Supplementary text and figures

The main manuscript describes the two hypothesized mechanisms that have dominated the study of the neuronal basis of attention. A third hypothesis has been recently proposed to explain the oft observed finding that attention decreases pair-wise noise correlations within a brain area. This hypothesis, which we call the internal states hypothesis[1], suggests that attention reduces uncontrolled fluctuations in the animals' cognitive states, allowing them to focus on their psychophysical task (Supplementary Figure 2A). In this scenario, the widely observed attention-related reduction in response variability in visual cortex[2-4] would be an epiphenomenon of decreased variability in cognitive states.

Interactions between brain areas do not support the internal states hypothesis

The internal states hypothesis provides a challenge for the approach of linking populations of visual neurons with behavior because it proposes that there is no link between attention-related changes in MT or other visual areas and performance. Instead, it posits that attention limits mind wandering, and the changes in performance and neuronal responses simply reflect improved stability in internal states[1]. There are two reasons this hypothesis seems unlikely to account for our data. First, the effects of spatial attention are spatially specific (e.g. correlated variability increases in one hemisphere while decreasing in the other, even when neurons in the two hemispheres are simultaneously recorded[2]), meaning that reductions in the variability of global cognitive processes like arousal and motivation are unlikely to account for the attention-related changes in visual cortex. Further, it is not obvious how reductions in fluctuations in internal states could account for the attention-related increases in firing rates observed in spatial attention studies like ours (Supplementary Figure 1), let alone the more complex firing rate changes associated with feature attention[5].

This hypothesis can also be addressed using a population-analysis approach, by using the responses of MT and SC neurons to attempt to quantify the variability in the animals' internal states. We reasoned that fluctuations in internal states would 1) often occur at timescales longer than the 400-600 ms between stimulus presentations in our task and 2) affect the covariability of neurons in many brain areas, including both MT and the SC. To test the hypothesis that attention affects uncontrolled fluctuations in internal states (Supplementary Figure 2), we created a procedure to identify slow fluctuations in population responses. We performed principal components analysis on population responses to the identical visual stimuli that occurred before the direction change on each trial (e.g. stimulus A in Figure 1) in each attention condition. Because the only variability in those population responses is internally generated, the first principal component (PC) represents the axis of greatest shared variability in the population of neurons in each brain area. We searched for slow fluctuations in internal states by measuring the auto- and cross-correlations in projections onto this first PC in each area.

The autocorrelation functions of projections onto the first PC show that there is indeed response variability in each area that fluctuates slowly and is reduced by attention (Supplementary Figures 2B, C), which is not as readily observable by computing noise correlations between pairs of neurons (Supplementary Figure 1C). However, the cross-correlation of projections onto the first PC in MT and the SC showed a qualitatively different time course than the autocorrelation function. Furthermore, not only did attention not reduce the covariability of these signals between these two areas (and presumably brain wide), attention *increased* the slow variability that is shared between areas (Supplementary Figure 2D and Supplementary Figure 1C). These results are in conflict with the idea that the attention-related decrease in covariability within each area is a byproduct of a decrease in uncontrolled fluctuations in internal states, because such a decrease should be brain-wide.

1 Ecker, A. S., Denfield, G. H., Bethge, M. & Tolias, A. S. On the Structure of Neuronal Population Activity under Fluctuations in Attentional State. *J Neurosci* **36**, 1775-1789, doi:10.1523/JNEUROSCI.2044-15.2016 (2016).

