**Current Biology, Volume *27***

## Supplemental Information

# Persistent Single-Neuron Activity during Working Memory in the Human Medial Temporal Lobe

**Simon Kornblith, Rodrigo Quian Quiroga, Christof Koch, Itzhak Fried, and Florian Mormann**

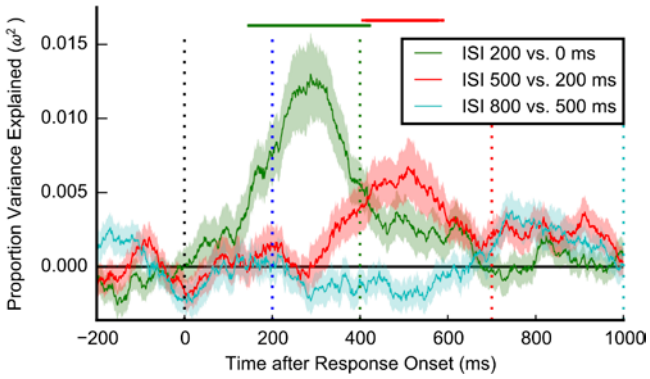


**Figure S1, related to Figure 2**: Interaction between Stimulus and ISI
Debiased proportion of variance explained by the interaction between stimulus and ISI (partial $\omega^2$; see Supplementary Experimental Procedures) for the same 107 units as in Figure 2. Colored dotted vertical lines indicate the onset of the next stimulus in the corresponding ISI conditions. Horizontal bars at the top indicate significant stimulus information at the given time point for the difference between the corresponding ISIs ($p < 0.05$ corrected for all time points by permutation test of mean $\omega^2$). Error bars are ± SEM.

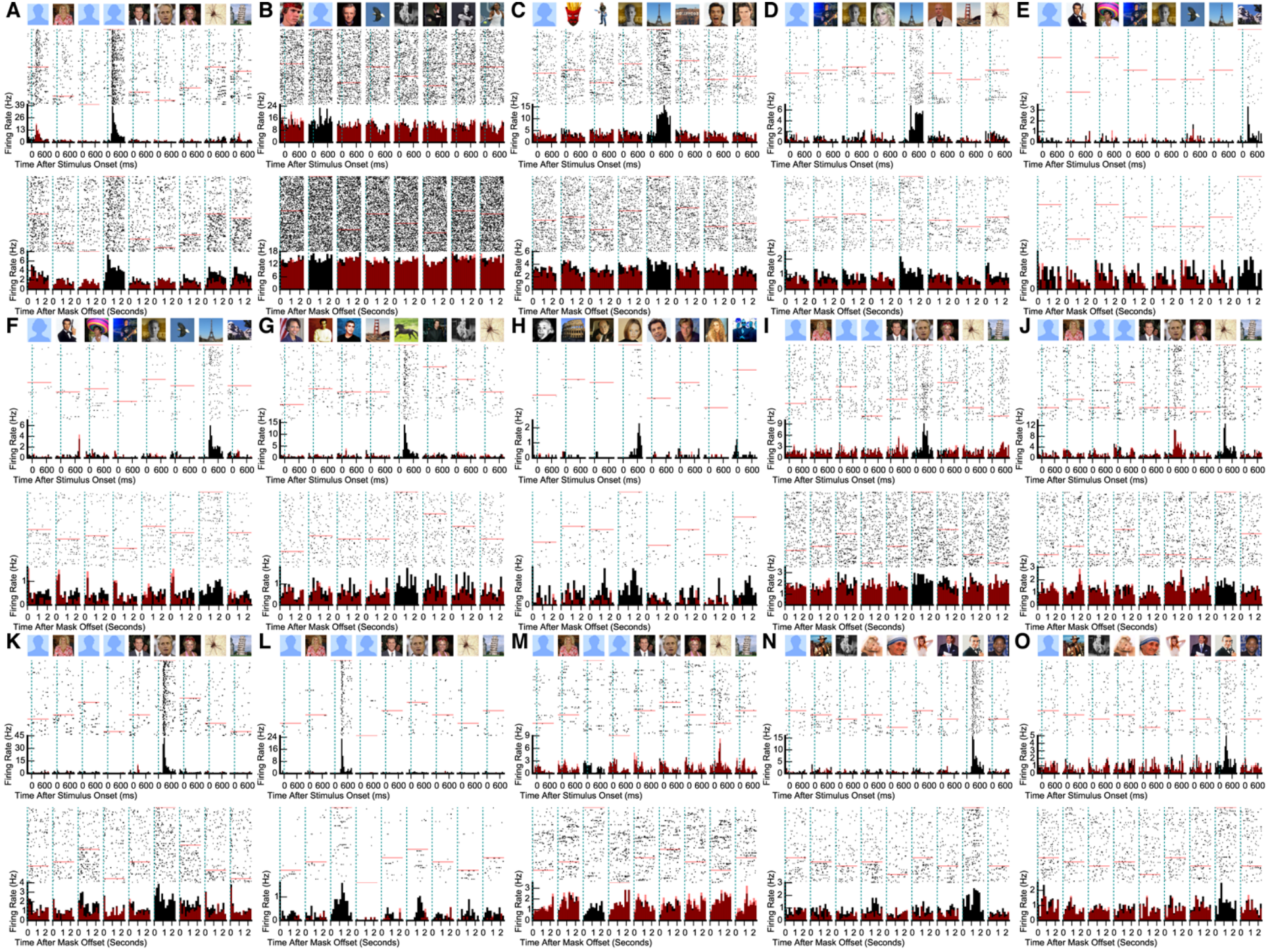
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Firing Rate (Hz)
Time After Stimulus Onset (ms)
Time After Mask Offset (Seconds)

