interspersed among clusters of non-spatially selective zones, the answer is: *not reliably.* We scanned through every RF-mapping experiment we collected, searching for evidence that some channels besides those in **Figure 2** might have manifested RFs briefly, perhaps for a few days. But there was no evidence except for one day out of 43 for monkey C, when a channel far away from the cluster (Ch. 55) showed this response pattern (**Response Fig. 9**):



Response Figure 9. Channel showing potential evidence of a transient RF.

While this position tuning was very compelling, it was not there on other days. So, we would not be comfortable describing it as an RF vs. a spurious finding. Overall, the manuscript analyses describe the most reliable RFs that we could find. So, what does the cluster analysis tell us? When starting the experiment, we had considered multiple alternative hypotheses with similar weights: as a multimodal region, either vlPFC sites would all have RFs, or a few of them would in a salt-and-pepper fashion (just like orientation-tuned neurons in mouse V1 cortex), or none of them would. We found that a few of them did, and they were clustered together. We think it would be prudent to limit our interpretation of the results until new experiments, probably with Neuropixels probes, bring more evidence to the table.

We thank the Reviewer for the prompt to re-examine these data. To provide a bit more context, we have added the following statement to the Results:

> The most likely location for the arrays was 8A, given their proximity to the ventral/dorsal aspect of the principal sulcus.

We have added a statement to the Discussion, describing the limitations of the study:

> *Limitations of the study.* One major limitation in our conclusions is the interpretation of the cluster of sites with reliable RFs. It is a possibility that this cluster reflects the border between two areas. While the literature shows no clear consensus on the areal borders within the region anterior to the arcuate sulcus and posterior/ventral to the principal sulcus, Petrides and Pandya (2001) illustrated that this area holds the confluence of area 8A, 45A, and 45B. The more dorsal aspect of this region encroaches onto the frontal eye field (FEF), although we strongly doubt that our recordings pertain this area, as FEF lies more along the anterior wall of the arcuate sulcus[30–32]. We also conducted a brief (three-day long) pilot experiment with microstimulation following the parameters and design of Schwedhelm *et al.* (2017)[33], but did not observe any eye movements, suggesting that we were not in FEF. Because our primary goal was to understand the visual encoding of vlPFC neurons, not map out cortical areas, this is a limitation that was built into our design, although the finding invites further investigation.

5) In the Experimental setup paragraph in the methods, it is mentioned that for most experiments animals were performing a passive fixations task. I am probably missing something because I thought in all experiments animals were performing a passive fixation task. Could you clarify that?

Yes, this is correct. The original phrase referred implicitly to the gaze-position switch experiments as being slightly different from our usual fixation task, but in fact, both are simple fixation tasks. We have changed the Methods to read

> In all experiments, the animals performed simple fixation tasks (i.e., holding their gaze on a 0.25°-diameter circle, within a ~1.2–1.5°-wide window on the center of the monitor for 2–3 s to obtain a drop of liquid reward (water or diluted fruit juice, depending on the subject's preference).

6) A reference to figure 1d, 2c and 5c is missing in the text

We have added references to figure 1d (now 1e), figure 2c, and 5c.

7) On lines 211-214 it is not clear what is being compared to what.

In the submitted manuscript, these lines were

> We also measured changes in overall site firing rate across different gaze directions (gain effects). We found small but significant effects (about 1-2 spikes/s). The median change in RF tuning curve amplitude across gaze direction was 1.56 ± 0.02 spikes/s for V4 sites, while the median gain effect for vlPFC sites was 0.41 ± 0.0003 spikes/s.

We re-wrote it as follows:

> We also measured changes in overall site firing rate across different gaze directions (gain effects). For each site, we computed the RF at different gaze positions and then compared their peak response values. We found small but significant differences between the RFs at different gaze positions (about 1-2 spikes/s). The median change in RF tuning curve amplitude across gaze direction was 1.56 ± 0.02 spikes/s for V4 sites, while the median gain effect for vlPFC sites was 0.41 ± 0.0003 spikes/s.

Reviewer #3 (Remarks to the Author):

Ventrolateral prefrontal cortex (VLPFC) is thought to be at the apex of the visual streams that converge in frontal cortex. This manuscript details a number of really quite interesting experiments that Rose and Ponce conducted that initially take a quite standard approach to characterizing the visual receptive field (RF) properties of VLPFC neurons and comparing these to V4 and central auditory parabelt neurons. Further, they investigate whether these RFs are retinotopically organized before conducting a really interesting analysis using cutting-edge machine learning driven prototype matching. This final piece of the manuscript is really quite exciting, and the results are rather unexpected; namely that image prototypes that maximally drive responding in VLPFC are often quite amorphous consisting of quite simple visual features.

This is a really quite an interesting manuscript, and the results should be of interest to those working in on the neurophysiology of frontal cortex as well as visual processing. The first part of the manuscript mapping RFs in VLPFC is very standard and robust. This is a good thing and is the foundation for the rest of the manuscript. The second half is more interesting as it is looking for what specific features drive VLPFC neuron responses. Some would have predicted that the prototypes that maximally drive VLPFC responses would have been more detailed.

While this is really quite interesting the biggest weakness of the study is the relatively coarse neurophysiology data that the authors have managed to obtain from their floating micro-arrays. On the one hand this could be seen as not so much of a problem as the main message of the paper relates to the visual tuning properties. On the other hand, the retinotopic mapping and ML-driven prototyping analysis is slightly diminished if the effects are being driven by multi-unit activity (at least I think that is what is happening in one of the monkeys) as opposed to single neurons. The point being that the conglomerated activity of many neurons will likely lead to more general/amorphous prototypes/RFs if individual neurons have very specific tuning profiles. I don't think that this necessarily means that the manuscript isn't sufficiently important, I'm just going to

need a little convincing about whether the effects can really be seen at the single neuron level. This main concern as well as a number of others fleshed out in more detail below:

1) As highlighted above I'm interested to know more about the distinction between true single neuron coding and coding by multi-unit/hash activity in the data. The first thing that the authors should do is provide a table accounting for how many single neurons and multi-unit activity they have recorded/analyzed. They allude to there being little difference between these measures, but I'd like to know what the numbers for each are, as this helps a reader to understand the robustness of the effects reported. Because the arrays are also fixed in place and thus less likely to move, it makes sense to also highlight if any of the single neurons recorded across days are thought to be the same (or not).

This is an important concern, and one that has driven our lab since we first launched this closed-loop approach. What happens when we perform activation maximization in single neurons, vs. multiunit channels, vs. groups of channels? Is it possible that, as the Reviewer suggests, prototypes are messy when they emerge from multiunit activity vs. single-units? If vlPFC neurons are less clustered by stimulus preference compared to visual cortex neurons, then one imagines how their collective tuning would create an amorphous visual mass, the equivalent of a raucous crowd of individuals, each speaking perfectly cogently at the same time, yet yielding an incomprehensible sound as a whole. This is a big question, and one we must answer in parts.

First, we offer the information requested by our Reviewer, accounting for single neurons and multi-unit activity. We populated **Supplemental Table 4** with numbers describing the number of single neurons and multi-unit activity prototypes that we recorded. To illustrate how these prototypes differed per signal area, we updated **Figure 5** with an extra panel showing prototypes yielded by signal type. We then provide further evidence of prototypes from single and multiunits from two other studies. Finally, we will provide a new analysis we conducted *in silico*, asking the question: is it possible to obtain a muddled prototype by randomly activating differently tuned single units?

1. **Prototypes as a function of PFC neuronal signal type**
2. **Prototypes as a function of visual cortex neuronal signal type**
3. ***In silico* simulation of multi-unit evolutions**

1. **Prototypes as a function of PFC signal type**

Here, we edited our Supplemental Table 4, which shows statistics on the image-synthesis experiments, noting signal type as well:

Supplemental Table 4. All prototype synthesis experiments targeting vlPFC sites in Monkeys C and D. Significance obtained through Wilcoxon rank sum test, with threshold of $P < 0.01$. Individual unit types are designated as follows: single unit (SU), distinct multiunit (MU), and whole-channel multiunit hash ("hash").

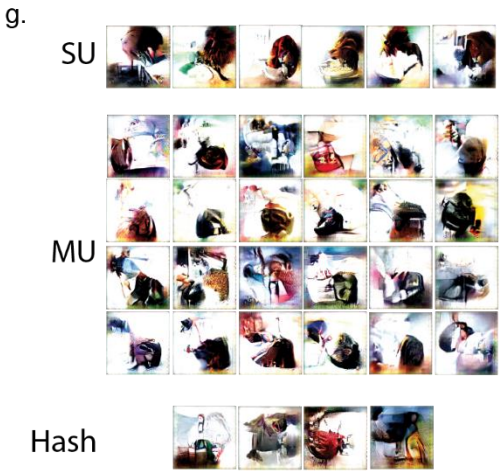
XDream success by site, $p < 0.01$. * **= within RF cluster**

Monkey C

Ch. number	nExp attempted	nExp successful	Success rate	Successful units	All units tested
Ch. 33*	2	0			1 MU, 1 hash
Ch. 34*	3	0			3 hash
Ch. 35*	7	3	42.9%	2 MU, 1 hash	2 distinct MU, 5 hash
Ch. 42*	5	2	40%	2 hash	5 hash
Ch. 43*	6	1	16.7%	1 hash	2 MU, 4 hash
Ch. 50*	2	0			2 well-isolated SU
Ch. 52	1	0			1 hash
Ch. 57*	10	1	10%	1 MU	5 MU, 5 hash
Ch. 58*	30	20	66.7%	20 MU	23 distinct MU, 7 hash
Ch. 59*	**6**	**0**			5 MU, 1 hash

Ch. 60*	**8**	**0**			6 MU, 2 hash
Monkey D					
Ch. number	nExp attempted	nExp successful	Success rate	Successful units	All units tested
Ch. 33*	2	0			1 MU, 1 hash
Ch. 34*	3	0			3 hash
Ch. 35*	7	3	42.9%	2 MU, 1 hash	2 distinct MU, 5 hash
Ch. 42*	5	2	40%	2 hash	5 hash
Ch. 43*	6	1	16.7%	1 hash	2 MU, 4 hash
Ch. 50*	2	0			2 well-isolated SU
Ch. 52	1	0			1 hash
Ch. 57*	10	1	10%	1 MU	5 MU, 5 hash
Ch. 58*	30	20	66.7%	20 MU	23 distinct MU, 7 hash
Ch. 59*	**6**	**0**			5 MU, 1 hash
Ch. 60*	**8**	**0**			6 MU, 2 hash

We updated this Table in the Supplemental Information. Next, we updated Figure 5 with an extra panel that shows the evolved prototypes:



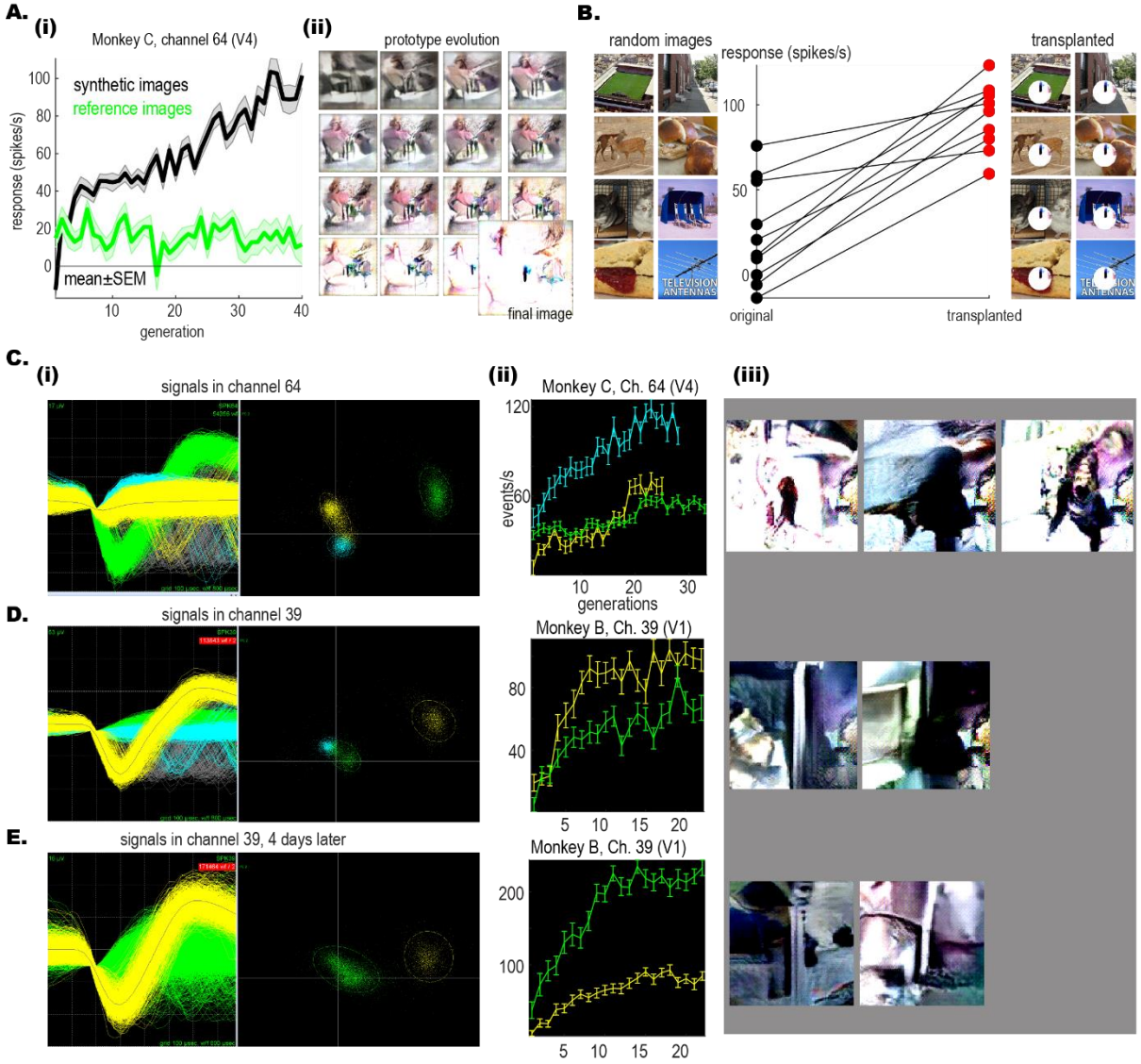
g. Prototypes plotted as a function of signal type.

Our working hypothesis is that signal type (single-unit- vs. multi-unit) did not lead to systematic differences in the complexity of tuning for vlPFC prototypes. The visible background clearing in many vlPFC prototypes did not appear to be dependent on extremely well-isolated single units contributing to the signal; we observed many prototypes driven by coarser multiunit signals that demonstrated comparable tuning properties to vlPFC single units. To make this point more convincingly, the top row shows the prototypes from the same single unit recorded over 7-10 days. We tested our working hypotheses using several more approaches with a larger dataset, as explored below.

2. Prototypes as a function of V4, IT signal type

We wanted to continue to explore the question of whether combining multiple single units can lead to degraded prototypes. This question prompted us to accelerate some analyses pertaining to the relationship between single units and the multiunit or hash activity within that channel (using previously collected data from V4 and IT in Rose *et al.,* 2021 and more recent unpublished experiments). When using chronic arrays, one

often finds multiunit and hash activity, but sometimes we also find separable signals that resemble single units. Over the years, we have had the opportunity to evolve prototypes from multiple signals in the same channel. We wanted to compare the prototypes of single units vs. multiunit/hash signals. What do we find? **Response Fig. 10** shows an example. Recently, we recorded from a channel in Monkey C (new array) targeting the lunate gyrus (V4). This channel showed one single unit and a hash signal. We recorded from the single unit on one day, and this unit's activity was maximized to a range of 100 spikes/s (**Response Fig. 10A(i)**). This unit had a very small RF; it began by creating a Gabor, then cleared the entire 3°-wide image and preserved a single vertical line (**Response Fig. 10A(ii)**). We confirmed this was the key activating feature because we "transplanted" this fragment to randomly selected images and increased their response across the board (**Response Fig. 10B**). Days later, we returned to this channel and found a new multiunit (**Response Fig. 10C**), so we evolved from the previous single unit, the new multiunit, and the hash. We found that the single unit returned the same vertical bar as the previous day; the multiunit returned a black area surrounding the location



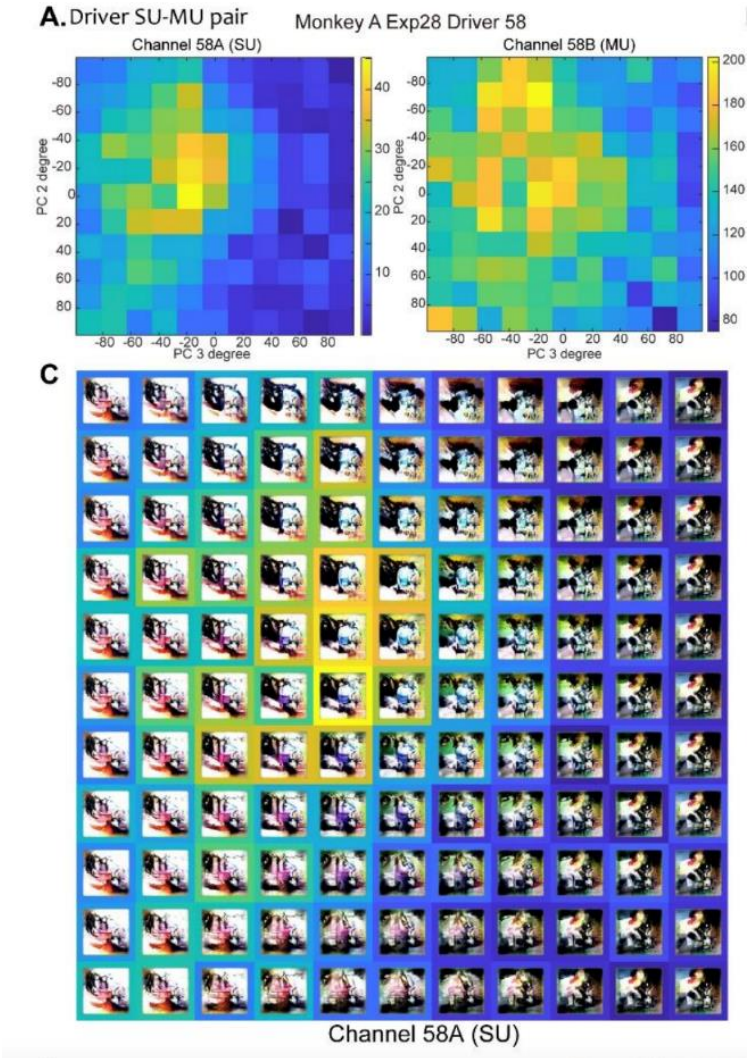
Response Figure 10. Examples of prototypes driven by single-unit signals and hash signals from the same channel. A. (i) Responses of single unit in monkey C, channel 64 (new array), to adaptive synthetic images (black) and static reference images (green). Curves show mean response ± SEM. **(ii)** Examples of the prototype taking form, and its final generation version ("final image"). **B.** Demonstration that the evolved prototype fragment identified both the location of the RF and a highly stimulating motif, as cutting-and-pasting this image region into randomly selected natural images increases their effectiveness. **C. (i)** Another experimental day with the SU in channel 64, this time accompanied by two signals including hash. All signals drove prototype evolution (**ii-iii**). **D.** Similar experiments using Channel 39 in monkey B (V1 single and multiunits/hash). This channel encompassed signals encoding different versions of the same vertically oriented contour.

of the bar, and the hash returned the black area containing a small, oriented fragment (**Response Fig. 10C, i-iii**). While analyses are ongoing, we interpret this hash signal seems like as a combination of local features.

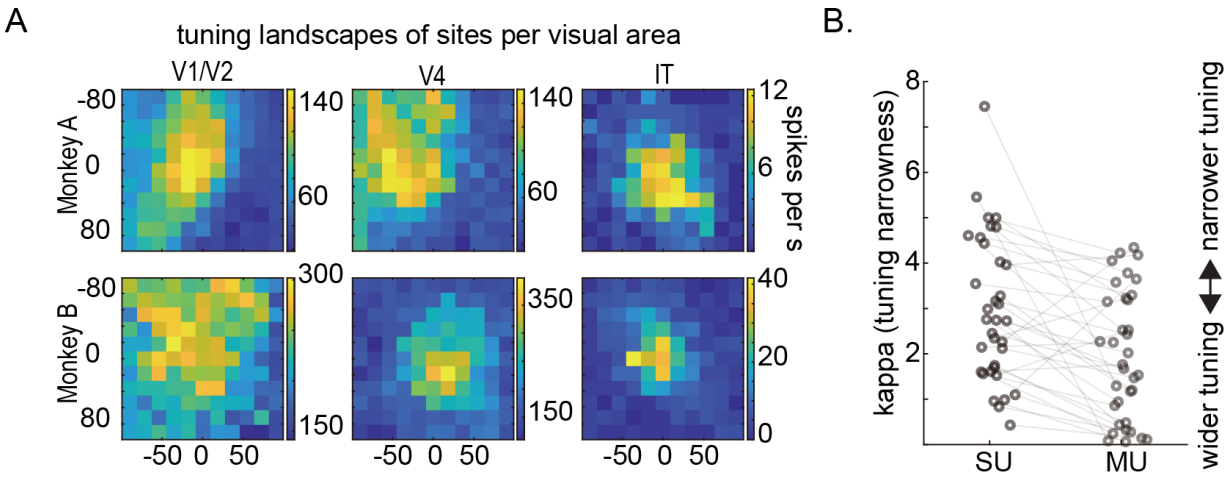
Another example comes from monkey B. We recorded a single unit and hash from Channel 39 (V1) four days apart. All evolutions resulted in lines as well (**Response Fig. 10D-E**). While we are working on another manuscript addressing these questions, our preliminary impression is that prototypes evolved from signals arising from the same channel tend to resemble each other more than the prototypes evolved from signals across channels. A similar working hypothesis is that the more similar the preferences of the channel's units, the more likely we are to evolve from the hash or multiunit. The evolution failure rate in IT is about 30% (see Rose *et al.* 2019, Nature Communications) and larger in PFC. This suggests that clustering of image preferences in less frequent in IT and PFC than in V1 or V4. So, this manuscript suggests that some PFC visual representations are clustered at the level of the local microelectrode recording span (~400 μm diameter).

