Peer Review File

Manuscript Title: Cortical reactivations predict future sensory responses

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

In this manuscript, Nguyen et al. propose that cortical reactivation events explain ongoing changes in population activity patterns, also referred to as representational drift, in the mouse visual cortex. The authors approach the topic by studying the dynamics of, and interactions between, evoked and spontaneously occurring activity patterns using two-photon calcium imaging. They recorded from thousands of neurons across several visual cortical areas, finding that (1) with repeated visual stimulus presentations, the neuronal representations of different stimuli became more dissimilar, and (2) following evoked activity, neurons showed reactivation (replay) events that reflected recent visual experience. The rate of these reactivations depended on local circuit activity, and their pattern was predictive of the eventual orthogonalized evoked response patterns that were observed later in the sessions. Based on their data, the authors argue that reactivations in visual cortex provide a template that instructs the neuronal population to orthogonalize evoked responses, supporting pattern separation.

The central theme in this study is the role of replay/reactivation in shaping future cortical response patterns. Replay has been extensively studied in hippocampal-entorhinal circuits, but much less in neocortex, even though that might be key to understand mechanisms for systems consolidation. The authors here tie together the topics of replay/reactivations and representational drift, giving it a refreshing spin and proposing that the purpose of the reactivations is to direct the ongoing changes in cortical activity patterns. Thereby, the study provides interesting insights into how reactivations could shape ongoing and learning-related changes of visual cortex neurons over longer periods of time. The manuscript is a pleasure to read, the experimental data appears of high quality and the analysis, including the detection of reactivation events, is done precisely and rigorously. However, I have concerns about the suggested causal relation between the observed reactivations and the representational drift, about the statistical power of control analyses and about how the overall findings relate to the literature on hippocampal replay/reactivation. Therefore, my enthusiasm for this manuscript has been tempered somewhat.

Main points

1. The authors argue that the reactivations are instructive in re-organizing the sensory population responses. However, I wonder if the results could also be explained by early in the session having an extra, stimulus non-specific, e.g. top-down attention/novelty response in addition to the evoked population response. Such a novelty/attention response could wear off with increasing trial number and the more orthogonal stimulus-specific responses would remain. Reactivations, on the other

hand, would only comprise the stable/sensory part of the representation, thus being more similar to, and seemingly predicting, the evoked representation later in the session. And the rate of reactivations could provide additional drive for scaling down the attention/novelty component. Such a mechanism could also fit in the framework suggested on lines 392-396, but would be conceptually very different from what is presented in the manuscript.

2. In my opinion, representational drift primarily refers to a neuronal population having a stable information content across time, but having the information carried by increasingly different sets of neurons. However, the data shown here suggests that the information content (discriminability) increases with time, thus not merely drifts.
The authors suggest that the changes in evoked responses are the result of "rapid plasticity that depends on the instantaneous content and rate of reactivations" (line 330). Following the presented line of thought, the evoked responses should then continue their representational drift on the next session, thus showing progressive changes not just within, but also across sessions. At the same time, it appears that the evoked patterns reach a steady-state of maximal orthogonalization within a session, and then re-initialize to non-orthogonalized states on the next session. The authors should provide more insight into this process by following the same neurons across sessions. In addition, I wonder if the steady-state would also be reached when stimuli are more similar to each other (the stimuli currently used seem orthogonal, but most natural visual experiences are not).

3. The behavioral tracking is relatively limited (eye tracking of one eye, brain motion, n=4), making it difficult to assess how arousal/behavioral changes covary with some of the reported effects. In addition, the stimulation paradigm and time course of the effects are somewhat reminiscent of SRP (Frenkel et al., Neuron, 2006) and the vidget (Cooke et al., Nat Neurosci, 2015). Finally, the authors suggest that stimulus salience plays a role (paragraph starting at line 182) and put some emphasis on "unexpected stimuli" inducing the reactivations (line 346). The role of stimulus salience and arousal/behavior in the observed representational drift should be investigated in more detail: besides increasing the n and improved video tracking, the authors could also think of experimentally manipulating arousal and/or stimulus expectation.

4. The authors link their findings to hippocampal reactivation/replay. To make this point, it should be investigated whether the cortical reactivation events align with the occurrence of sharp wave-ripple complexes.

5. Throughout the manuscript, the number of animals used is generally quite low, with n equaling often 4 and sometimes even fewer mice. This might be not so much of an issue when testing large effects, but certainly is a problem when no significant difference is observed. The low n results in low statistical power, making tests prone to type II errors. This could potentially result in the failure to detect significant differences in control analyses, influencing the overall conclusion of the study.

Minor
1. The authors pooled data from layer II to IV. Were there no differences between layers?
2. Line 131: Could more ensembles be detected when not only focusing on stimulus-driven neurons? Possibly ensembles of other, non-experimentally induced experiences, reactivating during the baseline session?

3. Why did the authors choose this specific retinotopically patterned stimulus?
4. Local inhibition reduced the stimulus reactivation rates, but the remaining reactivation rates were still much higher than baseline, why?
5. If optogenetic inhibition halts/slows the instructive effect of the reactivations, was this also accounted for in the model?
6. In the discussion the authors mention that the purpose of reactivations could be to "help enhance stimulus discrimination". Did the authors look at reactivations following learning? This might make the observed changes in activity patterns more interpretable.
7. The authors chose to use parametric tests throughout the manuscript, is this supported by any statistical indication?
8. Fig 1b: According to the legend, the white box and FOV image (left) should have the same size, but the scale bars indicate that the white box is about twice the size of the FOV.
9. Fig 3a and 3c: It would be useful to see the baseline values in the plots as well.

Referee #2 (Remarks to the Author):

This manuscript by Nguyen et al. examines patterns of activation and reactivation of neurons in the lateral visual cortex. Cortical glutamatergic neurons were imaged in head-fixed mice as they were presented with one of two stimuli, S1 and S2, interspersed by ~1min long inter-trial intervals, allowing for quantification of stimulus-evoked responses and reactivations during the ITI. The authors found individual neurons and population level responses that demonstrated stimulus-specific reactivations during the ITI. As has been previously reported, the authors found that reactivations are biased toward salient recent events. But departing from previous reports, the authors advance a hypothesis that the reactivations do not function in stabilizing sensory representations but rather predict and guide representational drift toward future states. Toward this, the authors demonstrate that "Decrease neurons" that eventually lose their sensory-evoked tuning, show a strong reduction in such tuning in the ITI periods, evident even on the very first ITI. However, such effects were less prominent for "Increase neurons". The authors then developed a model that takes in the stimulus evoked response on the first trial, and ITI responses across all trials, to accurately predict the eventual re-organization of evoked responses on the last trials. Overall, these findings are interesting and may be important for understanding how reactivations contribute to plasticity in the visual cortex. In its current form, however, the data to support the claim that reactivations predict future sensory-evoked reorganization are underdeveloped.
The major concern is that the reactivation results presented here could be a trivial result where neurons on average tend to have weak reactivations (strong drop in activity even by the first ITI), thus appearing as though the reactivation actively contribute to the reorganization and progressive loss of tuning of "Decrease" neurons. The small magnitude effects on "Increase" neurons compounds this problem. The results would be more convincing if the reactivation not only predicted loss of tuning, i.e. "Decrease" reorganizations, but also predicted active changes in tuning, such as from strongly S1 tuned to strongly S2 tuned, or even better if the stimulus set was larger (S1 to S5) where the increased dimensionality of the stimulus set and population responses would provide greater predictive power.
Related to the above, causal manipulations in the ITI are necessary to make strong claims about

whether reactivations actively lead to re-organization of evoked responses on later trials.

Other Points:

1. When justifying the scale of the imaging, it might be important to look at false positive/negative rates when selecting a random contiguous region of the overall imaging planes (rather than a random sample across the entire volume), to better simulate what a smaller ROI would not be able to capture. Perhaps reactivating cells are spatially correlated and the false positive rate would be lower.

2. When using a classifier to assign probabilities of reactivation patterns to synchronous activity (Extended data Figure 1A), the training set includes frames during stimulus presentation as well as frames during the ITI without synchronous activity. The authors pre-process the data to remove slow temporal correlations in the times between synchronous events because there still seems to be persistent activity patterns throughout the ITI in the raw data. If the classifier is trained only on frames during stimulus presentation and tested on every frame of the ITI, is the reactivation probability still significantly higher during synchronous events when compared to frame-shuffle data?

3. In the simultaneous imaging and optogenetic silencing experiments, 50% of trials are inhibited. How does this affect the learning rate/plasticity for these mice? It would be important to see the analyses in Figure 4 repeated for these inhibited mice. Specifically - does reducing the total number of reactivation events causally decrease the learning rate or change the number of Increase/Decrease neurons?

4. It appears as though “Decrease” neurons have a higher baseline activity level and “Increase” neurons have a lower baseline activity level compared to “No Change” neurons. Is this effect significant? Are there differences in activity before any stimuli are presented? Could the pattern of activity observed in Fig. 4C suggest that the ITI reactivations promote tendency toward a “homeostatic baseline” rather than toward future reorganization?

5. What was the distribution of neurons recorded across the mice? i.e., how many neurons did each mouse contribute to the “pool” of neurons examined here? What is the total number of neurons that were S1-responsive, S2-responsive, or responsive to both at the beginning and end of the sessions, and was this similar across mice? In general, total neuron counts (including increase/decrease neurons, etc.) should be included.

Referee #3 (Remarks to the Author):

In this exciting study, Nguyen and colleagues investigate cortical reactivation of sensory stimuli evoked activity patterns in the mouse lateral visual cortex with large-scale two-photon calcium imaging and optogenetic manipulations. The authors report that presentation of visual stimuli was followed by transient, stimulus-specific reactivations during the following minute. Optogenetic silencing of local circuits during sensory presentations abolishes reactivations. The authors

intriguingly find that the content of reactivation was different from preceding stimulus responses such that reactivations predict future stimulus responses. Based on these observations, the authors propose that reactivations guide representational drift supporting sensory discriminations.

Overall, this is an interesting and elegant study reporting potentially highly impactful observations related to offline cortical dynamics supporting sensory discrimination. The large-scale imaging approach is impressive as it uniquely allows the authors to identify cortical reactivation patterns. My comments below are aimed at improving the accessibility of this impressive study to the broader audience at Nature.

Major

1. Most reactivation seems to occur during the first 10-20 trials (Figure 1B). This early phase of stimulus presentation might be associated with the different behavioral states, particularly increased arousal, especially since the visual stimuli are not reinforced. This issue warrants a bit more scrutiny. Is there evidence and an effect of increased arousal at the beginning of each day of experiments resulting in higher baseline firing rates, especially in the first 10-20 trials to which the animals subsequently habituate? Especially given that 'Decrease' neurons account for most of the orthogonalization between the two stimuli (line 257). Variations in the animals' behavioral state and arousal may play an important role in trial-to-trial variations pushing cortical patterns away from the preferred manifold. This could be considered in the discussion (line 394), and there are a few places in the results section where this possibility should be considered:
a. Figure 1E: Only shows behavior during reactivations (which could be excluded based on behavioral states, line 966). Is there any evidence for increased arousal during stimulus during the early trials as opposed to the later ones? The authors could analyze pupil diameter, attempted running, and brain motion data across stimulus presentations.
b. Figure 3A, line 188: Is there a different model that fits the data better than linear regression? If a period of increased arousal transiently drives increased activity this might be expected to follow an exponential decay back to some baseline rate, which wouldn't be significantly fit by linear regression but is still significant. Is there a significant change in stimulus-evoked activity in the early template (first 3 stimuli) vs the late (last 3 stimuli)?
c. Figure 3A Inset: is there any evidence for individual differences in recording quality or mouse behavior that account for why the orange mouse has such a lower activity rate than the others (also see Figure 3C inset)?

2. The authors should provide a more detailed analysis to see if Increase and Decrease neurons differ in any physiological or anatomical properties (i.e., baseline firing, cortical layer location, subregion position within lateral visual cortex). Are there differences in functional connectivity structure between these functional cell types? In their previous study, the authors implemented an informative analysis of this matter (Sugden et al. Nat Neurosci, 2020). Thus it would also be informative to apply a similar analysis here. Relatedly, it would be informative to know if reactivations exhibit spatial drift over time across LVC areas and if reactivations are preferentially expressed in deeper layers or uniformly.

3. The study implicates a causal role of reactivation in driving representational drift. Of course, one

missing piece here is some local circuit manipulation to test this necessity more directly. I understand the authors' argument that such experiments (i.e., selective optogenetic perturbation of reactivation events in LVC) should most ideally be done in a closed-loop design, which is a challenging experiment. Nevertheless, I wonder if the temporal bias in reactivation, such that it peaks in the first 10-20 min and that it is also stronger during the first half of the ITI (Figure 1B), still provides a possible way to test for the idea that reactivation drives drift causally. Perhaps silencing can be targeted early during the session (i.e., first 10-20 min) and early during ITI (i.e., 20-30 sec)?

Minor:

1. The current study builds upon the group's previous work (Sugden et al. Nat Neurosci, 2020), so it would be relevant to add more discussion about the similarities and differences between the present and prior works. Especially since, in the current work, neutral stimuli without reinforcement elicit reliable reactivation, while in the previous study, it was mainly associated with reinforcement (related to major point 1 above).
2. The authors should consider acknowledging other studies in the hippocampus demonstrating that reactivation content significantly differs from online experience (e.g., Terada et al., Nature, 2021).
3. Line 131: How stable is the stimulus-preferring population between early and late trials (how similar are the top 5% responsive neurons in the early trials to the top 5% responsive in the late trials)? and what is the overlap between S1 and S2 preferring populations?
4. Figure 1E (bottom)/line 872: Reporting brain motion in terms of shifts without showing the corresponding 2D correlation assumes that all registrations were equally successful. What methods were used to check the quality of the motion correction or search for out-of-frame (z-axis) movement?
5. Figure 2: Since reactivation rate is related to the perception of the visual stimulus, are there poorly attended stimuli presentations coinciding with either behavioral markers of attention or stimulus-response strength that could be related to a deficit in reactivation rates on an individual trial basis?
6. Figure 2B, F: at which points (both in time relative to the first stimulus, left, and in time relative to each stimulus onset, right) is the reactivation rate significantly higher than the baseline rate?
7. Figure 2G: The inactivation abolishes the bias to the previous stimulus when there was inhibition, but is there a bias in reactivation rates toward the previous non-inhibited stimulus? Or does the silencing completely "reset" the reactivations back to random/baseline levels?
8. Figure 1B: It's unclear how the color scale "Azimuth" relates to the retinotopic mapping either an explanation should be added or a citation for Zhuang, J. et al. added to the main text (line 125) so it is clearer where to find this information.
9. Extended Data Figure 2C: The caption lists panels left and right, but they should be top and bottom.
10. Extended Data Figure 2C (bottom): It would be more straightforward to show the entire activity trace (e.g., Extended Data Figure 2I) and highlight the timepoints included in the reactivation analysis.
11. Figure 3 F, G, Extended Data Figure 3B, Figure 4I, J: Significance annotations in the upper margin of the figure, in some cases pointing to the legend box, are difficult to read.
12. Figure 3H: Report the correlation score so it is clear how this example relates to the mouse averages reported in Figure 3I.

Author Rebuttals to Initial Comments:

We thank all three reviewers for taking the time to carefully read our manuscript, for their appreciation of our study, and for their constructive comments. The experiments and analyses that we have performed as suggested by the reviewers have substantially improved our revised manuscript.

Each reviewer comment is shown below in **bold**, followed by our response.

All text changes made in the revised manuscript are colored blue. New figures added to the revised manuscript are labeled in blue.

In this response to reviewer comments, we labeled in blue the pointers to lines in the revised manuscript where new figures in the main text are discussed, or where other new text has been added to the revised manuscript. New figures that were not added to the main manuscript but that are shown in this response to reviewer comments are labeled in green, below.

Reviewer 1 (Remarks to the Author):

In this manuscript, Nguyen et al. propose that cortical reactivation events explain ongoing changes in population activity patterns, also referred to as representational drift, in the mouse visual cortex. The authors approach the topic by studying the dynamics of, and interactions between, evoked and spontaneously occurring activity patterns using two-photon calcium imaging. They recorded from thousands of neurons across several visual cortical areas, finding that (1) with repeated visual stimulus presentations, the neuronal representations of different stimuli became more dissimilar, and (2) following evoked activity, neurons showed reactivation (replay) events that reflected recent visual experience. The rate of these reactivations depended on local circuit activity, and their pattern was predictive of the eventual orthogonalized evoked response patterns that were observed later in the sessions. Based on their data, the authors argue that reactivations in visual cortex provide a template that instructs the neuronal population to orthogonalize evoked responses, supporting pattern separation.

