PNAS RESEARCH ARTICLE | NEUROSCIENCE OPEN ACCESS

Dynamic fading memory and expectancy effects in the monkey primary visual cortex

Yang Yiling[a], Johanna Klon-Lipok[b], Katharine Shapcott[a], Andreea Lazar[a], and Wolf Singer[a,b,c,1]

Edited by Robert Desimone, Massachusetts Institute of Technology, Cambridge, MA; received August 30, 2023; accepted January 10, 2024

In order to investigate the involvement of the primary visual cortex (V1) in working memory (WM), parallel, multisite recordings of multi-unit activity were obtained from monkey V1 while the animals performed a delayed match-to-sample (DMS) task. During the delay period, V1 population firing rate vectors maintained a lingering trace of the sample stimulus that could be reactivated by intervening impulse stimuli that enhanced neuronal firing. This fading trace of the sample did not require active engagement of the monkeys in the DMS task and likely reflects the intrinsic dynamics of recurrent cortical networks in lower visual areas. This renders an active, attention-dependent involvement of V1 in the maintenance of WM contents unlikely. By contrast, population responses to the test stimulus depended on the probabilistic contingencies between sample and test stimuli. Responses to tests that matched expectations were reduced which agrees with concepts of predictive coding.

working memory | primary visual cortex (V1) | non-human primates | expectancy effects | neural dynamics

Significance

Non-invasive imaging data from human subjects suggest that contents held in working memory (WM) can be read out from the primary visual cortex (V1). With unit recordings from monkey V1 engaged in a WM task, we confirm that information about the stimulus retained in WM (the sample) persists during the retention period in the population activity of V1 and can be enhanced by increasing V1 activity with unspecific visual stimulation. However, this trace of the sample is present also under passive viewing conditions that do not engage WM. We conclude that the persisting information is due to "fading memory", a hallmark of the reverberating dynamics of local recurrent networks. Still, WM could use this lingering trace in V1 if required.

Working memory (WM) refers to the ability to maintain and manipulate information in the absence of input. WM has traditionally been attributed to higher-order cortical areas, in particular, the prefrontal cortex (1–3) and more recently to cooperative processes across multiple brain areas (4–6). There is also evidence for a recruitment of primary sensory areas like V1 in visual WM processes (7–15). For example, information held in visual WM can be decoded from V1 activity (16–26); the volume (27) and activity (22, 28, 29) of V1 are positively correlated with behavioral performance in WM tasks. Early evidence suggests that WM is mediated by persistent firing during the delay period (1–3, 30–33). However, this view has been contested (34) because WM contents were decodable only from short, temporally segregated bouts of activity (35–37). Other evidence suggests the existence of covert, activity-independent traces (20, 38–44) that can be activated by "pinging" the brain with unspecific stimuli (20, 21) or transcranial magnetic stimulation (TMS) (39).

However, most of these studies used neuroimaging techniques in humans in order to retrieve the traces of WM contents, which limits the identification of the underlying neuronal signals. Thus, it is unclear whether the signals recorded from V1 that permit decoding of WM contents reflect reverberating activity ("fading memory") within the recurrent networks of lower visual areas (45, 46), or result from top–down projections that involve V1 in WM. Therefore, we set out to investigate at the neuronal level whether WM contents can be decoded from neuronal population activity in V1, whether pinging could revive WM traces, and whether persistent information about the stimulus could be attributed to fading memory in local circuits or showed the attention- and task-dependent properties of WM related activity.

Another goal of the present study was to investigate whether V1 responses are shaped by priors stored in memory. Responses of V1 neurons to external stimuli depend on both stimulus features and internal priors (47–50). Stimuli matching priors of Gestalt principles modify the synchronization patterns (51, 52), correlation structure (53), sequential activation (54), and response amplitude (55) of neuronal responses. Spatial predictability derived from stimulus context reduces firing rate and/or enhances oscillatory synchronization among neurons in V1 (51, 52, 56–59) and temporal predictability of stimuli suppresses V1 activation in humans (60, 61). These observations support the notion that the visual system learns the statistical regularities in sensory input to optimize its responses (48, 49, 62).

To pursue the above goals, we performed parallel multisite electrophysiological recording in awake monkey V1 while the animal performed a delayed match-to-sample (DMS) task. To test the possibility that WM content could serve as prior, we introduced probabilistic associations between stimulus-pairs in the DMS task in order to establish implicit predictions and investigated whether such priors modified V1 responses. We found that V1

Author affiliations: [a]Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt am Main 60528, Germany; [b]Max Planck Institute for Brain Research, Frankfurt am Main 60438, Germany; and [c]Frankfurt Institute for Advanced Studies, Frankfurt am Main 60438, Germany

Author contributions: Y.Y., K.S., A.L., and W.S. designed research; Y.Y., J.K.-L., and K.S. performed research; Y.Y., K.S., and A.L. contributed new analytic tools; Y.Y. analyzed data; and Y.Y. and W.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: wolf.singer@brain.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2314855121/-/DCSupplemental.

Published February 14, 2024.

population activity maintained an intrinsic, latent trace of the sample stimuli regardless of the need to actively engage WM. This trace could be reactivated and strengthened by an irrelevant, non-specific stimulus (20, 21). Once probabilistic priors were established for sample-test pairs, V1 responses to predicted test stimuli were reduced.

Results

We trained two monkeys to perform a DMS task that required stimulus encoding, retention of stimulus identity in WM, and a manual forced-choice response. During the task, the monkey fixated a spot at the center of the screen. Two stimulus images ("sample" and "test") were presented sequentially for 500 ms each, separated by a delay period of 1,500 ms (Fig. 1*A*). The animal had to report whether the two stimuli were the same ("match") or different ("nonmatch"), by pushing a mechanical lever forward or backward, respectively. The numbers of match and nonmatch trials were balanced in order not to bias the animal's behavioral response. Stimuli were standardized images (63, 64) of simple objects displayed on a gray screen (Fig. 1*B*). A set of three images were used in each session (counterbalanced for sample and test positions), and the set of images varied between sessions. The position and size of the stimuli were tailored for each monkey to cover the