This is all to address the fact that electrophysiology signals from within a channel tend to resemble each other, likely because of functional clustering (such as columns), which explains why prototypes from single- and multiunits are interpretable (more precisely, relatable to natural images) and not chaotic messes. This working hypothesis is also supported by a previous study comparing how single units and multiunits respond to image transformations in the generative network space. For example, in Wang and Ponce (2022), we showed the tuning landscape of single- and multi-units (a generalization of the classic *tuning curve*, only over multidimensional space), comparing their widths. We measured these tuning landscapes by evolving a single neuron's prototype, and then moving away from it slowly, in the generative network's latent (input) space. This revealed a shape of the neuron's tuning landscape that was Gaussian-like. We could then compare the same-channel multiunit tuning landscape shape (its peak and tuning width). They were similar to each other (**Response Fig. 11**).

Response Figure 11. Multidimensional tuning curves (tuning landscapes) of the ventral stream. **A.** In this example, a SU unit in channel 58 of monkey A (V4) was targeted for activity maximization, and images were sampled around the prototype (left, top). We also examined the tuning landscape of the multiunit/hash activity around the SU's prototype (right, top). **B.** Images driving the activity of the SU.



We also found that MU tuning landscapes were wider (**Response Fig. 12**):

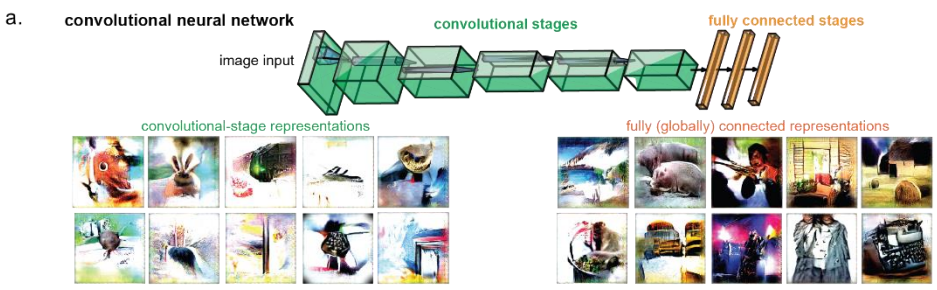


Response Figure 12. Tuning landscapes. A. Responses of neuronal sites in V1, V4, and IT (columns, spikes/s), in two monkeys (A, B, rows). After finding their prototype (peak of tuning function), we sampled the image space around this prototype by moving at a fixed rate away from the peak, in different directions. A Gaussian-like tuning function was evident. **B.** Tuning width measurements using the Kent function, which includes a parameter *kappa* that quantifies tuning width (larger value = narrower tuning), as a function of single unit (SU) vs. multi-unit/hash (MU).

So, our highest-ranked hypothesis is that mixing the selectivity of nearby units (as in the same channel) results in two outcomes: prototype evolution succeeds because the nearby units are similarly tuned, or prototype evolution fails because the nearby units are tuned to other features not covered by our current experimental design (such as direction of motion or depth) or because the units are not similarly tuned. So, what happens if we try to evolve from units having decidedly different preferences? We pursued this question *in silico*, as we present below.

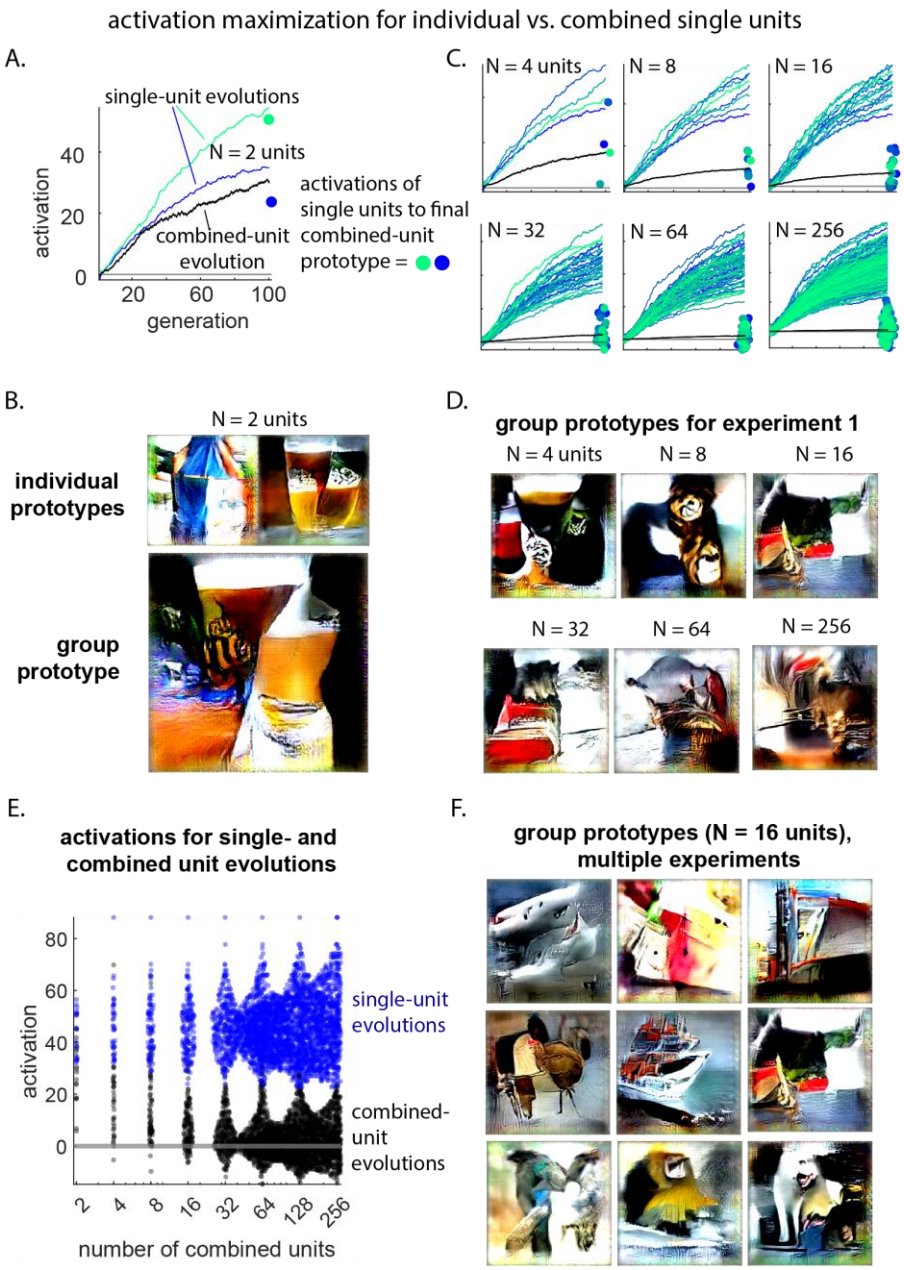
3. *In silico* simulation of multi-unit evolutions

We simulated evolutions as driven by convolutional neural network (CNN) hidden units ("single units," tested one at a time) vs. combinations of single-units. This follows a short study we did *in vivo* with monkeys A and B (Rose *et al.*, 2021). Years ago, early in our experimental research program, we thought it would be interesting to maximize the combined activity of multiple single units or channels, to see if we could create full objects or scenes. These experiments did not work — we noticed that when maximizing the activity of two channels c_1 and c_2, the channels would "fight" each other to guide the generator into different directions in its input latent space, and often the prototype would either resemble that of c_1 OR that of c_2, with only one being maximized. This puzzled us, because the maximization algorithm works well with both single- and multi-units, and even hash, and we concluded the only way this could be true is that in successful experiments, MU and hash signals originated from single-units with similar preferences. This clustering of preferences, scaling from micrometers to millimeters, is well-documented[34–36], so the units behind a multiunit signal had to be collaborating, not opposing each other. We moved in other directions and did not explore this fully, so this is an opportunity to test this. Here we can use standard CNNs, which have no topography or any kind of clustering, so they represent the ideal test case of units that are independently selective for different visual features. Specifically, in CNNs trained for category classification (via ImageNet), the 1000 output layer units acquire complex, scene-like representations, as shown in this Figure from our response to Reviewer 1:



To determine how these output units' prototypes might degrade by combining their responses, we conducted a series of activation-maximization experiments. The difference was that instead of activating one CNN unit at a time, we randomly sampled groups of units out of the population of 1000 and averaged their response, then tried to optimize images based on this "combined-unit" response.

Methods. We used AlexNet layer *fc8* (N = 1000); we defined a range of unit group sizes (G = powers of 2 from 2 to 256). For each experiment E_i where i = 1 to 10, a random order of the 1000 units was generated, and then we selected G_n units (starting with G_2 = 2), then started the evolution process, creating synthetic stimuli that maximize the averaged activation of the selected unit group. We then repeated for G_2, G_4, …, G_{256}, For each evolution we collected the mean mixed-population response; after each evolution converged to form a *mixed-group prototype* P_n, we measured the activation of individual units to one of these final generation (no. generations = 1000) mixed-group prototype examples. We then examined the group prototypes and how the individual units responded in each group size.



Response Figure 13. Combining single units during the activation maximization process. A. Activity of two units, as images are optimized just for each (color traces) or as images are optimized for both units concurrently (black line), over 100 iterations (generations). Dots show the final activation of the unit to the *combined-unit* prototype. **B.** The singly optimized images (*individual prototypes*) and the combined-unit (*group*) prototype. **C.** As in A but combining more units. **D.** Combined-unit prototypes for larger groups. **E.** Results aggregating multiple experiments, showing the activity (y-axis) of 256 units when images are optimized singly (blue dots), or when they are combined in increasing numbers (black dots, x-axis). **F.** Group prototypes across experiments.

