Probing the neural dynamics of mnemonic representations in humans

Wei Liu[1], Nils Kohn[1], Guillén Fernández[1]

1. Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands

Correspondence:

Wei Liu

Department of Cognitive Neuroscience

Donders Institute for Brain, Cognition and Behaviour

Radboud University Medical Centre

Trigon Building, Kapittelweg 29

6525 EN Nijmegen, The Netherlands

Tel: +31 (0)24 36 53276

E-mail: W.Liu@donders.ru.nl

Abstract (200/200 words)

Memories are not stored as static engrams, but as dynamic representations affected by processes occurring after initial encoding or even consolidation. How the modulation of memory traces after their formation is reflected in the neural activity during subsequent retrieval is currently not well understood. Using fMRI in 27 healthy human participants, we probed how neural representations of associative memories are dynamically modulated by two behavioral techniques that can either strengthen or weaken memories after encoding. Behaviorally, we demonstrated that, after an initial delay of 24 hours, associative memories can still be strengthened or weakened by repeated retrieval or suppression, respectively. Neurally, we show that repeated retrieval dynamically reduced activity amplitude in ventral visual cortex and hippocampus, but enhanced the distinctiveness of activity patterns in the ventral visual cortex. Critically, a larger reduction of activity amplitude in the ventral visual cortex associated with larger enhancement of distinctiveness of activity patterns in the same region. In contrast, repeated memory suppression was associated with reduced lateral prefrontal activity, but relative intact activity patterns. These results reveal dynamic adaptations of mnemonic representations in the human brains and how retrieval-related activity amplitude and distinctiveness of activation patterns change as a function of strengthening or weakening.

1. INTRODUCTION

Historically, memories were seen as more or less stable traces or engrams. After initial formation, memory traces are affected by consolidation leading to stabilization and weakening leading to forgetting (Ebbinghaus, 1885; Lashley, 1950; Müller and Pilzecker, 1900). However, contemporary research has provided ample evidence showing that memories continue to be dynamically adapted after initial encoding and consolidation and thus, can be modified by external factors beyond consolidation throughout their existence. For instance, retrieval practice can reinforce memory traces (Karpicke and Roediger, 2008), promote meaningful learning (Karpicke and Blunt, 2011), and protect memory retrieval against acute stress (Smith et al., 2016). In contrast, retrieval suppression can prevent unwanted memories to be retrieved (Anderson and Green, 2001), and reduce their emotional impact (Gagnepain et al., 2017). However, studies probing the neural basis of this dynamic process focused mostly on the mechanisms of modulation. How the modulation of memory traces after their formation is reflected in the neural activity during subsequent retrieval is currently not well understood. Here, we sought to characterize dynamically adapted mnemonic representations in humans by tracking the neural representations of the original memory. Memories were on the one hand reinforced by repeated memory retrieval and on the other hand, weakened by repeated memory suppression. This provides a window into the dynamically changing neural representations of memories.

The first challenge in tracking dynamically changing memory representations in humans is to characterize them noninvasively during memory retrieval. Memory representations and their dynamics can be potentially measured by changes in levels of activity amplitude (i.e., univariate analysis) or activation patterns (i.e., multivariate pattern analysis) based on the blood-oxygen-level-dependent (BOLD) signals using functional magnetic resonance imaging (fMRI). Successful retrieval of visual experiences is associated with increased activity amplitude in perceptual (Kosslyn et al., 1997; O'Craven and Kanwisher, 2000; Wheeler et al., 2000) and mnemonic regions (Kuhl et al., 2010; Shohamy and Wagner, 2008), which are also active during initial perception. Post-encoding processes like consolidation can have

78 further influence on retrieval-related activity amplitude: hippocampal activity continued to decrease,
79 whereas activity in the medial prefrontal region increased (Takashima et al., 2009, 2006). Although these
80 univariate analyses of activity amplitude have revealed the spatial-temporal features of human brain
81 networks involved in memory retrieval, multivariate pattern analysis (MVPA) (Cohen et al., 2017) of
82 activation patterns during memory retrieval can better capture individual memory representation (Xue,
83 2018). A large body of literature indicates that successful memory retrieval involves the reactivation of
84 local fine-grained activity patterns that were present when the stimulus was initially processed in stimulus-
85 specific perceptual and stimulus-general mnemonic regions (Chen et al., 2017; Lee et al., 2018; Polyn et
86 al., 2005; Wimber et al., 2015). To probe changing memory representations, we tracked retrieval-related
87 neural dynamics of both activity amplitude and activity patterns associated with repeated retrieval and
88 memory suppression.

89 Next, if both reactivated activity amplitude and patterns carry mnemonic information during retrieval, do
90 they jointly or independently change as a function of post-encoding modulation? Preliminary evidence
91 suggests that activity amplitude and activity patterns may carry complementary information (Jimura and
92 Poldrack, 2012), and evidence from visual expectation research supports this idea (de Lange et al., 2018;
93 Kok et al., 2012). Activity amplitude in V1 is suppressed by prior expectation, while the activity pattern of
94 the same area become more distinct, potentially carrying more fine-grained perceptual information (Kok
95 et al., 2012). Here, activity amplitude and pattern of the same region were separately modulated by
96 expectation. However, similar empirical evidence regarding the relationship between activity
97 amplitude/patterns and memory retrieval is lacking. Therefore, we reasoned that similar to the findings in
98 visual expectation, memory traces strengthened by repeated retrieval are accompanied by overall reduced
99 activity amplitude, yet more distinct activity patterns. In contrast, weakened memory traces may be
100 associated with opposite neural changes, that is, higher activity amplitude, but less distinct activity pattern.

101 The third challenge of measuring changing mnemonic representations is to disentangle neural activities
102 associated with memory cues presented at test and activity associated with reactivated mental images.

Processing of visual memory cues can elicit perception-related neural activity, for instance, in visual areas. At the same time, retrieval of vivid visual experiences can be associated with retrieval-related neural reactivation in the same regions. The challenge is to disentangle coexisting representations of visual and mnemonic information in the same region (Rademaker et al., 2019). One method to separate these two processes is to use two perceptual modalities (e.g. sounds as memory cues, and pictures as information to be retrieved)(Bosch et al., 2014). Here, using the same modality, we used highly similar memory cues across different memory associations. Thus, most of the variances in activity patterns of visual areas would be associated with retrieval-related neural reactivation. Furthermore, we reasoned that the behavioral modulation implemented would mainly alter retrieval-related instead of perception-related activity and patterns. We used a multivariate activity pattern-based index to quantify the fidelity of neural representation of retrieved "mental images". Across different memory associations, higher distinctiveness of neural representations (i.e., higher activity pattern variability) in visual or mnemonic regions is associated with more distinctive "mental images".

In sum, our primary goal is to reveal if two behavioral techniques (memory retrieval and suppression) differently modulate neural reactivation of associative memories, and if such modulation results in altered memory representations detected by fMRI. A better understanding of these neural dynamics would lead to deeper insights into how established memories are dynamically transformed by experience (Xue, 2018). To this end, 27 healthy participants underwent a two-session functional MRI (fMRI) experiment (Figure 1A). To localize areas for MVPA in the visual system, and to familiarize participants with the pictures of the to-be-remembered associations, we instructed participants to perform initially a familiarization task (Figure 1B) in which they sequentially viewed each of the 48 pictures used in the subsequent study phase. There, subjects intentionally memorized a series of 48 picture-location associations (Figure 1C) and returned 24h later for the fMRI session covering two experimental phases. A modulation phase (Think/No-Think paradigm (Benjamin J Levy and Anderson, 2012)) during which participants were cued to retrieve one-third of the associations (retrieval) or to avoid retrieving the associated images for another

third of associations (suppression). The remaining one-third of the associations were not cued during this phase and served as controls (Figure 1D). In the following test phase, subjects performed a final memory test during which all associations were cued once more to be retrieved (Figure 1E). We first investigated the possibility that associative memories can still be modulated after 24 hours. Behaviorally, we asked whether repeated retrieval and memory suppression would oppositely strengthen or weaken original memory traces. Next, using fMRI, we examined whether retrieval and suppression would modify neural reactivation of memories: first, we asked whether retrieval alters neural activity amplitude and pattern variability of memory representations in visual and mnemonic areas compared to control associations. And second, we tested whether memory suppression would modulate the same two neural measures, but in the opposite way.