**Figure S2, related to Figure 3:** Maintenance Period Activity for 15 Additional Units
Spike rasters and peri-stimulus time histograms for the sample (top) and maintenance periods (bottom) as in Figure 3, for 15 additional units selective during the maintenance period. Images at the top indicate presented stimuli, with images of lab personnel and images provided by the patient replaced with blue placeholders. A red line separates trials where the preferred stimulus at sample presentation was not presented (top) from trials where it was (bottom). Black histograms indicate average firing rate over all trials; red histograms indicate average firing rate after removing trials that included the preferred stimulus. **A-F**, units recorded from parahippocampal cortex. **G**, a unit recorded from entorhinal cortex. **H**, **I**, units recorded from the hippocampus. **J-O**, units recorded from the amygdala. While units in N and O were recorded from the same macroelectrode in the same session and respond to the same stimulus, they were recorded from different microwires, and the crosscorrelogram showed no peak at t = 0, suggesting that they reflect activity of different neurons.

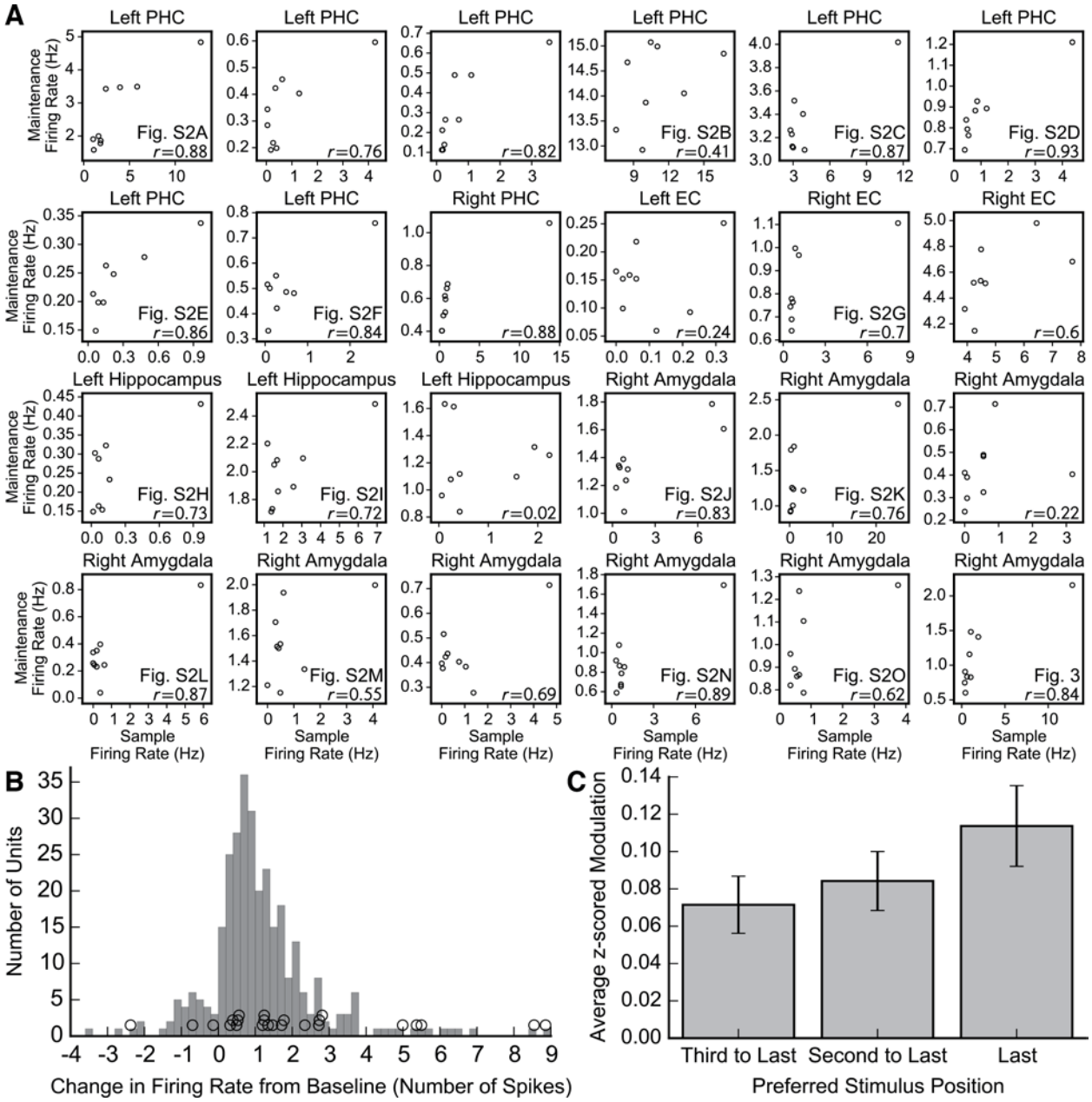


**Figure S3, related to Figure 3**: Comparison of Sample and Maintenance Period Activity
**A**, Scatterplots of the mean sample and maintenance period firing rates by stimulus for all 24 units selective during both periods. Each point represents a presented stimulus. The sample firing rate was computed over the time window from response onset to offset, determined as described in the Supplemental Experimental Procedures. The maintenance firing rate was computed over the window from 300 to 2400 ms after the presentation of the fixation cross. Text in bottom right indicates the figure showing rasters and peri-stimulus time histograms for the given unit (if applicable) and Pearson correlation between sample and maintenance firing rates. **B**, Histogram of the change in the number of spikes in the window between response onset and offset compared to the same window during the baseline period, for the stimulus eliciting the greatest change in each unit with a stimulus-selective visual response. Circles indicate units with stimulus-selective maintenance period activity. **C**, Average maintenance period modulation by the preferred stimulus, comparing trials when the preferred stimulus was presented third to last, second to last, or last in the trial to trials when it was not presented. While modulation was significant at all positions and greater at later positions, there was no significant difference among positions by repeated measures ANOVA. Error bars are ± SEM.

| Region | ISI = 0 | ISI = 200 ms | ISI = 500 ms | ISI = 800 ms | Average |
|---|---|---|---|---|---|
| Parahippocampal Cortex (n = 43) | 198 (89) | 199 (75) | 190 (81) | 207 (77) | 198 (67) |
| Entorhinal Cortex (n = 23) | 292 (62) | 252 (98) | 268 (50) | 276 (71) | 272 (57) |
| Hippocampus (n = 19) | 296 (68) | 285 (61) | 281 (60) | 280 (61) | 286 (42) |