Results. We found that CNN units acted in a way that resembled our *in vivo* observations. First, we found that it was possible to maximize the activity of combined units at the same time (**Response Figure 13A,B**), however, the mixed activation decreased as a function of the number of units combined. For example, on average, when fc8 units were activated separately, the mean activation value was 45.33±0.14; when pairs of units were combined, their mean activation was 29.9±2.8 or down to 66%; combining four units dropped activity to 43% (19.5±2.5), eight units, 31% (14.1±1.6); 16 units, 19% (8.6±0.9) and by 256 units, to 3% (1.4±0.1, **Response Figure 13C**). This was in a noiseless simulation: if we consider the response variability of real neurons, it is evident that evolutions would fail because the activation rise would be immediately swamped by noise. When concurrent activation maximization was successful, prototypes did not degrade, as much as they combined: for example, the mixed prototype of the blue *sarong* fc8 output unit (776) and the *beer glass* unit (442) showed a beer glass with a touch of blue (**Response Figure 13B**). As

more units were added, the group prototype became larger, covering the whole image as a texture, rather than becoming compressed into a black round object (**Response Figure 13C**). Across all experiments, the evolutions began to "fail" when N = 16 units were combined (**Response Figure 13E,F**), and their group prototypes were all extended and comprised multiple parts. Overall, we conclude that (a) successful evolutions combining units with different selectivity lead to extended prototypes that combine their individual preferred images, and that (b) mixing too many units results in failed evolutions. Because we were able to evolve from some PFC sites, both single- and multiunits, we believe that the MU prototype shapes are not the result of mixing differently tuned units, but from mixing the activity of similarly tuned neurons. On the other hand, it is still perfectly plausible that any channels that failed to yield an evolution comprised units with differently tuned neurons.

We found this critique to be motivating, and a lot of fun to address. We would like to share some of these results with the wider readership. We are adding a Supplemental Figure that includes the combined-unit analysis, updating the legend with the motivation and methods: the legend is written below.

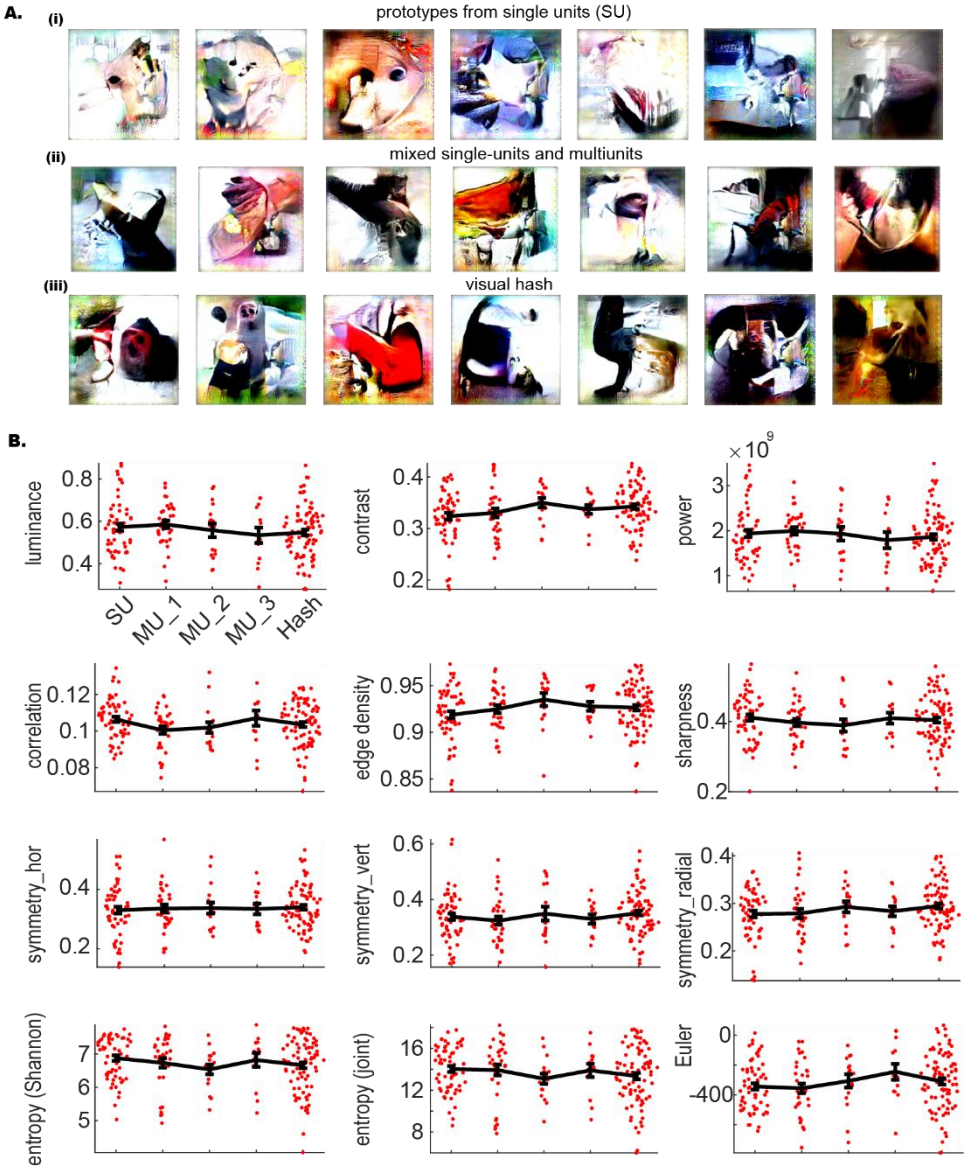
Supplemental Figure 3. Combining single units during the activation maximization process. We considered the possibility that the conglomerated activity of many neurons could lead to a more general/amorphous prototype, particularly if the individual neurons comprising the signal had specific and different tuning profiles. Our working hypothesis is that evolutions based on multiunit activity work largely (perhaps only) when the subjacent single units are similarly tuned. We have made relevant observations over several years. One of our previous studies showed that in evolutions involving a single electrode, single- and multiunit signals were strongly correlated in their tuning (Wang and Ponce, 2022, Tuning landscapes of the ventral stream). We have also tried to maximize the activity of different channels in two monkeys, with no clear results. However, to determine if our closed-loop activation-maximization paradigm works when units are tuned for different images, we conducted a simulation using artificial neural networks (ANNs). We used AlexNet layer fc8 (N = 1000); we defined a range of single (hidden) unit group sizes (G = powers of 2 from 2 to 256). For each experiment Ei where i = 1 to 10, we selected G units (starting with G = 2), then started the evolution process, creating synthetic stimuli that maximize the averaged activation of the selected unit group. We then repeated for G=2, G=4, …, G=256. For each evolution we collected the mean mixed-population response; after each evolution converged to form a mixed-group prototype Pn, we measured the activation of individual units to one of these final generation (no. generations = 1000) mixed-group prototype examples. We then examined the group prototypes and how the individual units responded in each group size.

We found that CNN units acted in a way that resembled our biological observations. First, we found that it was possible to maximize the activity of combined units at the same time (panel **A**, right), however, the mixed activation decreased as a function of the number of units combined. For example, on average, when fc8 units were activated separately, the mean activation value was 45.33±0.14; when pairs of units were combined, their mean activation was 29.9±2.8 or down to 66%; combining four units dropped activity to 43% (19.5±2.5), eight units, 31% (14.1±1.6); 16 units, 19% (8.6±0.9) and by 256 units, to 3% (1.4±0.1, Fig. Xa, left). This was in a noiseless simulation: if we consider the response variability of real neurons, it is evident that evolutions would fail because the activation rise would be immediately swamped by noise. When concurrent activation maximization was successful, prototypes did not degrade as much as combined: for example, the mixed prototype of the blue sarong fc8 output unit (776) and the beer glass unit (442) showed a beer glass with a touch of blue (panel **B**). As more units were added, the group prototype became larger, covering the whole image as a texture, rather than becoming compressed into a black round object (panel **C**). Across all experiments, the evolutions began to "fail" when N = 16 units were combined (panel **E**), and their group prototypes were all extended and comprised multiple parts. We conclude that (a) successful evolutions combining units with different selectivity lead to extended prototypes that combine their individual preferred images, and that (b) mixing too many units results in failed evolutions. Because we were able to evolve from some PFC sites, both single- and multiunits, we believe that the MU prototype shapes are not the result of mixing differently tuned units, but from mixing the activity of similarly tuned neurons.

A. Activity of two units, as images are optimized just for each (color traces) or as images are optimized for both units concurrently (black line), over 100 iterations (generations). Dots show the final activation of the unit to the *combined-unit* prototype. **B.** The singly optimized images (*individual prototypes*) and the combined-unit (*group*) prototype. **C.** As in A but combining more units. **D.** Combined-unit prototypes for larger groups. **E.** Results aggregating multiple experiments, showing the activity (y-axis) of 256 units when images are optimized singly (blue dots), or when they are combined in increasing numbers (black dots, x-axis). **F.** Group prototypes across experiments.

2) Following on from the above, the most interesting part of the manuscript for me are the experiments focusing on determining prototypes for neurons in VLPFC. However, I found it hard to know how consistent the results were for single neurons and multi-unit activity. The authors do, to their credit, highlight where there are differences between single units/hash/multi-units, but a little clarity/detail here would be good as it potentially alters the interpretation of the findings (see point above). One way to address this would be to include examples of prototypes from single units vs multiunits and if possible compute the degree of visual complexity for each and compare etc to see if there are differences. Again if there are differences it could indicate that single neurons have more visual like receptive fields whereas if there are not then I think it supports the authors current conclusions about VLPFC visual prototypes.

We see this as a related concern to the question above, which prompted another analysis. The question of single- vs. multi-unit activity and their associated prototypes is important and still underexplored, even in our own publications. Over the past few weeks of working on this response and revision, we have been motivated to create a new manuscript addressing this question. One of our new analyses examines years of prototypes and works through the statistics of prototypes from single units (SUs), multiunits (MUs), and hash. We asked if we could distinguish SU, MU, and hash prototypes by measuring over a dozen of descriptions, such as their luminance, contrast, overall Fourier frequency power, edge density, local-pixel correlations, entropy values, and more. We collected prototypes from three monkeys (A, B, C) in areas V1, V4, and IT (**Response Fig. 14A**) and ranked each evolution according to a five-point scale where 1 = single-unit, 5 = hash (impossible to isolate SUs), and 2-4 represent mixed single-unit activity where the lower the number, the easier it might be to isolate single units. We measured multiple statistics per prototype and aggregated them based on signal type. Unfortunately, we found no compelling statistical trends that prototypes varied across these metrics as function of signal type: for example, mean image luminance was 0.58±0.02 for SU prototypes, 0.56±0.02 for hash prototypes (ANOVA, unit F = 0.7, p = 0.68, after correction for multiple comparisons); edge density was 0.11±0.001 for SU prototypes, 0.10±0.001 for hash (ANOVA, F = 2.8, p = 0.65), and for Shannon entropy, 6.90±0.09 for SU prototypes, 6.49±0.10 for hash (ANOVA, F = 2.9, P = 0.65, **Response Fig. 14B**). Consistent with the previous results, we interpret this as more evidence that the global visual properties of the prototype are less related to the SU/MU nature of the signal, than it is to the joint visual preference of neuronal cortical clusters.