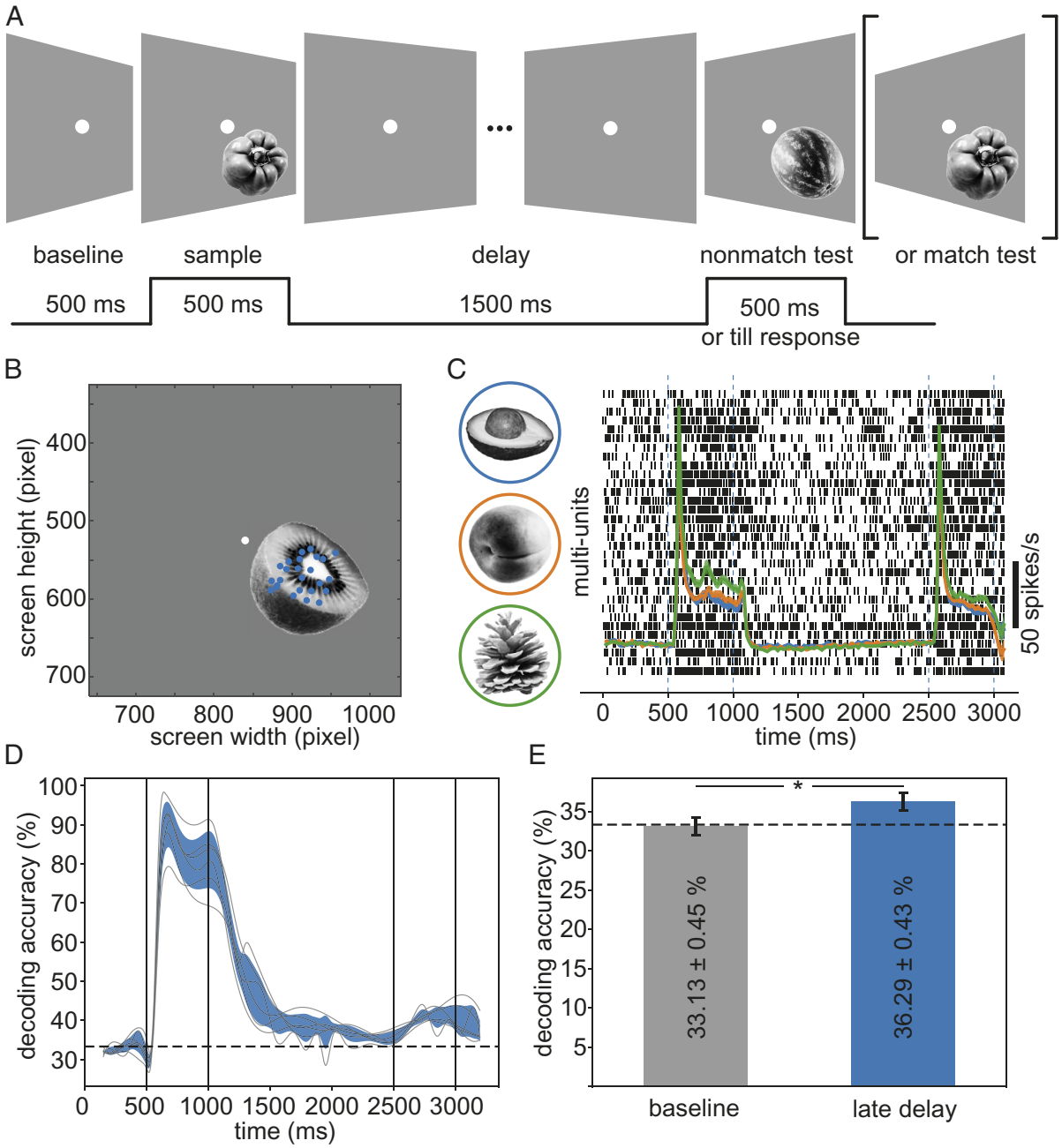

Fig. 1. Robust trace of stimulus-specific information during the delay period. (*A*) Task structure and trial time course. (*B*) Positions of the stimulus, fixation point (white dot), and RF (blue dots). (*C*) Raster plot of multi-unit activity in a single trial, overlaid with average population multi-unit firing rates for the three demo stimuli in match trials. Shades denote 95% CI (barely visible due to the large number of trials). (*D*) Time-resolved accuracy of decoding sample stimulus identity based on firing rate vectors. Gray traces: Results from individual sessions; blue shades: 95% CI around session average (n = 6). (*E*) Comparison of decoding accuracies between late delay period (2,000 to 2,500 ms) and pre-stimulus baseline (0 to 500 ms). Error bars: 95% CI. Error numbers in legends denote SEM.

ensemble of the receptive fields (RF) of the respective recording sites. For Monkey 1 (Fig. 1*B*), stimuli subtended 4.4° of visual angle, and their center was 2.36° lateral to the vertical meridian and 1.34° below the horizontal meridian. For Monkey 2, stimuli subtended 7.84° of visual angle, and their center was 4.05° lateral to the vertical meridian and 2.70° below the horizontal meridian. On average, Monkey 1 performed 78.1 ± 1.7% (SEM, n = 6 sessions) correct responses (*SI Appendix*, Fig. S1). The average reaction time for correct responses was 632.0 ± 174.7 ms (SD, median 586.4 ms. *SI Appendix*, Fig. S1). The RF positions and the behavioral performance for Monkey 2 are shown in *SI Appendix*, Fig. S2. Throughout the paper, analyses were always performed separately for each animal.

Fading Trace of Sample Stimulus. Multi-unit activity (MUA) was recorded from visual area V1 (left hemisphere) with a chronically implanted 32-channel microdrive system (Gray Matter Research, Bozeman, Montana, USA) while the animal performed the DMS task. The increased firing of neurons evoked by the sample stimulus rapidly decayed during the WM delay period to the pre-stimulus baseline level or even below (Fig. 1*C*). To compare the activity levels between the delay and baseline periods, we measured the spike counts (window size = 300 ms) for each channel and stimulus (pooled across sessions) at four time intervals in the delay period and compared the spike counts with those in the baseline period preceding the sample stimulus (*SI Appendix*, Fig. S3 for Monkey 1, *SI Appendix*, Fig. S4 for Monkey 2). Spiking activity dropped below the baseline level after the offset of the sample stimulus (1,300 to 1,600 ms: t = 13.91, $P < 0.01$; 1,600 to 1,900 ms: t = 15.64, $P < 0.01$; 1,900 to 2,200 ms: t = 7.51, $P < 0.01$. Paired t test, Monkey 1. Statistics for Monkey 2 in *SI Appendix*, Fig. S4) but recovered to a level slightly above baseline toward the end of the delay period (2,200 to 2,500 ms: t = –3.24, P = 0.00014). There was no clear indication for a sustained elevation of discharge rates during the delay period.

To examine the amount of stimulus-specific information in the population vector of responses to the sample stimulus, we trained decoders (linear discriminant analysis, LDA) at successive time points to predict the sample stimulus identity from the population firing rate vector (window size 100 ms, step size 50 ms). The decoding accuracy (Fig. 1*D* and *SI Appendix*, Fig. S4*B*) was highest during stimulus presentation (500 to 1,000 ms), decayed after stimulus offset but remained above chance level (33.3%, 3 stimuli per session) throughout the delay period. The average decoding accuracy (36.29 ± 0.43%, SEM, n = 6 sessions) for the sample stimulus in the last 500 ms of the delay period (from 2,000 to 2,500 ms) was still significantly above chance level (t = 6.26, P = 0.00153, t-test, two-sided unless noted otherwise), and higher than the baseline level (33.13 ± 0.45%, SEM, t = –3.68, P = 0.0142, paired t-test) which did not differ from chance (t = –0.42, P = 0.691). Similar results were obtained from three other sets of experiments which used different numbers of stimuli (*SI Appendix*, Fig. S5*A*) and also from Monkey 2 (*SI Appendix*, Fig. S4 *B* and *C*). To test further the robustness of the findings against the variation of stimuli, in a separate set of experiments, we used gratings as stimuli in the same DMS task. Here, the animal was required to discriminate the orientations of the sample vs. test gratings (spatial frequency 3 cycles per degree; 4 non-cardinal orientations per session; 6 sessions in total). Interestingly, in this set of experiments, the trailing sample stimulus information decayed rapidly to the baseline level (*SI Appendix*, Fig. S5*B*). However, in this DMS task with grating stimuli, the animal's performance (65.6 ± 1.5%, n = 6) was worse (t = 5.54, $P < 0.01$) than in the standard DMS task (78.1 ± 1.7%, n = 6), although still above chance level (t = 10.48, $P < 0.001$). As the simple grating stimuli could give rise to retinal afterimages, these results suggest that the trailing traces of the natural sample stimuli were not due to retinal adaptation. Thus, despite the low firing rates that did not differ much from baseline toward the end of the delay period, the population activity retained a low but robust trace of sample stimulus-specific information in the experiments where natural objects were used as stimuli.

To test whether the trailing stimulus-specific information was actually related to the WM task, we performed control experiments with passive viewing on an animal naive to the DMS task. Here, the two stimuli were shown at the same time points as in the DMS task, but the monkey was only required to attend to the fixation spot and was rewarded for detecting and responding to a color change of the fixation spot (*Methods*). The monkey had to push the lever forward or backward if the fixation point color changed to green or blue respectively. Nevertheless, stimulus-specific information about the irrelevant sample stimulus persisted throughout the delay interval (*SI Appendix*, Fig. S6), suggesting that the trailing stimulus information was not caused by the requirement to memorize the sample stimulus.