| Amygdala (n = 22) | 244 (72) | 248 (62) | 253 (50) | 208 (78) | 238 (39) |
| All (n = 107) | 245 (87) | 235 (82) | 236 (76) | 235 (80) | 238 (66) |

**Table S1, related to Figure 2**: Latency by Region and ISI
Mean latencies of selectivity computed by maximization of the Poisson likelihood ratio (see Supplemental Experimental Procedures) by region and ISI, for units shown in Figure 2. All values are in milliseconds; values in parentheses indicate standard deviation. The rightmost column indicates values for the mean latency across ISI conditions. Since the procedure used to compute these latencies uses all trials for each unit, these values are not directly comparable to previously reported single trial latencies [S1].

| Region | ISI = 0 | ISI = 200 ms | ISI = 500 ms | ISI = 800 ms | Average |
|---|---|---|---|---|---|
| Parahippocampal Cortex (n = 43) | 247 (219) | 363 (241) | 451 (267) | 468 (330) | 382 (209) |
| Entorhinal Cortex (n = 23) | 128 (59) | 200 (106) | 265 (167) | 299 (240) | 223 (96) |
| Hippocampus (n = 19) | 169 (82) | 237 (158) | 368 (236) | 306 (213) | 270 (138) |
| Amygdala (n = 22) | 231 (287) | 289 (101) | 437 (196) | 564 (315) | 380 (115) |
| All (n = 107) | 204 (199) | 290 (190) | 393 (238) | 423 (305) | 328 (173) |

**Table S2, related to Figure 2**: Duration by Region and ISI
Mean durations of selectivity computed by maximization of the Poisson likelihood ratio (see Supplemental Experimental Procedures) by region and ISI, for units shown in Figure 2. All values are in milliseconds; values in parentheses indicate standard deviation. The rightmost column indicates values for the mean duration across ISI conditions.

| Patient # | # of Units | # Vis. Sel. | # Maint. Sel. | % Vis. Sel. | % Maint. Sel. |
|---|---|---|---|---|---|
| 1 | 141 | 9 | 0 | 6% (3%-12%) | 0% (0%-34%) |
| 2 | 170 | 14 | 1 | 8% (5%-13%) | 7% (0%-34%) |
| 3 | 30 | 3 | 0 | 10% (2%-27%) | 0% (0%-71%) |
| 4 | 129 | 48 | 1 | 37% (29%-46%) | 2% (0%-11%) |
| 5 | 68 | 29 | 10 | 43% (31%-55%) | 34% (18%-54%) |
| 6 | 88 | 22 | 1 | 25% (16%-35%) | 5% (0%-23%) |
| 7 | 204 | 17 | 0 | 8% (5%-13%) | 0% (0%-20%) |
| 8 | 184 | 37 | 1 | 20% (15%-27%) | 3% (0%-14%) |
| 9 | 79 | 6 | 1 | 8% (3%-16%) | 17% (0%-64%) |
| 10 | 19 | 0 | 0 | 0% (0%-18%) | N/A |
| 11 | 84 | 20 | 1 | 24% (15%-34%) | 5% (0%-25%) |
| 12 | 61 | 6 | 1 | 10% (4%-20%) | 17% (0%-64%) |
| 13 | 170 | 41 | 3 | 24% (18%-31%) | 7% (2%-20%) |
| 14 | 151 | 31 | 4 | 21% (14%-28%) | 13% (4%-30%) |
| 15 | 93 | 6 | 0 | 6% (2%-14%) | 0% (0%-46%) |
| 16 | 47 | 5 | 0 | 11% (4%-23%) | 0% (0%-52%) |
| 17 | 52 | 10 | 0 | 19% (10%-33%) | 0% (0%-31%) |
| 18 | 37 | 8 | 0 | 22% (10%-38%) | 0% (0%-37%) |
| TOTAL | 1807 | 312 | 24 | 17% (16%-19%) | 8% (5%-11%) |

**Table S3, related to Figure 3:** Selective Units by Subject
Numbers and proportions of recorded units, visually selective units, and visually selective units with maintenance period selectivity for all 18 patients analyzed. Percentages in parenthesis indicate Clopper-Pearson 95% binomial confidence intervals.