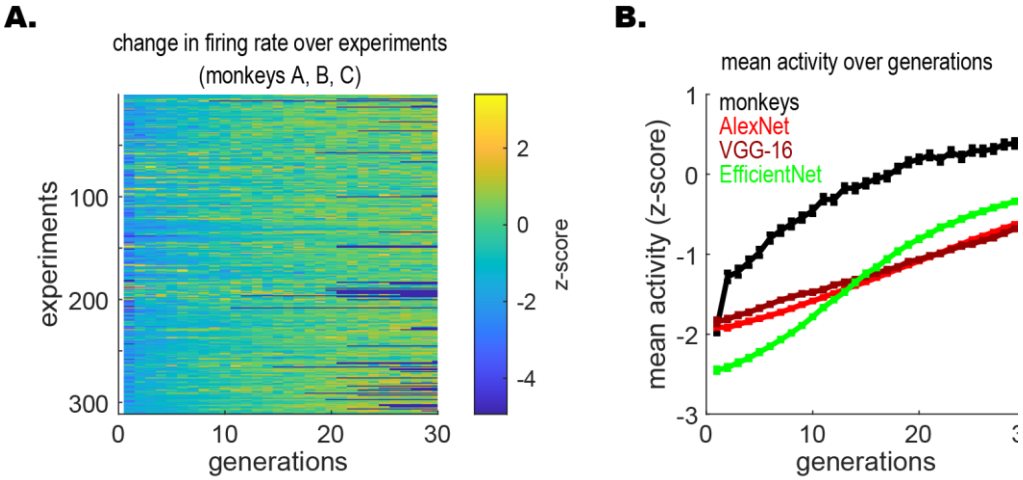
Response Figure 14. Low-level feature analyses of prototypes from single-units to visual hash.

A. Examples of images obtained from **(i)** SU recordings (where the signal is separable from the main channel hash in PCA space, and the signal shows a refractory period in the interspike interval distribution), **(ii)** from mixed multiunits, where multiple waveforms are separable from the main channel hash but their PCA clusters overlap, to **(iii)** hash signals, where only the main channel event threshold crossings are detectable, all across posterior IT, V4, and V1. **B.** Measurements of luminance, contrast, and other image-analysis metrics applied to the prototypes, as a function of their signal of origin.

Overall, we feel this is an important question that merits more exploration and a more extensive treatment, including new experiments. However, as this will require more work, we would like to pursue in an upcoming stand-alone manuscript. For now, we hope this provides some insights that address our Reviewer's concerns.

3) I'm also interested to know more about the methods for characterizing prototypes for VLPFC. The authors state that they went through 10 block iterations of images for each site/neuron/activity using the XDream approach. I wondered how this was arrived at and whether this parameter was based on determining prototypes in visual cortex where activity is more stereotyped/consistent to visual stimuli.

Thank you for this question. First, we would like to clarify that we went through dozens of iterations of images for each experiment: mean number of generations was 41.5±2.0 (SEM) for monkey C V4 evolutions, 36.0±1.4 for monkey C vlPFC evolutions, and 42.4±1.6 for monkey D vlPFC evolutions (we have added these values to **Supplemental Table 3**). We evaluated the success of the evolution using the first 10 blocks as a baseline. Over the years, we find that this is a crucial stage, and that successful evolutions begin to increase within the first 10-20 generations. To show this, below we include a figure showing the overall trends in mean activity during in 311 separate neurophysiology experiments, and several hundred other experiments simulated in convolutional neural networks of varying depths (AlexNet, VGG-16, and EfficientNet, **Response Fig. 15**). Empirically, we have found that it is not worthwhile waiting for too many more generations. Although we considered the possibility that some neurons or units might have a different "generational latency," where their activity would be more likely to rise after 20 or more iterations, we have not found this to be a reliable effect. We think this is a feature of the closed-loop algorithm itself, as the search algorithm (CMA-ES, covariance matrix adaptation evolutionary strategy) takes big steps within the image generator input space, so eventually, the search algorithm begins to move outside the most diverse image regions, which are near the center of the input space. Our article titled *A Geometric Analysis of Deep Generative Image Models and Its Applications* (Proc. International Conference on Learning Representations, 2021) describes the basis for these insights.



Response Figure 15. Evolution experiments showing mean firing rate activity over time, aligned to generation 0 (start of experiment). **A.** Individual monkey experiments. **B.** Activity averaged across experiments for monkeys and different artificial neural networks at the output layer.

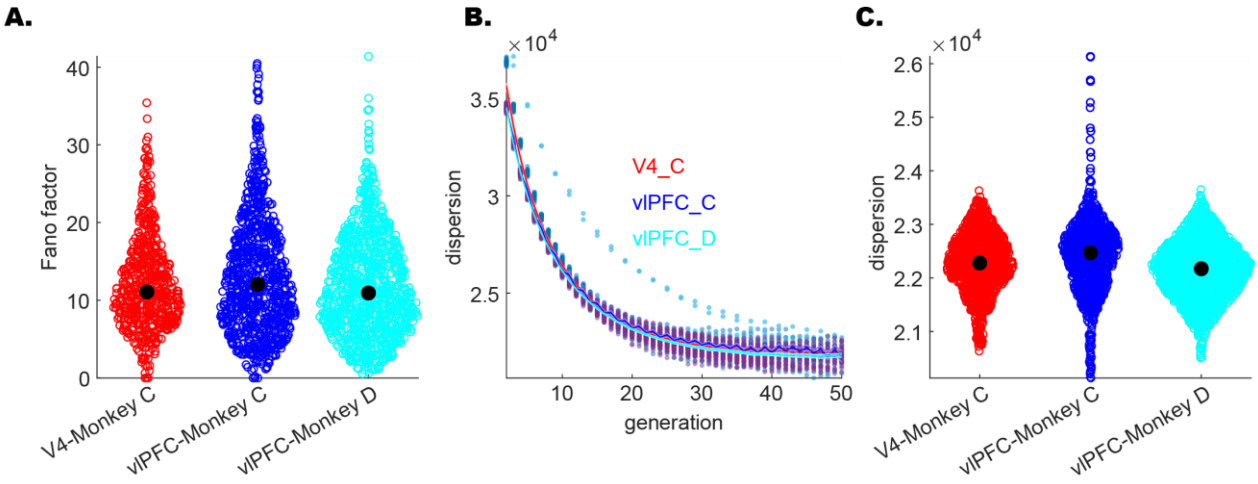
As an aside the 100ms image presentation seemed quite short given that it often takes ~100ms for PFC neurons to respond to visual images. My thought here is that prototypes in VLPFC may be more variable based on frontal cortex in general being more multi-sensory/exhibiting mixed selectivity than visual areas. I think that this is something that the authors should explore to highlight the differences between VLPFC and V4 neurons/sites. So, as the experiments cannot be performed again, how stochastic are the responses in VLPFC or V4 across the 10 block/iterations of the XDream procedure. I realize that the responses are changing as the images are updated, but I think that there could be interesting differences in how spiking activity converges on a stable response to the prototypes across the two areas across the blocks and that this should be reported here. Of course, if the prototypes converge at the same rate in V4/VLPFC then that would be an interesting finding in itself.

Our reviewer is right that we did not analyze stochasticity between V4 and VLPFC comprehensively, and on this request, we found that doing so did provide deeper insights. We conducted a few analyses. We

began by measuring the variability of the vlPFC and V4 sites during the evolution experiments. We used the Fano factor — the variance divided by mean of every site's spike-rate response distribution. Because every synthetic image was presented only once, we conducted this analysis using the reference images, which were shown once per generation (on average, 36-42 generations per experiment). The median Fano factor values per area were 11.1±0.3 (V4, monkey C, ±SE, interquartile range 7.38, sample size N = 558), 12.0±0.4 (vlPFC, monkey C, ±SE, interquartile range 10.56, sample size N = 838), and 11.0±0.2 (monkey D, ±SE, interquartile range 8.89, sample size N = 960, **Response Fig. 16A**). These values varied reliably across areas: a Kruskal-Wallis test revealed a difference in the median dispersion among the groups, $H(2) = 11.89$, $p = 2.6 \times 10^{-3}$. Yet, performing a two-sample test between V4 and vlPFC (monkey D) showed a *p* value of 0.24, highlighting no statistical difference between these groups. This suggests that the difference in Fano factor across areas is due to Monkey C vlPFC sites; as a group, they were more variable in their visual responses, and notably, this is a region that also showed more mixed auditory-visual selectivity than the monkey D vlPFC sites.

Next, we asked if there were differences in the variability of prototypes themselves, as a function of area. Based on one of our previous publications (Wang and Ponce, 2022, Tuning landscapes of the ventral stream), we know that the input codes to the generator (the *latent* codes or image "genes") map to image space, such that nearby points in the latent space correspond to similar images in pixel space. We leveraged this information to ask: how variable were the prototypes produced by V4 vs those produced by vlPFC? We measured the dispersion of latent vectors as a function of closed-loop iteration ("generation"). Dispersion was computed by obtaining the covariance matrix of all genes per generation (N = 40) and measuring the trace of the covariance matrix (the sum of the individual dimension variances). We saw that genetic dispersion started high at generation 2, then decreased and settled around the 30th generation (**Response Fig. 16B)**. To quantify this relationship, we fit an exponential decay function to these curves of the form $\mathrm{y(x)} = \mathrm{A} * e^{b*x} + \mathrm{c}$ to estimate the slope value *b*. We found that mean slope values differed per area: for V4, -0.135±0.001, and for vlPFC, -0.130±0.001 for both monkeys. This means that V4 sites were faster in reducing the variability in their prototypes, which we interpret to be a property of stronger visual selectivity. We performed a statistical test to determine if the difference between -0.135 and -0.130 was reliable. The statistical test was a randomization test. Specifically, to estimate the probability that the difference in slope between V4 and vlPFC could arise from the same underlying distribution, we combined the dispersion values (paired to their generation) from the V4 and vlPFC distributions, and then randomly sampled two groups from the mixed distributions, fitting each group with the exponential decay function, and measuring the difference between both slope values. We repeated this random re-sampling for 500 iterations and compared the actual observed difference to this mixed distribution. We found that the probability that the observed difference of 0.005 could arise from the null distribution of equal means was 3.2×10^{-2}. So, this is statistical evidence that V4 reduced the stochastic variability in prototype creation faster than vlPFC did.