Reactivation of Latent Memory by Impulse Stimulus. Modeling (41) and neuroimaging (20) studies reported that a global, unspecific activation of neuronal networks can reveal latent or covert traces of information held in memory. To examine whether such a manipulation could enhance sample-specific information during the delay interval, we modified the DMS task and inserted a full-screen impulse stimulus (100-ms duration, 100% intensity) in the delay interval (Fig. 2*A*). To indicate to the monkey that this intervening stimulus was irrelevant for the task, we presented different types of impulse stimuli: linear gratings (2 sessions with 0° orientation, 2 sessions with 90° orientation), a concentric grating (1 session), or a blank white stimulus (1 session). We also applied these stimuli in the passive viewing tasks, presenting one of the four stimuli in four different sessions. These stimuli were applied at either 500 ms or 1,000 ms after the offset of the sample stimulus (i.e., 1,500 ms or 2,000 ms in a trial. Fig. 2*B*). We then performed the same time-resolved decoding analysis as described above, pooling the results across all the different sessions (n = 10). As shown in Fig. 2*B* (*SI Appendix*, Fig. S7 for Monkey 2), decoding accuracy increased about 100 ms after these impulse stimuli and remained enhanced for 100 to 200 ms (at 1,700 ms: accuracy with impulse 43.11 ± 1.02% SEM, accuracy without impulse 35.38 ± 1.01%, t = 4.16, P = 0.0024; at 2,200 ms: accuracy with impulse 39.63 ± 0.52%, accuracy without impulse 34.25 ± 0.82%, t = 4.79, P = 0.00099, n = 10 sessions). The increase in decodability resembled closely in time the transient firing rate increase evoked by the impulse (Fig. 2*B*).

To investigate the time course of these impulse effects at higher temporal resolution, we systematically varied the timing of the impulse, increased the delay interval from 1,500 to 1,800 ms, reduced the duration of the sample stimulus from 500 to 400 ms and used only full screen white flashes of 100 ms duration (Fig. 2*C*). This allowed us to assess the impulse effects with a temporal resolution of 200 ms across a total of eight experimental sessions. Again, the impulse stimuli led to a transient enhancement of decodability of the sample stimulus, and this enhancement of decodability was closely related in time with the transient increase in firing rate (Fig. 2*C*). This impulse-related increase in decoding performance was also observed in the passive viewing experiments (*SI Appendix*, Fig. S7). These results suggest that the trailing trace of sample

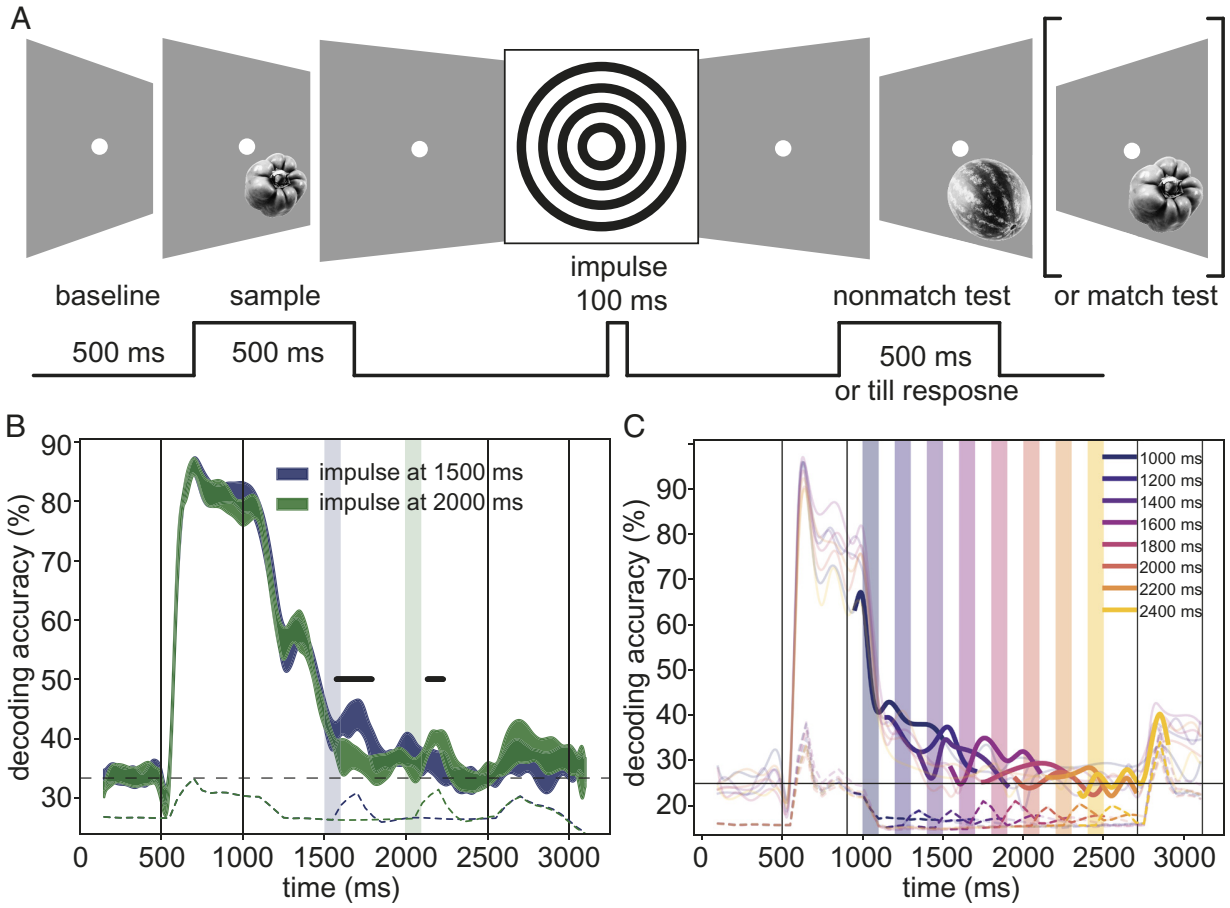

Fig. 2. Visual impulse stimulus enhances latent memory trace. (*A*) Modified DMS trial structure. A full-screen impulse stimulus (100 ms) is inserted in the delay period. (*B*) The accuracy of decoding the sample stimuli for two impulse conditions (blue: impulse at 1,500 ms; green: impulse at 2,000 ms). Shaded areas indicate 95% CI. Vertical blue and green bars mark the positions of the impulse stimuli at 1,500 and 2,000 ms, respectively. Dashed lines (color-coded) indicate average population firing rates. Vertical black lines flank the sample (500 to 1,000 ms) and test (2,500 to 3,000 ms) stimulus intervals. The horizontal black line marks chance level decoding accuracy. (*C*) Decoding accuracy of the sample stimuli. Similar to (*B*) but impulse stimuli were applied at systematically varied delays (colored bars). The colors of the traces correspond to the different flashes and highlight the changes in decoding accuracy induced by the flashes. Color-coded dashed lines indicate average population firing rates.

stimulus-specific information can be transiently enhanced as the V1 neurons are driven to fire by an unspecific impulse stimulus.

Enhanced stimulus trace with increased neuronal firing could simply be due to the fact that decodability of population vectors increases with discharge rate (45). However, this relationship does not always hold. In the DMS task (*SI Appendix*, Fig. S8) as well as its passive viewing version (*SI Appendix*, Fig. S9), stimuli evoked higher population firing rates when they were presented as sample rather than test. This difference was significant only in the last sustained response phase (300 to 350 ms after stimulus onset: $t = 5.67$, $P = 0.00238$, n = 6 sessions. t-test) but not during the response onset transient (e.g., 50 to 100 ms: $t = 1.15$, $P = 0.303$). However, the accuracy of decoding stimulus identity was higher for responses to the test than the sample stimulus, during both transient and sustained response phases (*SI Appendix*, Figs. S8 and S9). Reduced firing rate and better decodability to the test stimulus also held when we performed the same analyses on nonmatch trials only (*SI Appendix*, Fig. S10), to rule out potential effects of repeated exposure to the same stimuli as is the case in the match trials. As an additional control, we subtracted the mean firing rates at each time point from the rate vectors and decoded the sample stimulus identity from these mean-equalized (to zero) vectors, and the results remained the same (*SI Appendix*, Fig. S11). Thus, better decodability must have been due to other reasons than enhanced discharge rate.