## Supplemental Experimental Procedures

*Subjects*

18 patients with pharmacologically intractable epilepsy were implanted with chronic depth electrodes in order to localize the epileptogenic focus for possible resection. Each subject provided informed written consent. All studies conformed to the guidelines of the Medical Institutional Review Board of UCLA and the Institutional Review Board of Caltech.

*Electrophysiological recordings*

Signals were recorded from 9 platinum-iridium microwires (8 high-impedance recording electrodes and 1 low-impedance reference) protruding from the end of the depth electrodes. The voltage differential between the recording electrodes and reference was amplified, band pass filtered from 1 to 9000 Hz, and digitized at 27.8 kHz using a Neuralynx Cheetah system (Bozeman, MT). To identify single- and multi-unit activity, the digital signal was band-pass filtered between 300 and 3000 Hz. 64 samples surrounding deflections exceeding $5 \times 0.675$ mean absolute deviations were extracted as candidate spikes and sorted using Wave_clus [S2].

Sorted units were manually confirmed and classified as single units (SU), multi-units (MU), or artifacts based on spike shape and variance, peak-amplitude-to-noise level, the inter-spike interval (ISI) distribution of each cluster, and presence of a refractory period.

Data and analysis code are available online at https://github.com/simonster/Kornblith-et-al-2017-Current-Biology.

*Behavioral task*

Subjects performed a total of 43 behavioral sessions. In the morning of each recording day, subjects participated in an initial screening session in which they saw 6 repetitions of approximately 100 images of famous or personally known persons, animals, scenes and objects for one second. To ensure that subjects attended to the images, after each presentation, they signaled whether the preceding stimulus contained a face with a key press. After plotting the PSTHs for the 15 most promising image candidates for a response, separately for each unit, the experimenter browsed through these hundreds of PSTHs to manually select the 8 or 9 most promising response-eliciting images for the follow-up working memory experiment, run in the afternoon. We excluded two sessions, one in which the subject did not perform significantly better than chance, and one in which fewer stimuli (6) were presented than in all others, leaving 41 sessions for subsequent analysis.

The structure of the main behavioral task, derived from [S3], is shown in Figure 1A. At the start of each trial, subjects were instructed to fixate on a cross for 1 second. Subjects then saw 3 different stimuli (12 sessions) or 4 stimuli (29 sessions) for 200 ms each, with a blank delay separating each presentation. Possible inter-stimulus intervals were 0 ms, 200 ms, 500 ms, or 800 ms, so that the stimulus onset asynchrony was 200 ms, 400 ms, 700 ms, or 1000 ms. Delay times were constant throughout a single trial, and delay times and stimulus order were randomized such that every stimulus was presented 6 times for each possible presentation rate and position in the sequence. After the last delay, subjects saw a mask composed of images of faces for 200 ms, followed by a fixation cross. This cross remained on the screen for a "maintenance period" of 2.4 to 2.6 seconds. After this period elapsed, two stimuli appeared on the screen for 1500 ms, followed by a screen of text instructions to press the left or right arrow keys to signal which of the probe stimuli had been present in the earlier stimulus stream. Auditory feedback informed the subject whether their choice was correct. Subjects were instructed to "keep the items in memory for the duration of the maintenance period". They were not explicitly asked about which strategy they used. However, the experimenter noted if patients used explicit verbalization during the task. This was the case for only two patients (#2 and #14 in Table S3).

Subjects performed 192 trials at trial load of 4 (29 sessions), or 216 trials at a trial load of 3 (12 sessions), corresponding to 96 or 72 presentations of each stimulus. On average, subjects responded correctly on 80% of trials (inter-quartile range: 70% to 92%). In all analyzed sessions, subjects performed significantly better than chance ($p < 0.05$, binomial test).

Subjects performed both the screening and main tasks using a laptop while sitting in bed. Stimuli were presented at a distance of approximately 50 cm and subtended approximately 4 degrees of visual angle.

In the first session of the working memory task that a given patient performed, they had received equal exposure to all images at the time the task was run. Patients who performed more than one session of the working memory paradigm did this on different days, with typically one or two days of different, non-screening-related experiments (e.g. navigation tasks) in between. Each additional session was based on a separate screening. Only if the 8 or 9 stimuli used for subsequent sessions overlapped with those from preceding sessions had those overlapping images thus been shown more frequently (in the follow-up experiments on previous days) than the rest. However, this was the exception rather than the rule. Of the 340 images selected over the 41 analyzed sessions, only 45 were repeated from previous sessions. Within the 123 visually selective units recorded in a session where at least one stimulus had been repeated, the baseline z-scored stimulus modulation was no different for the repeated stimuli as compared to the non-repeated stimuli ($t(122) = -1.1$, $p = 0.26$, dependent samples t-test).