The central theme in this study is the role of replay/reactivation in shaping future cortical response patterns. Replay has been extensively studied in hippocampal-entorhinal circuits, but much less in neocortex, even though that might be key to understand mechanisms for systems consolidation. The authors here tie together the topics of replay/reactivations and representational drift, giving it a refreshing spin and proposing that the purpose of the reactivations is to direct the ongoing changes in cortical activity patterns. Thereby, the study provides interesting insights into how reactivations could shape ongoing and learning-related changes of visual cortex neurons over longer periods of time. The manuscript is a pleasure to read, the experimental data appears of high

quality and the analysis, including the detection of reactivation events, is done precisely and rigorously. However, I have concerns about the suggested causal relation between the observed reactivations and the representational drift, about the statistical power of control analyses and about how the overall findings relate to the literature on hippocampal replay/reactivation. Therefore, my enthusiasm for this manuscript has been tempered somewhat.

We appreciate the reviewer's recognition of the quality of our data/analysis and the value of our effort to bridge together findings related to different fields of systems neuroscience. We have addressed each of the reviewer's concerns in detail below.

Main points

1. The authors argue that the reactivations are instructive in re-organizing the sensory population responses. However, I wonder if the results could also be explained by early in the session having an extra, stimulus non-specific, e.g. top-down attention/novelty response in addition to the evoked population response. Such a novelty/attention response could wear off with increasing trial number and the more orthogonal stimulus-specific responses would remain. Reactivations, on the other hand, would only comprise the stable/sensory part of the representation, thus being more similar to, and seemingly predicting, the evoked representation later in the session. And the rate of reactivations could provide additional drive for scaling down the attention/novelty component. Such a mechanism could also fit in the framework suggested on lines 392-396, but would be conceptually very different from what is presented in the manuscript.

We agree with the reviewer that a possible driving factor could be top-down attention and/or novelty responses that wear off with increasing trial number, which could work in unison with reactivations to predict and/or drive changes in stimulus responses and representational drift across time. However, based on new analyses of the data and evidence derived from existing analyses, we believe the effects of a novelty signal that gradually decays across trials may be small relative to other potential effects of reactivations, for the following reasons:

1)

The overall magnitude of the stimulus-evoked response averaged across all neurons is stable across trials (Fig. 3c). Therefore, while the patterns of the evoked stimulus responses are changing across time, average stimulus responses do not indicate a decreasing global signal. However, Increase neurons in our dataset (those that increase their visual responses across trials) could compensate for any decrease in response in

other neurons related to a drop in novelty/attention (Fig. 3f). Thus, we ran additional analyses, below.

2)

We had identified subsets of neurons that show a significant decrease across trials in their responses to at least one of the two stimulus types (which we labeled Stimulus 1 Decrease neurons and Stimulus 2 Decrease neurons). Decrease neurons, as a population, develop increased stimulus selectivity across trials, reflected in a decrease in similarity of Stimulus 1 and Stimulus 2 response patterns across trials (Fig. 3g). This is because approximately twice as many Decrease neurons increased their selectivity to one or the other stimulus as compared to Decrease neurons that decreased their stimulus selectivity from early to late trials (Fig. 3h, discussed in Lines 293 and 295; by contrast, similar numbers of Increase neurons increased or decreased their stimulus selectivity). This is inconsistent with a novelty-related change in responsivity (which should not exhibit differential changes in the magnitude of responses to each stimulus across trials, as both stimuli are equally novel due to the randomization of trials, balanced numbers of presentations of each type, and matched low-level visual features such as contrast).

Below, we discuss why we do not believe that the increase in the overall stimulus-selectivity of Decrease neurons is largely driven by top-down attention and/or novelty responses, since the proportion of Decrease neurons that showed a similar decrease in stimulus-evoked responses to both S1 and S2 was quite low, only ~15% of all Decrease neurons.

3)

To quantify the effects of global novelty and/or top-down attentional effects on changes in stimulus responses across trials in our data, we performed the additional analyses below:

We identified Decrease neurons that had a similar absolute or percent decrease in activity for both stimuli (defined as those with no difference in their decrease in response to S1 or to S2 in either absolute units of normalized Ca^{2+}/s or in percent change in activity from early to late trials; $p>.05$, two-tailed paired t-test). These neurons could indeed be modulated by top-down attention and/or novelty, in line with the reviewer's hypothesis. Such neurons (which we refer to as Non-differential Decrease neurons) account for only ~15% of all Decrease neurons (Extended Data Fig. 6g, discussed in Line 255). Therefore, the vast majority of Decrease neurons selectively

decrease their response magnitude to one stimulus more than to the other stimulus across trials.

To examine the degree to which Non-differential Decrease neurons might affect the reorganization of sensory responses, we removed these Non-differential Decrease neurons from the response similarity analysis and found only very minor and non-significant changes relative to when including all neurons (Extended Data Fig. 6h, discussed in Line 255). When we removed Non-differential Decrease neurons from the analysis that previously showed representational drift of stimulus responses but more stable reactivation patterns across trials (Fig. 4b), we observed that Non-differential Decrease neurons don't contribute to overall changes in this drift across trials (Extended Data Fig. 9e, discussed in Line 330).

From these new analyses, we conclude that novelty and/or global top-down attention (potentially driven in part by Non-differential Decrease neurons whose activity profile is likely driven by general stimulus novelty) on stimulus response patterns and stimulus response orthogonalization across trials is minor. We now consider this question in detail and show these new results (itemized above) in the revised manuscript.

2. In my opinion, representational drift primarily refers to a neuronal population having a stable information content across time, but having the information carried by increasingly different sets of neurons. However, the data shown here suggests that the information content (discriminability) increases with time, thus not merely drifts. The authors suggest that the changes in evoked responses are the result of "rapid plasticity that depends on the instantaneous content and rate of reactivations" (line 330). Following the presented line of thought, the evoked responses should then continue their representational drift on the next session, thus showing progressive changes not just within, but also across sessions. At the same time, it appears that the evoked patterns reach a steady-state of maximal orthogonalization within a session, and then re-initialize to non-orthogonalized states on the next session. The authors should provide more insight into this process by following the same neurons across sessions. In addition, I wonder if the steady-state would also be reached when stimuli are more similar to each other (the stimuli currently used seem orthogonal, but most natural visual experiences are not).

We fully agree with the reviewer that it would be valuable to follow the same neurons across days both to further test our claim that the observed changes reflect representational drift, and to further test our claims about the relationship between representational drift and the content of reactivations.

We had originally collected much of our data with a strong attempt to image the same fields of view each day to maintain consistency. Thus, we now considered the possibility that we could track the same neurons across days and sessions for N = 5 mice.

We first attempted to use existing cross-day alignment algorithms such as CellReg (https://github.com/zivlab/CellReg), however, the algorithm did not work well with our large fields of view and datasets. Therefore, we decided to test out a recent and more powerful software package using machine learning methods in collaboration with Richard Hakim, a PhD student in the lab of Bernardo Sabatini at Harvard University. Our data (consisting of hundreds of thousands of cell masks derived from Suite2p's cell extraction algorithm) was used to develop this algorithm, which reliably tracks cells across large sets of days (see Methods, Lines 1393-1409, "ROICaT", unpublished). Using ROICaT (which assigns quality metrics to identify clusters of similar neurons across days), together with careful manual inspection, we were successful in aligning cells across 6 consecutive days in our datasets (Extended Data Fig. 7a-b, discussed in Lines 270 and 334). We limited the alignment to 6 days because all N = 5 mice had at least 6 consecutive days of imaging data and we wanted as high of a number of mice as possible for the cross-day analysis.

We then performed the suggested analysis of testing, now for the same set of neurons tracked across days, whether the changes in similarity between the two stimulus-evoked response patterns (S1 and S2), and in the patterns of responses to each individual stimulus, continue across days (a common signature of representational drift).

First, we found that the orthogonalization of stimulus responses (discriminability) continues to increase across days (Fig. 3d, discussed in Line 271). On subsequent days, the response similarity between S1 and S2 at the start of the day was generally lower than at the start of the previous day (Fig. 3d-e, discussed in Lines 271-272). Similarly, the response similarity between S1 and S2 at the end of each day was progressively lower than at the end of each previous day (Fig. 3d-e, discussed in Lines 271-272). However, at the start of each day, the correlation between S1 and S2 response patterns does partially reset and start higher than the end of the previous day (Fig. 3d and Extended Data Fig. 7c, discussed in Lines 271-274), demonstrating a "forgetting" effect, as the reviewer mentioned above.

We next used this new across-day tracking of neurons to examine across-day trajectories of the drift in the pattern of responses to each stimulus separately, and how they related to the content of reactivations of that same stimulus (Fig. 4). We first projected each stimulus response pattern and each reactivation pattern across Days 1-6

onto the axis of change in stimulus response pattern from the start to the end of Day 1 (Day-1-Early towards Day-1-Late; similar to the within-day axis of change in stimulus response pattern used in the original manuscript). This analysis showed that, *when considered along this particular axis*, the stimulus responses on subsequent days appear to drift towards or remain similar the Day-1-Late pattern (with some partial forgetting effect at the start of Day 2, Fig. 4c, discussed in Lines 338 and 340). Further, the stimulus reactivation patterns remain constant along this axis (no forgetting effect) not only on Day 1 but on all subsequent days (Fig. 4c, discussed in Lines 338 and 340) and reflect stimulus response patterns at the end of Day 1 (and Days 2-6). Thus, along this axis, the stimulus response patterns drift towards the reactivation pattern throughout Day 1, then partially reset at the start of subsequent days but drift back towards the constant reactivation pattern.

When we instead projected stimulus response patterns and reactivation patterns across days onto a different axis – the axis of change in pattern from Day 1 to Day 6 – we found that the stimulus responses do indeed continue to gradually drift across days from Day 1 to Day 6 (Fig. 4d, discussed in Line 342). By contrast, the reactivation patterns drift less and are more stable across days (Fig. 4d, discussed in Line 342). In fact, even on Day 1, the reactivation content along this axis already reflects the stimulus response pattern on Day 6.

Finally, because our main analyses in Fig. 4 assess changes in the stimulus response pattern from the start to the end of each separate day, pooled across days, we asked whether the original finding of stimulus response patterns drifting towards the same-day reactivation pattern, which we had previously shown to be true on average across all days, was also true when separately assessing each of the first 6 days of each experiment, using the same subset of neurons that could be tracked across all days. We confirmed that the main effects were visible on each individual day, and were stronger in the initial days (Extended Data Fig. 9c, discussed in Lines 316 and 320).

In summary, we now include three new analyses: 1) projection of all days onto the Day-1-Early to Day-1-Late axis of change in response pattern, 2) projection of all days onto the Day 1-to-Day 6 axis of change in response pattern, and 3) projection of all days onto their own within-day axis of change in response pattern. These across-day analyses of stimulus response and reactivation content solidifies our finding that reactivations are predictive of future stimulus responses, both within a day and across days.

Taken together, these across-day analyses have allowed us to confirm representational drift across days (a more conventional time scale in studies of representational drift). The new analyses have strengthened our conclusion that reactivation patterns are more

stable than stimulus response patterns, and that stimulus responses drift towards more stable reactivation patterns both within and across days. We have added these new and exciting results to the manuscript, and thank the reviewer for this suggestion.

In regards to the reviewers comment as to whether steady-state would also be reached when stimuli are more similar to each other, we do believe that steady-state would be reached in such a scenario. However, because the initial pattern similarity would be much higher, this would likely require more drift to reach steady state (assuming that pattern separation still occurs for such stimuli). This is an interesting future direction and we have mentioned this in our discussion (Lines 460-462).

3. The behavioral tracking is relatively limited (eye tracking of one eye, brain motion, n=4), making it difficult to assess how arousal/behavioral changes covary with some of the reported effects. In addition, the stimulation paradigm and time course of the effects are somewhat reminiscent of SRP (Frenkel et al., Neuron, 2006) and the vidget (Cooke et al., Nat Neurosci, 2015). Finally, the authors suggest that stimulus salience plays a role (paragraph starting at line 182) and put some emphasis on "unexpected stimuli" inducing the reactivations (line 346). The role of stimulus salience and arousal/behavior in the observed representational drift should be investigated in more detail: besides increasing the n and improved video tracking, the authors could also think of experimentally manipulating arousal and/or stimulus expectation.

We agree with the reviewer that the behavioral tracking analyses were relatively limited in the original manuscript. To this end, we tracked facial features from N = 3 mice during our reactivation experiments.

We used Facemap (https://github.com/MouseLand/facemap) to extract the location of 8 distinct features (three points on the nose area, the mouth, the lower lip, and three different whiskers) simultaneously across hours of imaging in addition to the existing eye tracking and brain motion data previously presented (Extended Data Fig. 4d-e, see Methods, discussed in Lines 208 and 1143-1152).

When we looked at the absolute movement of these 8 distinct features on the face of mice across hours of stimulus presentation, all 8 features showed (similar to the previous pupil analyses in Extended Data Fig. 4c) that facial motion was not significantly different across hours of stimulus presentation (Extended Data Fig. 4e, discussed in Line 208). Thus, we can now more confidently conclude that gradual changes in arousal state or motor behavior cannot explain the gradual changes in response patterns or in the rate of reactivations across the 2.5-hour sessions. In addition, we agree with the

reviewer that our stimulation paradigm and time course of the effects are reminiscent of SRP (Frenkel et al., Neuron, 2006) and the vidget (Cooke et al., Nat Neurosci, 2015) and we now have cited these studies in our manuscript (Line 119). However, we believe a unique feature of our paradigm is the very long (~58 second) inter-trial interval, which likely allows the brain to mostly revert to a baseline state that includes periods of low arousal and associated increases in cortical reactivations and hippocampal ripples (see below, Main point #4).

We have also improved the N in the analysis of brain motion and pupil across stimulus presentations and across inter-trial interval periods, by including a newly collected dataset from an additional mouse and by including datasets from the 50% of trials lacking optogenetic silencing of cortex, taken from the three additional mice in which a different subset of trials involved optogenetic silencing during stimulus presentation (all main findings were similar in all mice, see Reviewer 2, Other point #3, below). In total, this increased the number of mice included from N = 4 to N = 8 for these analyses.

We agree with the reviewer that experimentally manipulating arousal during stimulus presentation and/or stimulus expectation is an exciting possible direction. We therefore considered manipulating arousal levels. However, we were concerned that behaviorally elevating arousal levels prior to or during stimulus presentation (such as using an air puff or a loud noise) would be too long-lasting and thereby also decrease reactivation rates during the subsequent inter-trial interval, as reactivations rely upon the mouse being in a quiet-waking state with low arousal (Fig. 1e). To avoid this, we wondered whether we could briefly and locally simulate an arousal-like cortical state by stimulating locus coeruleus (LC) neurons projecting to lateral visual cortex, as LC noradrenergic neurons have a known role in mediating stimulus-evoked arousal and salience (Sara and Bouret 2012; Vazey et al., 2018). In initial experiments, we injected Cre-dependent tdTomato bilaterally into the LC of TH-Cre mice which express Cre specifically in neurons that contain tyrosine hydroxylase (which, in LC, release norepinephrine). We confirmed that LC sends dense projections to lateral visual cortex (Fig. R1, below):

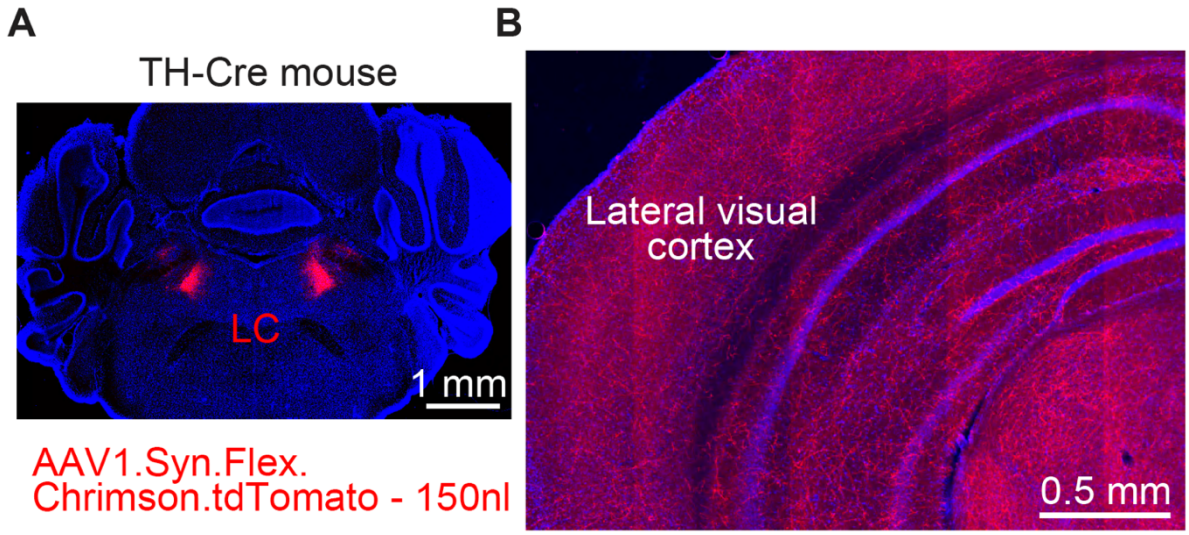

Figure R1. A) Example coronal slice of the mouse brain following bilateral expression of Cre-dependent tdTomato in LC neurons of a TH-Cre mouse. B) Example coronal slice showing LC axons broadly innervating lateral visual cortex.