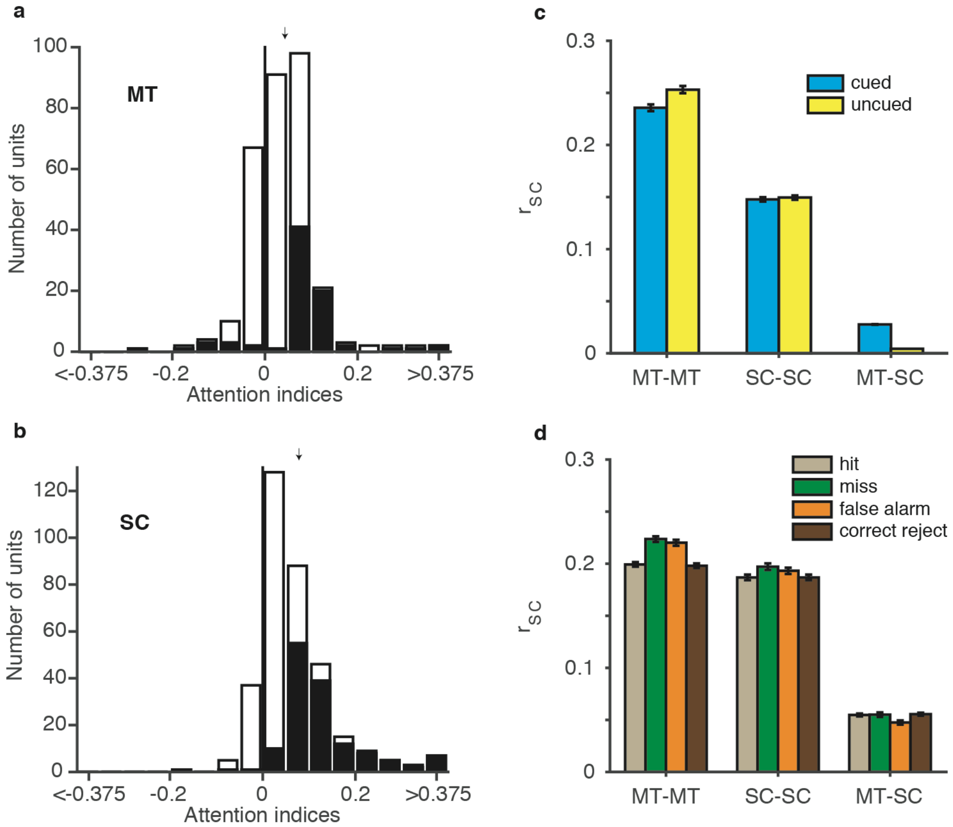
2 Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* **12**, 1594-1600 (2009).

3 Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J. & Cohen, M. R. Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463-465 (2018).

4 Mitchell, J. F., Sundberg, K. A. & Reynolds, J. H. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* **63**, 879-888 (2009).

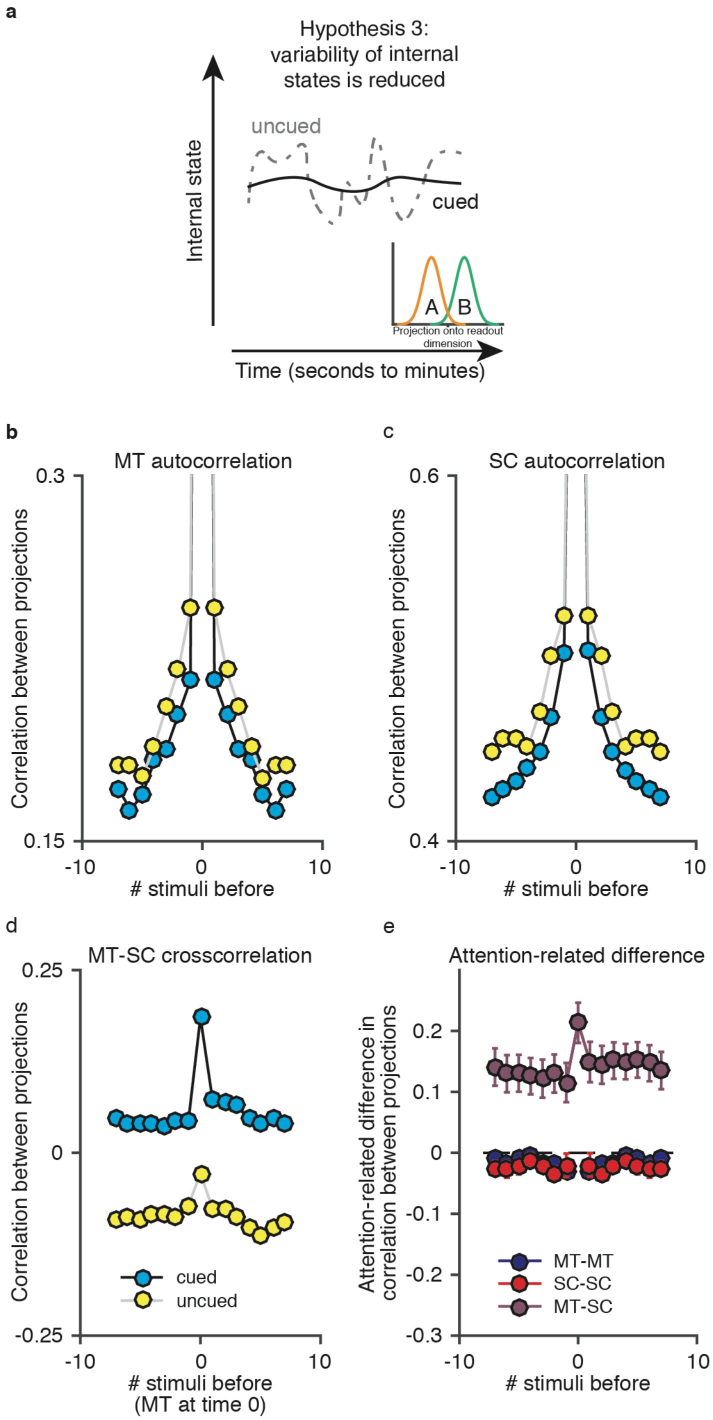
5 Treue, S. & Martinez-Trujillo, J. C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575-579 (1999).