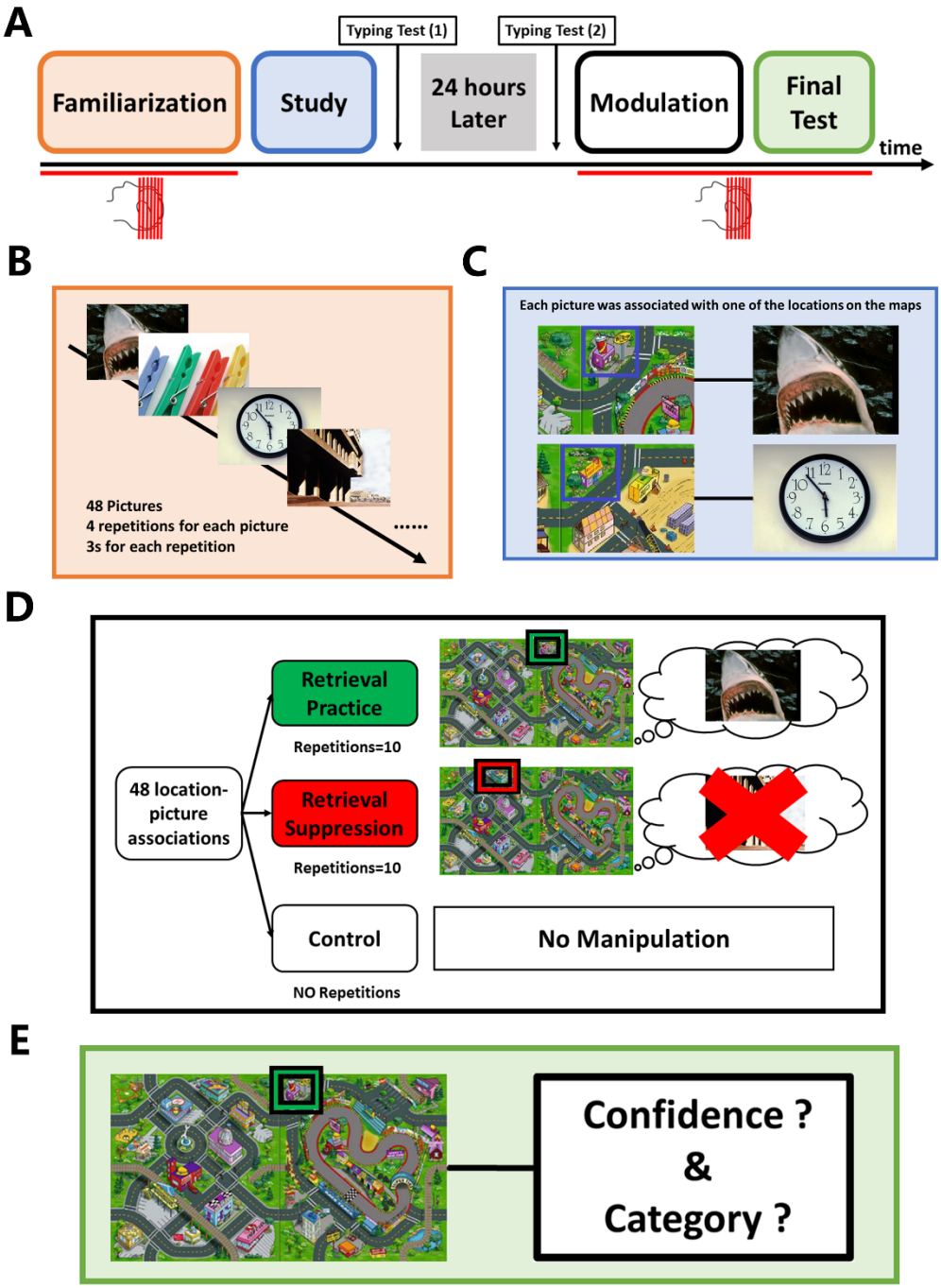

Figure 1 Schematic of the experiment design. (A) Timeline of the two-day experimental procedures. Red lines below the timeline indicate the tasks in the MRI scanner. **(B)** During the familiarization phase, all of the pictures of the to-be-remembered associations were randomly presented four times for the familiarization and estimation of picture-specific activation patterns. To keep participants focused, on each trial, they were instructed to categorize the picture shown as an animal, human, location, or object. **(C)** Study phase. Participants were trained to associate memory cues with presented pictures. **(D)** Modulation phase. After 24 hours, we used the Think/No-Think paradigm to modulate consolidated associative memories. Participants were instructed to actively retrieve associated pictures in mind ("retrieval"), or suppress the tendency to recall them ("suppression") according to the colors of the frames (GREEN: retrieval; RED: suppression) around locations. **(E)** Final memory test phase. Participants performed the final memory test after the modulation. For each of the 48 location-picture associations, locations were presented again, and participants were instructed to report the memory confidence and categorize the picture that came to mind.

2. RESULTS

2.1. Behavioral results

Pre-scan memory performance immediately after study and 24 hours later

On day1, participants were instructed to memorize a series of sequentially presented location-picture associations, for which 48 distinct pictures (photographs) were presented together with 48 specific locations on two cartoon maps (Study Phase; Figure 1C). We selected pictures from four categories consisting of animal, human, scene (e.g. train station), and object (e.g. pen and notebooks), so that memory performance could be assessed within the scanner by instructing participants to indicate the picture's category when cued by the map location during the final memory test on day2 (Figure 1E). Each location-picture association was presented twice during this study phase (For details, See Experimental design, Online Methods). Thereafter, to assess participants' immediate memory performance, all locations were highlighted sequentially, in a random order, and participants were instructed to briefly describe the associated picture by typing down one or two sentences (Two raters evaluated these answers independently with high Cohen's kappa coefficient (κ): 0.908 and 0.885 for day1 and day2 separately; For details, See Typing test phase, Online Methods). During the immediate typing test (day1), 88.01% of the associated pictures were described correctly (SD= 10.87%; range from 52% to 100%). Before scanning on day2, a second typing memory test was conducted to access delayed memory performance after a 24 hours consolidation period. Twenty-four hours later, participants still recalled 82.15% of all associations (SD = 13.87%; range from 50% to 100%). Although we observed less accurate memory 24 hours later ($t=4.73$, $p<0.001$) (Figure S1), participants could still remember most location-picture associations well.

Behavioral performance during the modulation phase

We used the think/no-think (TNT) paradigm (Modulation Phase; Figure 1D) with trial-by-trial reports of memory retrieval/suppression to modulate and monitor the individual memory trace of a particular association (Anderson et al., 2004; B. J. Levy and Anderson, 2012). During the modulation phase, one-

third of the associations belong to retrieval condition ("*Think condition*"), and another third of associations belong to the suppression condition ("*No-Think condition*"). The remaining one-third of associations were not presented during this phase ("*Control condition*"). After each retrieval trial or suppression trial, participants were instructed to use one of four response options (Never, Sometimes, Often, and Always) to access how frequent the associated picture was brought to mind when cued by the specific map location. For each association belongs to retrieval or suppression condition, the corresponding modulation was repeated for ten times throughout the experiment.

During retrieval trials, participants reported that associated pictures were successfully recalled on most of the trials (mean=84.05%, SD=11.79 %, range from 56.25% to 100%; Figure 2A). This number is close to the accuracy of the second typing test immediately before the modulation phase. Critically, we observed that with repeated attempts to retrieve, trial-by-trial retrieval frequency rating increased over repetitions, suggesting that pictures were more likely to stay stable ($F [9,26]=5.77$, $p<0.001$, $\eta^2 =0.182$; Figure 2B).

For the analyses of suppression trials, we excluded all location-picture associations which the participant could not describe correctly immediately before the modulation phase (Typing Test Day2). This approach controlled for individual differences in memory for associations that could interfere with the analysis of memory suppression. On suppression trials, participants reported that they successfully suppressed the tendency to recall the associated pictures in about half of the trials (mean=50.62%, SD=25.35%, range from 4% to 92.5%; Figure 2C). As shown in think/no-think literature before (B. J. Levy and Anderson, 2012), trial-by-trial intrusion frequency rating declined from the first to the tenth repetition ($F [9,26]=4.837$, $p<0.001$, $\eta^2 =0.157$; Figure 2D). These results suggest that participants were successful at retrieving or suppressing memory traces according to tasks instructions.

Memory performance during the final test

After the modulation phase, the final memory test was also performed while participants were scanned. On each trial, participants saw initially one of the 48 map locations serving as cues. Next, they rated the

198 confidence of their memory for the associated picture, and then classified the category of the recalled
199 picture (Figure 1E). Therefore, for each location-picture association, we assessed both subjective
200 (confidence rating) and objective (if they selected the correct category) memory. We also measured
201 reaction time (RT) of the category judgment as a proxy for the speed of memory retrieval.

202 During the final test, participants, on avenge, selected the correct category (chance level=1/4) for the
203 associated picture on 91.82% (SD = 6.05%; range from 70.83% to 100%) of the successfully recalled
204 associations of the typing test day2 (mean=39.43). We then examined how repeated retrieval and
205 suppression affected memory performance in the final test. We defined associations that were cued during
206 modulation to retrieval as *RETRIEVAL ASSOCIATIONS* and associations that were cued to suppression
207 retrieval as *SUPPRESSION ASSOCIATIONS*. The remaining third of associations were not cued during
208 modulation and regarded as *CONTROL ASSOCIATIONS*.

209 First, we compared the recall accuracies between three kinds of associations. Analysis of objective recall
210 accuracy after modulation showed no significant main effect of *modulation* ($F[2,26]=0.524$, $p=0.595$, η^2
211 $=0.02$; Figure 2E). Due to the lack of suppression-induced forgetting effect (lower accuracy for
212 *SUPPRESSION ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS*) at the group level, we further
213 performed correlational analysis to associate performance during the memory suppression and the final
214 memory test performance. More specifically, for each participant, we quantified the rate at which
215 intrusions declined over repetitions during the modulation phase (*intrusion slope score*: the more negative
216 the score is, more effective the participant can reduce intrusions; details in Online Methods) and
217 suppression-induced forgetting effect by subtracting the objective retrieval accuracy of *SUPPRESSION*
218 *ASSOCIATIONS* from *CONTROL ASSOCIATIONS* (*suppression score*: the more negative the score is,
219 more below-control forgetting on the final test). We found that participants who were more effective in
220 suppressing intrusions (higher intrusion slope score) during the modulation phase were the ones who show
221 larger suppression-induced forgetting effects ($r=0.411$, $p=0.03$; Figure2F), suggesting that successful

retrieval suppression was subsequently associated with suppression-induced forgetting. This correlation was also reported before in the think/no-think literature (B. J. Levy and Anderson, 2012).

Additionally, we investigated the effect of *modulation* on memory confidence and found a significant main effect (F [2,26]=5.928, p=0.005, η^2 =0.186; Figure 2G). Post-hoc analyses revealed higher recall confidence for *RETRIEVAL ASSOCIATIONS* compared to the *CONTROL ASSOCIATIONS* (t=3.35, p_{holm}=0.007) and a trend towards higher confidence compared to *SUPPRESSION ASSOCIATIONS* that just failed to reach our threshold for statistical significance (t=2.172, p_{holm}=0.07). Finally, we asked if modulation affected the speed of retrieval indexed by RT during the final test. Even though we did not find a significant main effect of modulation (F [2,26]=2.905, p=0.06, η^2 =0.10; Figure 2H), recall of *RETRIEVAL ASSOCIATIONS* was faster compared to the recall of *CONTROL ASSOCIATIONS* (t=-2.486, p=0.02).