We next attempted to manipulate LC input to lateral visual cortex to determine the effects of salience/arousal during stimulus presentation on the rate of inter-trial interval reactivations. If the visual stimulus was artificially made more salient/arousing by concurrently stimulating local LC axons in lateral visual cortex, would rates of subsequent reactivations of that visual stimulus increase?

To address this question, we expressed Chrimson (AAV1-syn-ChrimsonR-tdT) in LC neurons of EMX-1 Cre mice along with our usual preparation of Cre-dependent jGCaMP7s expression in excitatory neurons in lateral visual cortex. As we did not use TH-Cre mice for these experiments due to the interpretational difficulty of crossing these two Cre-expressing transgenic mice, TH-Cre and EMX1-Cre, we wanted to confirm that Chrimson-expressing axons in lateral visual cortex did indeed release norepinephrine (NE). This was important in order to be able to make any claims about arousal/salience and plasticity (NE acts on G protein-coupled adrenergic receptors in cortex). To do this, we photostimulated LC axons locally in lateral visual cortex while using the two-photon microscope to record from the green GRAB-NE sensor (AAV-syn-Ne2h, Feng et al., 2023) that was locally injected in lateral visual cortex of the same mice (we did not inject jGCaMP7s in these mice). Brief stimulation of LC axons in lateral visual cortex drove robust release of NE locally in lateral visual cortex for tens of seconds, Fig. R2, below):

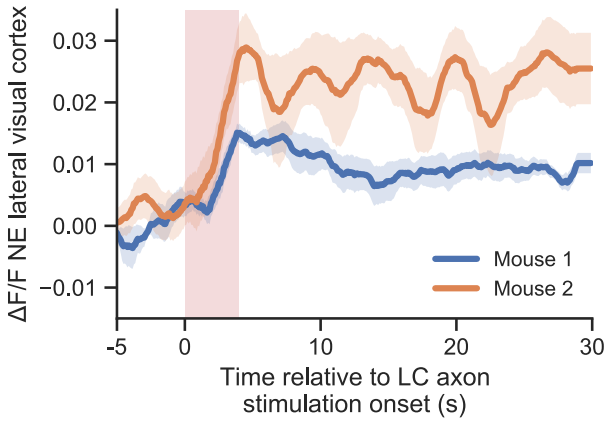

Figure R2. Stimulation of LC axons locally in lateral visual cortex drove an increase in fluorescence for tens of seconds of the NE sensor GRAB-NE2h that was injected in lateral visual cortex (N=2 mice, mean +/- SEM across stimulation trials).

Now that we had confirmed that the LC axons in lateral visual cortex release NE upon stimulation, we felt fairly comfortable using this protocol to artificially induce salience/arousal in mice during our experiments.

To artificially make the visual stimulus more salient/arousing, we decided to try two pairing protocols: (i) phasic LC stimulation for 2 seconds before stimulus onset (1 mW photostimulation at 16 Hz for 0.25 s, beginning at -1.75s and -0.25s before visual stimulus onset) or (ii) phasic stimulation during stimulus presentation (same photostimulation parameters but delivered at 0.25 s and 1.75 s after stimulus onset). We chose a phasic stimulus protocol because phasic but not tonic LC activity is evoked by intense or motivationally salient stimuli (Vazey et al, 2018). We chose to stimulate LC axons before stimulus onset to ensure sufficient time for NE to exert its effects (NE peaked several seconds post stimulation onset). We split the set of trials into thirds, with 1/3 of the trials being control (no stimulation) trials, 1/3 of the trials involving stimulation before stimulus onset, and the last 1/3 of trials involving stimulation during stimulus presentation.

In 7 sessions from 2 mice, we found that LC stimulation did not cause any obvious changes in reactivation rates across hours or within each trial (Fig. R3). Although stimulation of LC axons during stimulus presentation in lateral visual cortex somewhat induced higher levels of reactivations, the results were extremely noisy and not consistent enough to pursue further.

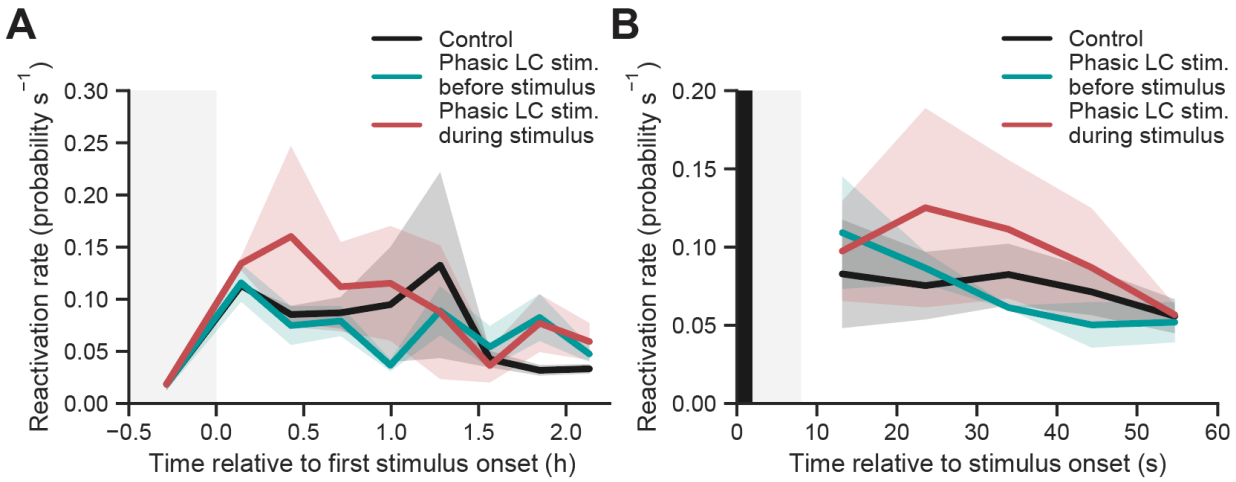

Figure R3. A) Reactivation rates across hours of imaging for three conditions: No stimulation trials, LC stimulation before stimulus onset, LC stimulation during stimulus presentation. B) Reactivation rates within a trial for three conditions: No stimulation trials, LC stimulation before stimulus onset, LC stimulation during stimulus presentation. N=2 mice, mean +/ SEM, 7 sessions.

Although we had predicted that LC axon stimulation locally in lateral visual cortex would increase rates of reactivations (the converse of shutting stimulus-evoked activity down in our Fig. 2 experiments which decreased reactivation rates), it is also known that arousal, depending on the level, can cause increases or decreases in learning and attention (McGinley et al., 2015). Too low, or too high, levels of norepinephrine-linked arousal are correlated with decreased performance. Here, we tried one stimulation protocol, but other protocols may have different or even opposite effects, depending on stimulation intensity as well as tonic vs. phasic stimulation (Vazey et al., 2018). Also, it is not clear that local release of NE would be sufficient to induce increases in reactivation rates, and perhaps such boosting in reactivation rates would require brain-wide actions of multiple neuromodulators. Future experiments should be able to examine the effects of LC stimulation more precisely by also using TH-Cre mice which will restrict expression in only LC neurons. Overall, these new experiments provided inconclusive/variable results, and we felt it was best not to include them in the revised manuscript to avoid confusion and to focus on the main conclusion of representational drift and its relation to the content of reactivations. We thank the reviewer for bringing this to our attention. We have toned down our findings and removed “unexpected stimuli” from the manuscript.

4. The authors link their findings to hippocampal reactivation/replay. To make this point, it should be investigated whether the cortical reactivation events align with the occurrence of sharp wave-ripple complexes.

We agree with the reviewer that it is useful to know whether reactivations in lateral visual cortex of awake mice during inter-trial intervals align with sharp wave-ripple events in the hippocampus. In a previous paper from the lab, it was found that reactivations in postrhinal cortex (POR, one of the four main subregions we image from) during a quiet waking state (after an hour of satiation that followed an operant visual conditioning paradigm) aligned with sharp wave ripple-band power (Sugden et al., 2020, see Fig. 2g of that manuscript). However, the reactivations we see in our data occur in a different context (minute-long inter-trial intervals between passive stimulus presentations), therefore it was important to confirm this finding in our context as well. While in the aforementioned study, we could insert acute laminar multielectrode probes followed by a long waiting period prior to recording, this was not possible in our study without deviating from the context in which the rest of the experiments were conducted. Thus, we used a new approach for chronic hippocampal ripple measurements during awake two-photon imaging in mouse visual cortex.

We developed a strategy to chronically implant electrodes into dorsal CA1 region contralateral to the imaging window, with a low-profile approach that didn't interfere with the headpost or imaging objective. We ran experiments in 5 mice. We confirmed the expected inverse relationship between ripple rates and pupil diameter in all 5 mice (McGinley et al., 2016), and also confirmed electrode placements in *post hoc* histology in a subset of mice. We then performed ripple measurements simultaneous with recordings of the activity of jGCaMP7s-expressing neurons in lateral visual cortex during the same stimulus protocol, as in other experiments in our manuscript. Similar to the previous study (Sugden et al., 2020), we found that ripple-band power in dorsal CA1 shows a sharp peak concurrent with cortical stimulus reactivations during the inter-trial interval following stimulus presentation (N = 5 mice, Fig. 1f and Extended Data Fig. 2a, discussed in Line 159). These effects were consistent both across mice and across 2-4 days per mouse. This was a nice confirmation that the stimulus reactivations we observe in lateral visual cortex are part of a more global event that includes similarly brief sharp-wave ripples in the hippocampus.

We were unfortunately unable to include these five new mice in other analyses beyond those described above, as these experiments differed in important ways. These included differences in extent of expression of jGCaMP7s and associated differences in the number of neurons recorded evenly throughout each region and depth. The emergence of stress responses inducing cloudy eyes also many times prevented us from running full sessions and/or enough sessions per mouse to be included in any further analysis.

5. Throughout the manuscript, the number of animals used is generally quite low, with n equaling often 4 and sometimes even fewer mice. This might be not so much of an issue when testing large effects, but certainly is a problem when no significant difference is observed. The low n results in low statistical power, making tests prone to type II errors. This could potentially result in the failure to detect significant differences in control analyses, influencing the overall conclusion of the study.

We agree with the reviewer that a low N results in low statistical power and makes tests prone to type II errors. First, although the number of mice was equal to N=4 mice for many of the analyses in the original manuscript, we did run, on average, 10 multi-hour sessions/mouse (total: 42 sessions). This greatly reduced the variability of each mouse-by-mouse estimate. Moreover, in many cases (e.g. Extended Data Fig. 3a-b and Extended Data Fig. 6c/f) we showed that effects were consistent within each daily session (see also now Extended Data Fig. 9c, discussed in Lines 316 and 320). We were wary of doing statistics per session rather than per mouse, which can be prone to Type I errors due to false assumptions about independence across sessions. While we have used generalized mixed-effects models in some previous papers, based on this reviewer's feedback, we decided to instead try to increase the number of mice. A major challenge we encountered in these revision efforts was a high attrition rate due to a multitude of factors. Our revision experiments to increase N had a very low success rate (<10%) due to high quality metrics that we required for broad expression of jGCaMP7s across a 3-mm window and 3 planes achieving >5,000 neurons (which we showed previously using drop-out analyses was important to accurately and sensitively detect reactivations, Extended Data Fig. 2h-i, discussed in Line 172). Further, potentially due to changes in the microscope rooms outside of our control, mice behaviorally did not exhibit relaxed states as frequently, and only a small subset of the mice remained in low-stress, quiet waking states with periods of low pupil-linked arousal during inter-trial intervals, which were required for expression of reactivations across hours each day, and across many days of imaging. Nevertheless, we added one additional mouse that met all these criteria, for an N of 5 mice. In addition, and based on the suggestions of Reviewer 2, we reran most of the analyses in the original manuscript on the additional 3 mice that underwent cortical inhibition during 50% of stimulus presentations (Fig. 2d-g). By focusing on the other 50% of control trials in which the cortex was not silenced, we found that the vast majority of the results were similar for these Stimulus inhibition mice and for the Non-inhibition mice (Extended Data Fig. 6 and Extended Data Figs. 8-10). Because of the high similarity between these two datasets, we combined them in the main figures and showed them separately in supplementary figures (Extended Data Fig. 6 and Extended Data Figs. 8-10). The one consistent difference between these two datasets was that the stimulus responses on control trials were higher in Stimulus

inhibition mice, likely because the neurons hadn't been driven by a stimulus in the last 120 seconds on average (versus in the last 60 seconds for the Non-inhibition mice). This difference doesn't affect any of the conclusions in our paper, and in fact further supported our conclusions as explained in the response to Reviewer 2, Other point #3, below).

With the addition of the Stimulus inhibition mice, we now have a total of N = 8 mice for most of the analyses, which increased statistical power and thereby reduced any risk of Type II errors. Moreover, the consistency of main results in each of these 8 mice is, in our view, among the most consistent we have observed in our previous work and in the literature. For example, a central finding of our work is summarized in Fig. 4e-f. Here, we show that Decrease neurons – neurons that decrease their stimulus response during a session – tend to participate less than average in reactivations early in a session, while Increase neurons tend to participate more than average in reactivations early in a session. Both of these findings were evident in all 8 mice, and were separately observed for analyses of Stimulus 1 responses and reactivations, and for analyses of Stimulus 2 responses and reactivations.

Minor

1. The authors pooled data from layer II to IV. Were there no differences between layers?

The vast majority of data was collected from layer 2/3. We included layer IV in the text as a possible location of imaging because we recorded from a lower depth of 280 microns from the brain surface. This depth is technically considered lower layer 2/3 but in extreme cases, could have been upper layer IV as the imaging window placement compresses the brain tissue. However, upon inspection of previous histology, we now have changed this text to indicate that we estimate that we recorded in layer 2/3 (Line 129).

All mice were imaged from 130-280 microns from the brain surface across 3 planes typically spaced ~60 μm apart. We have now re-analyzed our results using cells originating from the uppermost plane (~156 microns deep on average) and separately for cells originating from the lowermost plane (~266 microns deep on average). We found no difference in the participation in reactivations for neurons imaged at these two depths (Extended Data Fig. 2e, discussed in Line 165). We also found no major differences in the decrease in response similarity across trials (Extended Data Fig. 6d, discussed in Line 249) or in the drift of stimulus response patterns towards reactivation patterns (Extended Data Fig. 9d, discussed in Line 324), when these analyses were

restricted to neurons at either depth. We have added these new results to the manuscript.

2. Line 131: Could more ensembles be detected when not only focusing on stimulus-driven neurons? Possibly ensembles of other, non-experimentally induced experiences, reactivating during the baseline session?

This was an interesting but difficult question to address fully since we are not able to train classifiers on these other non-S1-related and non-S2-related experiences. When imaging from ~6,900 neurons in total, reactivating ensembles are a fraction of all neurons and therefore, a template is needed in order to identify them. Further analyses indicated that if we did not present S1 and S2 to mice to identify S1 and S2 ensembles, we would not be able to identify S1 and S2 reactivation patterns with high sensitivity. We can however attempt to identify epochs of highly synchronous activity of all ~6,900 neurons on average. Whether these epochs of highly synchronous activity of ~6,900 neurons reflect reactivations is unclear. We used the same prior that selected moments during the ITI with synchronous activity across subsets of stimulus-driven neurons, but this time including all ~6,900 recorded neurons. We excluded all time points when an S1 or S2 reactivation occurred, in order to focus on other potential reactivation patterns.

We found that on average, we could detect ~5 synchronous events per minute during the baseline period prior to any stimulus presentations (Fig. R4a, below).