*Permutation test for stimulus selectivity*

We determined whether cells exhibited any stimulus selectivity using the Poisson generalized linear model (GLM) equivalent of a one-way ANOVA. In GLM formalism, the test statistic is the difference between the deviance of an intercept-only GLM ($D_{\text{null}}$) and the deviance of a GLM incorporating the effect of stimulus ($D_{\text{full}}$), analogous to the difference between the total and residual sums of squares in an ANOVA. Alternatively, it is 2 times the log of the ratio of likelihood of the data with different parameters for different stimuli to the likelihood with the same parameter for all stimuli, when all parameters are estimated by maximum likelihood:

$$D_{\text{null}} = -2\log\mathcal{L}\left(k_*\middle|\bar{k}\right) + 2\sum_{i=1}^{n}\sum_{j=1}^{m}\log\mathcal{L}\left(k_{ij}\middle|k_{ij}\right)$$

$$D_{\text{full}} = -2\log\mathcal{L}\left(k_*\middle|\overline{k_1},\ldots,\overline{k_n}\right) + 2\sum_{i=1}^{n}\sum_{j=1}^{m}\log\mathcal{L}\left(k_{ij}\middle|k_{ij}\right)$$

$$D_{\text{null}} - D_{\text{full}} = 2\left(\log\mathcal{L}\left(k_*\middle|\overline{k_1},\ldots,\overline{k_n}\right) - \log\mathcal{L}\left(k_*\middle|\bar{k}\right)\right) = 2\log\left(\frac{\mathcal{L}\left(k_*\middle|\overline{k_1},\ldots,\overline{k_n}\right)}{\mathcal{L}\left(k_*\middle|\bar{k}\right)}\right)$$

$$= 2\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\log\mathcal{L}\left(k_{i*}\middle|\,\bar{k}_i\right) - \log\mathcal{L}\left(k\middle|\bar{k}\right)\right)$$

$$= 2\sum_{i=1}^{n}\sum_{j=1}^{m}\left(-\bar{k}_i + k_{ij}\log\bar{k}_i - \log k_{ij}! — \bar{k} + k_{ij}\log\bar{k} - \log k_{ij}!\right)$$

$$= 2\sum_{i=1}^{n}\sum_{j=1}^{m}k_{ij}\left(\log\bar{k}_i - \log\bar{k}\right)$$

where $\mathcal{L}(y|\theta)$ denotes the likelihood of the data under a Poisson distribution with the given the parameters, $n$ is the number of stimuli, $m$ is the number of trials per stimulus, $k_*$ is the vector of spike counts for all presentations, $k_{i*}$ is the vector of spike counts for presentations of stimulus $i$, $k_{ij}$ is the spike count for presentation $j$ of the stimulus $i$, $\bar{k}_i$ is the mean spike count for stimulus $i$, and $\bar{k}$ is the overall mean spike count. Since the mean is the maximum likelihood estimator of the parameter of a Poisson distribution, these means maximize the likelihoods in these equations for the given spike counts.

This is the standard test statistic for this GLM, and its motivation comes from the Neyman-Pearson lemma, which states that a likelihood ratio test is the most powerful test at a given significance level. Because a GLM estimates the relevant parameters by maximum likelihood, and we further compare the test statistic against a permutation distribution rather than a distribution known *a priori*, the test does not perfectly satisfy this lemma, but our simulations show that it achieves greater power for Poisson-distributed data than alternative statistics, and it provides accurate p-values even if spike counts are not Poisson. The statistic is related to the difference in Akaike information criterion (AIC) or Bayesian information criterion (BIC) between a Poisson model with different parameters for each stimulus and a model with the same parameter for all stimuli by a constant that depends on the total number of stimuli, and to the relative likelihood of the models by a strictly monotonically increasing function. Thus, any of these alternative statistics would give identical results for the permutation test we describe, and for the measurements of latency and duration described below.

Rather than select a single time window in which to compute this statistic, we computed the statistic over all possible time windows at least 20 ms in length between 100 and 1000 ms after stimulus onset, in increments of 10 ms, and performed a permutation test of the maximum. We first binned spikes into 90 bins of 10 ms width between 100 and 1000 ms after stimulus onset. We computed the deviance statistic above for all time windows from bin *i* to bin *j*, where $j > i$. The total number of windows is $\sum_{i=1}^{n-1}(n-i) = n(n-1)/2$, or 4005 when $n = 90$ bins. We computed a p-value as the proportion of the maxima of the statistic across time windows for each of 10,000 permutations generated by shuffling stimulus labels that were greater than the maximum of the statistic across time windows for the observed data. If the original data are permuted, this test rejects the null hypothesis at the specified α with probability no greater than α, but possibly with a lower probability, due to the discrete nature of the data.

*Neuronal response latency and duration*

To analyze latency and duration, we included only the 107 units that carried significant stimulus information in all four ISI conditions at $\alpha = 0.05$ per the permutation test described above. Significantly more units were selective at ISI = 200 ms compared with ISI = 0 (15% [264/1807] vs. 10% [186/1807], $p < 10^{-7}$, exact McNemar test), and at ISI = 500 ms compared with ISI = 200 ms (17% [300/1807] vs. 15% [264/1807], $p = 0.02$), but not at ISI = 800 ms compared with ISI = 500 ms (17% [300/1807] vs. 18% [328/1807], $p = 0.08$). Power to detect a response may be lower at lower ISIs because of increased response variability in the 100 to 1000 ms window in which we search for responses due to presentation of subsequent stimuli within this window, or differences in firing rate modulation in the optimal time window, perhaps due to the response duration effects we describe below. Performing the following analyses on all 312 units that were stimulus-selective at $\alpha = 0.001$ when pooling trials across all ISI conditions introduced additional noise into our measures of latency and durations, but did not substantively change our results.