Reduced Firing to Stimuli Matching Priors. After the animal had learned the DMS task, we investigated whether the animal could learn implicit probabilistic associations between sample and test stimuli and use this information in the DMS task to predict the nature of the test stimulus given a particular sample. Predicted stimuli evoke smaller responses than unexpected stimuli (56, 57, 61). Therefore, we examined whether the same holds for test stimuli that were predicted with high or low probability by the sample stimuli. To this end, we modified the DMS task by introducing probabilistic pairing between sample and test stimuli in the nonmatch condition, such that the sample stimulus would predict with variable probability the upcoming test stimulus. Specifically, as shown in Fig. 3*A*, in nonmatch trials (50% of all trials), the sample stimulus was followed by one of the two test stimuli with either high (40%, e.g., onion to kiwi, apple to paprika) or low (10%, e.g., onion to paprika, apple to kiwi) probability. The occurrence of sample–test pairs in the two probability conditions was counterbalanced. The other 50% of trials were match trials: Each sample stimulus was followed by itself. In each session, we used two stimuli as sample and another two stimuli as test. The same set of four stimuli was used for the experimental sessions performed within a week to permit enough repetitions for the learning of the probabilistic association but different sets of stimuli were used for sessions in different weeks.

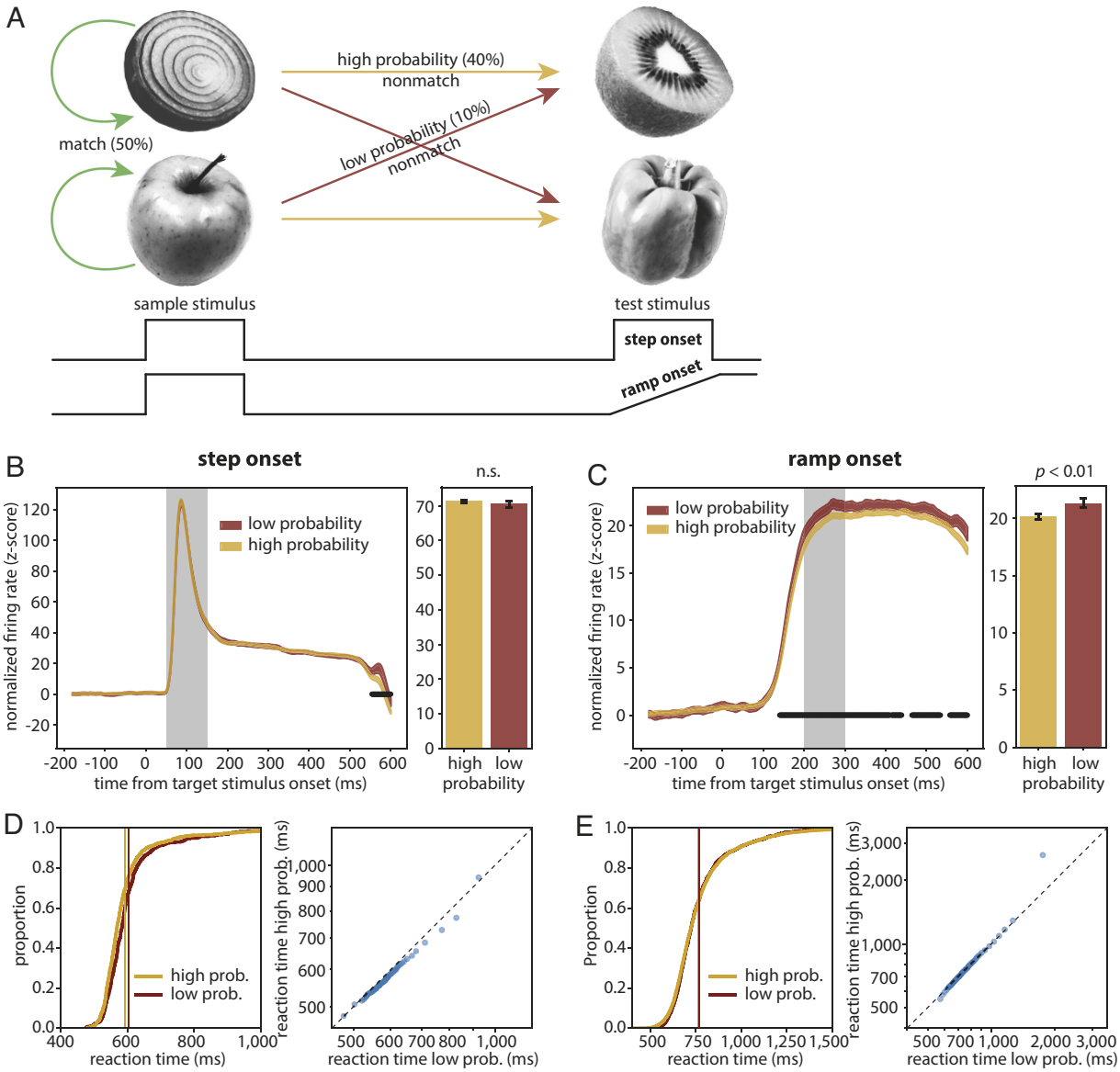

Fig. 3. Neural and behavioral results for learning implicit probablistic associations. (*A*) Task structure. (*B*) Traces: Normalized firing rate responses to high- and low-probability test stimuli in step onset condition. Bar plots: Average firing rates (measurement interval marked in gray). Width of the traces and error bars represent 95% CI. The black line marks statistically significant differences between high- and low- probability conditions. Note: The spurious difference at 550 to 600 ms is caused by the fact that trials are cut off at behavioral response time and that response time is shorter for high-probability condition. (*C*) Same convention as (*B*) but for ramp onset condition. (*D*, *Left*) Cumulative density functions of reaction times in step onset condition. Vertical lines mark average reaction time. (*Right*) Quantile–quantile plot between reaction times for low (abscissa) and high (ordinate) probability conditions. The diagonal dashed line marks identity. (*E*) Same convention as (*D*) but for the ramp onset condition.

To test whether neuronal responses to the test stimulus differed in high vs. low probability conditions, we computed test-evoked population firing rates (z-scored to pre-sample baseline per channel, summed across channels and normalized for pooling of trials from different sessions). We found no difference between high and low probability conditions (Fig. 3*B*. Average firing rate within 50 to 150 ms after test onset: 71.31 ± 0.24 (SEM) for high probability, 70.62 ± 0.47 for low probability; $t = 1.28$, $P = 0.20$). One reason for the lack of differences could be that subtle effects might have been overridden by the sharp initial transient responses caused by sudden stimulus onset. To examine this possibility, we slowly ramped up the intensity of the test stimulus (Fig. 3*A*), expecting that such gradual visual input would dampen the abrupt increase in firing rate. The duration of the ramp was kept fixed for each session, but we varied this parameter across sessions (Monkey 1: 500 ms, 8 sessions; 2,000 ms, 4 sessions; 3,000 ms, 8 sessions. Monkey 2: 2,000 ms, 14 sessions). Interestingly, in the experiments with ramping stimuli, for Monkey 1 the firing rate responses were weaker ($t = -4.84$, $P = 1.34 \times 10^{-6}$, all ramp durations combined, n = 20) to the test stimuli associated with high (20.15 ± 0.10 SEM, averaged within 200 to 300 ms after test onset) than low (21.33 ± 0.20) probability (Fig. 3*C*). This difference was already evident in the early response phase (Fig. 3*C*). As sanity check, we compared the firing rates evoked by the same stimuli when they were in the sample position. There were no differences in high and low probability conditions (*SI Appendix*, Fig. S12*A*). To examine whether this probability-dependent difference in firing rate to the test stimulus was related to learning, we stratified the data into early (first experience with each set of stimuli and their associations) and late (last experience with the

same sets of stimuli) sessions (200 trials in each of the early and late sessions, respectively, to equalize sample size), and found that the reduction of responses to high-probability test stimuli was only present in the late sessions (*SI Appendix*, Fig. S12 *C* and *D*, 22.43 ± 0.33 for high probability, 24.078 ± 0.64 for low probability, $t = -2.08$, $P = 0.038$) and not in the early sessions (19.93 ± 0.43 for high probability, 21.31 ± 0.87 for low probability, $t = -1.31$, $P = 0.19$). The probability-dependent response difference in late rather than early sessions suggest the possibility that this effect was due to learning. However, we were unable to reproduce this result in Monkey 2 (*SI Appendix*, Fig. S13*A*).