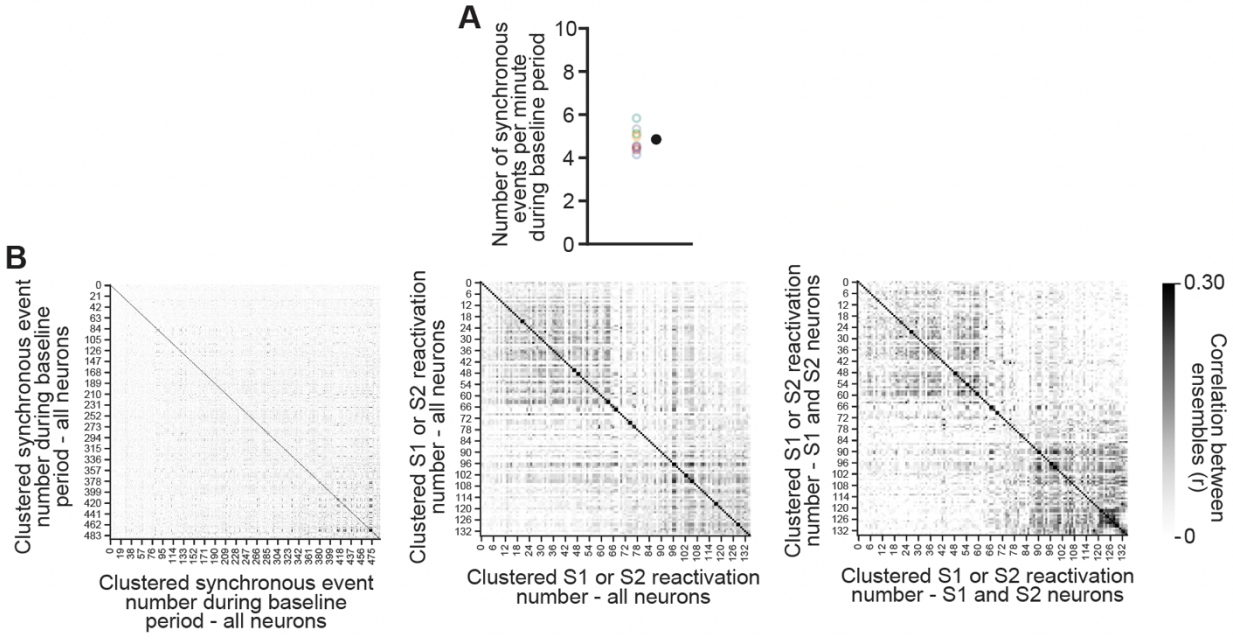

Figure R4. A) Number of synchronous events per minute during baseline period, excluding times of S1 or S2 reactivations. N = 8 mice. B) Correlation matrix between ensembles. Each correlation matrix is clustered using hierarchical clustering, which sorts the matrix based on events that are most similar to each other. Left: clustered correlation matrix of synchronous events during baseline period using the activity of all neurons and excluding S1 or S2 reactivation times. Middle: clustered S1 or S2 reactivations using the activity of all neurons. Right: clustered S1 or S2 reactivations using only the activity of S1 and S2 neurons (as done in the manuscript).

We wondered whether these events contained repeated instances of the same patterns of active neurons (indicating a putative ensemble), or whether each pattern was different. To do this, we clustered the events based on similarity using hierarchical clustering, and found that these events did not cluster into discrete ensembles (Fig. R4b, left). In contrast, when we clustered S1 and S2 reactivations, we found 2 distinct groups (Fig. R4b, middle), which became even more evident when we only analyzed S1- and S2-responsive neurons instead of all neurons (Fig. R4b, right), similar to the inclusion criterion for pattern analyses used in the manuscript.

These new analyses suggest that non-stimulus-related synchronous events involving subsets of neurons do occur during the baseline period in a daily recording, prior to delivery of any stimuli. However, these events potentially reflect a diverse array of visual and/or non-visual experiences, rather than stereotyped repetition of one or a few dominant patterns.

Because this analysis was difficult to interpret due to the lack of a training set analogous to the visual evoked response patterns, and because these findings did not affect our main conclusions, we elected to omit this analysis from the revised manuscript.

3. Why did the authors choose this specific retinotopically patterned stimulus?

We chose to use two checkerboard stimuli, with each rectangle in the checkerboard containing band-pass filtered spatiotemporal white noise, for multiple reasons:

1)

When we designed these experiments, our priority was to understand the detailed content of reactivations, and not just whether reactivations occur above chance levels. Our goal was to record cellular activity patterns during presentation of two distinct visual stimuli in a manner that would allow maximal decoding of whether even a degraded pattern of ongoing activity was similar to one or the other stimulus-evoked response pattern, or

neither. Thus, our choice of spatially distinct stimuli enhanced our ability to classify patterns as more similar to S1 or to S2.

2)

Bandpass filtered noise drives many more neurons compared to other stimuli such as drifting gratings (see Fig. R5, below), consistent with previous studies demonstrating the superior efficacy of this stimulus in driving many neurons with diverse tuning within a subregion of visual space (Niell et al., 2008). This was very important, as the number of false negative and false positive reactivation classifications decreases substantially as the number of visually driven neurons used by the classifier is increased (Extended Data Fig. 2h-i, discussed in Line 172). We have now added motivation for the choice of this stimulus in the Methods (Lines 1063-1064).

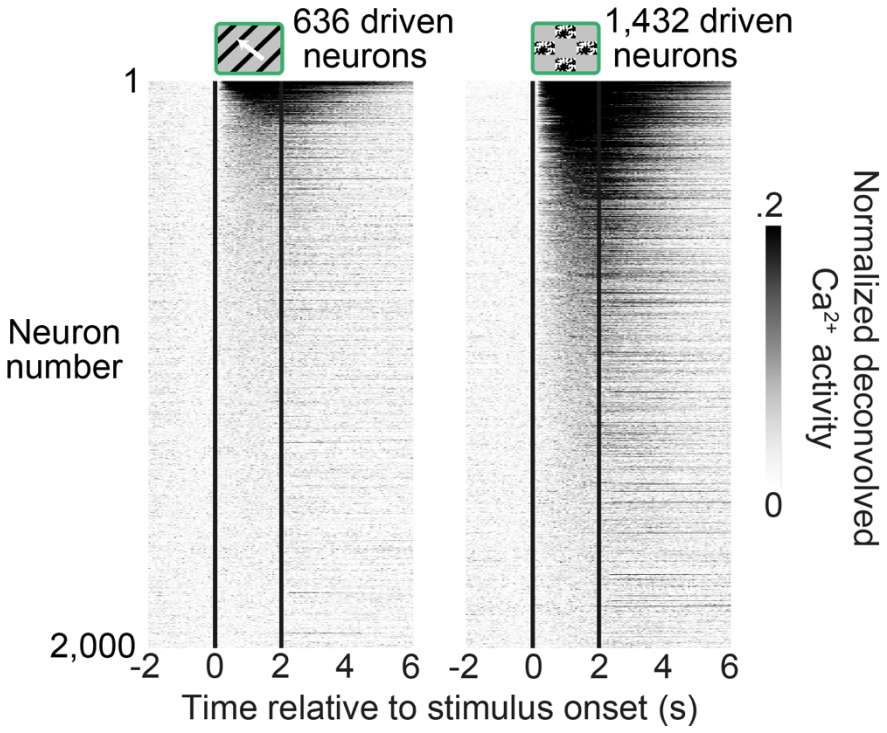

Figure R5. Lateral visual cortex neuronal responses to presentation of a drifting grating (left) or a bandpass-filtered noise stimulus (right) for two seconds. Bandpass-filtered noise stimuli drove over 2 times greater activity in lateral visual cortex than drifting gratings. Neurons are sorted by max response during stimulus presentation. While gratings at different orientations and spatial/temporal frequencies would likely drive other neurons, we specifically focused on the best two stimuli that could drive many neurons, as our long ITIs precluded the use of many stimulus types in the same session.

4. Local inhibition reduced the stimulus reactivation rates, but the remaining reactivation rates were still much higher than baseline, why?

We considered several possible explanations as to why the remaining reactivation rates remained higher than baseline after local inhibition of stimulus responses:

1)

We first considered whether reactivation rates may not fully decay back to baseline levels at the end of the previous 60-s non-inhibition control trial (specifically, trials without local inhibition of stimulus responses that preceded an inhibition trial). In this way, the elevated reactivation rates induced by the stimulus presented during the previous control trial may persist, even following the subsequent transient inhibition. However, when we plotted these control trials that are followed by inhibition trials, the rate of reactivations did not continue to decay throughout the inter-trial-interval following stimulus inhibition, but rather reached a plateau level that was higher than the baseline reactivation rate (Extended Data Fig. 5e, discussed in Line 229). Relatedly, the reactivation *bias* towards the specific stimulus presented during the proceeding control trial was no longer present for reactivations detected during the subsequent inter-trial interval that followed stimulus inhibition (Extended Data Fig. 5g, discussed in Line 231). Taken together, this suggests that the hypothesis that stimulus-specific reactivations induced by the previous stimulus persist into the ITI following stimulus inhibition trial is unlikely to explain why the reactivation rates during inhibition trials were still much higher than baseline.

2)

We then considered whether this above-baseline level of reactivation rates could instead potentially be due to the fact that, in our studies, the pupil area during the quiet waking baseline period is higher than during the inter-trial interval following inhibition (Extended Data Fig. 4c). Reactivation events occur during states associated with low pupil area (Fig. 1e). However, even when we subselected epochs with similar pupil area during the baseline period and during ITI periods from stimulus inhibition trials, the rates of reactivations remained higher during these ITI periods (Extended Data Fig. 5f, discussed in Line 229). Therefore, a lower arousal state does not explain why reactivation rates during the ITI periods following stimulus inhibition are higher than during the baseline period.

3)

Inputs from other brain areas could drive some of the reactivations in lateral visual cortex following stimulus inhibition (Zutshi and Buzsaki, 2023). Reactivations are likely global events, recruiting distributed brain regions (e.g. hippocampus, Fig. 1f, discussed in Line 159) which could also drive reactivations locally in lateral visual cortex. However, testing this hypothesis is beyond the scope of this study.

5. If optogenetic inhibition halts/slows the instructive effect of the reactivations, was this also accounted for in the model?

Surprisingly, although we inhibited half of the stimulus responses during our optogenetic inhibition experiments (which reduced the rates of reactivations on these trials relative to control trials, Fig. 2f), the average reactivation rates in experiments involving optogenetic silencing of 50% of stimulus responses were similar to those in Non-inhibition mice (Extended Data Fig. 5h, discussed in Line 265). This may be due to the consistently stronger stimulus-evoked response magnitudes on control trials in inhibition mice (Extended Data Fig. 6e-f, discussed in Line 263), which may have induced this higher rate of subsequent reactivations, since trial-to-trial response magnitude does correlate with subsequent reactivation rate (Extended Data Fig. 3f, discussed in Line 264). These increased stimulus response magnitudes may be due to the fact that the time since the cortex has experienced strong sensory input is, on average, 120 seconds in the inhibition experiments, as compared to 60 seconds in normal experiments.

Any differences in reactivation rates over time are inherently accounted for in our model due to a recursive term that updates the model each time there is a reactivation (rather than once per trial). This allows the model to account for the slowing down of reactivation rates over the length of the session (Fig. 2b, left) and can partly explain the more rapid drift in stimulus response patterns and stimulus orthogonalization at the start of each session. Specifically, the increased rate of reactivations on early trials causes the model to iterate more frequently, inducing more plasticity and more rapid drift in the stimulus pattern towards the reactivation pattern (Fig. 4g-h). This was confirmed when we now reanalyzed Figs. 3 and 4 with the mice that involved stimulus inhibition (see response to Reviewer 2, Other point #3). We found that the vast majority of the results were recapitulated in the Stimulus inhibition mice. We were able to model representational drift using only the reactivation rate and content just as well in Stimulus inhibition mice as in Non-inhibition mice (Extended Data Fig. 10c, discussed in Line 386).

6. In the discussion the authors mention that the purpose of reactivations could be to "help enhance stimulus discrimination". Did the authors look at

reactivations following learning? This might make the observed changes in activity patterns more interpretable.

We unfortunately did not look at reactivations following classical cue-outcome learning. However, changes in stimulus responses similar to those we have observed in our study, such as decreases in response similarity (pattern separation), have been linked to subsequent increases in sensory discrimination following learning (Rust and Dicarlo, 2010; Schmid et al., 2023). We now discuss these studies and highlight this topic as a valuable direction for future studies (Lines 245, 457, and 460-462).

7. The authors chose to use parametric tests throughout the manuscript, is this supported by any statistical indication?

We performed Shapiro-Wilk tests on all of our data to test for normality. We were unable to reject the hypothesis that our distributions resemble a normal distribution, and we therefore felt comfortable using parametric tests which assume that the data is normally distributed. This also increases our statistical power as mentioned by the reviewer above in Major point #5. We have added this information to the Methods (Line 1451).

8. Fig 1b: According to the legend, the white box and FOV image (left) should have the same size, but the scale bars indicate that the white box is about twice the size of the FOV.

In Fig. 1b, the image on the left is a zoomed-in example subregion of the field of view and not of the entire FOV. The purpose is to clearly show that we are imaging at single cell resolution. We apologize for the confusion and have made this clear in Fig. 1b by annotating ‘example subregion’ next to the zoomed-in subregion of the field of view and by clearly stating this in the figure legend.

9. Fig 3a and 3c: It would be useful to see the baseline values in the plots as well.

We added the baseline value for Fig. 3a (now Fig. 3c). For Fig. 3c (now Fig. 3b), it is not possible to plot a baseline value for this analysis as it examines similarity between two visually-evoked patterns, which are not assessed during the 30-minute baseline period.

Reviewer 2 (Remarks to the Author):

This manuscript by Nguyen et al. examines patterns of activation and reactivation of neurons in the lateral visual cortex. Cortical glutamatergic neurons were imaged in head-fixed mice as they were presented with one of two stimuli, S1 and S2, interspersed by ~1min long inter-trial intervals, allowing for quantification of

stimulus-evoked responses and reactivations during the ITI. The authors found individual neurons and population level responses that demonstrated stimulus-specific reactivations during the ITI. As has been previously reported, the authors found that reactivations are biased toward salient recent events. But departing from previous reports, the authors advance a hypothesis that the reactivations do not function in stabilizing sensory representations but rather predict and guide representational drift toward future states. Toward this, the authors demonstrate that "Decrease neurons" that eventually lose their sensory-evoked tuning, show a strong reduction in such tuning in the ITI periods, evident even on the very first ITI. However, such effects were less prominent for "Increase neurons". The authors then developed a model that takes in the stimulus evoked response on the first trial, and ITI responses across all trials, to accurately predict the eventual re-organization of evoked responses on the last trials. Overall, these findings are interesting and may be important for understanding how reactivations contribute to plasticity in the visual cortex. In its current form, however, the data to support the claim that reactivations predict future sensory-evoked reorganization are underdeveloped.

We thank the reviewer for recognizing the potential importance of our findings.

We believe that our data presentation and nomenclature may have led to some confusion and partial misinterpretation of some of the results in our study. We will attempt to address concerns as best we can. We apologize in advance for any confusion we inadvertently caused. We have made many changes to the manuscript to increase clarity.

The major concern is that the reactivation results presented here could be a trivial result where neurons on average tend to have weak reactivations (strong drop in activity even by the first ITI), thus appearing as though the reactivation actively contribute to the reorganization and progressive loss of tuning of "Decrease" neurons. The small magnitude effects on "Increase" neurons compounds this problem.

We interpret the reviewer's major concern as: stimulus reactivation patterns are a weaker, scaled-down version of the stimulus-evoked response patterns, with nearly uniformly weaker activity across all neurons participating in a reactivation than in a stimulus response pattern. Therefore, because stimulus-driven neurons are less active during stimulus reactivations than during stimulus-evoked responses, it will appear as if reactivations are predictive (or contribute to) of the loss of activity in Decrease neurons that decrease their stimulus-evoked responses across the session.

In summary, we believe the reviewer is stating that the reactivation patterns are identical, albeit scaled down, versions of the stimulus-evoked patterns, and therefore seemingly predictive of the loss of activity in Decrease neurons in future stimulus responses.

If the above was true, then if we scaled up, by a constant scale factor across all neurons, the activity levels during stimulus reactivations to match those of stimulus-evoked responses, then reactivations should no longer be predictive of the loss of activity in Decrease neurons in future stimulus responses. To do this, we scaled the reactivation activity level to be the same as the stimulus activity level (Extended Data Fig. 9g). We multiplied all reactivation activity by ~1.3x, in order to match that of the stimulus-evoked response activity on average across stimulus-driven neurons (referred to now as 1.3x-scaled reactivations).