For each unit, we measured neuronal response latency and duration by finding the contiguous time window that maximized the Poisson log likelihood ratio statistic described in the previous section. Given the relationship between this statistic and AIC/BIC, this procedure corresponds to selecting the time window that yields the maximum amount of information regarding the stimulus. We binned responses into 1 ms bins between 1 and 1000 ms after stimulus onset and computed the log likelihood ratio statistic above for all possible 498,501 time windows from bin *i* to bin *j*, where $j > i$. We defined the response onset and offset as the shortest window at which the log likelihood reached its maximum. We used the measured latencies to align individual units in Figures 2C and S1, to test proportions of units with significant selectivity in predefined windows following stimulus offset, to determine preferred stimuli, and to measure response modulation at presentation. When measuring response latencies and durations for individual ISI conditions, we included bins from 1 to 1500 ms after stimulus onset, to account for the possibility of long-lasting responses at long ISIs.

Since this procedure combines information across multiple presentations, the estimated latencies are likely shifted toward earlier times compared to the single trial latency estimates reported in [S1]. For example, in the case of a noiseless unit (i.e., a unit with a baseline firing rate of zero), this procedure would select the earliest spike to the stimulus on any of the 72 or 96 trials as the onset latency, while the single-trial procedure from [S1] would select the median time to first spike across stimulus presentations. However, our procedure is well-suited to responses where modulation to the preferred stimulus (or stimuli) is low, since the corresponding change in firing rate may not be measurable on individual trials. Additionally, the present methodology provides a duration, which is impossible to establish on a single-trial basis, given that the duration of the first burst need not correspond to the duration of stimulus-selective activity.

In the text, we report that latencies varied significantly by region but not by ISI. However, these estimates are based on all stimulus presentations. Since we would expect that any change in latency with ISI would occur only for stimulus presentations after the first in a trial (i.e., only for stimuli presented soon after the preceding stimulus), we additionally computed latencies using only stimulus presentations after the first in a trial. The effect of region remained significant ($F(3,103) = 7.0$, $p = 0.0002$, two-way ANOVA with repeated measure of ISI), and the effect of ISI remained insignificant ($F(3,309) = 1.8$, $p = 0.14$).

*Proportion of variance explained by stimulus for time point and ISI*

For each of the 107 units selected as described in the previous section, we computed partial $\omega^2$ to determine the proportion of variance explained. $\omega^2$ is an alternative to the common $\eta^2$ statistic that is less biased by sample size and the number of parameters included in the model, and is defined as:

$$\omega_p^2 = \frac{df_{\text{effect}} \times (MS_{\text{effect}} - MS_{\text{error}})}{df_{\text{effect}} \times MS_{\text{effect}} + (N - df_{\text{effect}}) \times MS_{\text{error}}}$$

$$df_{\text{effect}} = p_{\text{full}} - p_{\text{null}}$$

$$MS_{\text{effect}} = \frac{RSS_{\text{null}} - RSS_{\text{full}}}{df_{\text{effect}}}$$

$$MS_{\text{error}} = \frac{RSS_{\text{full}}}{N - p_{\text{full}}}$$

where $p_{\text{full}}$ is the number of predictors of the full model, $p_{\text{null}}$ is the number of predictors in the null model (i.e., the nuisance parameters), $RSS_{\text{full}}$ is the sum of the squared residuals of the full model, $RSS_{\text{null}}$ is the sum of squared residuals of the null model, and $N$ is the total number of trials.

To see how long information persisted, we computed partial $\omega^2$ for a model including terms for the current stimulus, previous stimulus, and next stimulus versus a model including terms for only the previous and next stimulus for spike counts over 200 ms bins, with the zero point for each unit set as the response latency as defined in the previous section. Due to task structure, the previous and next stimuli could not be the same as the current stimulus. Thus, we included those stimuli in the null model to ensure that any variance explained corresponded to information about what the presented stimulus was, and not what the previous and next stimuli were not. The null model was thus:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j I\left(s_i^{\text{prev}} = j\right) + \sum_{j=1}^{m} \beta_{m+j} I(s_i^{\text{next}} = j) + \epsilon_i$$

where $y_i$ is the firing rate in the time window being modeled, relative to stimulus presentation $i$; $s_i^{\text{prev}}$ is the number of the stimulus presented before stimulus presentation $i$, or 0 if first in the trial; $s_i^{\text{next}}$ is the number of the stimulus presented after stimulus presentation $i$, or 0 if last in the trial; $m$ is the number of stimuli used in the session (i.e., 8 or 9); $I(x)$ is 1 if condition $x$ is true, or 0 if it is false; and $\epsilon_i$ is the error term. The full model was:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j I\left(s_i^{\text{prev}} = j\right) + \sum_{j=1}^{m} \beta_{m+j} I(s_i^{\text{next}} = j) + \sum_{j=2}^{m} \beta_{2m+j} I(s_i^{\text{current}} = j) + \epsilon_i$$

where $s_i^{\text{current}}$ is the stimulus at presentation $i$. Thus, the $\omega^2$ value we compute is a debiased estimate of the partial proportion of variance explained by the presented stimulus in a three-way ANOVA that also includes effects of the previous and next stimuli, using type II sums of squares.