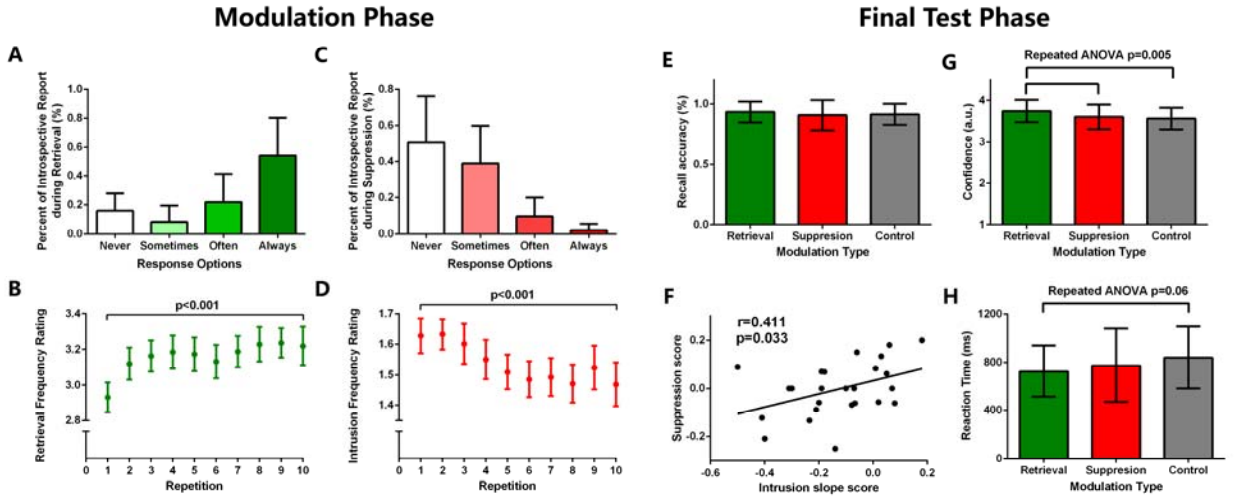

Figure 2 Behavioral performance during modulation and final test phase. (A) Percentage of the trial-by-trial introspective report during the retrieval trials. For most of the retrieval trials (mean=84.05%, SD=11.79 %), associated pictures were successfully recalled (sometimes+often+always). **(B)** With repeated retrieval attempts, associated pictures were more likely to stay in mind stably (F=5.77, p<0.001). **(C)** Percentage of the trial-by-trial introspective report during the suppression trials. During half of the suppression trials (mean=50.62%, SD=25.35%), participants successfully suppressed the tendency to recall the associated pictures (never). **(D)** As the number of repetition of suppression increase, the possibility of suppression failure declined (F=4.837, p<0.001). (E) There is no effect of retrieval or suppression on the accuracy of the categorization during the final test (p=0.595). **(F)** Participants who are more effective in reducing suppression failures (more negative the *Intrusion Slope Score*) were the ones who show more evidence suppression-induced forgetting (more negative the *Suppression Score*). **(G)** For *RETRIEVAL ASSOCIATIONS*, participants reported higher subjective confidence compared to *SUPPRESSION ASSOCIATIONS* (t=2.172, p holm=0.07) , and *CONTROL ASSOCIATIONS* (t=3.35, P_{holm}=0.007). **(H)** For *RETRIEVAL ASSOCIATIONS*, participants spent less time during categorization compared to the *CONTROL ASSOCIATIONS* (t=-2.486, P=0.02), and the effect between three conditions tend to be significant (F=2.905, p=0.06). a.u= arbitrary unit.

2.2 fMRI results

We aimed to assess differences in activity amplitude and activity patterns that are associated with the reported behavioral effects of repeated retrieval and memory suppression. For the analysis of activity amplitude, we used the standard whole-brain univariate General Linear Model (GLM) analysis, contrasting *RETRIEVAL ASSOCIATIONS* or *SUPPRESSION ASSOCIATIONS* with *CONTROL ASSOCIATIONS* separately.

For the analysis of activity patterns, we used a multivariate activity pattern-based neural index to measure the fidelity of pattern reactivation during retrieval. We assumed that the ventral visual cortex (VVC) and the hippocampus would demonstrate neural reactivation of picture-specific activity patterns during retrieval. First, using the data from the familiarization task, we identified areas (Figure 3A) within the VVC that are sensitive to picture-specific information during perception and demonstrated the neural pattern reactivations during memory retrieval 24 hours later (Details See Supplemental Texts and Figure S2-S4). The VVC mask together with the anatomically-defined bilateral hippocampus mask (Figure 3B) were regarded as regions-of-interest (ROIs). Second, trail-by-trail voxel-wise activity patterns of ROIs during modulation and the final memory test were extracted. To quantify distinctiveness of retrieved "mental images", we calculated the activation pattern variability using Pearson correlation (Figure 3C). We hypothesized that high activity pattern variability (low pattern similarity measured by R-values) reflects neural reactivations of distinctive "mental images". Here, we first report the effects of repeated retrieval on activity amplitude and activity patterns, and then the counterparts related to suppression.

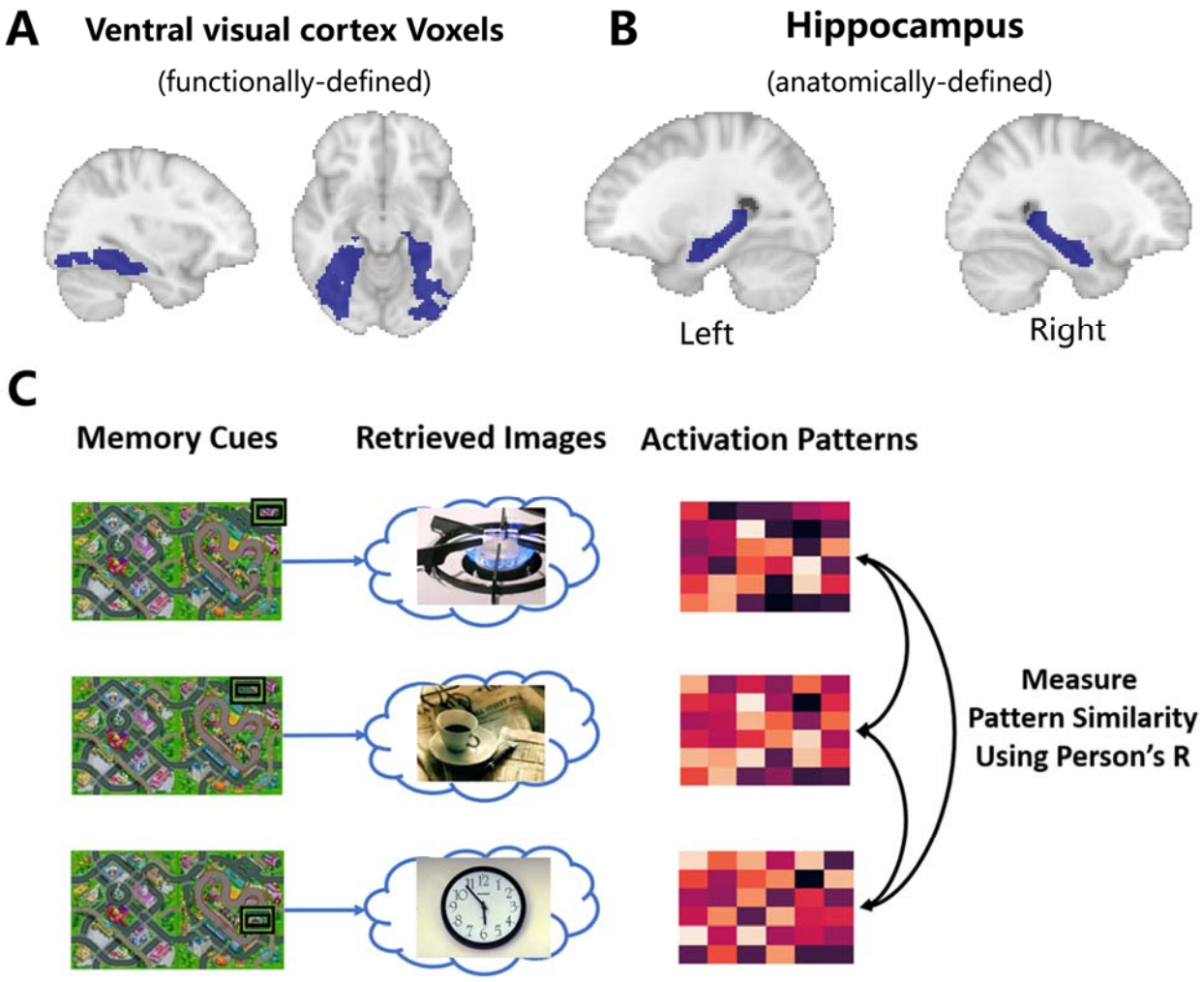

Figure 3 Regions-of-interest (ROI) and rationale of the activity pattern variability analysis. (A) Functionally-defined voxels within the ventral visual cortex (VVC). We identified voxels whose activation patterns can be used to differentiate pictures that were processed during the familiarization phase and were reactivated during successful memory retrieval during the final test (Details in Supplemental Materials). **(B)** Anatomically-defined bilateral hippocampus ROI. **(C)** During the final test, "mental images" were retrieved based on highly similar memory cues (different locations within maps were cued). We derived activation patterns for each memory retrieval trials based on fMRI data, and then quantify the pattern variability across trials using Person's r. Lower the similarity measure (r-value), higher the pattern variability.

2.2.1 Repeated retrieval leads to reduced activity amplitude, but more distinct activity patterns

Repeated retrieval reduces the activity amplitude in the visual cortex and hippocampus

Compared to *CONTROL ASSOCIATIONS*, retrieval of *RETRIEVAL ASSOCIATIONS* was associated with less activation in medial occipital cortex, fusiform gyrus, supplementary motor area (SMA),

anterior/medial cingulate cortex (MCC), left precentral gyrus, precuneus, bilateral insula, and bilateral inferior frontal gyrus (IFG) (voxelwise $P_{uncorrected}<0.001$, $p_{FWE\text{-}cluster}<0.05$; Figure 4A; Figure S5; Table S1). Most of these regions are located within the anatomically-defined VVC. The whole-brain analysis did not show an effect of retrieval on the activity amplitude of hippocampal voxels under the same threshold. However, ROI-based analysis of hippocampal signal found reduced activity when retrieving *RETRIEVAL ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS* (left hippocampus: t=-2.43, p=0.022; Figure 4E; right hippocampus: t=-2.18, p=0.038; Figure 4G).

Next, we confirmed that the observed activity reduction in VVC and bilateral hippocampus is related to a linear decrease in activity with repeated retrieval using the data from the modulation phase. Specifically, we extracted the beta coefficient from the VVC cluster defined by the *RETRIEVAL vs CONTROL* contrast (as shown in Figure 3A) and the hippocampus for each run of the modulation phase and tested for the change in activity amplitude across runs. We found reduced VVC over repeated retrieval attempts ($F [4, 25]=6.95$, $p<0.001$, $\eta^2 =0.218$; Figure 4B). Similarly, for the bilateral hippocampus, we observed a trend toward a gradual decrease of hippocampal signal across repetitions (left hippocampus: $F [4, 25]=2.39$, $p=0.056$, $\eta^2 =0.087$; right hippocampus: $F [4, 25]=2.22$, $p=0.072$, $\eta^2 =0.082$).