Finally, does the speed at which V4 reduces prototype variability mean that the *final* prototypes were more or less stochastic than those in vlPFC? To answer this question, we measured the mean dispersion per generation over the flat regions of the dispersion over generation curves (above generation 20, see **Response Fig. 16B**), and compared these values across areas. For V4 sites, the median dispersion value was 22279±22, for vlPFC (monkey C), 22468±18, and for vlPFC (monkey D), 22172±17 (**Response Fig. 16C**). A Kruskal-Wallis test revealed a statistical difference among the groups, H(2) = 133.29, p = 1.1x10^{-29}. So, the V4 and vlPFC monkey D prototypes were both less variable than the vlPFC monkey C sites, and this is likely related to the Fano factor result from above. So, we conclude that the final prototype variability was stable, but it depended on the relative variability of individual sites.



Response Figure 16. Stochasticity in prototypes across areas. **A.** Fano factor for individual sites in V4 and vlPFC (red: monkey C V4, blue: monkey C vlPFC, cyan: monkey D vlPFC). Each color point shows an individual site/experiment. The large black points show the median value for the distribution. **B.** Dispersion plotted as a function of generation for all experiments in V4, vlPFC (monkeys C and D). Each point is the mean dispersion per generation per experiment. **C.** Median dispersion value per experiment per site, as function of area (color). The large black points show the median value for the distribution.

As predicted by the Reviewer, was an informative set of analyses for us, so we are adding these results to the manuscript. Specifically, we are adding a summary of these results to the section *Stable visual encoding in vlPFC without training, category, or semantics*:

In the main Results, we added this:

To examine the differences in prototypes across areas, we focused on variability. We began by measuring the variability of the vlPFC and V4 sites during the evolution experiments using the Fano factor — the variance divided by mean of every site's spike-rate response distribution. Because every synthetic image was presented only once, we conducted this analysis using the reference images, which were shown once per generation. The median Fano factor values per area were 11.1±0.3 (V4, monkey C, ±SE, interquartile range 7.38, sample size N = 558), 12.0±0.4 (vlPFC, monkey C, ±SE, interquartile range 10.56, sample size N = 838), and 11.0±0.2 (monkey D, ±SE, interquartile range 8.89, sample size N = 960). These values varied reliably across areas: a Kruskal-Wallis test revealed a difference in the median dispersion among the groups, H(2) = 11.89, p = 2.6x10^{-3}. Yet, performing a two-sample test between V4 and vlPFC (monkey D) shows a p value of 0.24, highlighting no statistical difference between these groups. This suggests that the difference in Fano factor across areas is due to Monkey C vlPFC sites; as a group, they were more variable in their visual responses, and notably, this is a region that also showed more mixed auditory-visual selectivity than the monkey D vlPFC sites.

Next, we asked if there were differences in the variability of prototypes themselves, as a function of area. Based on one of our previous publications, we know that the input codes to the generator (the *latent* codes or image "genes") map closely to image space, such that nearby points in the latent space correspond to similar images in pixel space. We leveraged this information to ask: how variable were the prototypes produced by V4 vs those produced by vlPFC? We measured the dispersion of latent vectors as a function of closed-loop iteration (generation). Dispersion was computed by obtaining the covariance matrix of all genes per generation (N = 40) and measuring the trace of the covariance matrix (the sum of the individual dimension variances). We saw that genetic dispersion started high at generation 2, then decreased and settled around the 30th generation. To quantify this relationship, we fit an exponential decay function to these curves of the form $\mathrm{y(x)} = \mathrm{A} * e^{b*x} + \mathrm{c}$ to estimate the slope value *b*. We found that mean slope values were different: for V4, -0.135±0.001, and for vlPFC, -0.130±0.001 for both monkeys. This means that V4 sites were faster about reducing the variability in their prototypes, which we interpret to be a property of their stronger visual selectivity. We performed a randomization test to estimate the probability of that the observed difference could arise from the same distribution and found this was $p = 3.2\text{x}10^{-2}$ (see Methods). So, we conclude that V4 reduced the stochastic variability in prototype creation faster than vlPFC did. However, this did not mean that the final prototypes in vlPFC were more reliably more stochastic than those in V4: the median prototype dispersion after responses converged was 22279±22 for V4, 22468±18 for vlPFC (monkey C), and 22172±17 for vlPFC (monkey D). A Kruskal-Wallis test revealed a statistical difference among the groups, $H(2) = 133.29$, $p = 1.1\text{x}10^{-29}$. So, the V4 and vlPFC monkey D prototypes were both less variable than the vlPFC monkey C sites, and this was likely related to the Fano factor result from above. So, we conclude that the prototype variability was stable, but it depended on the relative variability of individual sites, even to natural (reference) images.

This was added to the Methods:

Variability in prototypes. We estimated differences in the variability of prototypes as a function of area. For each evolution, we measured the dispersion of latent vectors within each generation. Dispersion was computed by obtaining the covariance matrix of all genes per generation (N ~ 40) and measuring the trace of the covariance matrix (the sum of the individual dimension variances). We saw that genetic dispersion started high and then decreased and settled around the 30th generation. To quantify this relationship, we fit an exponential decay function to these curves of the form $\mathrm{y(x)} = \mathrm{A} * e^{b*x} + \mathrm{c}$ to estimate the slope value *b*. To determine if the differences in slopes across areas could arise from the same distribution, we performed a randomization test. Specifically, to estimate the probability that the difference in slope between V4 and vlPFC could arise from the same underlying distribution, we combined the dispersion values from the V4 and vlPFC distributions, and then randomly sampled two groups from the mixed distributions, fitting each group with the exponential decay function, and measuring the difference between both slope values. We repeated this random re-sampling for 500 iterations and compared the actual observed difference to this mixed distribution.

4) The point that the authors make in the discussion that VLPFC neurons can drive adaptive image generators like XDream should not be understated. It also made me wonder if the authors might want to connect to the literature that has emphasized the attentional functions of VLPFC (see Rushworth et al., J Neuroscience, 2005 and related articles).

Thank you, this was an interesting study that was not part of our original background literature. The Rushworth *et al.* study focused on the roles of ventral and orbital prefrontal cortex in decision-making, specifically how these area's potential roles in attention related to action selection. They tested this by training monkeys to perform an action selection task, where each monkey was shown one colored object (copied in two locations), and the monkey was required to touch one of two alternative response boxes (left or right), based on the image identity. The interesting manipulation is that in some trials, the object images could be within the response boxes, and in other trials, the object images were gradually moved away from the response boxes. In other trials, distractor targets were introduced. They found that monkeys took longer to touch a response box if the object images were far away from the boxes, but that distractors did not have a reliable effect on the animals' responses. Then, they lesioned PFC (ventrolateral and orbital) in a subset of animals and determined the

effects on the performance. Interestingly, they found that the lesioned animals developed larger error rates in the appropriate selection of actions given visual stimuli, and this error rate was worse with larger spatial separations between object and response box. This article potentially demonstrates one reason for our findings: for PFC, it might be more efficient to select actions when visual stimuli are spatially associated with the response. In our study, we found that subsets of neurons in vlPFC encoded for coarse shapes independently of task training, suggesting that these neurons reflect a natural association between this specific visual shape and a motor movement — most likely, a saccade, since the vlPFC neurons' receptive fields were eccentric. This raises the possibility that, for this given location, some shapes are more likely to evoke saccades than others, and that the more similar the visual stimulus is to the encoded shape, the more likely the neuron is likely to influence the motor outcome. We are pursuing these experiments using a new set of animals, and we are excited to see the potential results.

We have added this to the Discussion section:

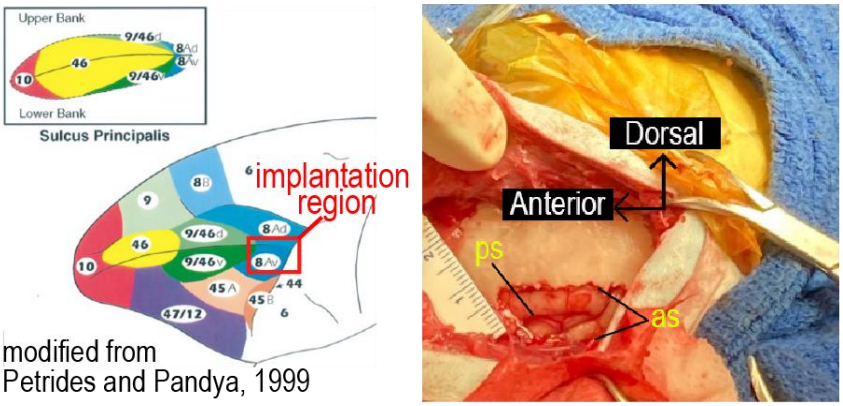
> One interesting possibility is that the purpose of visual encoding in PFC is to facilitate visually driven-action selection: studies such as Rushworth *et al.* (2005) have shown that ventrolateral and orbital PFC are important for linking spatially localized visual information (*e.g.,* object identity) with motor response selection (touching a box on the left or the right). Thus, for PFC, it might be more efficient to select actions when visual stimuli are spatially linked with the response. We found that subsets of neurons in vlPFC encoded for coarse shapes independently of task training, suggesting that these neurons reflect a natural association between this specific visual shape and a motor movement — most likely, a saccade, since the vlPFC neurons' receptive fields were eccentric.

Minor
The references to figures in the text are a little scarce throughout the manuscript. For instance, in the section describing the positional tuning of RF's in VLPFC, there were no call outs to figures that illustrated the effects being described. Consider including these.

Thank you, we have added all the missing call outs. Other Reviewers also noted this.

Could the authors provide a little more information about the exact location of the arrays in VLPFC? There are a number of cytoarchitectonic areas that have been identified in this location and it would be good to provide detail on which the authors think the arrays are in.

Happy to do so. In both monkeys, we exposed the principal sulcus and the arcuate sulcus surgically, then we aimed to implant the array close to the ventral/dorsal border marked by the principal sulcus, favoring the ventral aspect of the gyrus further from all blood vessels (**Response Fig. 17**).



Response Figure 17. (left) Areal map based on cytoarchitectonic parcellation of lateral PFC, from Petrides and Pandya (1999). **(right)** Intraoperative image of craniotomy for monkey C, indicating the principal sulcus (PS) and the arcuate sulcus at the margin of the craniotomy.

According to Petrides and Pandya (1999), our target location would have likely reached 8A, possibly at the border of 45A/45B/8A. We have added the following statement to the Results:

> The most likely location for the arrays was 8A, given their proximity to the ventral-dorsal border marked by the principal sulcus.

We thank our Reviewers for their time reading this work and helping us improve our contributions to the neuroscience community.