To test for significance, we computed $\omega^2$ for each time point between -200 ms and 1000 ms after response onset for 1000 permutations of all trials at the given ISI. For each permutation, we took the mean of $\omega^2$ across all units, and the maximum over all time points. We calculated p-values as the proportion of permutation maximum $\omega^2$ values that were greater than the mean $\omega^2$ value at the given time point.

To test for differences between conditions, we computed partial $\omega^2$ combining trials from a given ISI and the next longest ISI, i.e., 0 vs. 200 ms, 200 ms vs. 400 ms, and 500 ms vs. 800 ms. We compared a model including terms for and interactions between the ISI condition and the current, previous, and next stimuli against a model with terms for and interactions between ISI and the previous and next stimuli, along with terms for the current stimulus and ISI condition. Thus, the $\omega^2$ value we report for the interaction is a debiased estimate of the partial proportion of variance explained by the interaction between the ISI condition and presented stimulus in a four-way ANOVA that includes effects of ISI condition; presented, previous, and next stimuli; and terms for stimulus-ISI interactions. We constructed a permutation distribution by permuting spike counts within presented stimulus conditions but not between ISI conditions, and computed p-values as above.

To control for the possibility that ISI effects are driven by changes in the response to the preceding stimulus, rather than persistence of the response to the presented stimulus, we additionally performed this analysis using only the first stimulus presentation on any given trial. The stimulus/ISI interaction between the 200 ms and 0 ISI conditions remains significant in the contiguous window from 211 to 380 ms after response onset, with a similar time course to that in Figure S1. The stimulus/ISI interaction was not significant between the 500 and 200 ms ISI conditions, but this probably results from increased noise, since this analysis discards between ⅔ and ¾ of all data acquired.

*Number of units encoding stimulus information in specific time windows at specific ISIs*
To determine whether a given unit encoded stimulus information at a specific time point and ISI, we computed the Poisson log likelihood ratio as above, and calculated the p-value as the proportion of 10,000 permutations that had a log likelihood ratio greater than or equal to the observed value. As above, due to task structure, the next stimulus could not be the same as the current stimulus. Thus, we held the next stimulus constant when constructing permutations, so that it would explain an equal amount of variability in the firing rate in the permutations.

*Tests for selectivity during the maintenance period*
We assessed maintenance period selectivity by computing the F statistic for a linear model of the firing rate over the window from 300 to 2400 ms following the start of the maintenance period as the sum of coefficients representing the contributions of each stimulus that was previously presented on a given trial. This model is:

$$y_i = \sum_{j=1}^{m} \beta_{i-1} I(j \in s_i) + \epsilon_i$$

where $y_i$ is the firing rate during the maintenance period on trial $i$, $s_i$ is the set of stimuli presented on trial $i$, $m$ is the number of stimuli used in the session (i.e., 8 or 9), $I(x)$ is 1 if condition $x$ is true or 0 if it is false, and $\epsilon_i$ is the error term. This model was fit by least squares and compared to an intercept-only model. The F statistic is:

$$F = \frac{\left(\frac{RSS_{\text{null}} - RSS_{\text{full}}}{m-1}\right)}{\left(\frac{RSS_{\text{full}}}{n-m}\right)}$$

where, *n* is the number of trials (i.e., 192 or 216), $RSS_{\text{full}}$ is the sum of squared residuals from the model above, and the null model is the intercept-only model, so:

$$RSS_{\text{null}} = \sum_{i=1}^{n} (y_i - \bar{y})\text{^}2$$

where $\bar{y}$ is the mean firing rate across all trials.

Comparing the F statistic against an F distribution only produces a valid p-value when the sum of squared residuals follow a chi-squared distribution. If the data are highly non-normal, as may be the case for units with very low firing rates, this assumption is violated. We wanted to make sure that the effects we report are real, and not due to inappropriate model assumptions. Thus, we computed an accurate p-value by calculating the F statistic after permuting the dependent variable (the firing rate during the maintenance period) between different trials, and determining the proportion of 100,000 permutation F statistics that were greater than or equal to the F statistic computed on the unpermuted data.

13% (42/312) of units selective at presentation were selective during the maintenance period at α = 0.05; 8% (24/312) were selective at α = 0.01; and 4% (13/312) were selective at α = 0.001. Thus, relaxing α beyond 0.01 did not identify substantially more significant units than expected by chance.