To test whether probabilistic association between sample and test stimuli had an effect on the animal's behavior, we analyzed the animal's reaction (lever pressing) time after test stimulus appearance. For Monkey 1, when the test was presented with step onset, the reaction time was shorter for high (592.33 ± 98.97 ms, SD) than low (604.53 ± 96.15 ms) probability test stimuli (Fig. 3*D*, $t = -2.60$, $P = 9.29 \times 10^{-3}$, *t*-test on log-transformed reaction time to ensure normality). However, when the test stimulus was presented with ramp onset, there was no difference in reaction time in the two probability conditions (Fig. 3*E*, high probability 767.03 ± 178.87 ms, low probability 766.93 ± 165.30 ms; $t = -0.51$, $P = 0.61$, *t*-test). The same results held for the stratification in early (high probability 941.05 ± 266.52, low probability 936.35 ± 239.92; $t = -0.028$, $P = 0.98$) and late (high probability 75.56 ± 149.75, low probability 744.28 ± 105.91; $t = 1.61$, $P = 0.11$) sessions (*SI Appendix*, Fig. S12 *E* and *F*). Between early and late sessions, the animal's response accuracy (*SI Appendix*, Fig. S12*B*) increased from 72.0 ± 2.8% (SEM) to 94.8 ± 2.2%, and reaction time decreased from 913.5 ± 269.7 ms (SD) to 713.8 ± 167.6 ms. For Monkey 2, we only used ramped test onset, and found that reaction times were slightly shorter for high (487.44 ± 121.04 ms) than low (495.51 ± 132.75 ms) probability tests (*SI Appendix*, Fig. S13*B*), but the difference was not statistically significant ($t = -1.73$, $P = 0.084$, *t*-test). However, if we first calculated the average reaction time per session for high and low probability conditions, respectively, and then performed pair-wise statistics across sessions, Monkey 2 also seemed to respond faster to high (488.31 ± 36.67 ms) than low (496.50 ± 37.93 ms) probability tests (*SI Appendix*, Fig. S13*D*, $t = -2.72$, $P = 0.0176$, paired *t*-test). This pairwise comparison revealed that the average reaction time per session was systematically shorter for the high probability test, although pooling data from all sessions did not uncover any statistically significant differences. Therefore, it seems that the animals may have learned and took advantage of the pairing between sample and test stimuli and responded faster to high-probability test stimuli.

Discussion

In this study, we trained monkeys to perform a WM task and investigated the effect of both task-related and task-irrelevant factors on V1 neural activity. We found that V1 neurons did not show persistent spiking activity during the memory retention interval. However, decoding analysis revealed that the population vector of spiking activity contained a robust trace of the information about the sample stimulus despite the low firing rate. This trailing memory trace was apparently not caused by the specific demands of the WM task because it was also present in the passive viewing tasks. The trace of the preceding sample stimulus could be reactivated by an impulse stimulus that was unrelated to the sample but enhanced firing rates. Moreover, we found in the second series of experiments, that the amplitude of responses to the test stimulus depended on its expected probability and was reduced when the test matched prior expectation.

Before discussing these findings, a few methodological considerations are warranted. To examine the role of V1 in WM, we as well as other investigators relied on decoding methods to extract WM content from V1 activity. However, decodability of WM content from V1 activity does per se not imply an involvement of V1 in WM. A better strategy would be to examine whether decoders could differentiate between the items that are required to be retained in WM (e.g., test) and those that are cued to be ignored ("distractor"). Such a distractor design is common in human psychophysics but challenging for non-human primates. But even then, improved decodability for WM items could reflect top–down effects related to predictive coding or feature-specific attention rather than an involvement in the maintenance of WM contents. These problems could in principle be overcome with loss of function experiments, i.e., transient inactivation of V1 during the retention interval (65). Another problem is that we were not able to determine whether the lingering memory traces in the DMS task and the passive viewing control had the same format or structure. This question could be resolved by performing transfer decoding to examine whether decoders trained on the DMS task could generalize to the passive viewing task and vice versa. Unfortunately, this was not possible because we used different sets of stimuli across sessions and the recorded signals likely drifted over days. Moreover, the passive viewing condition may not be an ideal control for the lack of WM engagement because the monkeys could have used different strategies from what we expected. We attempted to reduce this possibility by applying the passive viewing task in a monkey that was not trained on the WM task. However, we cannot exclude completely the possibility that the animal nevertheless paid attention to the task-irrelevant stimuli or stored them in WM. Furthermore, the response modifications associated with test probability were not replicable in the second animal. Possible reasons are fewer sessions and less ramp variations for Monkey 2 than Monkey 1, different eccentricity of recording sites (RFs) and sampling bias. However, the results from Monkey 1 were robust and the stratification test provided unequivocal evidence for a learning-dependent process. Finally, our failure to retrieve WM-related information from V1 spiking activity may have been due to limitations of the decoder. We used a linear decoder which may have missed information contained in higher-order correlations (53) or the temporal order of responses (54).

Despite the low activity during the delay interval, the population firing rate vector contained information specific for the sample stimulus. The fact that this information was present in both DMS and passive viewing tasks makes it unlikely that it is related to an intentional effort to remember the sample. It is also unlikely that the lingering stimulus trace reflects a retinal afterimage, because stimulus contrast was low. Moreover, grating stimuli, which are typically used to induce afterimages, did not produce such lingering stimulus traces. The fact that the weaker traces after gratings were associated with worse behavioral performance compared to conditions with natural stimuli might be taken as evidence that lingering V1 activity is actually involved in the maintenance of WM. However, the results of the passive viewing task obtained in a naive animal do not support this assumption. Therefore, we favor the interpretation that the lingering trace reflects a form of fading memory that is maintained by the intrinsic dynamics of recurrent cortical networks (45, 46).

If so, this raises the question of why information about complex natural stimuli persists longer than information about the orientation of gratings. Our results let it appear unlikely that this is simply due to a prolongation of reverberating responses to natural stimuli. Another possibility is that natural scene stimuli engage a larger

network of recurrently coupled visual areas than gratings because they also match priors stored in the functional architecture of higher cortical areas (53). This would imply that cooperation among multiple visual areas can enhance the persistence of stimulus-specific correlation structures in V1 activity. This interpretation is supported by the recent observation that stimulus-specific information persists longer in population responses recorded from areas V1 and V4 if responses are evoked by natural scene stimuli rather than by manipulated stimuli in which certain statistical regularities of natural stimuli were removed (54). Also in these experiments, there was no simple relation between discharge rates and decodability. The interpretation is also consistent with the finding that the cortex has a hierarchy of intrinsic timescales, with primary sensory and prefrontal areas exhibiting shorter and longer timescales, respectively (66). The differences in intrinsic timescales may result from the differences in cell type and receptor composition, neuronal excitability, and dendritic morphology between different cortical areas (67–70). The engagement of higher-order sensory areas by natural stimuli may extend the timescale of information retention in V1. Taken together, fading memory is not a simple consequence of prolonged reverberation, a possibility worth further examination.