According to the reviewer's major concern, if we matched for average activity levels between reactivations and stimulus-evoked response patterns, then the 1.3x-scaled reactivations should be the same as the stimulus-evoked responses and therefore, scaled reactivations would not be predictive of any loss of activity of Decrease neurons. However, this was not the case.

During early trials in a session, in which population activity patterns during reactivations and during stimulus-evoked responses were believed to be different (Fig. 4b), we subtracted the 1.3x-scaled reactivation patterns from the stimulus-evoked response patterns (Fig. 4f). *We found that Decrease neurons still exhibited relatively lower activity during the 1.3x-scaled reactivations than during the stimulus-evoked responses. No Change neurons had equal activity between scaled reactivations and stimulus-evoked responses. Increase neurons exhibited more activity in 1.3x-scaled reactivations than in stimulus-evoked responses.* This demonstrates that the *pattern* of the reactivations themselves are different than the *pattern* of the stimulus-evoked responses during early trials. *Decrease neurons participate relatively less in reactivations than in stimulus-evoked responses, even after we scaled the mean population activity levels up to be just as high as during stimulus-evoked responses.* Therefore, the pattern of the reactivations, and not their mean activity levels, is predictive of future changes in the pattern of stimulus-evoked responses. This prediction of the future response pattern involves significant over-participation of Increase neurons in reactivations, as well as significant under-participation of Decrease neurons (Fig. 4f). We further note that, despite the predictive power of reactivations even after the first few trials, the actual changes in stimulus response patterns occur gradually and only plateau later in the 2.5-hour session (Fig. 3b and Fig. 4b), and also that the mean stimulus-evoked response

across all driven neurons is stable throughout the session (as Increase neurons compensate for the decreased response in Decrease neurons, Fig. 3c/f).

The above finding that the pattern and not simply the activity level during early reactivations is predictive, on a cell-by-cell basis, of future changes in stimulus evoked activity several hours later was also confirmed using a separate analysis that more directly addresses the difference in pattern between early stimulus responses and early reactivations (Fig. 4e). Here, we divided the mean activity of Decrease neurons by the mean activity of No Change neurons during individual stimulus responses or reactivation patterns. We found that Decrease neurons were relatively less active than No Change neurons in early reactivations than in early stimulus-evoked responses. This analysis does not depend on the scaling operation described above, or on differences in overall activity levels during stimuli vs. reactivations. By contrast, Increase neurons that later increase their stimulus responses hours later were relatively more active than No Change neurons in early reactivations than in early stimulus-evoked responses (Fig. 4e).

In the model we use to predict future stimulus response patterns using only the stimulus response pattern on the first trial and subsequent reactivations (Fig. 4g-j), we used the 1.3x-scaled reactivations to predict future stimulus responses. This removed any effects related to any differences in population mean activity levels during reactivations vs. during stimuli, and solely focused on the differences in patterns. When we focused on the patterns alone, we found that reactivation patterns were predictive of both representational drift and of the decrease in similarity of responses to the stimuli that we have observed in Decrease neurons but not Increase neurons (Fig. 4h-j).

We hope that the above explanations provide clarity and address the reviewer's major concern**: the reactivation results presented here *are not* a trivial result where neurons on average tend to have weak reactivations (strong drop in activity even by the first ITI), thus appearing as though the reactivation actively contributes to the reorganization and progressive loss of tuning of "Decrease" neurons.**

We apologize for any confusion we may have caused due to miscommunication on our part. We have now revised our manuscript to make these points as clear as possible (Lines 347-364).

As a side note: we were confused as to why the reviewer referred to Decrease neurons as having a loss of tuning in their comments. It may well have been a miscommunication on our part or differences in nomenclature. If the reviewer refers to tuning simply as the activity of a neuron during the stimulus presentation, then we hope

our responses above have provided clarity on this matter. However, it is possible that the reviewer was instead or also referring to “tuning” in regards to a neuron’s selectivity in responding more to one stimulus versus the other. We have now added analyses showing that, indeed, Decrease neurons increase their stimulus tuning (selectivity), or their relative magnitude of responses to one versus the other stimulus, on average (Fig. 3h, discussed in Lines 293 and 295). This at least partly explains the decreasing similarity of the S1-evoked and S2-evoked population activity patterns across the session. Relatedly, the nomenclature in the previous Fig. 3f-g and text, may have been confusing. We have now modified our analysis to be simpler and clearer. We now instead quantified the proportions of Increase, Decrease, or No Change neurons that exhibited an increase or decrease in stimulus selectivity (where selectivity for S1 or S2 is calculated as $(S1_{response} - S2_{response}) / (S1_{response} + S2_{response})$) from early to late trials (Fig. 3h, discussed in Lines 293 and 295). This now more clearly shows that twice as many Decrease neurons significantly increase their selectivity as those that significantly decrease their selectivity. This bias towards increased selectivity was not evident in Increase neurons or No Change neurons. This supports the idea that Decrease neurons are especially important in driving reductions in the similarity of S1 and S2 population response patterns (Fig. 3b/g, captured by the model using only 1.3x-scaled reactivation patterns, see Fig. 4i-j). We hope this analysis addresses any remaining confusion.

The results would be more convincing if the reactivation not only predicted loss of tuning, i.e. “Decrease” reorganizations, but also predicted active changes in tuning, such as from strongly S1 tuned to strongly S2 tuned, or even better if the stimulus set was larger (S1 to S5) where the increased dimensionality of the stimulus set and population responses would provide greater predictive power.

The reviewer mentions being able to predict changes in tuning such as from strongly S1 tuned to strong S2 tuned. In our manuscript, we have focused solely on how S1 reactivations predict changes in S1 stimulus-evoked responses, and how S2 reactivations predict changes in S2 stimulus-evoked responses. To predict changes in tuning preference, such as from S1 to S2, we don’t think it is possible using solely S1 or S2 reactivations alone, as S1 reactivations can’t predict S2 stimulus-evoked changes and visa-versa. However, using both S1 and S2 reactivations, we should be able to predict active changes in tuning.

Therefore, in response to this reviewer’s comment, we posed the question: do neurons that have changes in tuning preference (i.e., from S1 to S2) differentially participate in certain reactivations (i.e., participate more in S2 reactivations relative to S1 reactivations)? For an S1 neuron to become an S2 neuron, that neuron either must lose more of their response to S1, gain more response to S2, or both. In all three cases,

such neurons should participate less in S1 reactivations than S2 reactivations (and vice versa). We found that neurons that changed their stimulus tuning from being significantly driven only by S1 to only by S2 participated in S1 reactivations less than S2 reactivations (and the inverse for neurons that changes from S2- to S1-driven neurons) (Extended Data Dig. S9i, discussed in Line 364). Therefore, by considered both types of stimulus reactivations, we can predict active changes in tuning preference. We have added this new finding to our manuscript.

We agree with the reviewer that if the stimulus set was larger (5 stimuli instead of 2 for example) and reactivations still predicted representational drift of all 5 stimuli independently, that it would provide even stronger evidence. However, we elected for a design with two stimuli to maximize classifier sensitivity and to maximize our statistical sensitivity to gradual changes in the pattern of these two stimulus responses within each individual session, as well as now tracking responses of the same neurons across sessions (Fig. 3d-e, discussed in Lines 271-272, Fig. 4c-d, discussed in Lines 338-342, Extended Data Fig. 7a-c, discussed in Lines 270 and 334). We have added these points to the Discussion (Lines 460-462) regarding the benefits of expanding the stimulus set in future experiments.

Related to the above, causal manipulations in the ITI are necessary to make strong claims about whether reactivations actively lead to re-organization of evoked responses on later trials.

We agree with the reviewer that causal manipulations during the ITI are necessary to make strong claims about the functional role of reactivations in the re-organization of stimulus-evoked responses and in representational drift.

Reviewer 3 raised a similar point, copied below for reference:
"The study implicates a causal role of reactivation in driving representational drift. Of course, one missing piece here is some local circuit manipulation to test this necessity more directly. I understand the authors' argument that such experiments (i.e., selective optogenetic perturbation of reactivation events in LVC) should most ideally be done in a closed-loop design, which is a challenging experiment. Nevertheless, I wonder if the temporal bias in reactivation, such that it peaks in the first 10-20 min and that it is also stronger during the first half of the ITI (Figure 1B), still provides a possible way to test for the idea that reactivation drives drift causally. Perhaps silencing can be targeted early during the session (i.e., first 10-20 min) and early during ITI (i.e., 20-30 sec)?"

We agree with the general strategy suggested by Reviewer 3 to use our optogenetic inhibition approach to inhibit spontaneous activity during the ITI for extended periods of time. The hypothesis we could test in this way would be*:* if we inhibit activity during the ITI by sustained optogenetic activation of PV+ interneurons, thereby inhibiting ongoing excitatory activity including stimulus reactivations, do we see a decrease in the change of response similarity between stimuli and in representational drift? This would provide causal evidence for the role of reactivations in re-organization and drift.

We set out (after consultation with the editor) to inhibit activity during most of each ITI (42 seconds total) in order to disrupt reactivations from occurring during this period. We felt it would be important to inhibit the ITI period for longer than Reviewer 3 suggested, in order to inhibit as many reactivations as possible and thereby maximize our chances observing an effect. For each experiment, we set out to inhibit every ITI on manipulation days, and to compare with control days in which no inhibition occurs. We felt comfortable comparing effect sizes across days but within the same mice because the change in response similarity and representational drift on odd vs. even days exhibited no difference in our existing dataset (Fig. R6):

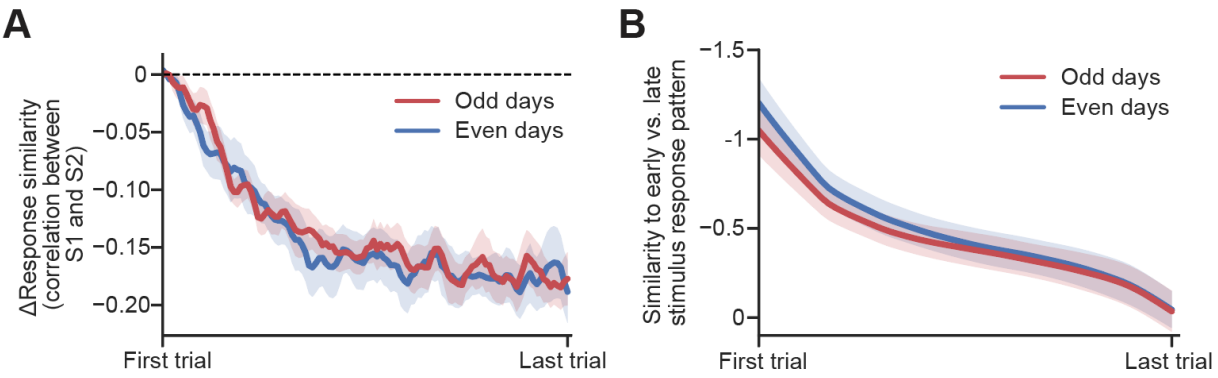

Figure R6. A) Change in response similarity from the first trial of the day to the last trial of the day, plotted separately for odd and even days. N = 8 mice, mean +/- SEM across mice. B) Similarity to early vs. late stimulus response pattern from the first trial of the day to the last trial of the day, plotted separated by odd vs even days. N = 8 mice, mean +/- SEM across mice.

In the original manuscript, we had performed silencing for the 4 seconds surrounding stimulus presentation via optogenetic activation of PV+ interneurons through the imaging objective. We made local viral injections of the red-shifted opsin Chrimson under the S5E2 enhancer, which selectively targets parvalbumin interneurons, along with Cre-dependent jGCaMP7s in Emx1-Cre mice. We had validated that our approach silenced >90% of sensory-evoked responses in excitatory neurons over this 4 second period, with minimal rebound effects after termination of photoinhibition, and with no

acute effects on arousal assessed by tracking pupil area (Fig. 2d-e and Extended Data Fig. 5a-d).

As a key step towards validating the method for silencing the ITI, we first assessed whether optogenetic inhibition of ongoing activity for 42 seconds would persistently inhibit activity of lateral visual excitatory cortical neurons. We tested this approach in 5 new mice. Optogenetic stimulation of PV+ interneurons persistently suppressed lateral visual cortex excitatory neurons for all 42 seconds in all 5 mice (Fig. R7a, below). Activity levels following ITI suppression returned to pre-stimulation levels within seconds (Fig. R7a, below).

We next assessed how the intensity of red light used for optogenetic inhibition modulated the degree of suppression. Optogenetic photostimulation of PV+ neurons at 1 mW suppressed activity in excitatory neurons less than when using 10 mW in all three mice where this was tested (Fig. R7b). Thus, we concluded that in order to more completely prevent reactivations and ongoing activity during the inter-trial interval, 10 mW of power was necessary.

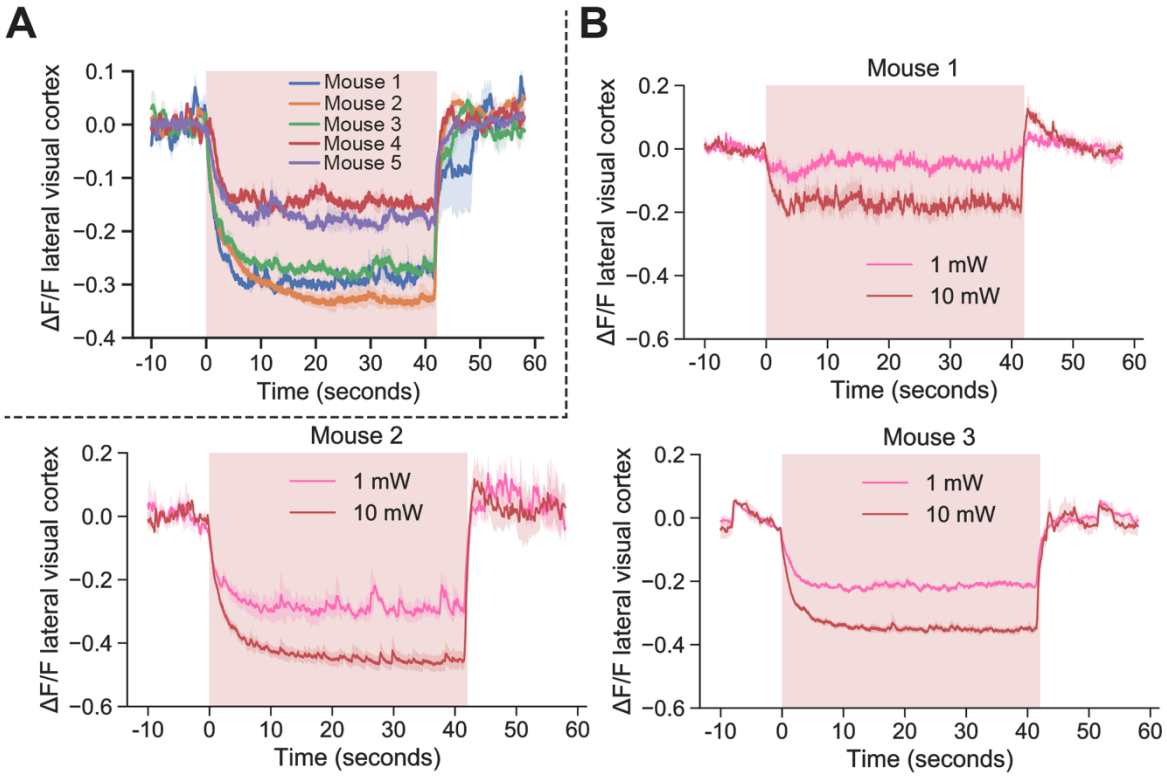

Figure R7. A) Optogenetic stimulation of PV interneurons at 10 mW for 42 seconds (red shading) persistently suppressed lateral visual cortex excitatory neurons. Following this inhibition, activity of lateral visual cortex excitatory neurons returned back to baseline within seconds. N = 5 mice. B) Same as A but using either 1 mW or 10 mW of power (N

= 3 mice). Lateral visual cortex excitatory neurons were suppressed to a greater extent by 10 mW vs. 1 mW photostimulation.

We next attempted, in several mice, to run a full experimental session with optogenetic inhibition of excitatory neurons in lateral visual cortex during each ITI, to prevent reactivations from occurring. We ran sessions without silencing on odd days, and sessions with ITI inhibition on even days.