We also tested for maintenance period selectivity by comparing firing rates during maintenance periods when the subject was maintaining the stimulus that elicited the strongest absolute modulation at presentation against maintenance periods when the subject was not. This comparison is potentially more powerful, since it has only a

single free parameter instead of a parameter for each stimulus. However, it assumes conserved selectivity during sample and maintenance. Since neurons could show temporal autocorrelation in spiking over the course of a trial, we estimated the preferred stimulus as the stimulus eliciting the strongest absolute modulation on a separate set of trials (every fourth trial) that was subsequently excluded in the analysis of maintenance period activity. Comparing firing rates between trials when the preferred stimulus was presented and trials when it was not, 6% of visually selective units (19/312) responded differently ($\alpha = 0.01$, Mann-Whitney U test), a comparable proportion to the 8% identified by the permutation F test described above ($p = 0.30$, exact McNemar test). We defined the maintenance period modulation as:

$$\mathrm{sgn}\left(\bar{x}_{\mathrm{sample}} - \bar{x}_{\mathrm{baseline}}\right)\left(\frac{\bar{x}_{\mathrm{present}} - \bar{x}_{\mathrm{absent}}}{s_{\mathrm{baseline}}}\right)$$

where $\bar{x}_{\mathrm{sample}}$ is the mean firing rate to the preferred stimulus in the window from response onset to response offset on the ¼ trials used to determine the preferred stimulus; $\bar{x}_{\mathrm{baseline}}$ is the mean firing rate in an equal sized window in the baseline period on these trials; $\bar{x}_{\mathrm{present}}$ is the mean firing rate from 300 to 2400 ms after the start of the maintenance period on the subset of the remaining ¾ trials when the preferred stimulus was presented; $\bar{x}_{\mathrm{absent}}$ is the mean firing rate from 300 to 2400 ms after the start of the maintenance period on the subset of the remaining ¾ trials where the preferred stimulus was not presented; and $s_{\mathrm{baseline}}$ is the standard deviation of the firing rate over the 1000 ms baseline period estimated over all trials. If firing rates are equal during the maintenance period between trials when the preferred stimulus was presented and trials when it was not, the expected value of this statistic will be zero.

We used a one-sample t-test to determine whether the population mean was significantly non-zero, but all reported results were also significant with similar p-values if compared against zero using a nonparametric bootstrap test. Additionally, our finding of modulation in units without significant maintenance period activity when tested individually remained significant when excluding both the 24 units selective by the permutation F test described in the text and the additional 5 units that were significant by a Mann-Whitney U test comparing modulation between trials when the preferred stimulus was presented and trials when it was not (modulation = 0.05, $t(282) = 4.7$, $p < 10^{-5}$).

*Test for a link between neural activity and behavior*

We tested for a link between neural activity and behavior by first determining each unit's preferred stimulus as the stimulus that elicited the strongest absolute modulation at presentation in the window from response onset to response offset, as determined using the procedure above. We then compared firing rates between trials where the subject correctly selected the image as having been previously presented against trials where the subject incorrectly selected the alternative image. We defined the behavioral modulation as:

$$\mathrm{sgn}\left(\bar{x}_{\mathrm{sample}} - \bar{x}_{\mathrm{baseline}}\right)\left(\frac{\bar{x}_{\mathrm{correct}} - \bar{x}_{\mathrm{incorrect}}}{s_{\mathrm{baseline}}}\right)$$

where $\bar{x}_{\mathrm{sample}}$ is the mean firing rate to the preferred stimulus at sample in the optimal window; $\bar{x}_{\mathrm{baseline}}$ is the mean firing rate in an equal sized window in the baseline period; $\bar{x}_{\mathrm{correct}}$ is the mean firing rate on incorrect trials where the preferred stimulus was presented and probed; and $s_{\mathrm{baseline}}$ is the standard deviation of the firing rate over the 1000 ms baseline period.

When assessing the difference in firing rates between trials when the preferred stimulus was probed and correctly selected versus trials when it was probed but the alternative stimulus was incorrectly selected, we determined the preferred stimulus at presentation using all trials, since there is no reason to expect a difference in temporal autocorrelation between the two groups. As above, we used a one-sample t-test to determine whether the population mean was significantly non-zero, but all reported results were also significant with similar p-values if compared against zero using a nonparametric bootstrap test.

We also tested individual units for significant difference in firing rate between trials when the preferred stimulus was probed and correctly selected versus trials when it was probed but the alternative stimulus was incorrectly selected. The proportion of units was not significant in any epoch tested. During the maintenance period, only 0.7%

(2/274) showed a significant difference at $\alpha = 0.01$ ($p = 0.76$, binomial test). However, power to detect effects at an individual unit level is very low, since we obtained a median of only 3 incorrect trials and 19 correct trials. While direct power calculation for a Mann-Whitney U test is impossible without auxiliary assumptions, with a t-test, assuming a standardized effect size of 0.1-0.2 (comparable to the mean population-level modulation of 0.12 reported in the text), the power to detect a difference at the individual unit level with the median sample sizes above at $\alpha = 0.01$ is 0.011-0.013, i.e., on average we should reject the null hypothesis 1.1-1.3% of the time. For enough individual units to be significant to reject the null hypothesis of $\alpha = 0.01$ in a binomial test at $p < 0.05$, 6/274 units, or 2.2%, would need to show an effect. Thus, significant effects at the individual unit level are not expected given the population level effect size we report.

## Supplemental References


[S1] Mormann F, Kornblith S, Quiroga RQ, Kraskov A, Cerf M, Fried I, et al. Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. J Neurosci 2008;28:8865–72.

[S2] Quiroga RQ, Nadasdy Z, Ben-Shaul Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 2004;16:1661–87.

[S3] Sternberg S. High-Speed scanning in human memory. Science 1966;153:652–4.