Repeated retrieval dynamically enhances the distinctiveness of activity patterns in the visual cortex, but not hippocampus

We next examined whether the reduced activity amplitude was associated with reduced or enhanced distinctiveness of activity patterns during the final memory test. Focusing on the identified VVC areas and hippocampus (Figure 3A and 3B), we calculated the trial-by-trial activity pattern variability for *RETRIEVAL ASSOCIATIONS* and *CONTROL ASSOCIATIONS* separately. Results show that retrieval-related activity patterns for *RETRIEVAL ASSOCIATIONS* have increased variability in VVC compared to *CONTROL ASSOCIATIONS* (t=2.3, df=26, p=0.029; Figure 4C). However, we did not observe a similar effect in the hippocampus (left hippocampus: t=-0.91, df=26, p=0.36, Figure 4F; right hippocampus: t=-0.456, df=26, p=0.65 ; Figure 4H). To test the robustness of increased pattern variability for *RETRIEVAL*

ASSOCIATIONS, we performed the same contrast based on (1) all associations instead of only remembered association, the VVC areas defined by (2) different thresholds and (3) different classification labels. All control analyses yield the same result as the reported main analysis (*See Supplemental Text and Figure S6-S8*).

To characterize the dynamic modulation of activity pattern variability in the VVC, we further applied the same variability analysis to each run of the modulation phase and analyzed these pattern variability values using a 2×5 ANOVA (*modulation*; *run*; Figure 4D). We saw a significant main effect of the *run*, reflecting that pattern variability of the VVC increased with repetitions ($F [4,25]=10.55$, $p<0.001$, $\eta^2 =0.297$). We also saw a main effect of *modulation*, reflecting that pattern variability of the *RETRIEVAL ASSOCIATIONS* is consistently higher than the variability of *SUPPRESSION ASSOCIATIONS* ($F [1, 25]=23.77$, $p<0.001$, $\eta^2 =0.487$). The interaction between modulation and runs just failed to be significant ($F [4, 25]=2.427$, $p=0.053$, $\eta^2 =0.089$). This pattern of results suggests that increased pattern variability is not only the result of repetition: Even though memory cues of *SUPPRESSION ASSOCIATIONS* have also been presented ten times during the modulation, repeated retrieval more effectively enhanced pattern distinctiveness compared to suppression. We did not perform the same dynamic analysis to the hippocampal activity patterns because no effect was found in the final memory test.

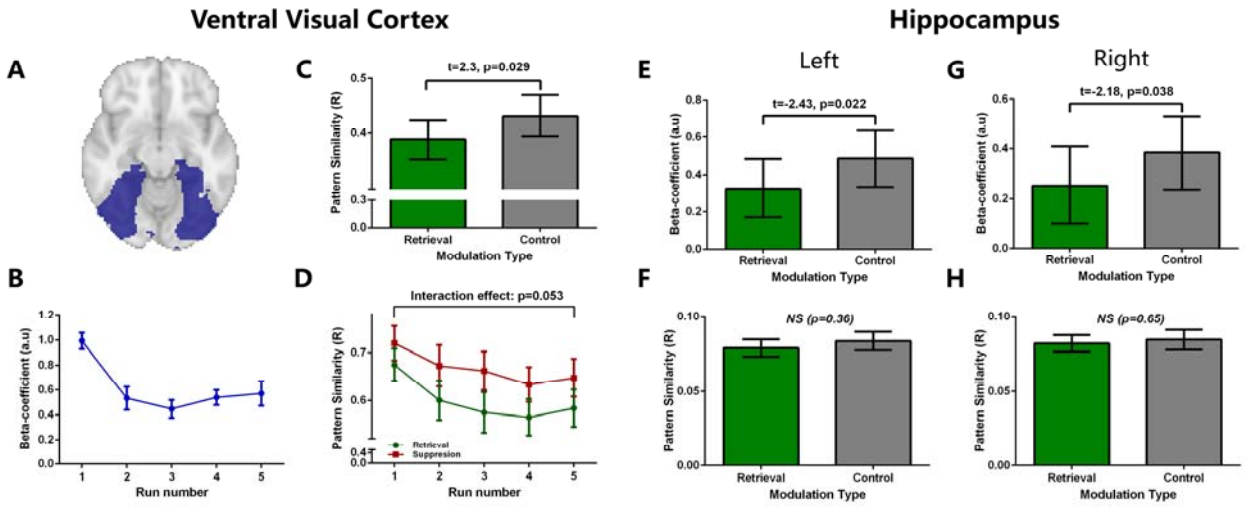

Figure 4 Repeated retrieval dynamically modulated activity amplitude and patterns variability. (A) During the final test, compared to *CONTROL ASSOCIATIONS*, *RETRIEVAL ASSOCIATIONS* was associated with lower activity amplitude in several brain regions, largely located within the ventral visual cortex (VVC)(whole-brain visualization in Figure S5). (B) The same VVC cluster showed decreased activity amplitude over repetitions of retrieval during the modulation phase. (C) Higher activation pattern variability (lower pattern similarity) in the VVC for *RETRIEVAL ASSOCIATIONS* compared to the *CONTROL ASSOCIATIONS* during the final test (t=2.3, p=0.029). (D) Dynamically enhanced pattern variability in the VVC. For both *RETRIEVAL ASSOCIATIONS* and *SUPPRESSION ASSOCIATIONS*, VVC's pattern variability increased over repetitions during the modulation (F=11.12, p<0.001). However, repeated retrieval tends to more effectively enhance pattern variability compared to suppression (F=2.42, p=0.053). (E) Reduced left hippocampal activity amplitude for *RETRIEVAL ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS* during the final test (t=-2.43, p=0.022). (F) No differences in left hippocampal activity pattern variability between *RETRIEVAL ASSOCIATIONS* and *CONTROL ASSOCIATIONS* (p=0.36) during the final test. (G) Reduced right hippocampal activity amplitude for *RETRIEVAL ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS* during the final test (t=-2.18, p=0.038). (H) No differences in right hippocampal activity pattern variability between *RETRIEVAL ASSOCIATIONS* and *CONTROL ASSOCIATIONS* (p=0.65) during the final test.

Reduced activity amplitude associated with enhanced distinctiveness of activity patterns in the visual cortex

Our activity amplitude and activity pattern variability analysis independently demonstrate that repeated retrieval dynamically reduced activity amplitude, but enhances the distinctiveness of activity pattern in the VVC. However, we performed these analyses at the whole-brain level and ROI level separately. To further confirm the dissociation between the effects of repeated retrieval on activity amplitude and pattern variability, we restricted our analyses to the same set of voxels within the VVC (Figure 5A), which showed reduced activity amplitude during the final test and, at the same time carries picture-specific information during perception and retrieval. Similar to our main analysis (as shown in Figure 4A and 4B), we found reduced activity amplitude (t=5.823, df=26, p<0.001; Figure 5B), but increased distinctiveness

of activity patterns (lower pattern similarity: t=2.887, df=26, p=0.007; Figure 5C) in the same set of VVC voxels. Critically, participants who showed a larger reduction in activity amplitude were more likely to show a larger increase in the distinctiveness of patterns (r=0.524, p=0.005; Figure 5D).

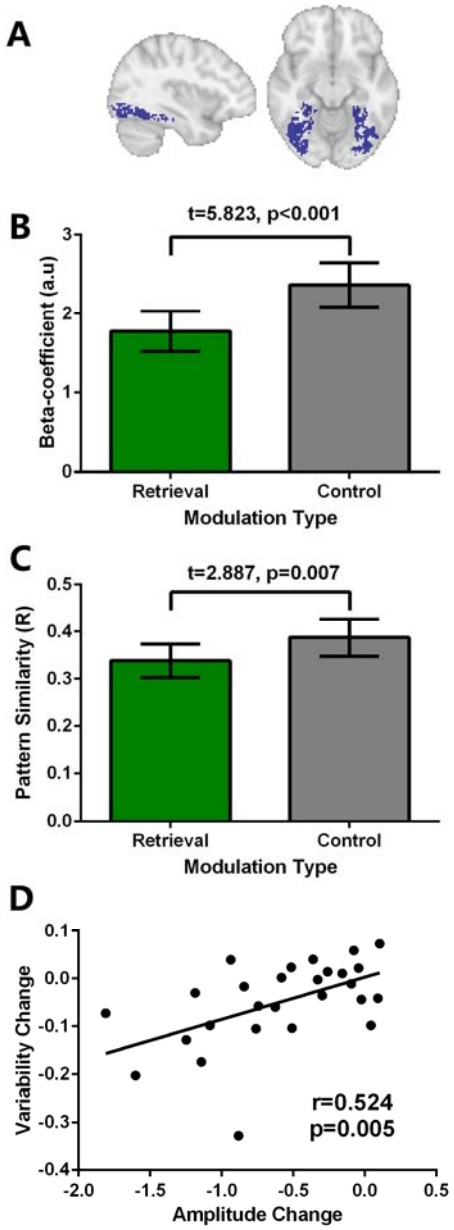

Figure 5. Measuring both activity amplitude and distinctiveness of activity patterns in the ventral visual cortex. (A) We identified overlapping voxels between activity amplitude and activation pattern analyses. First, these voxels showed reduced activity amplitude for *RETRIEVAL ASSOCIATIONS* during the final test. Second, these voxels demonstrated picture-specific activation patterns during both perception and retrieval. (B) Reduced activity amplitude of these voxels for *RETRIEVAL ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS* during the final test (t=5.82, p<0.001). (C) Higher distinctiveness of activation patterns (lower pattern similarity) of these voxels for the *RETRIEVAL ASSOCIATIONS* compared to the *CONTROL ASSOCIATIONS* during the final test (t=2.88, p=0.007). (D) Across participants, the extent of activity amplitude reduction positively correlated with pattern distinctiveness enhancement (r=0.524, p=0.005).

2.2.2 Retrieval suppression was associated with reduced of the lateral prefrontal activity

Weaker lateral prefrontal cortex (LPFC) activation as the result of retrieval suppression

The contrast between retrieval of *SUPPRESSION ASSOCIATIONS* and *CONTROL ASSOCIATIONS* during the final test revealed decreased activation of only one cluster in the left LPFC (x=−52,y=38, z=16, Z_{peak}=4.09, size=1320 mm^3; Figure 6A). We did not find any significant effect of retrieval suppression on hippocampal activity amplitude in the whole-brain or the ROI analysis (left hippocampus: t=-1.14, df=26, p=0.26; right hippocampus: t=-0.81, df=26, p=0.43).