Information could also be stored in stimulus-specific synaptic modifications that persist without requiring any sustained activity (41, 44). The finding that intervening impulse stimuli that transiently increased firing rates enhanced decodability agrees with this interpretation. In simulations, latent synaptic memory traces could be reactivated by unspecific network-wide stimulation (41). Likewise, in experiments with human subjects, TMS or strong visual stimulation revived latent contents of WM (20, 39). In our study, the reactivation of stimulus-specific response vectors was similar in the WM tasks and the passive viewing controls. This suggests that the mechanism responsible for the fading memory trace can be activated by passive exposure and does not involve attention, which is in line with results of experiments on fading memory performed under anesthesia (45). However, this does not imply that the lingering trace cannot be exploited by WM when required, nor does it exclude an influence of WM content on V1 processes. Identifying the site and mechanism of storage of WM contents during natural behavior and under impulse stimulation is beyond the scope of the present study and awaits further investigation.

The stimuli appearing in the test position evoked lower firing rates but were more decodable than in the sample position, in contrast to previous reports where higher firing rates improved decodability (45). The present results bear similarities with bottom–up mechanisms, such as adaptation and repetition suppression. However, several observations render classical adaptation unlikely. For briefly presented stimuli (400 to 500 ms) as used for our sample stimuli, adaptation acts mainly on the early transient response component, has only a weak or no influence on the late response phase (71–74) and vanishes within a few hundred milliseconds (71, 73, 75). By contrast, in our experiments, only the late response phase was attenuated. Moreover, the effects were the same in nonmatch trials where the test stimulus was preceded by a different, therefore non-adapting, sample stimulus. Repetition suppression is also an unlikely explanation for the results, for the following reasons: i) Unlike in area IT (76), repetition suppression in V1 builds up only over multiple trials (77), and thus on a longer timescale than in our experiments; ii) reduced firing to the test stimulus was also observed in nonmatch trials where the test stimulus did not repeat the sample stimulus; and iii) repetition suppression in both V1 and IT affects the early response phase (76, 77), whereas here only the late response phase was attenuated. Another reason for the attenuation of responses to the test stimulus could be the predictability of the time of appearance and/or the need to respond to it. Since the trial timing was fixed, the monkey could predict precisely when the test stimulus would appear. Expectation and predictability have been shown to reduce neuronal firing (61, 78–82), to improve stimulus encoding (60, 83), and to enhance gamma synchronization (58, 84, 85) in sensory areas. Notably, Todorovic and de Lange (79) showed that, in line with our results, expectation-dependent suppression influenced the late response component (100 to 200 ms), whereas repetition suppression acted on the early component (40 to 60 ms). In sum, we favor the interpretation that the differences in sample- vs. test-evoked responses result from temporal expectation.

The effect of expectation on V1 activity is also evident in the DMS task in which we varied the probability with which a sample stimulus predicted a particular test stimulus. The probabilistic association between sample–test pairs was designed to establish an internal prior which rendered the sample stimulus predictive of the upcoming test stimulus. We found that test stimuli of high-probability evoked reduced firing rates as compared to the low-probability stimuli. Because the same set of test stimuli was used in the two probability conditions and the stimulus set varied across sessions, test stimulus-specific differences in firing rate are unlikely to explain the effect. Moreover, the probability-dependent modification of firing rate emerged only in late but not in early sessions. As adaptation and repetition-dependent effects are unlikely (see above), these observations suggest learned predictability as a likely cause for the dependence of firing rate on the probability of the test stimulus. Although our data do not allow us to identify the site where the prior for the prediction is generated, the ramping paradigm revealed that the prior-associated effect kicked in already in the initial response phase, suggesting fast access to information about the nature of the expected stimulus. The reduction of responses to predictable stimuli is in agreement with reports of reduced activation of the visual cortex by stimuli that comply with predictions (49, 56, 57, 60, 61, 78, 81, 86). This suppression has been interpreted in the context of predictive coding as a mechanism to facilitate perceptual inference (47, 48, 50, 62), to reduce redundant signals originating from compressible stimuli (56–58) and to improve stimulus representation (60, 83).

In summary, the intrinsic dynamics of early visual areas are capable of maintaining re-activatable traces of complex visual stimuli. We propose as the most likely mechanism the fading memory that is characteristic of recurrent networks. These lingering memory traces do not seem to depend on active, attention-dependent WM processes but could of course support WM if required. Ample evidence indicates that the responses of V1 neurons depend to a large extent on the match between sensory evidence and priors. Some of these priors have been acquired during evolution, are complemented by experience-dependent developmental pruning and perceptual learning, and are stored in the functional architecture of the visual cortex (reviewed in Singer, 48). Other priors are derived from the actual context in which stimuli are presented (56, 57, 87). Our results indicate that also predictions derived from learned associations impact responses in V1. Although initial acquisition of these associations between temporally distant stimuli very likely involves WM, once established, these associations must be stored in long-term memory. As our results indicate, this covert knowledge about the likelihood of the appearance of a particular stimulus is available in the primary visual cortex. We consider it unlikely that the association between sample and test is formed in V1 and therefore favor the interpretation that the priors set up dynamically by the sample stimulus are conveyed to V1 by top–down projections.

Methods

Behavioral Task. As previously described (88), results presented here were obtained from two adult rhesus monkeys (*Macacca mulatta*. Monkey 1: male, 11 kg, 12 y old. Monkey 2: female, 9 kg, 17 y old). All experimental procedures were in compliance with the German and European regulations on laboratory animal protection and welfare and were approved by the local authority (Regierungspräsidium Darmstadt). The monkey was seated 60 cm in front of a screen (Samsung SyncMaster 2233RZ; 120 Hz refresh rate) inside a dark booth. The monkey initiated a trial by fixating at a white fixation dot displayed at the center of the screen, and had to maintain fixation on the fixation dot until the trial ended. The eye position was monitored with the EyeLink tracker (SR Research, Ottawa, ON, Canada).

In the DMS task, two stimuli were presented sequentially and the monkey had to report whether the two stimuli were the same (match) or different (nonmatch). A trial started with a fixation period of 500 ms, during which the screen was blank. Then, the first stimulus (sample) was presented for 500 ms, followed by a delay period of 1,500 ms after which the second stimulus (test) was presented. The test stimulus disappeared once the monkey responded and kept on for maximally 500 ms in case of delayed responses. The monkey had to respond by moving a two-way mechanical lever; forward in match trials and backward in nonmatch trials. Monkeys were not rewarded for responding swiftly. The number of match and nonmatch trials was pseudo-randomized to be equal. A correct response was rewarded with a drop of water or juice. If the monkey broke fixation (1.3° around fixation point), or moved the lever before test onset, the trial was aborted.

The trial structure of the passive viewing control experiment mimicked that of the DMS task. In the passive viewing task, the monkey had to maintain fixation and respond to a color change of the fixation spot in order to be rewarded. During fixation, the stimuli were presented as in the DMS task but were irrelevant to the animal. The fixation dot changed its color to either green or blue, requiring forward or backward moves of the lever, respectively.

Stimulus Design. As previously described (88), the stimuli were images of single isolated objects with transparent background. The visible region of the image was normalized to equal pixel intensity (0.5) and root-mean-square contrast (0.275) and covered the classical RF of the recorded multi-units. The image size was 160 × 160 pixels (4.4° visual angle) for Monkey 1 and 250 × 250 pixels (6.9° visual angle) for Monkey 2. The images were shown at 50% transparency (alpha = 0.5) to reduce the potential influence of visual adaptation. The background color of the display screen was 0.5 gray level throughout the experiment. In the standard DMS task, a set of three stimuli were used in each session. The set of images varied between sessions. Each image could appear in both the sample and test positions, at pseudo-randomized equal probability.

In the probabilistic DMS task, a set of four stimuli were used in each session, and the pairing between sample and test stimuli was additionally manipulated. Only two of the four stimuli could appear as sample (and as test in match trials), and the other two stimuli only appeared as test stimuli (i.e., only appeared in nonmatch trials). In nonmatch trials (50% of all trials; the other 50% are match trials), the sample stimulus was followed by one of the two test stimuli with either high (40%) or low (10%) probability. The occurrence of sample–test pairs in the two probability conditions was counterbalanced, such that in nonmatch trials both test stimuli appeared at equal probability and only their probabilistic pairing with the proceeding sample stimulus was shuffled.