Although we were successful in inhibiting activity during the inter-trial interval, we noticed that following optogenetic inhibition, there was persistently elevated activity above levels observed on control sessions from the same mice (Fig. R8 and Fig. R9, below, N = 2 mice; note: ongoing activity levels [deconvolved calcium activity s^{-1}] were very consistent across control mice, Extended Data Fig. 4b). Following 42 seconds of inhibition, subsequent levels of ongoing activity were ~2.5x higher than during control days (Fig. R8, below). This elevated ongoing activity persisted for at least 20 seconds, through the next stimulus presentation period (Fig. R8 and Fig. R9, below). We could not avoid this confound by reducing the power to 1 mW, as levels of ongoing activity following 1 mW photoinhibition were still ~2x higher than during control days (Fig. R8 and Fig. R9, below). These results are consistent with some previous studies showing a persistent elevation in ongoing activity following a sustained experimental inhibition (e.g. Nelson et al., 2003), which could be due to depolarization block (Herman et al., 2014).

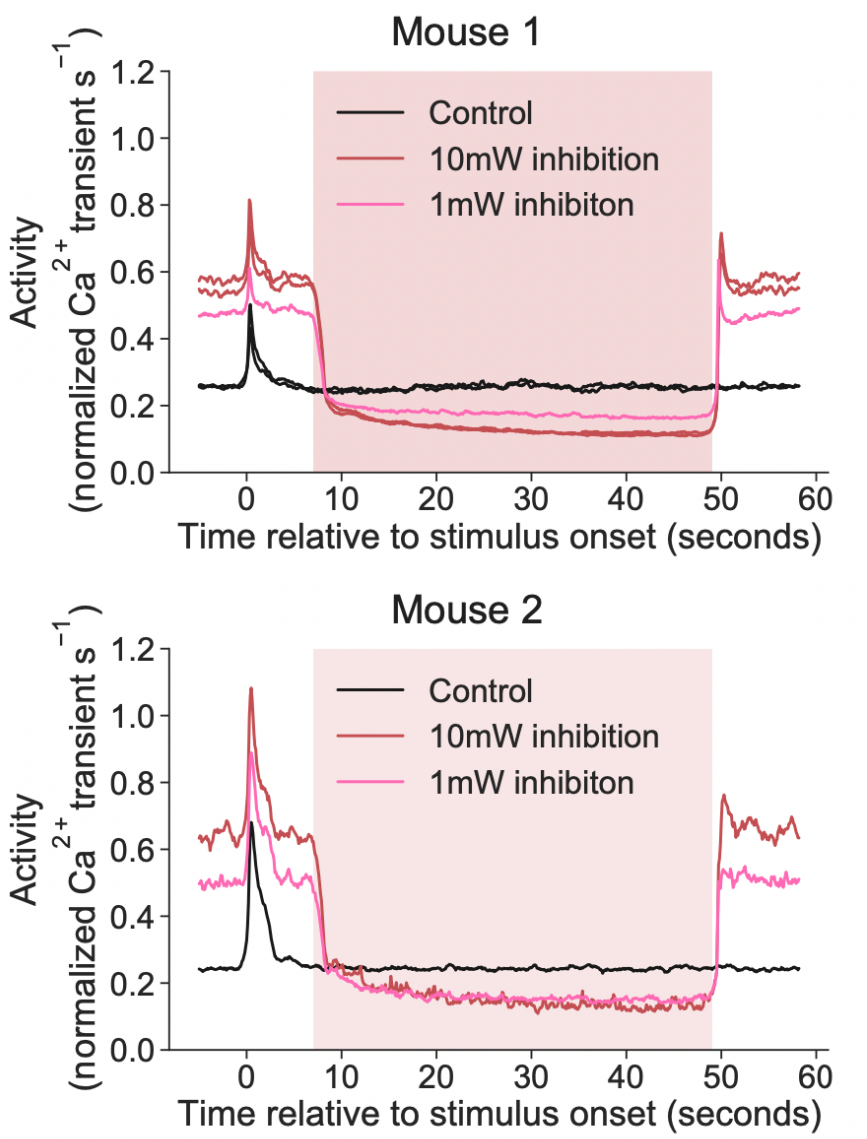

Figure R8. Activity during control sessions and sessions involving optogenetic inhibition of the ITIs at 1 mW and 10 mW. In N = 2 mice, 10 mW inhibition induced heightened persistent activity for tens of seconds following the end of the optogenetic inhibition. 1 mW inhibition had a similar but slightly weaker effect. In Mouse 1, 10 mW and control sessions were each run on multiple days (order: control, inhibition, control, inhibition), and illustrate how reliable this rebound excitation effect is within a mouse.

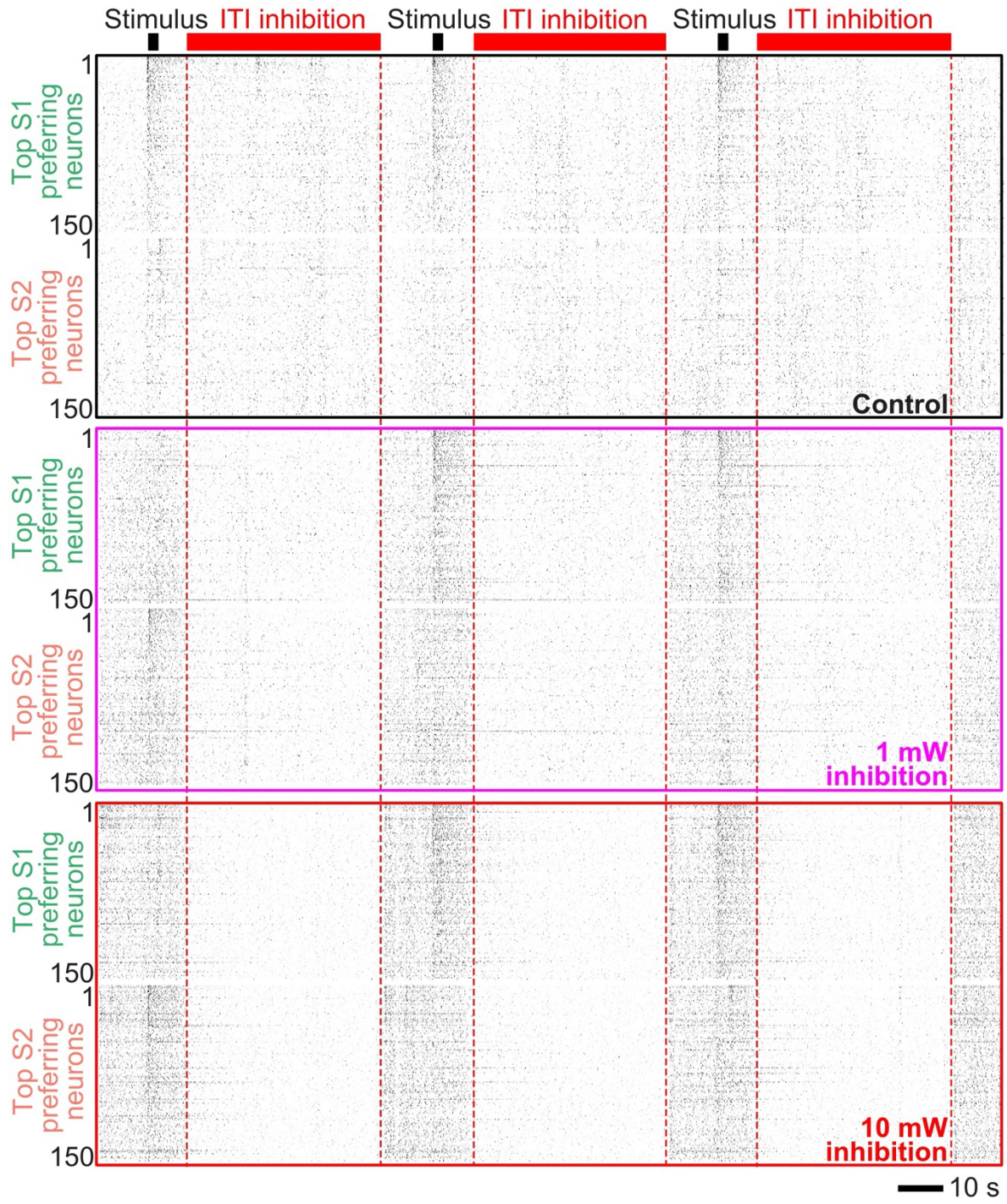

Figure R9. Example raster plots of ongoing deconvolved calcium activity (Fig. 1d) during stimulus presentation (black solid square, top) and during the inter-trial interval for 3 conditions: control trials (top), inter-trial interval inhibition trials using 1 mW of power (middle), and inter-trial interval inhibition trials using 10 mW of power (bottom). Note that as inhibition strength increases, inter-trial interval activity during inhibition

decreases but subsequent ongoing activity following the termination of inhibition increases drastically and persistently.

In summary, we observed a substantial rise in activity post-inhibition in sessions with 42-s of inhibition during the ITI (though not following 4-s of inhibition, Extended Data Fig. 5d), which persisted for tens of seconds and into the following trial (thus affecting neurons at the time of the next stimulus). Thus, we felt uncomfortable pursuing this paradigm further to examine the effects of inhibiting the ITI and associated reactivations on changes in stimulus response patterns. Indeed, the single-trial stimulus-evoked activity patterns are clearly much noisier trial-to-trial in the ITI inhibition context due to the elevated ongoing activity now being on par with the magnitude of the stimulus-evoked response (which itself didn't change very much, Fig. R8). Thus, despite a major effort to get to this point (the five mice with proper expression of both Chrimson and jGCaMP required many more attempted experiments in additional mice due to the combination of technical and surgical challenges), we unfortunately could not address the question of causality.

We considered instead playing loud noises or presenting the mouse with mild air puffs during the inter-trial interval in order to prevent a quiet waking state from occurring, which would have likely prevented most reactivations from occurring. However, non-visual stimuli that increase ITI arousal levels would likely also increase arousal during subsequent stimulus presentations, which would in turn likely affect visual cortical plasticity.

We did not intend to suggest, in the original manuscript, that reactivations are the main driver of the reorganization of visual evoked responses. Rather, we only tried to argue that several lines of evidence from our work suggested that they may play a role, including our model that used the reactivation patterns alone to quantitatively predict later changes in stimulus response patterns surprisingly well, even on a trial-to-trial level (after filtering out the slow changes in response magnitude at timescales of 20 trials or longer, Fig. 4h, *inset*). Because we were unable to prove that reactivations causally drive representational drift, we have tried to further temper this statement through the manuscript, including in the abstract. We now increase the focus on the finding that over- or under-participation of neurons in reactivations early in a session can predict which neurons will gradually increase or decrease their visual responses over the next two hours. We have solidified this finding substantially, including characterizing the drift of stimulus response patterns towards reactivation patterns not only within a session, but now also across sessions by tracking the same neurons over time. We feel confident that our data supports the conclusion that we are able to *predict* (and "may contribute to", Line 63, Abstract) future changes in stimulus response pattern, even

though we have not been able to prove that it is a major *drive*r of future changes in stimulus response patterns despite an intense effort. We have largely used the word 'predict' in the Introduction and Results sections. We have raised the potential for causality only when discussing the model at the end of the last figure and in the Discussion, as an exciting future direction when closed-loop experiments become more feasible, as noted by Reviewer 3.

Other Points:

1. When justifying the scale of the imaging, it might be important to look at false positive/negative rates when selecting a random contiguous region of the overall imaging planes (rather than a random sample across the entire volume), to better simulate what a smaller ROI would not be able to capture. Perhaps reactivating cells are spatially correlated and the false positive rate would be lower.

We reran the analysis but this time including neurons from a random contiguous subregion of the overall imaging volume, to simulate a smaller field of view. We randomly picked one neuron and analyzed all neurons in a region defined by a disc tangential to the cortical surface and surrounding that neuron. This allowed us to subsample neurons from a contiguous region. To include progressively more neurons, we expanded the radius of the disc. We found that, similar to when we randomly sub-selected neurons across brain areas in our initial analysis, the false positive and false negative rates increased to a similar degree when analyzing smaller and smaller numbers of neurons selected from any given subregion (Extended Data Fig. 2i, discussed in Line 172). We have added this new analysis to the manuscript.

2. When using a classifier to assign probabilities of reactivation patterns to synchronous activity (Extended data Figure 1A), the training set includes frames during stimulus presentation as well as frames during the ITI without synchronous activity. The authors pre-process the data to remove slow temporal correlations in the times between synchronous events because there still seems to be persistent activity patterns throughout the ITI in the raw data. If the classifier is trained only on frames during stimulus presentation and tested on every frame of the ITI, is the reactivation probability still significantly higher during synchronous events when compared to frame-shuffle data?

We interpret the reviewer's general question here as, "*What are the reactivation probabilities outside of our defined moments of synchronous activity of subsets of stimulus-driven neurons during the ITI after stimulus presentations? We removed slow temporal correlations in the original classification but, if we had not pre-processed the*

data and had included the rest of the ongoing activity during the ITI, do we also observe high rates of reactivations during these other ITI moments?"

As the reviewer mentions, our training set for our multinomial logistic regression-based classifier uses activity patterns from three sets of frames: those during the presentation of S1 stimuli, those during the presentation of S2 stimuli, and those during epochs in each ITI in which synchronous activity of the set of neurons driven by S1 and/or S2 is low along with the 30-minute baseline period prior to stimulus presentation. This third set of frames served to provide a large alternative set of patterns that, in initial analyses, we found were very rarely similar to the S1- and S2-evoked patterns.

To address the reviewer's question, we trained a new classifier in which we removed the set of patterns from non-synchronous ITI periods from the training set (synchronous ITI periods are defined as moments of elevated activity across top stimulus-driven neurons), so that we could test for S1 and S2 reactivations during these non-synchronous ITI periods. Note that, while the reviewer mentioned training on only S1 and S2 trials, a third set of other patterns is essential, as the probability of a pattern resembling either an S1 pattern, an S2 pattern, or a 'other pattern' must add up to one. Thus, we now used only activity patterns during the 30-minute baseline period prior to any stimulus presentation as the set of 'other patterns' for training.

When we tested this new classifier on all of the raw, unfiltered ITI data (without any temporal filtering or subselection of synchronous frames), we found that the reactivation rates were significantly higher during previously defined synchronous periods involving stimulus-driven neurons than during other ITI periods (~0.8 s^{-1} vs. ~0.1 s^{-1}; Extended Data Fig. 1g, discussed in Line 146). This was a nice confirmation that the synchronous epochs we defined using stimulus-driven neurons during the ITI indeed have high stimulus content specificity. We have added this to the manuscript.

We note that this increased similarity of ongoing patterns of activity to the previous experienced stimulus during periods of synchronous ongoing activity is similar to the observations of hippocampal replay being most prominent during hippocampal ripples, which reflect synchronous activity of many excitatory neurons. Perhaps relatedly, we now show that our S1 and S2 visual cortical reactivations during synchronous periods in the ITI are indeed correlated with ripple-band power in the hippocampal local field potential (Fig. 1f, discussed in Line 159).

3. In the simultaneous imaging and optogenetic silencing experiments, 50% of trials are inhibited. How does this affect the learning rate/plasticity for these mice? It would be important to see the analyses in Figure 4 repeated for these

inhibited mice. Specifically - does reducing the total number of reactivation events causally decrease the learning rate or change the number of Increase/Decrease neurons?

We compared the reactivation rate for Non-inhibition mice vs. Stimulus inhibition mice, in which 50% of trials in each session involved stimulus inhibition, but not inhibition of the subsequent ITI. Surprisingly, we found that the total reactivation rate did not change for Non-inhibition mice vs. Stimulus inhibition mice (Extended Data Fig. 5h, discussed in Line 265). Below, we present a possible explanation for this: first, we found that stimulus-evoked responses on the trials lacking optogenetic silencing in Stimulus inhibition mice were larger than responses in Non-inhibition mice (Extended Data Fig. 6e-f, discussed in Line 263), potentially related to the fact that the average duration of time since the most recent, non-inhibited stimulus is 120 seconds in these Stimulus inhibition mice (whose stimulus responses were inhibited on a random 50% of trials), compared to 60 seconds in the Non-inhibition mice. Second, we found that trials with stronger stimulus responses (averaged across all stimulus-driven neurons) were correlated with higher reactivation rates during the subsequent ITI (Extended Data Fig. 3f, discussed in Line 264), even though the mean stimulus-evoked response across a session remained constant on average. This may have resulted in higher reactivation rates following control trials in Stimulus inhibition mice than when considering all trials in Non-inhibition mice (Extended Data Fig. 5h, discussed in Line 265). Taken together, these two findings may explain why reactivation rates overall are the same for Non-inhibition vs. Stimulus inhibition mice.

We repeated all of the analyses in Fig. 4, and also in Fig. 3 for completeness, separately when using the data from the Stimulus inhibition mice (N = 3; we excluded analysis of activity patterns during the 50% of stimulus presentations with concurrent inhibition). Because the average rates of reactivations were similar to those of Non-inhibition mice, and our model is dependent on the rate of reactivations, we predicted that the results in Fig. 3 and Fig. 4 would be largely similar between Non-inhibition mice vs. Stimulus inhibition mice.