To characterise dynamical activity changes in the left LPFC, we extracted beta values from the cluster for each modulation run and found a decrease of the amplitude of retrieval suppression from the first run to the fourth run during the retrieval of *SUPPRESSION ASSOCIATIONS* (F [3, 25]=2.98, p=0.036, η^2 =0.107). However, we found an unexpected activation increase from the fourth to the fifth run, and if we combined data from all five runs, the effect is not significant anymore (F [4, 25]=2.03, p=0.09, η^2 =0.075; Figure 6B)

Intact neural representations of after memory suppression

Although we did not find evidence for an effect of retrieval suppression on VVC or hippocampal activity amplitude, we still examined if retrieval suppression modulated activity patterns in the VVC or hippocampus. Pattern variability analysis revealed no significant difference between *SUPPRESSION ASSOCIATIONS* and *CONTROL ASSOCIATIONS* in all regions investigated (VVC: t=-1.035, df=26, p=0.31; Figure 6C; left hippocampus: t=-0.75, df=26, p=0.43 ; Figure 6D; right hippocampus: t=-.010, df=26, p=0.92 ; Figure 6E)

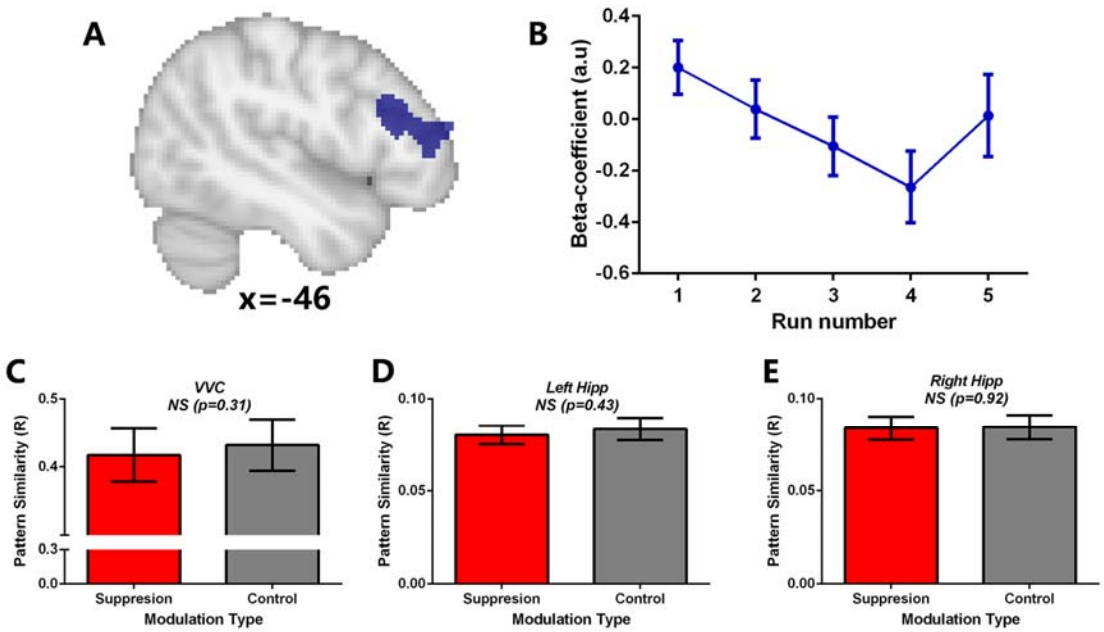

Figure 6 Repeated suppression disengaged lateral prefrontal cortex (LPFC) during subsequent memory retrieval. (A) During the final memory test, we found lower activity amplitude in the left LPFC for *SUPPRESSION ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS*. (B) During the modulation, the activity amplitude of the same LPFC cluster tended to decreased over repetitions (from run1 to run4, p=0.03, from run1 to run5, p=0.09). (C) No differences in VVC's activity pattern variability between *SUPPRESSION ASSOCIATIONS* and *CONTROL ASSOCIATIONS* (p=0.31) during the final test. (D) No differences in left hippocampal activity pattern variability between retrieval *SUPPRESSION ASSOCIATIONS* and *CONTROL ASSOCIATIONS* (p=0.43) during the final test. (E) No differences in right hippocampal activity pattern variability between *SUPPRESSION ASSOCIATIONS* and *CONTROL ASSOCIATIONS* (p=0.92) during the final test.

3. DISCUSSION

Memories are not stored as stable engrams, but flexible representations that can be modified throughout their existence. Behaviorally, our results demonstrate that, after an initial delay of 24 hours, repeated retrieval strengthened memories further, indexed by higher recall confidence and shorter reaction times. In turn, successful memory suppression during modulation was subsequently associated with lower memory performance. Neurally, we show that retrieving strengthened memories is associated with reduced activity amplitudes, but enhanced distinctiveness of pattern reactivations in the ventral visual cortex, while retrieving suppressed memories is associated with reduced activation in the left lateral prefrontal cortex.

First, our behavioral results revealed that associative memories could still be strengthened or suppressed by repeated retrieval or suppression separately after initial consolidation. The beneficial effect of retrieval practice on the subsequent retrieval is well established (Karpicke and Blunt, 2011; Karpicke and Roediger, 2008; Karpicke and Roediger III, 2007; Smith et al., 2016). In this study, memory accuracy was already near to ceiling level after consolidation, and thus we did not find higher recall accuracy of *RETRIEVAL ASSOCIATIONS* compared to *CONTROL ASSOCIATIONS*. But we observed higher memory confidence and shorter reaction times during memory retrieval, suggesting strengthened associative memories after this modulation. Corroborating the behavioral effect for the final memory test, we also found that repeat retrieval of certain memories increased their tendency to remain stable in mind during the modulation phase. In contrast, we only found a modest effect of retrieval suppression. There are at least two possible reasons for this: first, due to extensive training during encoding and/or the nature of our picture-location associations, recall accuracy for all conditions was close to ceiling level. This made suppression of associations difficult. Second, the suppression-induced forgetting effect is smaller when memories have been consolidated (Liu et al., 2016). Thus, in line with previous studies, suppression-induced forgetting may have not emerged at the group level (Gagnepain et al., 2017; Liu et al., 2016). Critically, we replicated two findings in the memory suppression literature to confirm that our memory suppression modulation was still effective. First, when unwanted memories were suppressed repeatedly, their tendency to intrude was reduced (Benoit et al., 2015; Gagnepain et al., 2017; Hellerstedt et al., 2016; B. J. Levy and Anderson, 2012; van Schie and Anderson, 2017). Second, the extent of this reduction correlated with subsequent suppression-induced forgetting effects across participants (B. J. Levy and Anderson, 2012).

For *RETRIEVAL ASSOCIATIONS*, fMRI revealed a dynamic process based on decreased retrieval-related activity amplitude and enhanced distinctiveness of activity patterns. Recently, Antony and colleagues proposed that retrieval acts as a fast route to memory consolidation (Antony et al., 2017), and provided initial evidence for the rapid creation of neocortical memory traces during retrieval practice (Ferreira et al., 2019). Similar to their results, we also found that hippocampal activity decreased across repeated retrieval,

425 even though our modulation targeted already consolidated associative memories. These results are
426 consistent with decreasing retrieval-related hippocampal activity over consolidation (Takashima et al.,
427 2009, 2006). Extending these hippocampal findings, we revealed that repeated retrieval-induced neural
428 dynamics was associated with both reduced activity amplitude and increased activity pattern
429 distinctiveness in relevant perceptual regions. This pattern of results is in line with our knowledge about
430 how expectations shape brain responses. Expected stimuli reduce overall activity amplitude, a
431 phenomenon termed "expectation suppression"(Summerfield et al., 2008; Summerfield and De Lange,
432 2014). Underlying neural representations become more distinct over the course of repetition (de Lange et
433 al., 2018; Kok et al., 2012). Our findings suggest that this principle holds for modulation of memory
434 representation and not only for visual expectation. During retrieval of strengthened memories, redundant
435 neural activity is suppressed and only the fine-grained neural patterns remain that enable more distinctive
436 memory representations with higher fidelity. This explanation is further supported by the correlation
437 between changes in activity amplitude and distinctiveness of activity patterns across participants when we
438 restricted our analyses to the same cluster of visual processing voxels. One may argue that the observed
439 neural changes may just be associated with repeated visual processing of memory cues, but not
440 specifically with repeated retrieval. Our fMRI results from the modulation phase challenge this argument.
441 We did not only find reduced activity amplitude and more distinct activity patterns during the final
442 memory test, but also demonstrated that these neural changes gradually emerged across repeated retrieval
443 during the modulation phase. A similar effect was not found during repeated suppression, even though
444 memory cues were also presented repeatedly. Therefore, the observed neural changes cannot be simply
445 explained by repeated processing of memory cues at the perceptual level, and thus they are more likely to
446 be the result of repeated retrieval.

447 For *SUPPRESSION ASSOCIATIONS*, we observed lower LPFC activity amplitude, but relatively intact
448 activity patterns in visual and mnemonic areas during subsequent retrieval. Active memory suppression
449 during retrieval is proposed to be partially supported by inhibitory control mechanisms mediated by the

lateral prefrontal cortex (Anderson and Hanslmayr, 2014; Guo et al., 2018). During retrieval suppression, LPFC is typically activated (Anderson et al., 2004; Guo et al., 2018; B. J. Levy and Anderson, 2012), but it showed gradually decreasing activity amplitude from the early suppression attempts to the later trials of suppression (Depue et al., 2007). Consistent with this pattern of temporal change, we found a similar decrease in LPFC activity amplitude across suppression attempts during the modulation phase and lower activity amplitude during the subsequent retrieval (i.e., final memory test). Together with the trial-by-trial intrusion frequency rating during modulation, this activity decrease across suppression attempts may suggest less inhibitory control demands when suppressing increasingly weakened memory traces. The observed reduction in LPFC activity during the subsequent retrieval might be a long-lasting effect of the progressively reduced activity amplitude after repeated memory suppression. The lower LPFC activity during the final memory test might be a long-lasting effect of the progressively reduced activity amplitude after repeated memory suppression. Another interesting observation related to memory suppression is that we found weak evidence for suppression-induced changes in activity pattern variability of both perceptual and mnemonic areas during the final memory test. Even though the brain regions, in particular for LPFC-hippocampal top-down modulation circuit, involved in memory suppression have been examined (Anderson and Hanslmayr, 2014; Guo et al., 2018), the changes in neural representations of individual memory trace underlying the suppression-induced forgetting effect remained understudied. One possibility is that, similar to competition-induced forgetting, the cortical neural representation of individual memory is suppressed, leading to forgetting (Wimber et al., 2015). Another complementary explanation is that although underlying neural representations remain intact, memory retrieval operation is impaired, causing difficulties when recalling stored information. Our results provide support for the latter explanation by showing diminished LPFC activation when retrieving previously suppressed memories. LPFC not only plays a critical role in inhibitory control over retrieval suppression (Anderson et al., 2016) but also supports cognitive control processes needed to retrieve episodic memories (Badre and Wagner, 2007; Spaniol et al., 2009). The observed reduction in LPFC activity during the final memory test might suggest insufficient cognitive control resources to facilitate retrieval. It is noteworthy that our pattern analysis was

restricted to remembered associations after the memory suppression modulation. Future studies with stronger suppression-induced forgetting effect can directly compare the activity pattern variability between still-remembered associations and forgotten associations.