Supplementary Figure 1



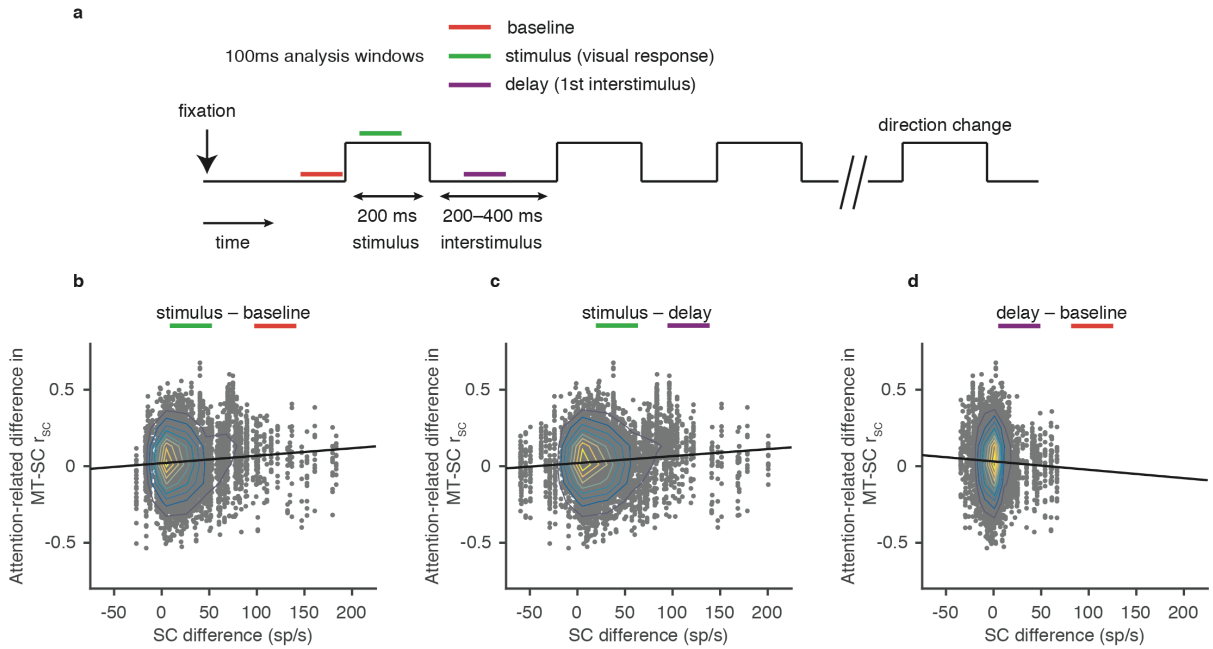
Supplementary Figure 1. Effects of attention on common analyses of individual units and pairs of units (A) Attention increases firing rates in MT, quantified as the difference in firing rates in the different attention conditions divided by the sum. Units with significant differences in average responses for the two conditions are specified by black bars (t-test, $p<.05$). This distribution (mean = 0.04, median = 0.04) is significantly different from zero (Wilcoxon signed rank test, $p< 10^{-21}$). (B) Same as A, for SC data. This distribution (mean = 0.073, median = 0.05) is significantly different from zero (Wilcoxon signed rank test, $p< 10^{-43}$). (C) Within and between area noise correlations calculated from spike counts during stimulus presentations that preceded successful maintenance of fixation from trials that ended with either a hit or miss or were a successful catch trial. Attention decreases average correlations within MT (Wilcoxon signed rank test, $p< 10^{-12}$), not in the SC (Wilcoxon signed rank test, $p=0.8$) and increases them between the two areas (Wilcoxon signed rank test, $p< 10^{-41}$). Error bars are standard error of the mean. (D) Within and between area noise correlations calculated from spike counts that immediately preceded different behavioral outcomes during cued trials. Misses and false alarms are associated with higher correlations within MT (t-test, $p<10^{-3}$) and SC (t-test, $p<10^{-3}$) but not between the two areas (t-test, $p=0.23$). Error bars are standard error of the mean.