References

1. Rose, O., Johnson, J., Wang, B. & Ponce, C. R. Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. *Nat. Commun. 2021 121* **12**, 1–16 (2021).
2. Wang, B. & Ponce, C. R. Tuning landscapes of the ventral stream. *Cell Rep.* **41**, 111595 (2022).
3. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci.* **115**, E9015 LP-E9024 (2018).
4. Bardon, A., Xiao, W., Ponce, C. R., Livingstone, M. S. & Kreiman, G. Face neurons encode nonsemantic features. *Proc. Natl. Acad. Sci.* **119**, (2022).
5. Vinken, K., Prince, J. S., Konkle, T. & Livingstone, M. S. The neural code for “face cells” is not face-specific. *Sci. Adv.* **9**, eadg1736 (2023).
6. Van Essen, D. C., Anderson, C. H. & Felleman, D. J. Information processing in the primate visual system: an integrated systems perspective. *Science* **255**, 419–23 (1992).
7. Cao, R. *et al.* Feature-based encoding of face identity by single neurons in the human amygdala and hippocampus. 2020.09.01.278283 Preprint at https://doi.org/10.1101/2020.09.01.278283 (2022).
8. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8619–24 (2014).
9. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71**, 856–67 (1994).
10. Mishkin, M., Ungerleider, L. & Macko, K. Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* **6**, 414–417 (1983).
11. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–25 (1999).
12. Serre, T. Deep Learning: The Good, the Bad, and the Ugly. *Annu. Rev. Vis. Sci.* **5**, 399–426 (2019).
13. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).
14. Ungerleider, L. G., Gaffan, D. & Pelak, V. S. Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. *Exp. Brain Res.* **76**, 473–484 (1989).

15. Scalaidhe, S., Wilson, F. & Goldman-Rakic, P. Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: evidence for intrinsic specialization of neuronal coding. *Cereb. Cortex N. Y. N 1991* **9**, 459–475 (1999).
16. Riley, M. R., Qi, X.-L. & Constantinidis, C. Functional specialization of areas along the anterior–posterior axis of the primate prefrontal cortex. *Cereb. Cortex* **27**, 3683–3697 (2017).
17. Cong, S. & Zhou, Y. A review of convolutional neural network architectures and their optimizations. *Artif. Intell. Rev.* **56**, 1905–1969 (2023).
18. Erhan, D., Bengio, Y., Courville, A. & Vincent, P. Visualizing Higher-Layer Features of a Deep Network. *Tech. Rep. Univeristé Montr.* (2009).
19. Pu, S., Dang, W., Qi, X.-L. & Constantinidis, C. Prefrontal neuronal dynamics in the absence of task execution. *bioRxiv* 2022.09.16.508324 (2022) doi:10.1101/2022.09.16.508324.
20. Suzuki, H. & Azuma, M. Topographic studies on visual neurons in the dorsolateral prefrontal cortex of the monkey. *Exp. Brain Res.* **53**, 47–58 (1983).
21. Mikami, A., Ito, S. & Kubota, K. Visual response properties of dorsolateral prefrontal neurons during visual fixation task. *J. Neurophysiol.* **47**, 593–605 (1982).
22. Viswanathan, P. & Nieder, A. Comparison of visual receptive fields in the dorsolateral prefrontal cortex and ventral intraparietal area in macaques. *Eur. J. Neurosci.* **46**, 2702–2712 (2017).
23. Viswanathan, P. & Nieder, A. Visual Receptive Field Heterogeneity and Functional Connectivity of Adjacent Neurons in Primate Frontoparietal Association Cortices. *J. Neurosci.* **37**, 8919–8928 (2017).
24. Olah, C., Mordvintsev, A. & Schubert, L. Feature Visualization. *Distill* **2**, e7 (2017).
25. Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z. & Connor, C. E. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* **11**, 1352–1360 (2008).
26. Goodfellow, I. *et al.* Generative Adversarial Nets. in *Advances in Neural Information Processing Systems* (eds. Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q.) vol. 27 (Curran Associates, Inc., 2014).
27. Dosovitskiy, A. & Brox, T. Generating Images with Perceptual Similarity Metrics based on Deep Networks. *Adv. Neural Inf. Process. Syst. NIPS* (2016).

28. Brock, A., Donahue, J. & Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *7th Int. Conf. Learn. Represent. ICLR 2019* 1–35 (2018).
29. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* vols 2019-June 4396–4405 (IEEE Computer Society, 2019).
30. Robinson, D. A. & Fuchs, A. F. Eye movements evoked by stimulation of frontal eye fields. *J. Neurophysiol.* **32**, 637–648 (1969).
31. Bruce, C. J., Goldberg, M. E., Bushnell, M. C. & Stanton, G. B. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J. Neurophysiol.* **54**, 714–734 (1985).
32. Schall, J. D. Neuronal activity related to visually guided saccades in the frontal eye fields of rhesus monkeys: comparison with supplementary eye fields. *J. Neurophysiol.* **66**, 559–579 (1991).
33. Schwedhelm, P., Baldauf, D. & Treue, S. Electrical stimulation of macaque lateral prefrontal cortex modulates oculomotor behavior indicative of a disruption of top-down attention. *Sci. Rep.* **7**, 17715 (2017).
34. Fujita, I., Tanaka, K., Ito, M. & Cheng, K. Columns for visual features of objects in monkey inferotemporal cortex. *Nature* **360**, 343–6 (1992).
35. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci. Off. J. Soc. Neurosci.* **17**, 4302–11 (1997).
36. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–4 (2006).

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

I appreciate the effort that the authors put into responding to all my comments. Particularly, the response about the novelty of the study and the changes the authors introduced in the introduction to support the motivation and goal of the study have substantially improved the original text. However, I am still not convinced that the recording approach and the data yield can support several of their conclusions.

Although the authors claim that older studies also reported small proportions of neurons with specific functional properties, my initially expressed concern was more about the small sample rather than the small proportions. Finding 2 out of five sites (not necessarily neurons) doing something is inconclusive and the analysis should not go into the paper. Sample sizes smaller than 10 sites, using a small array limited in a small patch of cortex within the large PFC with no clear anatomical localization (or other functional features that could help characterize the region) I am afraid are not reliable for drawing conclusions about PFC functional properties. This together with the relatively low recording quality can introduce confounding factors that I am afraid cannot be easily controlled by offline analysis. This was one of my major concerns from the beginning and it still holds. The authors responded that their sampling is well-within the range of classic studies of visual and prefrontal cortex but this is an unfair statement considering the detailed single unit past studies that would sample dozens of neurons to end up reporting 10-20 neurons with specific properties. Moreover, it is clear from the authors' responses that several of the questions they ask cannot be answered with the existing dataset including the dependence on gaze position and the clustering of sites with spatial tuning.

Considering the other points of my original review, I still think that the authors could have done a better job with the analysis. For example, they could have used similar window sizes for the assessment of visual responses by comparing the visual stimulation period to the initial fixation period in each trial. Moreover, the one-way ANOVA used for the assessment of gaze gain fields is still hard to conceive despite the explanations and additions. Why not just use a two way ANOVA with factors gaze position and stimulus position and assess the depndence on each factor and their interaction? The question is whether the visual response and RFs change with gaze position and this should clearly answer it.

The important and perhaps counterintuitive finding that PFC sites can synthesize highly activating stimuli using image generators is both novel and intriguing. However, it would make more sense if this could be compared to the visual cortex by assessing what proportion of sites in different areas have this capacity, differences in the organization of neurons with this capacity in different areas,

differences in the prototypes etc. Most functions are distributed in the brain, but it is quite clear that not all areas do the same job. I think any finding in this direction appropriately supported by the data would be pushing forward our understanding. I am very sympathetic to the authors work and effort and I appreciate the enthusiasm they share but I am afraid the end result is not as convincing as I would expect it to be for publication in Nature Communications.

Reviewer #3 (Remarks to the Author):

I thoroughly enjoyed reading the authors very extensive and thoughtful responses to my comments. I didn't expect such a robust set of analyses to my queries and the additional CCN modeling was quite an impressive approach to a tricky question. No further comments or questions.

We are grateful to our Reviewers for their insightful feedback. We are happy for another opportunity to continue this conversation, as it has consistently led to improvements of this manuscript.

In this document, the original reviewer comments will be in Verdana, dark red font. Our responses will be in Arial, black. All changes to the manuscript and figures will be listed here for the convenience of the reviewers, and changed regions in the main manuscript will be in Arial, blue font.

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

I appreciate the effort that the authors put into responding to all my comments. Particularly, the response about the novelty of the study and the changes the authors introduced in the introduction to support the motivation and goal of the study have substantially improved the original text. However, I am still not convinced that the recording approach and the data yield can support several of their conclusions. Although the authors claim that older studies also reported small proportions of neurons with specific functional properties, my initially expressed concern was more about the small sample rather than the small proportions. Finding 2 out of five sites (not necessarily neurons) doing something is inconclusive and the analysis should not go into the paper. Sample sizes smaller than 10 sites, using a small array limited in a small patch of cortex within the large PFC with no clear anatomical localization (or other functional features that could help characterize the region) I am afraid are not reliable for drawing conclusions about PFC functional properties. This together with the relatively low recording quality can introduce confounding factors that I am afraid cannot be easily controlled by offline analysis. This was one of my major concerns from the beginning and it still holds. The authors responded that their sampling is well-within the range of classic studies of visual and prefrontal cortex but this is an unfair statement considering the detailed single unit past studies that would sample dozens of neurons to end up reporting 10-20 neurons with specific properties. Moreover, it is clear from the authors' responses that several of the questions they ask cannot be answered with the existing dataset including the dependence on gaze position and the clustering of sites with spatial tuning.

We agree with the Reviewer that the *Results* section assessing the gaze-position properties of vlPFC sites makes a key conclusion based on just two out of five multiunit sites in vlPFC. These sites were measured in only one monkey. Previously, we have stated the perspective that if potentially underpowered studies (due to low electrode counts) can robustly replicate phenomena across animals, then the *a priori* problem of statistical power is obviated. This is why we are comfortable publishing studies using 32-96-channel arrays, as have other teams in hundreds of studies in visual and cognitive neurophysiology. But in this case, the Reviewer is right that we did not replicate this particular experiment in both animals; so, we agree that we should at least remove these experiments from the main section. We would prefer to move it to the Supplemental Information section, not keep it from being seen by the community. While few, these were rigorously performed experiments: even *one* neuron in *one* monkey, if tested with appropriate controls, provides information that could be of benefit to the field. This is also consistent with the practice of open science. So, in response to this concern, we have removed this section from the main Results and attached them to a new **Supplemental Figure 2**. Further, to prevent new readers from potentially overinterpreting the conclusions, we added the following statement to the legend:

> These experiments were performed late in the lifetime of the arrays, so only vlPFC sites from one monkey passed criteria for examination (see Methods), and we could not obtain viable RFs from the second monkey's vlPFC arrays, only from its V4 array. These results are proffered as a proof-of-concept only.

Overall, it might serve the readers of these reviews to emphasize a key trade-off that should inform future studies: chronic FMAs are a popular, well-validated neurophysiology technique with alternative goals to single-electrode physiology, as it trades off maximizing the number of sampled single-units for the ability to record multiunit sites with high stability over time. We believe that both techniques yield distinct—yet complementary—results that are equally vital to understanding how the primate PFC contributes to visual recognition and visually-guided behavior.