Taken together, our results provide behavioral and neural evidence for dynamically adapting representations of episodic memories separately modulated by repeated retrieval and suppression. We found that repeated retrieval reduced activity amplitude, but increased the distinctiveness of activation patterns in visual areas. We propose that lower overall activity amplitude, but higher distinctiveness of cortical reinstatement is evidence for more distinct neural representations of associative memories strengthened by repeated retrieval. In contrast, for resilient memories, repeated memory suppression dynamically reduced lateral prefrontal activity without altering the distinctiveness of pattern reactivations.

4. Materials and Methods

4.1 Participants

Thirty-two right-handed, healthy young participants aged 18-35 years were recruited from the Radboud Research Participation System. They all had corrected-to-normal or normal vision and reported no history of psychiatric or neurological disease. All of them are native Dutch speakers. Two participants were excluded from further analyses due to memory performance lower than the chance level, three participants were excluded from all of the analyses because of excessive head motion during scanning, and. Neuroimaging data of one participant was partly used: he/she was excluded from the analysis of the modulation phase (Think/No-Think paradigm) due to the head motion only during this task, while his/her data during other tasks were included in the analyses. Finally, 27 participants (16 females, age=19-30, mean=23.41, SD=3.30) were included in the analyses of the final test phase, and 26 participants (15 females, age=19-30, mean=23.51, SD=3.30) were included in the analyses of modulation phase. We used Dutch-version of Beck Depression Inventory (BDI) (Roelofs et al., 2013) and State-Trait Anxiety Inventory (STAI) (van der Bij et al., 2003) to measure the participants' depression and anxiety level. All of our participants scored normally in BDI and STAI. Furthermore, because of the two-session design (24 hours' interval) of the study, we used the Pittsburgh sleep quality index (PSQI) (Buysse et al., 1989) to measure the sleep quality between two scanning sessions. No participants reported abnormal sleep-related behaviors or significantly less than normal sleep time. The experiment was approved by, and conducted in accordance with requirements of, the local ethics committee (Commissie Mensgebonden Onderzoek region Arnhem-Nijmegen, The Netherlands) and the declaration of Helsinki, including the requirement of written informed consent from each participant before the beginning of the experiment.

4.2 Materials

Locations and maps

We used 48 distinctive locations (e.g. buildings, bridges) drawn from two cartoon maps as the memory cues in our study. The maps are not corresponding to the layout of any real city in the world and participants have never been exposed to the maps before the experiment.

Pictures

48 pictures (24 neutral and 24 negative pictures) from the International Affective Picture System (IAPS) (Lang et al., 1997) were used in this study, and these pictures can be categorized as one of four following groups: animal (e.g. cat), human (e.g. reading girl), object (e.g. clock) or location (e.g. train station). Category information was used for the following memory-based category judgment test. We did not report the valence-related results in this study. All images were converted to the same size and resolution for the experiment with their original colors.

Picture-location associations

Each picture was paired with one of the 48 locations to form the picture-location association. We (W.L and J.V) carefully screened all the associations to prevent the explicit semantic relationship between picture and location (e.g. lighter- fire department). All 48 picture-location association were divided into three groups for different types of modulation (See Modulation Phase).

4.3 Experiment design

Overview of the design

This study is a two-session fMRI experiment, with the 24 hours interval between two sessions (Figure 1A). Day1 session consists of the familiarization phase (Figure 1B), study phase (Figure 1C), and immediate typing test. Day2 session consists of the second typing test, modulation phase (Figure 1D), and final memory test (Figure 1E). Among these phases, the familiarization, modulation, and final memory test phase was performed in the scanner, while study phase and two typing tests were performed in the behavioral lab.

544 *Familiarization phase*

545 The first task in the scanner for our participants was the familiarization phase. This phase was
546 conducted to obtain the picture-specific brain responses to all of the 48 pictures, measured by the Blood-
547 Oxygen-Level Dependent (BOLD) activity patterns. The second purpose of the task is to let participants
548 become familiar with the pictures to be associated with locations later. There are in total four functional
549 runs of the familiarization. For each run, each picture was shown once for 3s. The order of the
550 presentation was pseudorandom and pre-generated by self-programmed Python code. The dependence
551 between the orders of different runs was minimized to prevent potential sequence-based memory encoding.
552 To keep participants focused during the task, we instructed them to categorize the presented picture via the
553 multiple-choice question with four options (animal, human, object, and location). We used an exponential
554 inter-trial intervals (ITI) model (mean=2s, minimum=1s, maximum=4s) to generate the ITIs between trials.
555 Participants' responses were recorded by an MRI-compatible response box.

556 *Study phase*

557 Each picture-location association was presented twice in two separate runs. During each study trial,
558 the entire map was first presented for 2.5s, then one of the 48 locations was highlighted with a BLUE
559 frame, for 3s, and finally, the picture and its associated location were presented for 6s. Participants were
560 encouraged to use both the relative position of the memory cues on the maps and the appearance of the
561 highlighted areas to facilitate association learning. We pre-generated a pseudorandom order of the trials to
562 minimize the similarity between the order used in familiarization and the order used in the study phase.

563 *Typing test phase*

564 Immediately after the study phase, participants performed a typing test (day1) assessing picture-
565 location association learning. Each location was presented again (4s) in an order which differs from the
566 study phase, and participants had maximally 60s to type on the standard keyboard to describe the
567 associated picture. Twenty-four hours later (day2), participants performed the typing test again at the same

behavioral lab. The procedure is identical to the immediate typing test, but with a different order of the trials.

Modulation phase

The modulation phase is the first task participants performed on the day2 MRI session. We used the think/no-think (TNT) paradigm with the trial-by-trial self-report measures to modulate established associative memories. The same paradigm has been used in previous neuroimaging studies, and the self-report does not affect the underlying memory control process (Anderson et al., 2004; B. J. Levy and Anderson, 2012). Forty-eight picture-location associations were divided into three conditions. One-third of the associations (16 associations) were assigned to retrieval ("Think") condition, one-third of the associations were assigned to suppression ("No-Think") condition, and the remaining one-third of the associations were assigned to control condition. The assignment process was counterbalanced between participants. Therefore, at the group level, for each picture-location association, the possibility of belonging to one of the three conditions of modulation is around 33%. Associations that belong to different conditions underwent different types of modulation during this phase. Locations which belong to the control condition were not presented during this phase. For retrieval trials, locations were highlighted with the GREEN frame for 3s, and participants were instructed to recall the associated picture quickly and actively and to keep it in mind until the map disappeared from the screen. For the suppression trials, locations were highlighted with the RED frame for 3s, and our instruction for participants is to prevent the potential memory retrieval and try to keep an empty mind. We also told the participants that they should not close their eyes or pay attention to other things outside the screen during the presentation of memory cues. After each retrieval or suppression trial, participants have maximum 3s to report their experience during the cue presentation. Specifically, they answered a multiple-choice question with four options (*Never, Sometimes, Often, and Always*) by pressing the button on the response box to indicate whether the associated picture entered their mind during that particular trial.

The modulation phase consisted of in total of five functional runs (64 trials per run). In each run, 32 locations (half retrieval trials, and half suppression trials) were presented twice. Therefore, for each memory cue that not belongs to the control condition, it was presented ten times during the entire modulation phase. Again, we pre-generated the orders of the presentation to prevent the similar order sequences across five modulation runs. Between each trial, fixation was presented for 1-4s (mean=2s, exponential model) as the ITI.

Final test phase

After the modulation phase, participants performed the final memory test within the scanner. All 48 locations (including both the retrieval/suppression associations as well as control associations) were presented again by highlighting a certain part of the map with a BLUE frame. During the presentation (4s), participants were instructed to recall the associated picture as vividly as possible and keep the mental image in their mind. Critically, visual inputs during this phase were highly similar across trials and participants because the entire maps were always presented, just with different locations highlighted. Next, participants were asked to give the responses on two multiple-choice questions within 7s (3.5s for each question): (1) “how confident are you about the retrieval?” They responded with one of the four following options: Cannot recall, low confident, middle confident and high confident. (2) “Please indicate the category of the picture you were recalling?” They also had four options to choose from (Animal, Human, Object, and Location).

4.4 Behavioral data analysis

Familiarization phase

In this study, we did not focus on the accuracy of the category judgement because categorization could be a subjective process. We mainly used the responses from participants to control for the effect of subjective category categorization on the following memory performance evaluation. Specifically, for a picture, if the participant consistently labels it across four repetitions as a different category compared to

our predefined labels, we will generate an individual-specific category label for that participant and then use his/her own category label for this picture to evaluate their responses on the final test.

Typing test

Participants' answers were evaluated by two native Dutch experimenters (S.M and J.V). The general principle is that if the answer contains enough specific information (e.g. a little black cat), to allow the experimenter to identify the picture from the 48 pictures used, it will be labelled as a correct answer. On the contrast, if the answer is not specific enough (e.g. a small animal), then it will be labelled as an incorrect answer. Two independent assessors evaluated answers from the two typing test phase. We used Cohen's kappa coefficient (κ) to measure intra-rater reliability of their evaluations. In general, κ lager than 0.81 suggests almost perfect rater reliability. After the immediate typing test, we only invited participants with at least 50% accuracy to the day2 experiment. For the typing test 24 hours later, participants' responses were evaluated by the same experimenter again. We do not have accuracy requirement for day2 typing test, all of the participants (if they meet other quality control criteria) were analysed regardless of the accuracies. Based on the participants' responses on the typing testing 24 hours later, for each participant, we identified picture-location associations that he/she did not learn or already forgot. These associations were not considered in the following behavioral and neuroimaging analyses because participants have no memory associations to modulate. We calculated the average accuracies for the immediate typing test and typing test 24 hours later and investigated the delay-related decline of memory performance using the paired t-test.