Electrophysiology. As previously described (88), both monkeys were chronically implanted in the left hemisphere over V1 with a microdrive system (Gray Matter Research) which had 32 individually movable microelectrodes. Data acquisition was performed using the TDT system (Tucker-Davis Technologies, Alachua, FL, USA). The signal was amplified and digitalized at 25 kHz (TDT PZ5 NeuroDigitizer). This raw signal was bandpass-filtered between 300 and 4,000 Hz to extract multi-unit spiking activity (MUA), and low pass-filtered at 300 Hz and down-sampled with a decimation factor of 24 to about 1 kHz to retrieve the local field potential. MUA was isolated using the online detection algorithm in OpenEx software (Tucker-Davis Technologies). Events crossing a threshold of 4 times the SD of the filtered spiking band activity were considered as spikes and analyzed further.

Data Analysis. All decoding analyses were based on LDA classifiers. As previously described (88), firing rates across channels were treated as independent variables, i.e., predictors, with repeated measurements across trials. To test for stimulus-specificity, image identity was used as class label. For time-resolved decoding, independent classifiers were trained at successive time points. In most decoding analyses, firing rate was calculated by binning spikes in moving windows of 100 ms and steps of 50 ms. For finer timescale comparison of decoding sample- vs. test-evoked activity, we used smaller windows (50 ms) and step sizes (10 ms). Decoding accuracy for each session was measured by averaging the cross-validated classification performance over 20 repeated stratified sampling of the dataset (20 folds). Decoding accuracy values were averaged over sessions.

Data, Materials, and Software Availability. Data used in this study have been deposited in a public repository (89).

ACKNOWLEDGMENTS. We thank the Ernst Strüngmann Foundation, Max Planck Society, the Human Frontier Science Program (HFSP RGP0044/2018), and the Deutsche Forschungsgemeinschaft (DFG Reinhart Koselleck Project 325248489) for supporting this work. We are grateful to Rosanne Rademaker and Michael Wolff for sharing helpful comments on the manuscript.

1. J. M. Fuster, G. E. Alexander, Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
2. S. Funahashi, C. J. Bruce, P. S. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
3. K. Kubota, H. Niki, Prefrontal cortical unit activity and delayed alternation performance in monkeys. *J. Neurophysiol.* **34**, 337–347 (1971).
4. S. L. Brincat *et al.*, Interhemispheric transfer of working memories. *Neuron* **109**, 1055–1066.e4 (2021).
5. R. M. G. Reinhart, J. A. Nguyen, Working memory revived in older adults by synchronizing rhythmic brain circuits. *Nat. Neurosci.* **22**, 820–827 (2019).
6. I. E. J. de Vries, H. A. Slagter, C. N. L. Olivers, Oscillatory control over representational states in working memory. *Trends Cogn. Sci.* **24**, 150–162 (2020).
7. A. H. Lara, J. D. Wallis, The role of prefrontal cortex in working memory: A mini review. *Front. Syst. Neurosci.* **9**, 173 (2015).
8. K. K. Sreenivasan, C. E. Curtis, M. D'Esposito, Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18**, 82–89 (2014).
9. M. D'Esposito, B. R. Postle, The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* **66**, 115–142 (2015).
10. M. D'Esposito, From cognitive to neural models of working memory. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **362**, 761–772 (2007).
11. J. M. Scimeca, A. Kiyonaga, M. D'Esposito, Reaffirming the sensory recruitment account of working memory. *Trends Cogn. Sci.* **22**, 190–192 (2018).
12. J. T. Serences, Neural mechanisms of information storage in visual short-term memory. *Vis. Res.* **128**, 53–67 (2016).
13. T. Pasternak, M. W. Greenlee, Working memory in primate sensory systems. *Nat. Rev. Neurosci.* **6**, 97–107 (2005).
14. T. van Kerkoerle, M. W. Self, P. R. Roelfsema, Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nat. Commun.* **8**, 13804 (2017).
15. H. Supèr, H. Spekreijse, V. A. Lamme, A neural correlate of working memory in the monkey primary visual cortex. *Science* **293**, 120–124 (2001).
16. E. F. Ester, J. T. Serences, E. Awh, Spatially global representations in human primary visual cortex during working memory maintenance. *J. Neurosci.* **29**, 15258–15265 (2009).
17. S. A. Harrison, F. Tong, Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**, 632 (2009).
18. J. T. Serences, E. F. Ester, E. K. Vogel, E. Awh, Stimulus-specific delay activity in human primary visual cortex. *Psychol. Sci.* **20**, 207–214 (2009).
19. R. L. Rademaker, C. Chunharas, J. T. Serences, Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat. Neurosci.* **22**, 1336–1344 (2019).
20. M. J. Wolff, J. Jochim, E. G. Akyürek, M. G. Stokes, Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20**, 864–871 (2017).
21. M. J. Wolff, J. Ding, N. E. Myers, M. G. Stokes, Revealing hidden states in visual working memory using electroencephalography. *Front. Syst. Neurosci.* **9**, 123 (2015).
22. E. F. Ester, D. E. Anderson, J. T. Serences, E. Awh, A neural measure of precision in visual working memory. *J. Cogn. Neurosci.* **25**, 754–761 (2013).
23. S. M. Emrich, A. C. Riggall, J. J. Larocque, B. R. Postle, Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *J. Neurosci.* **33**, 6516–6523 (2013).
24. S. J. D. Lawrence *et al.*, Laminar organization of working memory signals in human visual cortex. *Curr. Biol.* **28**, 3435–3440.e4 (2018).
25. T. B. Christophel, P. Iamshchinina, C. Yan, C. Allefeld, J. D. Haynes, Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* **21**, 494–496 (2018).
26. E. S. Lorenc, K. K. Sreenivasan, D. E. Nee, A. R. E. Vandenbroucke, M. D'Esposito, Flexible coding of visual working memory representations during distraction. *J. Neurosci.* **38**, 5267–5276 (2018).
27. J. Bergmann, E. Genc, A. Kohler, W. Singer, J. Pearson, Neural anatomy of primary visual cortex limits visual working memory. *Cereb. Cortex* **26**, 43–50 (2016).
28. P. Iamshchinina, T. B. Christophel, S. Gayet, R. L. Rademaker, Essential considerations for exploring visual working memory storage in the human brain. *Vis. Cogn.* **29**, 425–436 (2021).