Supplementary Figure 2



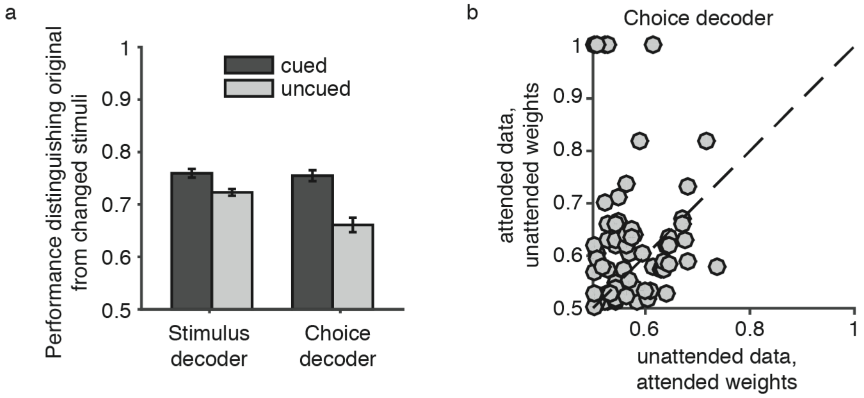
Supplementary Figure 2. Attention has opposite effects on slow fluctuations in neuronal population responses within and across areas. (A) Hypothesis 3 proposes that attention-related changes in visual cortex are epiphenomenal. Instead, it suggests that attention reduces uncontrolled variability in the animals' internal states, which might produce less variable neuronal population responses and therefore more separable projections onto the readout dimensions. (B, C) Autocorrelations between projections onto the first PCs of population responses to repeated presentations of the same visual stimulus in (B) MT, and (C) the SC. The x-axis plots time lag in units of stimulus presentations (400-600 ms; see Supplementary Text). (D) Cross correlation between projections onto the first PCs in MT and the SC (same data and plotting conventions as in B and C). (E) Attention-related difference in autocorrelation or cross correlations between the projections in the previous plots. Error bars represent standard error of the mean. Attention was associated with a statistically significant decrease in autocorrelation overall (t-tests, p<.05) in both areas and in 11/15 individual MT data sets and 9/15 SC data sets (t-tests, p<0.05 with a Bonferroni correction) and a significant increase in cross correlation overall (t-test, p<.001) and in 11/15 individual data sets.

Supplementary Figure 3



Supplementary Figure 3. Relationship between SC responses during different task epochs and attention-related correlation changes with MT. (A) Schematic of task timing depicts the three 100ms epochs used to count spikes in SC units. The baseline period began 100ms before the first stimulus appeared, which is after stable fixation had been acquired. The stimulus period was shifted 30 ms after the appearance of the visual stimulus, to account for the earliest visual latencies observed in the SC. The delay period began 100ms after the first stimulus turned off and always ended prior to the onset of the second stimulus. (B) Attention-related changes in MT-SC r_{SC} plotted against the difference between each SC unit's response during the stimulus and baseline periods. There are multiple MT-SC correlation differences measured for each SC unit. Correlations between MT and SC were calculated using the same data and methods as Supplementary Figure 1C (Pearson correlation, rho=0.087, $p< 10^{-12}$). Isolines depicting the decile boundaries are overlaid over the individual data points. (C) Similar to B, but data are now sorted by the difference between each SC unit's response during the stimulus and delay periods (Pearson correlation, rho=0.092, $p< 10^{-14}$). (D) Similar to B, but data are now sorted by the difference between each SC unit's response during the delay and baseline periods (Pearson correlation, rho=-0.042, $p< 10^{-4}$).

Supplementary Figure 4



Supplementary Figure 4. Effects of attention on the stimulus information that can be decoded from small populations of V4 neurons is similar to MT. (A) Ability of a cross-validated linear decoder to distinguish the original from changed stimuli (intermediate change amount) for both Stimulus and Choice decoders (no SC data was available). Error bars represent SEM. Attention significantly affected the performance of both the Stimulus and Choice decoders (t-test, p<0.05), but the attention-related improvement in the Choice decoder was greater than in the Stimulus decoder (paired t-tests, p<0.05). (B) We conducted the same weight swapping analysis described in the main text (Figure 3B), which demonstrated that decoding performance was typically better using the V4 responses from the cued condition and the Choice decoder weights from the uncued condition (y-axis) than using the V4 responses from the uncued condition and the Choice decoder weights from the cued condition (x-axis; paired t-test, p<.05).