Modulation phase

Responses during the modulation phase were analyzed separately for retrieval trials and suppression trials. We first calculated the percentage of each option (never, sometimes, often, and always) chosen across 160 retrieval trials and 160 suppression trials for each participant without considering ten repetitions. Next, we quantified the dynamic changes in task performance across repetitions (runs). Before

the following analyses, we coded the original categorical variable using numbers (Never-1; Sometimes-2; Often-3; Always-4). For all the established picture-location associations, we calculated their average retrieval frequency rating (based on retrieval trials) and intrusion frequency rating (based on suppression trials) on each repetition. We assumed that retrieval frequency rating indicates the retrieval quality of the memory retrieval (the higher the rating, stronger and more vivid the retrieved pictures) and intrusion frequency rating reflects the success of memory control (lower the rating, more successful the suppression is). We used a repeated-measures ANOVA to model the change of retrieval and intrusion rating across repetitions to test if the repeated attempt to retrieval or suppress a memory trace could strength or weaken the associations. Additionally, to quantify individual differences in memory control efficiency (B. J. Levy and Anderson, 2012), we calculated the *intrusion slope score* for each participant. Using all the intrusion rating of suppression trials, we used the linear regression to calculate the slope of intrusion rating across ten repetitions during the modulation phase for each participant. The increasingly negative slope score reflects better memory control at preventing associated pictures come into awareness.

Final test phase

For each trial of the final test, we calculated both the subjective memory measure based on the confidence rating (1,2,3,4) and objective memory measure based on the memory-based category judgment (correct/incorrect). Also, we extracted the reaction times (RT) of the memory-based category judgments to represent the speed of memory retrieval. Subjective memory measure, objective memory measure, and retrieval speed were averaged across trials within each condition of modulation (retrieval, suppression, and control) for each participant. To investigate the effect of types of modulation on the subjective, objective memory, and retrieval speed, we performed a repeated-measure ANOVA to detect within-participants' differences between *RETRIEVAL ASSOCIATIONS*, *SUPPRESSION ASSOCIATIONS*, and *CONTROL ASSOCIATIONS*. To assess individual differences in suppression-induced forgetting, we calculated the *suppression score* by subtracting the objective memory measure of retrieval suppression

associations ("no-think" items) from control association. Participants showed more forgetting as the result of suppression had more negative suppression scores.

Combinatory analysis of modulation and final test phase

To replicate the relationship between the memory control efficiency during the TNT task and suppression-induced forgetting effect during later retrieval reported before (B. J. Levy and Anderson, 2012), we correlated suppression scores with intrusion slope scores across all the participants. Notably, sample size (N=27) of this cross-participant correlational analysis is modest. But it is just a secondary analysis for the replication attempt.

4.5 fMRI data acquisition and pre-processing

Acquisition

MRI data were acquired using a 3.0 T Siemens PrismaFit scanner (Siemens Medical, Erlangen, Germany) and a 32 channel head coil system at the Donders Institute, Centre for Cognitive Neuroimaging in Nijmegen, the Netherlands. For each participant, MRI data were acquired on two MRI sessions (around 1 hour for each session) with 24 hours' interval. We used three types of sequences in this study: (1) a 3D magnetization-prepared rapid gradient echo (MPRAGE) anatomical T1-weighted sequence with the following parameters: 1 mm isotropic, TE = 3.03 ms, TR = 2300 ms, flip angle = 8 deg, FOV = 256 × 256 × 256 mm; (2) Echo-planar imaging (EPI)-based multi-band sequence (acceleration factor=4) with the following parameters: 68 slices (multi-slice mode, interleaved), voxel size 2 mm isotropic, TR = 1500 ms, TE = 39 ms, flip angle =75 deg, FOV = 210 × 210 × 210 mm; (3) magnitude and phase images were collected to correct for distortions (voxel size of 2 × 2 × 2 mm, TR = 1,020 ms, TE = 12 ms, flip angle = 90 deg).

During the day1 session, anatomical T1 image was acquired firstly, followed by the field map sequence. Before the four EPI-based pattern localization runs, 8 minutes resting-state data was acquired from each participant using the same sequence parameters. We did not present any results based on the

688 collected resting-state data in this study. Day2 session began with field map sequence. Then six EPI-based
689 task-fMRI runs (five runs of the modulation phase and one run of the final test phase) were acquired using
690 the same sequence.

691 *Preprocessing of neuroimaging data*

692 All functional runs underwent the same preprocessing steps using FEAT (FMRI Expert Analysis Tool)
693 Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl)(Jenkinson et al., 2012). In
694 general, the pipeline was based on procedures suggested by Mumford and colleagues
695 (http://mumfordbrainstats.tumblr.com) and the article that introduced the ICA-based strategy for
696 Automatic Removal of Motion Artifacts (ICA-AROMA) (Pruim et al., 2015). The first 4 volumes of each
697 run were removed from the 4D sequences for the stabilization of the scanner. The following pre-statistics
698 processing was applied; motion correction using MCFLIRT (Jenkinson et al., 2002); field
699 inhomogeneities were corrected using B0 Unwarping in FEAT; non-brain removal using BET (Smith,
700 2002); grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor. We
701 used different spatial smoothing strategies based on the type of analysis the data will be used for. For data
702 to be used in univariate analyses, we applied a 6mm kernel. However, for data to be used in multivariate
703 pattern analyses, no spatial smoothing was performed to largely keep the voxel-wise pattern information.
704 In addition to the default FSL motion correction algorithm, we used ICA-AROMA to further remove the
705 motion-related spurious noise, and chose the results from the “non-aggressive denoising” algorithm for the
706 following analyses. Prior to time-series statistical analyses, highpass temporal filtering (Gaussian-
707 weighted least-squares straight line fitting with sigma=50.0s) was applied.

708 Registration between all functional data, high-resolution structural data, and standard space was
709 performed using the following steps. Firstly, we used the Boundary Based Registration (BBR) (Greve and
710 Fischl, 2009) to register functional data to the participant’s own high-resolution structural image. Next,
711 registration of high resolution structural to standard space was carried out using FLIRT (Jenkinson et al.,
712 2002; Jenkinson and Smith, 2001) and was then further refined using FNIRT nonlinear registration

(Andersson et al., 2007). Resulting parameters were used to align maps between naive-space and standard space and back-projected region-of-interests into naïve space.

4.6 Anatomical Region-of-Interest (ROI) in fMRI analyses

Based on previous neural reactivation studies which also used visual stimulus (Lee et al., 2018; Polyn et al., 2005; Wimber et al., 2015), we hypothesized that ventral visual cortex (VVC) and hippocampus might carry picture-specific and category-specific information of the memory contents. Therefore, we chose VVC and the hippocampus as the ROIs in our fMRI analyses. All ROIs were first defined in the common space and back-projected into the participant's naïve space when necessary using parameters obtained from FSL during registration.

We defined anatomical VVC ROI based on the Automated Anatomical Labeling (AAL) human atlas which is implemented in the WFU pickatlas software (http://fmri.wfubmc.edu/software/PickAtlas). The procedure was used before in a previous neural reactivation study conducted by Wimber and colleagues (Wimber et al., 2015). Brain regions including bilateral inferior occipital lobe, parahippocampal gyrus, fusiform gyrus, and lingual gyrus were extracted from the AAL atlas and combined to the VVC mask. The VVC mask was mainly used as the boundary to locate visual-related voxels in the following neural pattern analyses.

The hippocampal ROI was defined using a bilateral hippocampus mask within the AAL provided by WFU pickatlas software. To yield better coverage of participants' anatomies, the original mask was dilated by a factor of 2 in the software. Admittedly, the generation of the hippocampus via the AAL atlas is less precise compared to the subject-specific hippocampus segmentation algorithm such as FSL's automatic subcortical segmentation protocol (FIRST) (Khan et al., 2008). The reason why we still use the AAL atlas is that our following analyses included both within-participant and cross-participant analyses. The hippocampal ROIs were mainly used for hypothesis-driven fMRI data analyses if the whole-brain analyses and correction do not reveal cluster(s) within the hippocampus.

737 4.7 Univariate Generalized Linear Model (GLM) analyses of response amplitude

738 *GLM analyses of neuroimaging data from the final test phase*

739 To investigate how different modulations (retrieval/suppression) modulate the subsequent univariate
740 activation, we ran the voxel-wise GLM analyses of the final test run. All time-series statistical analysis
741 was carried out using FILM with local autocorrelation correction (Woolrich et al., 2001) using FEAT. In
742 total, six regressors were included in the model. We modelled the presentation of memory cues (locations)
743 as three kinds of regressors (duration=4s) based on their modulation history (retrieval, suppression, or
744 control). To account for the effect of unsuccessful memory retrieval, we separately modelled the location-
745 picture associations which they cannot recall as a separate regressor. Lastly, the button press was modelled
746 as two independent regressors (confidence and category judgment). All the trails were convolved default
747 hemodynamic response function (HRF) within the FSL.

748 We conducted two planned contrasts (retrieval vs control and suppression vs control) first at the
749 subject space and then aligned resulting statistical maps to MNI space using the parameters from the
750 registration. These aligned maps were used for the group-level analyses and corrected for multiple
751 comparisons using default cluster-level correction within FEAT (voxelwise $Z>3.1$, cluster-level $p < .05$
752 FWER corrected). All of the contrasts were first conducted at the whole-brain level. Then, given the role
753 of the hippocampus in the task, we further extracted beta values from the hippocampal ROIs and
754 compared them for the same contrasts (retrieval vs control and suppression vs control).

755 *GLM analyses of neuroimaging data from the modulation phase*

756 We ran the voxel-wise GLM analyses for each modulation run separately. In total, three regressors
757 were included in the model. We modelled the presentation of the memory cues (location) as two kinds of
758 regressors (duration=3s) according to their modulation instruction (retrieval or suppression). Button press
759 was modelled as one independent regressor. In addition, if applicable, location-picture associations that
760 our participants could not recall were modelled as a regressor. Based on contrast-defined ROI (e.g.

retrieval vs control, threshold Z>2.3), or anatomical-defined ROI (e.g. hippocampus), we extracted beta values of these ROIs from whole-brain maps of each modulation run separately. We investigated repetition-related changes in beta values using the Repeated ANOVA for retrieval and suppression condition separately. No multiple comparison correction was used to control for the number of ROIs involved, and we reported raw p-values for each ROI analysis.

4.8 Multivariate pattern analyses of brain activation patterns

Activation Pattern estimation

All preprocessed (unsmoothed) familiarization, modulation, and final test functional runs were modelled in separate GLMs in each participant's naïve space. For each trial within familiarization, we generated a separate regressor using the onset of picture presentation and 3s as the duration. At the same time, we generated one regressor for different button presses of the category judgment to control for the motor-related brain activity. In total, 49 regressors were included in the model. This procedure led to a separate statistical map (t-values) for each trial. Similarity, for each modulation and final test run, we generated a separate regressor using the onset of the presentation of location (memory cue) and 3s as the duration. However, button presses were not included in the model because they may potentially carry ongoing memory-related information. Similarity, we got a separate t map for each modulation or test trial.