29. P. Iamshchinina, T. B. Christophel, S. Gayet, R. L. Rademaker, Understanding how analysis choices are essential for the meaningful interpretation of visual working memory data. *J. Vis.* **21**, 2721–2721 (2021).
30. S. Kornblith, R. Quian Quiroga, C. Koch, I. Fried, F. Mormann, Persistent single-neuron activity during working memory in the human medial temporal lobe. *Curr. Biol.* **27**, 1026–1032 (2017).
31. J. Kaminski *et al.*, Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat. Neurosci.* **20**, 590–601 (2017).
32. C. Constantinidis *et al.*, Persistent spiking activity underlies working memory. *J. Neurosci.* **38**, 7020–7028 (2018).
33. M. Haller *et al.*, Persistent neuronal activity in human prefrontal cortex links perception and action. *Nat. Hum. Behav.* **2**, 80–91 (2018).
34. M. Lundqvist, P. Herman, E. K. Miller, Working memory: Delay activity, yes! persistent activity? Maybe not. *J. Neurosci.* **38**, 7013–7019 (2018).
35. M. Lundqvist, P. Herman, M. R. Warden, S. L. Brincat, E. K. Miller, Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nat. Commun.* **9**, 394 (2018).
36. M. Lundqvist *et al.*, Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–164 (2016).
37. R. Romo, C. D. Brody, A. Hernandez, L. Lemus, Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
38. D. Trubutschek, S. Marti, H. Ueberschar, S. Dehaene, Probing the limits of activity-silent non-conscious working memory. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 14358–14367 (2019).
39. N. S. Rose *et al.*, Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
40. M. G. Stokes, "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci.* **19**, 394–405 (2015).
41. G. Mongillo, O. Barak, M. Tsodyks, Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
42. Y. Sugase-Miyamoto, Z. Liu, M. C. Wiener, L. M. Optican, B. J. Richmond, Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Comput. Biol.* **4**, e1000073 (2008).
43. F. Fiebig, A. Lansner, A spiking working memory model based on hebbian short-term potentiation. *J. Neurosci.* **37**, 83–96 (2017).
44. M. A. Erickson, L. A. Maramara, J. Lisman, A single brief burst induces GluR1-dependent associative short-term potentiation: A potential mechanism for short-term memory. *J. Cogn. Neurosci.* **22**, 2530–2540 (2010).
45. D. Nikolić, S. Hausler, W. Singer, W. Maass, Distributed fading memory for stimulus properties in the primary visual cortex. *PLoS Biol.* **7**, e1000260 (2009).
46. D. V. Buonomano, W. Maass, State-dependent computations: Spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).
47. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
48. W. Singer, Recurrent dynamics in the cerebral cortex: Integration of sensory evidence with stored knowledge. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101043118 (2021).
49. F. P. de Lange, M. Heilbron, P. Kok, How do expectations shape perception? *Trends Cogn. Sci.* **22**, 764–779 (2018).
50. L. Aitchison, M. Lengyel, With or without you: Predictive coding and Bayesian inference in the brain. *Curr. Opin. Neurobiol.* **46**, 219–227 (2017).
51. C. M. Gray, W. Singer, Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 1698–1702 (1989).
52. C. M. Gray, P. Konig, A. K. Engel, W. Singer, Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
53. M. Bányai *et al.*, Stimulus complexity shapes response correlations in primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2723–2732 (2019).
54. Y. Yiling *et al.*, Robust encoding of natural stimuli by neuronal response sequences in monkey visual cortex. *Nat. Commun.* **14**, 3021 (2023).
55. M. K. Kapadia, M. Ito, C. D. Gilbert, G. Westheimer, Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron* **15**, 843–856 (1995).
56. C. Uran *et al.*, Predictive coding of natural images by V1 firing rates and rhythmic synchronization. *Neuron* **110**, 1240–1257.e8 (2022).
57. A. Peter *et al.*, Surface color and predictability determine contextual modulation of V1 firing and gamma oscillations. *Elife* **8**, e42101 (2019).
58. M. Vinck, C. A. Bosman, More gamma more predictions: Gamma-synchronization as a key mechanism for efficient integration of classical receptive field inputs with surround predictions. *Front. Syst. Neurosci.* **10**, 35 (2016).
59. W. E. Vinje, J. L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
60. P. Kok, J. F. Jehee, F. P. de Lange, Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron* **75**, 265–270 (2012).
61. A. Alink, C. M. Schwiedrzik, A. Kohler, W. Singer, L. Muckli, Stimulus predictability reduces responses in primary visual cortex. *J. Neurosci.* **30**, 2960–2966 (2010).
62. W. Singer, A. Lazar, Does the cerebral cortex exploit high-dimensional, non-linear dynamics for information processing? *Front. Comput. Neurosci.* **10**, 99 (2016).
63. M. B. Brodeur, K. Guerard, M. Bouras, Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS One* **9**, e106953 (2014).
64. M. B. Brodeur, E. Dionne-Dostie, T. Montreuil, M. Lepage, The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS One* **5**, e10773 (2010).
65. R. L. Rademaker, V. G. van de Ven, F. Tong, A. T. Sack, The impact of early visual cortex transcranial magnetic stimulation on visual working memory precision and guess rate. *PLoS One* **12**, e0175230 (2017).
66. J. D. Murray *et al.*, A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* **17**, 1661–1663 (2014).
67. J. I. Luebke, Pyramidal neurons are not generalizable building blocks of cortical networks. *Front. Neuroanat.* **11**, 11 (2017).
68. S. Torres-Gomez *et al.*, Changes in the proportion of inhibitory interneuron types from sensory to executive areas of the primate neocortex: Implications for the origins of working memory representations. *Cereb. Cortex* **30**, 4544–4562 (2020).
69. J. M. Amatrudo *et al.*, Influence of highly distinctive structural properties on the excitability of pyramidal neurons in monkey visual and prefrontal cortices. *J. Neurosci.* **32**, 13644–13660 (2012).
70. M. Wang *et al.*, NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* **77**, 736–749 (2013).
71. C. A. Patterson, S. C. Wissig, A. Kohn, Distinct effects of brief and prolonged adaptation on orientation tuning in primary visual cortex. *J. Neurosci.* **33**, 532–543 (2013).
72. J. R. Muller, A. B. Metha, J. Krauskopf, P. Lennie, Rapid adaptation in visual cortex to the structure of images. *Science* **285**, 1405–1408 (1999).
73. N. J. Priebe, M. M. Churchland, S. G. Lisberger, Constraints on the source of short-term motion adaptation in macaque area MT. I. The role of input and intrinsic mechanisms. *J. Neurophysiol.* **88**, 354–369 (2002).
74. Y. Liu, S. O. Murray, B. Jagadeesh, Time course and stimulus dependence of repetition-induced response suppression in inferotemporal cortex. *J. Neurophysiol.* **101**, 418–436 (2009).
75. K. Cohen-Kashi Malina, M. Jubran, Y. Katz, I. Lampl, Imbalance between excitation and inhibition in the somatosensory cortex produces postadaptation facilitation. *J. Neurosci.* **33**, 8463–8471 (2013).
76. E. K. Miller, R. Desimone, Parallel neuronal mechanisms for short-term memory. *Science* **263**, 520–522 (1994).
77. A. Peter *et al.*, Stimulus-specific plasticity of macaque V1 spike rates and gamma. *Cell Rep.* **37**, 110086 (2021).
78. T. Meyer, C. R. Olson, Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19401–19406 (2011).
79. A. Todorovic, F. P. de Lange, Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *J. Neurosci.* **32**, 13389–13395 (2012).
80. C. Wacongne *et al.*, Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20754–20759 (2011).
81. C. M. Schwiedrzik, W. A. Freiwald, High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron* **96**, 89–97.e4 (2017).
82. G. G. Parras *et al.*, Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nat. Commun.* **8**, 2148 (2017).
83. A. H. Bell, C. Summerfield, E. L. Morin, N. J. Malecek, L. G. Ungerleider, Encoding of stimulus probability in macaque inferior temporal cortex. *Curr. Biol.* **26**, 2280–2290 (2016).
84. A. K. Engel, P. Fries, W. Singer, Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat. Rev. Neurosci.* **2**, 704–716 (2001).
85. B. Lima, W. Singer, S. Neuenschwander, Gamma responses correlate with temporal expectation in monkey primary visual cortex. *J. Neurosci.* **31**, 15919–15931 (2011).
86. C. Summerfield, F. P. de Lange, Expectation in perceptual decision making: Neural and computational mechanisms. *Nat. Rev. Neurosci.* **15**, 745–756 (2014).
87. A. Lazar, C. Lewis, P. Fries, W. Singer, D. Nikolic, Visual exposure enhances stimulus encoding and persistence in primary cortex. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105276118 (2021).
88. Y. Yiling, J. Klon-Lipok, W. Singer, Joint encoding of stimulus and decision in monkey primary visual cortex. *Cereb. Cortex* **34**, bhad420 (2023), 10.1093/cercor/bhad420.
89. Y. Yiling, J. Klon-Lipok, K. Shapcott, A. Lazar, W. Singer, Neuronal (V1) and behavioural data from two monkeys performing a delayed match-to-sample (DMS) task and a passive viewing control task. German Neuroinformatics Node (G-Node) Infrastructure (GIN). https://gin.g-node.org/YSquare/DMS-V1. Deposited 30 January 2024.