As we hypothesized, we found that the results in Fig. 3 and Fig. 4 using stimulus response patterns from the control trials in Stimulus inhibition mice were extremely similar both quantitatively and qualitatively to those from Non-inhibition mice. All major conclusions that we reported in Fig. 3 and Fig. 4 in the original manuscript were confirmed in the Stimulus inhibition mice, including: (i) response similarity between S1 and S2 response patterns showing a decrease over trials, due to activity changes in Decrease Neurons (Extended Data Fig. 6a-f and Extended Data Fig. 8a/g, discussed in Lines 248, 259, 280, and 286), (ii) the content of reactivations predicting future changes

in stimulus response patterns (Extended Data Fig. 9b, discussed in Lines 316 and 320), (iii) the pattern of early stimulus reactivations systemically differing from the pattern of early stimulus responses (Extended Data Fig. 9f/h, discussed in Lines 352 and 359), and (iv) the ability of our simple model involving iterative updating of estimated response patterns each time a reactivation occurs to capture the decrease in similarity of S1 and S2 response patterns in Fig. 3 and drift of the stimulus response patterns towards reactivation patterns in Fig. 4 (Extended Data Fig. 10c-f, discussed in Lines 386, 392, 396, and 398).

Because the results were so similar between these two groups of mice, we felt comfortable grouping these results in the main figures. However, we wanted to also show that there was a difference in the average magnitude of responses to a stimulus in the non-inhibition experiments vs. the stimulus inhibition experiments, even though it did not affect the conclusions of our study. Therefore, we plotted most of Fig. 3 and Fig. 4 separately for Non-inhibition and for Stimulus inhibition mice, and added them to the manuscript (Extended Data Fig. 6 and Extended Data Figs. 8-10).

4. It appears as though "Decrease" neurons have a higher baseline activity level and "Increase" neurons have a lower baseline activity level compared to "No Change" neurons. Is this effect significant? Are there differences in activity before any stimuli are presented? Could the pattern of activity observed in Fig. 4C suggest that the ITI reactivations promote tendency toward a "homeostatic baseline" rather than toward future reorganization?

We now compared the activity for these three neuron classes during the baseline period before any stimuli are presented. We found that all three cell classes exhibited similar levels of baseline activity (Extended Data Fig. 8c, discussed in Lines 282-283). We have added this to the manuscript.

The homeostatic baseline idea is interesting. We agree it is possible that the pattern in Fig. 4c (now Fig. 4e) could be a result of ITI reactivations gradually promoting a shift towards an intermediate level of responsiveness to a repeatedly presented stimulus. This would also be consistent with the observations above that Increase and Decrease neurons start off with stronger and weaker responses, respectively, than No Change neurons, but end up with more similar response magnitudes, on average, at the end of the session (Fig. 3f; note that we now show quantitatively that Decrease neurons have significantly higher stimulus-evoked activity at the beginning of the day compared to No Change neurons, and Increase neurons have significantly lower stimulus-evoked activity at the beginning of the day compared to No Change neurons). This is somewhat similar to the firing rate homeostasis proposal from Turrigiano, Van Hooser and

colleagues (Hengen et al., 2013, now cited in the manuscript, Line 420), except that in our case the homeostasis is at the level of a regularization of each neuron's evoked response magnitude, rather than of ongoing activity levels.

However, a homeostatic rule alone would not explain the changes in *response selectivity* to the two stimuli from early to late trials that we observe in Decrease neurons (Fig. 3h, discussed in Lines 293 and 295). In order to exhibit changes in stimulus selectivity, the stimulus-evoked responses must become more different for S1 vs. S2 across time (i.e. not driven to be the same for both S1 and S2 trials, as would occur based on a homeostatic mechanism alone). In turn, these changes in stimulus selectivity from early to late trials underlie the orthogonalization of S1 and S2 response patterns (Fig. 3g; despite the mean responses across neurons during S1 and S2 remaining constant throughout the session, on average, Fig. 3c).

We believe that promoting a tendency towards a homeostatic baseline response magnitude can work in unison with a differential change in responses to the two stimuli. For example, if pattern separation between two stimuli is achieved by differentially decreasing the responses to one or the other stimulus in initially strongly driven Decrease neurons, then it may be that initially weakly driven Increase neurons compensate via a non-selective increase in responses to both stimuli (Fig. 3h, discussed in Lines 293 and 295). In this way, the mean response across neurons is preserved across trials and sessions, as we observe and as we have observed previously (McGuire et al., 2022). Further, such a homeostatic contribution might ensure that neurons maintain a large dynamic range of stimulus sensitivities. We have added these interesting ideas to the Discussion (Lines 420-422).

5. What was the distribution of neurons recorded across the mice? i.e., how many neurons did each mouse contribute to the "pool" of neurons examined here? What is the total number of neurons that were S1-responsive, S2-responsive, or responsive to both at the beginning and end of the sessions, and was this similar across mice? In general, total neuron counts (including increase/decrease neurons, etc.) should be included.

We have now quantified the counts of total neurons, numbers of S1-responsive, S2-responsive, and both-responsive neurons over the entire session and separately at the beginning and end of the session, as well as counts of Increase, Decrease, and No Change neurons (Extended Data Fig. 1b and Extended Data Fig. 8b, discussed in Lines 133-134 and 282-283). We also found that the numbers and proportions of S1-responsive, S2-responsive, and both-responsive neurons were similar across mice

(Extended Data Fig. 1b, discussed in Lines 133-134). We have added this to the manuscript.

Reviewer 3 (Remarks to the Author):

In this exciting study, Nguyen and colleagues investigate cortical reactivation of sensory stimuli evoked activity patterns in the mouse lateral visual cortex with large-scale two-photon calcium imaging and optogenetic manipulations. The authors report that presentation of visual stimuli was followed by transient, stimulus-specific reactivations during the following minute. Optogenetic silencing of local circuits during sensory presentations abolishes reactivations. The authors intriguingly find that the content of reactivation was different from preceding stimulus responses such that reactivations predict future stimulus responses. Based on these observations, the authors propose that reactivations guide representational drift supporting sensory discriminations.

Overall, this is an interesting and elegant study reporting potentially highly impactful observations related to offline cortical dynamics supporting sensory discrimination. The large-scale imaging approach is impressive as it uniquely allows the authors to identify cortical reactivation patterns. My comments below are aimed at improving the accessibility of this impressive study to the broader audience at Nature.

We are happy to hear that the reviewer is excited by the observations in our study and their potential impact on the field. We have addressed all comments below in order to make our study more accessible.

Major

1. Most reactivation seems to occur during the first 10-20 trials (Figure 1B). This early phase of stimulus presentation might be associated with the different behavioral states, particularly increased arousal, especially since the visual stimuli are not reinforced. This issue warrants a bit more scrutiny. Is there evidence and an effect of increased arousal at the beginning of each day of experiments resulting in higher baseline firing rates, especially in the first 10-20 trials to which the animals subsequently habituate? Especially given that 'Decrease' neurons account for most of the orthogonalization between the two stimuli (line 257). Variations in the animals' behavioral state and arousal may play an important role in trial-to-trial variations pushing cortical patterns away from the preferred manifold. This could be considered in the discussion (line 394), and

there are a few places in the results section where this possibility should be considered:

We agree with the reviewer that the consideration of arousal during early trials needs to be more fleshed out. We have added more on this in the main text (Lines 193-208) as well as by performing the analyses as suggested below:

a. Figure 1E: Only shows behavior during reactivations (which could be excluded based on behavioral states, line 966). Is there any evidence for increased arousal during stimulus during the early trials as opposed to the later ones? The authors could analyze pupil diameter, attempted running, and brain motion data across stimulus presentations.

We agree that increased arousal during stimulus presentation in early vs. late trials could potentially contribute to differences in reactivation rates.

When we analyzed pupil diameter and brain motion during stimulus presentation, we found that, on average, both were extremely stable and flat across all trials, including the earliest trials in the session (Extended Data Fig. 4a, discussed in Line 202). This was in line with our analysis of pupil and activity levels during the inter-trial interval periods (Extended Data Fig. 4b-c).

Overall, the arousal state of the mouse remained unchanged, on average, throughout the recording session. We attribute this to the fact that mice were heavily habituated to the imaging set-up before we began experiments. An additional contributing factor may be that, in contrast to many rodent studies involving frequent stimulus presentations and rewards (and associated satiation), our brief stimuli are presented only rarely, at one-minute intervals, and are not associated with rewards or punishments. We have now also added face tracking to our manuscript, and like pupil area, dynamics of various facial features (nose, mouth, whiskers) remain unchanged across hours of imaging (Extended Data Fig. 4d-e, discussed in Line 208).

Altogether, these findings suggest that arousal during stimulus presentation is unable to explain the 4-fold increased reactivation rates at the beginning of the session and subsequent gradual decline in reactivation rates. However, we did observe shorter time-scale trial-to-trial fluctuations in arousal levels throughout the session, and we found that pupil area during stimulus presentation was positively correlated with reactivation rate during the subsequent ITI (Extended Data Fig. 3e, discussed in Line 199). We have added text discussing this point to the manuscript. We could not analyze attempted

running, but such attempts to run should correlate with brain motion, which we have reported.

b. Figure 3A, line 188: Is there a different model that fits the data better than linear regression? If a period of increased arousal transiently drives increased activity this might be expected to follow an exponential decay back to some baseline rate, which wouldn't be significantly fit by linear regression but is still significant. Is there a significant change in stimulus-evoked activity in the early template (first 3 stimuli) vs the late (last 3 stimuli)?

We tested whether a linear or exponential function would fit the data in Fig. 3a (now Fig. 3c) and Fig. 3c (now Fig. 3b) better, by modeling an exponential vs. linear function to the data. We determined goodness of fit by using the R-squared value.

The exponential function fit the data better than the linear function in Fig. 3c (now Fig. 3b), but the linear function fit the data slightly better in Fig. 3a (now Fig. 3c, see Fig. R10, below).

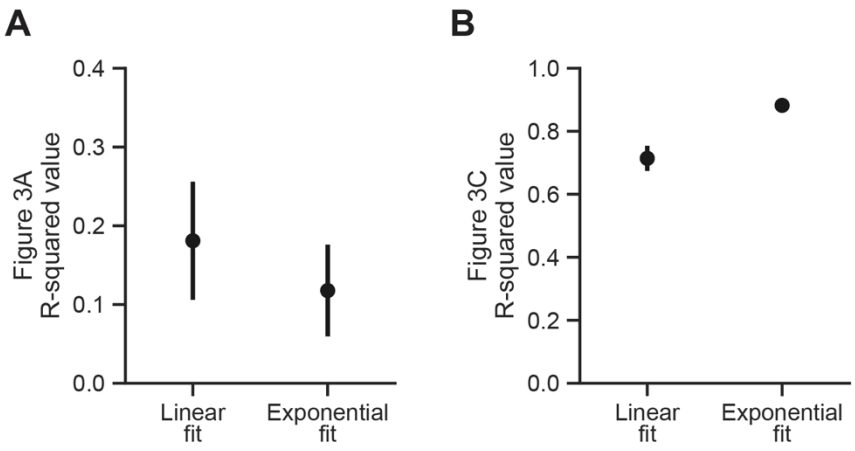

Figure R10. A) R-squared value for linear vs. exponential fit for the data in the previous Fig. 3a (N = 8 mice). B) R-squared value for linear vs. exponential fit for the data in the previous Fig. 3c (N = 8 mice).

We ran the statistics as the reviewer suggested, by comparing the first 3 and last 3 trials for the data in Fig. 3a (now Fig. 3c), and found that the mean response was not significantly different (two-tailed paired t-test), consistent with the linear fit being better than the exponential fit. We did not include this statistical test in the manuscript to keep the statistical tests consistent between panels. Please note that the small decrease in mean response across the first several trials was previously based on the average of

four mice but is now based on the average of eight mice (each of these eight mice separately replicated the main conclusions of the paper). With the addition of the extra mice, the mean population activity rate doesn't significantly drop from the first to last trials, but we agree there is a slight drop during initial trials (please see updated Fig. 3c). We thank the reviewer for raising this point.

c. Figure 3A Inset: is there any evidence for individual differences in recording quality or mouse behavior that account for why the orange mouse has such a lower activity rate than the others (also see Figure 3C inset)?

With the additional of 4 more mice to the dataset (1 new mouse and 3 Stimulus inhibition mice, control trials only), the previously orange mouse, now the gray mouse, does not seem to be a major outlier in what is now Fig. 3c (formerly Fig. 3a) and Extended Data Fig. 6f (formerly Fig. 3c inset). However, this mouse still exhibits slightly lower levels of stimulus-evoked activity than other mice. Extended Data Fig. 4a-c (discussed in Lines 202-205) show that behavioral and arousal changes aren't different in the gray mouse vs. other mice. The field of view location for the gray mouse is also similar as in other mice. However, the number of neurons imaged in the gray mouse was lower than in all other mice (~5,000 neurons for the gray mouse vs. ~7,000 neurons for other mice; Extended Data Fig. 1b, discussed in Line 134). The expression of jGCaMP7s was lower in the gray mouse. With lower expression of jGCaMP7s, it is possible that the measurement sensitivity is lower in this mouse, which would explain a lower activity rate. Although this mouse could be an outlier in this way, we felt comfortable keeping it in the dataset as the dataset had a sufficiently high number of driven neurons that, based on our other analyses of false-positive and false-negative reactivation rates, indicated that the error rates were sufficiently small (Extended Data Fig. 2h-i, discussed in Line 172). Please note that all major findings evident in other mice were also observed in the gray mouse despite this slightly lower activity rate.

2. The authors should provide a more detailed analysis to see if Increase and Decrease neurons differ in any physiological or anatomical properties (i.e., baseline firing, cortical layer location, subregion position within lateral visual cortex). Are there differences in functional connectivity structure between these functional cell types? In their previous study, the authors implemented an informative analysis of this matter (Sugden et al. Nat Neurosci, 2020). Thus it would also be informative to apply a similar analysis here.

Increase and Decrease neurons do not differ in baseline firing (Extended Data Fig. 8c, discussed in Lines 282-283) or in cortical layer location within layer 2/3 (Extended Data Fig. 8e, discussed in Lines 282-283), and only showed slight but non-significant

differences in subregion position within lateral visual cortex (Extended Data Fig. 8d, discussed in Lines 282-283). Area P has a slightly lower relative proportion of Increase neurons. We have also analyzed all of these aspects for No Change neurons and added them to the same plots.

We applied a similar analysis as in the previous paper from our lab (Sugden et al., 2020) for functional connectivity comparisons between these cell classes. To measure the trial-to-trial stimulus response fluctuations between neurons (noise correlations) in our dataset, we subtracted out the correlation derived from a general increase (or decrease) of activity across trials in Increase (or Decrease) neurons (see Methods). This correlation would be large and would overwhelm the noise correlations, which are smaller trial-to-trial fluctuations. We found that there were no differences in functional connectivity, as measured by pair-wise noise correlations, between pairs of neurons within No Change, Increase, and Decrease neuron groups (Extended Data Fig. 8f, discussed in Lines 282-283).

Relatedly, it would be informative to know if reactivations exhibit spatial drift over time across LVC areas and if reactivations are preferentially expressed in deeper layers or uniformly.

We found that reactivations shift slightly (center of mass change of ~100 μm in recordings from a field of view spanning several millimeters) in the posterior direction over time (Extended Data Fig. 2f, discussed in Line 166). This posterior drift indicates a slight shift towards lower-order visual cortices and away from certain tactile and auditory sensory association cortices.

We also now show that neurons participating in reactivations were not preferentially located in deeper vs. more superficial depths within layer 2/3 (Extended Data Fig. 2e, discussed in Line 165).

3. The study implicates a causal role of reactivation in driving representational drift. Of course, one missing piece here is some local circuit manipulation to test this necessity more directly. I understand the authors' argument that such experiments (i.e., selective optogenetic perturbation of reactivation events in LVC) should most ideally be done in a closed-loop design, which is a challenging experiment. Nevertheless, I wonder if the temporal bias in reactivation, such that it peaks in the first 10-20 min and that it is also stronger during the first half of the ITI (Figure 1B), still provides a possible way to test for the idea that reactivation drives drift causally. Perhaps silencing can be targeted early during the session (i.e., first 10-20 min) and early during ITI (i.e., 20-30 sec)?