Considering the other points of my original review, I still think that the authors could have done a better job with the analysis. For example, they could have used similar window sizes for the assessment of visual responses by comparing the visual stimulation period to the initial fixation period in each trial.

We hope that we made some progress on this point, by providing additional analyses on the parametric variation of window sizes in our previous revision. There, we confirmed that too-small window sizes can lead to small changes in overall population results, as expected by the fact that small window sizes have fewer spiking responses. However, the overall conclusions were the same. As reported (line 88),

> To control for baseline window sizes, we also measured responsiveness of PFC sites in separate experiments, using 5-10°-wide images at locations near the population RF. We focused on the change in activity at -150 to 0 ms before image onset vs. 50 to 200 ms after. We found that out of 434 recordings across days, 41.9% and 40.6% of sites showed responsiveness (using a shorter baseline window resulted in estimates of 46.6% and 43%).

In this revision, we will further tell the reader about this control analysis in the "Visual responsiveness" section of the *Methods* section:

> Visual responsiveness. To determine whether each vlPFC site had a significant change in firing rate in response to stimulation in different regions of the visual field, we divided the first 200 ms following image onset into time windows: the early window comprised the first 1:30 ms after image onset, while the late window period comprised 50-150 ms. We computed the mean firing rate in each time window per experiment, then performed a two-sided Student's t-test between the mean firing rates during the baseline windows and the evoked windows for each site. Finally, we corrected all resulting p-values for multiple comparisons using false-discovery rate. We also repeated this analysis with a longer time window for the baseline activity, collected over the fixation period.

Overall, we take this point seriously, so in this revision, we performed additional analyses to control for the length of the baseline window and its effects on major results. We focused on the reported percentage of image-selective sites at the $p < 0.04$ level (as reported in Supplemental Table 2) and did four variations. We measured the site activity during the visual stimulation period to four baseline periods: (1) the period [-149 0]-ms preceding the image onset, (2) the period [0 50]-ms after image onset, (3) the period when only the fixation spot was present on the screen, and in the final condition, (4) we simply did not remove a baseline and just counted events during the stimulation period. We present the results below.

While we see some variation in the final statistics, the overall takeaway is that image selectivity increases when baseline activity is *not* subtracted (this is because the baseline activity estimate is itself a noisy random variable, and that introduces uncertainty). To this point, shorter windows also reduce the estimated number of selective sites. Subtracting the mean activity within a 150-ms window occurring right before image onset or during fixation results in comparable estimates. In all cases, the overall conclusions were held: high frequency of selectivity in V4, none in CPB, and intermediate in PFC.

	Total recordings (32 unique sites/array)	**No. with RFs within 4.0° of stimulus center**	**No. selective at P<0.04**	**No. selective at P<0.04**	**No. selective at P<0.04**	**No. selective at P<0.04**
			Baseline window r/t image onset: -149–0 ms	Baseline window r/t image onset: none	Baseline window r/t image onset: 0–50 ms	Baseline window during fixation period (150 ms)
Monkey C						
V4	1568	208	88.9%	90.4%	89.4%	89.9%
vlPFC	1425	79	12.7%	21.5%	13.9%	13.9%
Monkey D						
CPB	1198	32	0.0%	0.0%	0.0%	0.0%
vlPFC	1252	139	23.0%	35.3%	21.6%	25.9%

We updated Supplemental Table 2 as above for our readers to examine. I think this has convinced us that the removal of baseline activity from the stimulation period may not be as helpful in the context of chronic

arrays, probably because the overall white band activity is stable. We will re-examine this practice going forward, and we thank the Reviewer for prompting this analysis.

Moreover, the one-way ANOVA used for the assessment of gaze gain fields is still hard to conceive despite the explanations and additions. Why not just use a two way ANOVA with factors gaze position and stimulus position and assess the depndence on each factor and their interaction? The question is whether the visual response and RFs change with gaze position and this should clearly answer it.

We appreciate the Reviewer's patience as we work to clarify the analysis as best as possible. First, as mentioned above, in this revision, we removed this analysis from the main Results section, moved it to a Supplemental Figure, and in the legend, we explicitly caution our readers about the limitations of the study. Further, we agree that the proposed ANOVA is a simple statistical test that would strongly complement the current analyses. Here is how we carried it out. As a reminder, we measured each array site's RF during three different gaze conditions (gaze fixation at the center of the screen, fixation to the left, and fixation to the right). If a site showed a statistically robust RF during the center gaze condition, the channel was considered "live," and we pursued further analysis. We collected each site's responses to individual stimulus presentations and created a table containing a stimulus position and a gaze condition, then ran a two-factor ANOVA assessing the statistical significance of the stimulus position on firing rate, the gaze position, and the interaction between both. The key statistic was the interaction term: did the site alter its RF position on the screen, tuning based on gaze? For a classically visual area such as V4, we would expect that to be the case. Based on five sessions involving monkey C, at a highly strict probability threshold ($P = 1x10^{-7}$), we found that in V4, 38 sites showed a statistically reliable RF, and of these, 65.8% showed an interaction between RF and gaze. We interpret this to be a noise ceiling, given the retinocentric nature of V4 RFs. Monkey C PFC array was not functional. In contrast, monkey D's CPB array showed zero sites with statistically reliable RFs. Its PFC array showed 10 sites with robust RFs (also $P = 1x10^{-7}$), and of these 50% showed an interaction with gaze position. Consistent with our previous analyses of this data, allowing noisier channels to contribute by lowering the P threshold to 0.06 reduces the interaction to 30%. We interpret these results as suggesting the most robust RFs in V4 and PFC are retinocentric, that is, anchored to the position of the fovea.

We have amended the Supplemental Information pertaining to this analysis as follows:

> Analyzing responses of sites across days, we applied a two-way ANOVA analysis for each site (with factors of stimulus position and gaze position). In V4, 38 site-days showed a statistically reliable RF, and of these, 65.8% showed an interaction between RF and gaze V4 ($P < 1x10^{-7}$). In CPB, no sites showed statistically significant RFs. The remaining PFC array showed 10 site-days with robust RFs (also $P < 1x10^{-7}$), and of these 50% showed an interaction with gaze position; these values decreased as the P-value threshold was loosened, down to 30% of PFC sites with $P = 0.06$. Overall, we interpret these results as suggesting the most robust RFs in V4 and PFC were retinocentric, that is, anchored to the position of the fovea. These experiments were performed late in the lifetime of the arrays, so only vlPFC sites from one monkey passed criteria for examination (see Methods), and we could not obtain viable RFs from the second monkey's vlPFC arrays, only from its V4 array. These results are proffered as a proof-of-concept only.

The important and perhaps counterintuitive finding that PFC sites can synthesize highly activating stimuli using image generators is both novel and intriguing. However, it would make more sense if this could be compared to the visual cortex by assessing what proportion of sites in different areas have this capacity, differences in the organization of neurons with this capacity in different areas, differences in the prototypes etc. Most functions are distributed in the brain, but it is quite clear that not all areas do the same job. I think any finding in this direction appropriately supported by the data would be pushing forward our understanding. I am very sympathetic to the authors work and effort and I appreciate the enthusiasm they share but I am afraid the end result is not as convincing as I would expect it to be for publication in Nature Communications.

We fully agree that future work will need to build upon this initial finding that PFC sites are able to generate highly activating stimuli *de novo*, as this present study opens a new line of investigation that needs to address the proportions and subregions of PFC that are driven by low-level visual attributes. We do not believe these findings fully solve the question; in fact, this study raises more questions —we consider that a sign of a successful discovery.

However, these new lines of investigation are predicated on demonstrating initial proof that subregions within the primate PFC can be robustly driven by low-level image generators, not just by trained visuocognitive tasks or complex images of faces. We hope that the novelty of this discovery launches further investigations in other labs, and it has done in our own.

Reviewer #3 (Remarks to the Author):

I thoroughly enjoyed reading the authors very extensive and thoughtful responses to my comments. I didn't expect such a robust set of analyses to my queries and the additional CCN modeling was quite an impressive approach to a tricky question. No further comments or questions.

We would like to thank our Reviewer again for their comments and feedback. We are delighted to have successfully addressed your initial concerns, and we greatly appreciate the time and energy invested in the review.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Thanks for giving me the opportunity to cross-comment for Reviewer 2. Reviewer 2 asked critical questions, and his/her suggestions greatly improved the quality and clarity of this manuscript. I also learned a lot from his insightful questions. Needless to say, I cannot represent Reviewer 2, but I just try my best to give my comments.

1. Regarding the concerns about 'small sample size' and 'low recording quality', the authors have removed related results to the supplemental information section and added explanations in the main text to prevent overinterpreting their conclusions. I think the authors have addressed the concerns properly.

2. For the suggestion about 'better data analysis', the authors have followed the suggestion and did the new analysis. The new results from this analysis were added in the supplemental information section. The new results do show some differences, but the conclusions stay the same. I think this issue is addressed.

3. For the suggestion about 'using two-way ANOVA', the authors followed the suggestion. The main conclusions from the two-way ANOVA analysis stay the same. Related results and explanations are added to the supplemental information section. I think this issue is also addressed.

4. Regarding the concerns about 'not convincing' to be published in Nature Communications. Although this study mainly focused on PFC, which seems not 'complete' enough, I agree with the authors that this initial discovery can encourage more investigations as Reviewer 2 suggested. I think publishing this study in high-impact journals, like Nature Communications, will strongly encourage related research.

Thank you to our Reviewer for their valuable time. Reviewer comments will be in **bold**, and our responses in regular Aptos text.

Reviewer #1 (Remarks to the Author):

Thanks for giving me the opportunity to cross-comment for Reviewer 2. Reviewer 2 asked critical questions, and his/her suggestions greatly improved the quality and clarity of this manuscript. I also learned a lot from his insightful questions. Needless to say, I cannot represent Reviewer 2, but I just try my best to give my comments.

1. Regarding the concerns about 'small sample size' and 'low recording quality', the authors have removed related results to the supplemental information section and added explanations in the main text to prevent overinterpreting their conclusions. I think the authors have addressed the concerns properly.

2. For the suggestion about 'better data analysis', the authors have followed the suggestion and did the new analysis. The new results from this analysis were added in the supplemental information section. The new results do show some differences, but the conclusions stay the same. I think this issue is addressed.

3. For the suggestion about 'using two-way ANOVA', the authors followed the suggestion. The main conclusions from the two-way ANOVA analysis stay the same. Related results and explanations are added to the supplemental information section. I think this issue is also addressed.

4. Regarding the concerns about 'not convincing' to be published in Nature Communications. Although this study mainly focused on PFC, which seems not 'complete' enough, I agree with the authors that this initial discovery can encourage more investigations as Reviewer 2 suggested. I think publishing this study in high-impact journals, like Nature Communications, will strongly encourage related research.

We appreciate these comments, and we are happy that they find that our responses clarifying. We hope that this study will encourage future research into the visual processing properties of prefrontal cortex; we are certain to continue these studies into the future.