We agree with the reviewer and have addressed this thoroughly in our response to Reviewer 2 who had a similar comment: **"Related to the above, causal manipulations in the ITI are necessary to make strong claims about whether reactivations actively lead to re-organization of evoked responses on later trials."** Please see our response to this Major Comment from Reviewer 2 above, including new Figs. R6-R9 in this response to reviewer comments. We have copied the above reviewer comment in that response as well. In short, we tried a modified version of the experiment suggested by the reviewer in several mice. However, despite many months of effort dedicated to these experiments, we found that the persistent rebound activity inherent to sustained silencing of cortex for tens of seconds via activation of PV+ interneurons prevented us from establishing an acceptable paradigm to test for causality. We have therefore toned down the strength of any suggestions of causality.

Minor:

1. The current study builds upon the group's previous work (Sugden et al. Nat Neurosci, 2020), so it would be relevant to add more discussion about the similarities and differences between the present and prior works. Especially since, in the current work, neutral stimuli without reinforcement elicit reliable reactivation, while in the previous study, it was mainly associated with reinforcement (related to major point 1 above).

In the Sugden et al. 2020 paper, there was a neutral stimulus (and associated neutral stimulus reactivations) that was presented along with positively and negatively reinforced stimuli. However, it is hard to tell whether a neutral stimulus is truly neutral in such a context, as it could signal a lack of expected outcome. We have added Discussion comparing our current study with the previous study in the lab performed by Sugden et al., 2020 and to highlight how our paradigm with neutral stimuli drives robust reactivations without reinforcement (Lines 156-159 and 415-417).

We also now show that, similar to Sugden et al., 2020, cortical reactivations couple to dorsal CA1 hippocampal ripple-band power (Fig. 1f and Extended Data Fig. 2a, discussed in Line 159).

2. The authors should consider acknowledging other studies in the hippocampus demonstrating that reactivation content significantly differs from online experience (e.g., Terada et al., Nature, 2021).

We apologize for having missed this and have added Terada et al., Nature, 2021 and other relevant citations to our study (Gupta et al., 2010; Swanson et al., 2020). We have added text acknowledging these studies (Lines 103-108 and 427-429).

3. Line 131: How stable is the stimulus-preferring population between early and late trials (how similar are the top 5% responsive neurons in the early trials to the top 5% responsive in the late trials)? and what is the overlap between S1 and S2 preferring populations?

About 50% of neurons remain stable at the top 5% of responsive neurons from early to late trials (Extended Data Fig. 1i, discussed in Line 152). We expected some neurons to have changing stimulus-evoked activity profiles from early to late trials, which is the reason why we trained 3 separate classifiers (early, middle, and late) to classify reactivations. With three different templates for classification, we could identity reactivation patterns that look like any of the possible stimulus-evoked response pattern templates (see Methods, Extended Data Fig. 1d).

We measured the overlap between S1- and S2-preferring neurons using 1) a binary metric relying on a statistical non-parametric test to determine significance and to describe neurons as responsive to S1, S2, or both stimuli, and 2) by estimating the response selectivity index of each neuron individually, where selectivity for S1 or S2 is calculated as $(S1_{response} - S2_{response}) / (S1_{response} + S2_{response})$. We found that, using the binary metric, there are equal proportions of S1-preferring neurons, S2-preferring neurons, and neurons that have significant responses to both S1 and S2 (Extended Data Fig. 1b, discussed in Line 134). Similarly matched levels of selectivity to the S1 and S2 stimuli was observed when using the selectivity index metric in Extended Data Fig. 1c (discussed in Line 134), where this distribution was plotted for all mice. We observed a continuous distribution of relative responses to S1 and S2, with many neurons exhibiting more selective or less selective responses to S1 or S2 (Extended Data Fig. 1c, discussed in Line 134). This followed a normal distribution. It was important for us to choose stimuli which drove selective responses to S1 or S2 in a large proportion of neurons in order to maximize classification sensitivity for specific stimulus reactivation patterns. Based on the reviewer's suggestion, we have added these new analyses to the manuscript.

4. Figure 1E (bottom)/line 872: Reporting brain motion in terms of shifts without showing the corresponding 2D correlation assumes that all registrations were equally successful. What methods were used to check the quality of the motion correction or search for out-of-frame (z-axis) movement?

We agree with the reviewer that showing the corresponding 2D correlation is useful in determining the quality of motion correction. We now report the phase correlation (of frames relative to the reference frame) version of the previous Fig. 1e, bottom (now Extended Data Fig. 2b). We found that the phase correlation in the 40 seconds surrounding a reactivation is extremely stable (Extended Data Fig. 2c, discussed in Line 160). The phase correlation drops very slightly after reactivation onsets. We believe this is because the evoked reactivation pattern is stereotyped and the correlation to the reference frame therefore is not an average of random activity patterns, in contrast to the surrounding frames.

We also confirmed the quality of motion correction by looking at raw vs. motion-corrected movies across time. See linked example movies of raw and motion correction versions:

Raw movie:
https://drive.google.com/file/d/1CWFTXF3kVnKkcxfLo3c0g1tb9O1sfQ4T/view?usp=drive_link

Motion corrected movie:
https://drive.google.com/file/d/1Zw1nvXgT63IwE9sd1jBhbtfHKl7PDwyX/view?usp=drive_link.

We unfortunately did not collect z-stacks during the experiments to search for out-of-frame / z-axis movement. However, we did track z-axis movement for a typical recording to determine the relationship between z-axis movement and x-y-axis movement (which we have tracked for all experiments). We did this by collecting a z-stack at the end of the session and finding the imaging frames that correlate best with frames in the z-stack. We found that large z-axis movements were also accompanied by large x-y-axis movement (Fig. R11). Unlike the 3-micron resolution that we had with x-y-axis movement, we had lower resolution with z-axis movement (~10 microns) and so we could only track larger changes. In our reactivation analyses, we excluded frames during the ITI any detectable x-y-axis movement at all which should also remove large z-axis movements.

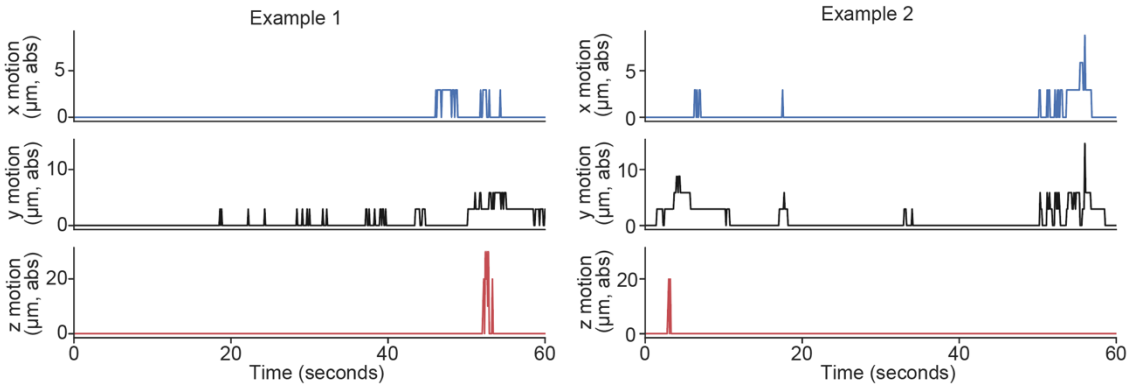

Figure R11. Two examples of x- (top), y- (middle), and z- (bottom) axis movement across 60 seconds of imaging. Large z-axis movements were accompanied by large x-y-axis movements.

5. Figure 2: Since reactivation rate is related to the perception of the visual stimulus, are there poorly attended stimuli presentations coinciding with either behavioral markers of attention or stimulus-response strength that could be related to a deficit in reactivation rates on an individual trial basis?

A behavioral marker of arousal, which may include an increase in attention to sensory stimuli, is pupil area (higher pupil area can relate to increased attention, Wierda et al., 2012). We quantified the relationship between pupil area and the rate of reactivations across trials. We found that there was a significant positive correlation between the pupil area during stimulus presentation and reactivation rate during the subsequent ITI (Extended Data Fig. 3e, discussed in Line 199). Note that this is not due to a systematic change in pupil area across the session, as peri-stimulus pupil area was constant on average across the session in our protocol. Thus, these findings suggest a trial-to-trial effect of stimulation during elevated arousal/attention on the subsequent reactivation rate.

We also performed a similar analysis for stimulus response strength and found that stimulus response strength was also positively correlated with the subsequent reactivation rate (Extended Data Fig. 3f, discussed in Line 199).

Therefore, signatures of poorly attended to stimuli such as low pupil area and low stimulus response strength are correlated with lower subsequent reactivations rates. This is interesting, as it might suggest that offline learning rates are lower for weaker or less attended stimuli. We have now reported and discussed these findings in the manuscript.

6. Figure 2B, F: at which points (both in time relative to the first stimulus, left, and in time relative to each stimulus onset, right) is the reactivation rate significantly higher than the baseline rate?

We performed statistics to show the time points for which the reactivation rates are significantly higher than baseline in Fig. 2b and Fig. 2f, as indicated using stars at the top above each point (Fig. R12, below). Due to multiple comparisons correction for numerous statistical tests (up to 17 statistical tests in one single panel), some of the points were not significant, although visually they are clearly above the baseline rate and were almost all significant before multiple comparisons correction (Fig. R12). We worried that this point would cause misunderstanding in our manuscript, so we elected to omit it but have shown it below:

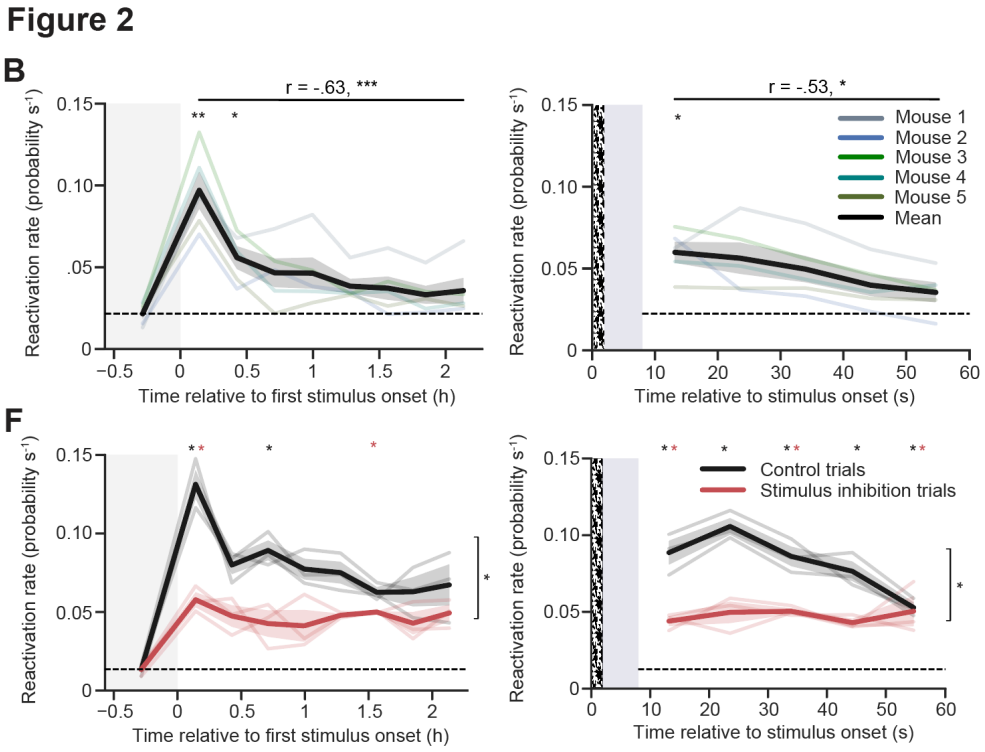

Figure R12. B) Same as Fig. 2b but with stars indicating points where reactivation rates were significantly above baseline, after correction for multiple comparisons. F) Same as Fig. 2f but with stars indicating points where reactivation rates were significantly above baseline.

7. Figure 2G: The inactivation abolishes the bias to the previous stimulus when there was inhibition, but is there a bias in reactivation rates toward the previous non-inhibited stimulus? Or does the silencing completely "reset" the reactivations back to random/baseline levels?

We found that the bias in reactivations produced from the previous non-inhibition trial does not seem to continue into the following inter-trial interval that follows stimulus inhibition (Extended Data Fig. 5g, discussed in Line 231). It should be noted however that the bias had already decayed to a low level by the end of the preceding trial (Extended Data Fig. 5g, discussed in Line 231). We also found that the reactivation rates plateau following the inhibition and do not continue to decrease, but instead reach a steady-state level that is above the rate observed in baseline activity prior to presentation of the first stimulus (Extended Data Fig. 5e, discussed in Line 229). Thus, inhibition does not 'reset' the reactivation rate back to baseline levels, and the rates continue to be much greater than baseline, even when we match for pupil area (Extended Data Fig. 5f, discussed in Line 229, see also response to Reviewer 1, Minor point #4). It is possible that, in addition to the biased reactivations that depend on activation of neurons within lateral visual cortex, a balanced number of reactivations of each stimulus may also be driven by other regions upstream of lateral visual cortex which becomes more apparent during stimulus inhibition trials (Zutshi and Buzsaki, 2023).

8. Figure 1B: It's unclear how the color scale "Azimuth" relates to the retinotopic mapping either an explanation should be added or a citation for Zhuang, J. et al. added to the main text (line 125) so it is clearer where to find this information.

We have cited Zhuang, J, et al. in the main text (Line 126) to make it clearer where to find the information for retinotopy. We apologize for any confusion.

9. Extended Data Figure 2C: The caption lists panels left and right, but they should be top and bottom.

We apologize for this typo and have corrected this.

10. Extended Data Figure 2C (bottom): It would be more straightforward to show the entire activity trace (e.g., Extended Data Figure 2l) and highlight the timepoints included in the reactivation analysis.

We agree with the reviewer that showing the entire activity trace and timepoints of reactivation analysis would be more informative in Extended Data Fig. 2c, bottom (now Extended Data Fig. 4b-c, discussed in Lines 204-205). We now show the entire trace (including stimulus presentation period) for both the trace of average activity throughout a trial and for the average peak normalized pupil area throughout a trial. We also highlighted the timepoints included for reactivation analysis.

11. Figure 3 F, G, Extended Data Figure 3B, Figure 4I, J: Significance annotations in the upper margin of the figure, in some cases pointing to the legend box, are difficult to read.

We removed Fig. 3f-g, Extended Data Fig. 3b, and Fig. 4i-j as we felt these figures caused more confusion than clarity in our manuscript, which may have also confused Reviewer 2 (see Reviewer 2's first Major concern, above). We instead have simplified this by quantifying the amount of Increase, Decrease, or No Change neurons that exhibited an increase or decrease in stimulus selectivity (see our response to this reviewer's Minor point #3 above for how we calculate stimulus selectivity) from early to late trials (Fig. 3h, discussed in Lines 293 and 295). This new plot now makes it clear that a greater proportion of Decrease neurons increase their stimulus response selectivity from early to late trials, while this was not the case for Increase neurons. This supports the idea that Decrease neurons are especially important in driving a reduction in the similarity of population responses to the two stimuli.

12. Figure 3H: Report the correlation score so it is clear how this example relates to the mouse averages reported in Figure 3I.

We have added the correlation score to this panel (now Fig. 3i).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have fully addressed my questions and concerns. This is an impressive piece of work, providing important novel insights for understanding replay and representational drift in the neocortex.

Referee #2 (Remarks to the Author):

I appreciate the significant effort that went into this revision. My comments below are simply an attempt to clarify my previous points.

The trivial result I suggested has indeed been addressed by the authors, but more broadly, the question is whether the “representational drift” that the authors observe from early to late trials is simply that the late trials are a subset of the neurons that were active in early trials (rather than drifting to a new representation). Given that the “Decrease” neurons are driving the orthogonalization, it appears that what the authors are studying is progressive sparsification of sensory representations (which has been extensively reported and studied mechanistically), not representational drift. Reviewer 1 suggests a similar scenario, where the very first stimulus responses are a noisy superset, which are then refined in subsequent trials to become a sparser “stimulus-specific subset” of the original representation. Importantly, such refinement of sensory representations has been described in the literature, even when reactivations were not present.

My second main point, and related to the above, was to show through causal manipulations of the ITI that reactivations are important for this “drift”. The technical challenges here are very understandable but perhaps there are simple ways to convince oneself that these reactivations are important— what if the behavior was performed without ITIs (and thus without reactivations), would one still see the same sort of orthogonalization and sparsification of sensory representations, and might this actually happen even faster in the absence of reactivations?

Referee #3 (Remarks to the Author):

The authors have addressed most of my critique with great care and depth. While the authors' effort to carry out 'causal' manipulation experiments has remained inconclusive at this stage due to technical challenges, I still support the publication of the study.