ROI-based trial-by-trial pattern variability analysis on the modulation and final test data

Representation similarity analysis (RSA) (Cohen et al., 2017) was used to calculate trial-by-trial pattern variability within particular types of test trials (e.g. recall of associations belongs to the *RETRIEVAL ASSOCIATIONS*). Because of the nature of the within-participant analysis here, to improve the pattern variability estimation, all the calculations were based on activation patterns on the native space.

Firstly, we analyzed the multivariate activation patterns on the final test. The identified VVC voxels were transformed from the standard space to native space and then used as the mask to extract single-trial activation patterns to vectors and z-scored. We excluded all the trials with the incorrect memory-based

category judgement, then divided remaining trials into three conditions based on their modulation history (e.g. retrieval practice or retrieval suppression). Next, for activation patterns of trials within the same condition, we calculated neural pattern variability using Pearson correlations between all possible pairs of trials within the group. The calculations led to three separate correlation matrices for three types of test trials for each participant. Finally, we used the mean value of all of the r values located at the left-triangle to represent the neural pattern variability of the condition (higher the r-value, lower the pattern variability). All the mean r values were Fisher-r-to-z transformed before the following statistical analyses. To investigate if different modulations have different effects on memory representation during the final test, we performed two planned within-participant comparisons: [1] *RETRIEVAL ASSOCIATIONS* vs *CONTROL ASSOCIATIONS;* [2] *SUPPRESSION ASSOCIATIONS* vs *CONTROL ASSOCIATIONS*

Next, we used the same approach to analyze the modulation data, in order to track the dynamic change of memory representation. For each presented location, activity patterns were extracted using the same mask from five modulation runs. Similarly, within-condition (retrieval or suppression) trial-by-trial pattern variability was calculated for each condition and each run. The dynamic change was modelled using the condition by run interaction using the ANOVA analysis.

4.9 Data and code availability.

Custom code used and datasets generated and/or analysed during the current study are available from the corresponding author upon request or via the Open Science Framework (OSF)(https://osf.io/ucty2/).

References

Anderson MC, Bunce JG, Barbas H. 2016. Prefrontal--hippocampal pathways underlying inhibitory control over memory. *Neurobiol Learn Mem* **134**:145–161.

Anderson MC, Green C. 2001. Suppressing unwanted memories by executive control. *Nature* **410**:366–369. doi:10.1038/35066572

Anderson MC, Hanslmayr S. 2014. Neural mechanisms of motivated forgetting. *Trends Cogn Sci* **18**:279–292.

Anderson MC, Ochsner KN, Kuhl B, Cooper J, Robertson E, Gabrieli SW, Glover GH, Gabrieli JDE. 2004. Neural systems underlying the suppression of unwanted memories. *Science (80-)* **303**:232–235.

Andersson JLR, Jenkinson M, Smith S, others. 2007. Non-linear registration aka Spatial normalisation FMRIB Technial Report TR07JA2. *FMRIB Anal Gr Univ Oxford*.

Antony JW, Ferreira CS, Norman KA, Wimber M. 2017. Retrieval as a fast route to memory consolidation. *Trends Cogn Sci* **21**:573–576.

Badre D, Wagner AD. 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* **45**:2883–2901.

Benoit RG, Hulbert JC, Huddleston E, Anderson MC. 2015. Adaptive top--down suppression of hippocampal activity and the purging of intrusive memories from consciousness. *J Cogn Neurosci* **27**:96–111.

Bosch SE, Jehee JFM, Fernández G, Doeller CF. 2014. Reinstatement of associative memories in early visual cortex is signaled by the hippocampus. *J Neurosci* **34**:7493–7500.

Buysse DJ, Reynolds III CF, Monk TH, Berman SR, Kupfer DJ. 1989. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res* **28**:193–213.

Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, Hasson U. 2017. Shared memories reveal shared structure in neural activity across individuals. *Nat Neurosci* **20**:115.

Cohen JD, Daw N, Engelhardt B, Hasson U, Li K, Niv Y, Norman KA, Pillow J, Ramadge PJ, Turk-Browne NB, others. 2017. Computational approaches to fMRI analysis. *Nat Neurosci* **20**:304.

de Lange FP, Heilbron M, Kok P. 2018. How Do Expectations Shape Perception? *Trends Cogn Sci* **22**:764–779. doi:10.1016/j.tics.2018.06.002

Depue BE, Curran T, Banich MT. 2007. Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. *Science (80-)* **317**:215–219.

Ebbinghaus H. 1885. Über das gedächtnis: untersuchungen zur experimentellen psychologie. Duncker & Humblot.

Ferreira CS, Charest I, Wimber M. 2019. Retrieval aids the creation of a generalised memory trace and strengthens episode-unique information. *Neuroimage*. doi:10.1016/j.neuroimage.2019.07.009

Gagnepain P, Hulbert J, Anderson MC. 2017. Parallel regulation of memory and emotion supports the suppression of intrusive memories. *J Neurosci* 2716–2732.

Greve DN, Fischl B. 2009. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**:63–72.

Guo Y, Schmitz TW, Mur M, Ferreira CS, Anderson MC. 2018. A supramodal role of the basal ganglia in memory and motor inhibition: Meta-analytic evidence. *Neuropsychologia* **108**:117–134. doi:10.1016/j.neuropsychologia.2017.11.033

Hellerstedt R, Johansson M, Anderson MC. 2016. Tracking the intrusion of unwanted memories into awareness with event-related potentials. *Neuropsychologia* **89**:510–523.

Jenkinson M, Bannister P, Brady M, Smith S. 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**:825–841.

Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. 2012. Fsl. *Neuroimage* **62**:782–790.

Jenkinson M, Smith S. 2001. A global optimisation method for robust affine registration of brain images. *Med Image Anal* **5**:143–156.

Jimura K, Poldrack RA. 2012. Analyses of regional-average activation and multivoxel pattern information

tell complementary stories. *Neuropsychologia* **50**:544–552.
Karpicke JD, Blunt JR. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science (80-)* **331**:772–775.
Karpicke JD, Roediger HL. 2008. The critical importance of retrieval for learning. *Science (80-)* **319**:966–968.
Karpicke JD, Roediger III HL. 2007. Repeated retrieval during learning is the key to long-term retention. *J Mem Lang* **57**:151–162.
Khan AR, Wang L, Beg MF. 2008. FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping. *Neuroimage* **41**:735–746.
Kok P, Jehee JFM, De Lange FP. 2012. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* **75**:265–270.
Kosslyn SM, Thompson WL, Alpert NM. 1997. Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *Neuroimage* **6**:320–334.
Kuhl BA, Shah AT, DuBrow S, Wagner AD. 2010. Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nat Neurosci* **13**:501.
Lang PJ, Bradley MM, Cuthbert BN. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent Study Emot Atten* **1**:39–58.
Lashley KS. 1950. In search of the engram. *Physiol Mech* 454.
Lee S-H, Kravitz DJ, Baker CI. 2018. Differential Representations of Perceived and Retrieved Visual Information in Hippocampus and Cortex. *Cereb Cortex* 1–10. doi:10.1093/cercor/bhy325
Levy Benjamin J, Anderson MC. 2012. Purging of memories from conscious awareness tracked in the human brain. *J Neurosci* **32**:16785–94. doi:10.1523/JNEUROSCI.2640-12.2012
Levy B. J., Anderson MC. 2012. Purging of Memories from Conscious Awareness Tracked in the Human Brain. *J Neurosci* **32**:16785–16794. doi:10.1523/JNEUROSCI.2640-12.2012
Liu Y, Lin W, Liu C, Luo Y, Wu J, Bayley PJ, Qin S. 2016. Memory consolidation reconfigures neural pathways involved in the suppression of emotional memories. *Nat Commun* **7**:13375.
Müller GE, Pilzecker A. 1900. Experimentelle beiträge zur lehre vom gedächtniss. JA Barth.
O'Craven KM, Kanwisher N. 2000. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J Cogn Neurosci* **12**:1013–1023.
Polyn SM, Natu VS, Cohen JD, Norman KA. 2005. Category-specific cortical activity precedes retrieval during memory search. *Science (80-)* **310**:1963–1966.
Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. 2015. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage* **112**:267–277.
Rademaker RL, Chunharas C, Serences JT. 2019. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat Neurosci* **22**:1336–1344. doi:10.1038/s41593-019-0428-x
Roelofs J, van Breukelen G, de Graaf LE, Beck AT, Arntz A, Huibers MJH. 2013. Norms for the Beck Depression Inventory (BDI-II) in a large Dutch community sample. *J Psychopathol Behav Assess* **35**:93–98.
Shohamy D, Wagner AD. 2008. Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron* **60**:378–389. doi:10.1016/j.neuron.2008.09.023
Smith AM, Floerke VA, Thomas AK. 2016. Retrieval practice protects memory against acute stress. *Science (80-)* **354**:1046–1048.
Smith SM. 2002. Fast robust automated brain extraction. *Hum Brain Mapp* **17**:143–155.
Spaniol J, Davidson PSR, Kim ASN, Han H, Moscovitch M, Grady CL. 2009. Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia* **47**:1765–1779.
Summerfield C, De Lange FP. 2014. Expectation in perceptual decision making: neural and computational mechanisms. *Nat Rev Neurosci* **15**:745.
Summerfield C, Trittschuh EH, Monti JM, Mesulam M-M, Egner T. 2008. Neural repetition suppression reflects fulfilled perceptual expectations. *Nat Neurosci* **11**:1004.
Takashima A, Nieuwenhuis ILC, Jensen O, Talamini LM, Rijpkema M, Fernández G. 2009. Shift from

hippocampal to neocortical centered retrieval network with consolidation. *J Neurosci* **29**:10087–10093.

Takashima A, Petersson KM, Rutters F, Tendolkar I, Jensen O, Zwarts MJ, McNaughton BL, Fernandez G. 2006. Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proc Natl Acad Sci* **103**:756–761.

van der Bij AK, de Weerd S, Cikot RJLM, Steegers EAP, Braspenning JCC. 2003. Validation of the dutch short form of the state scale of the Spielberger State-Trait Anxiety Inventory: considerations for usage in screening outcomes. *Public Health Genomics* **6**:84–87.

van Schie K, Anderson MC. 2017. Successfully controlling intrusive memories is harder when control must be sustained. *Memory* **25**:1201–1216.

Wheeler ME, Petersen SE, Buckner RL. 2000. Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proc Natl Acad Sci* **97**:11125–11129.

Wimber M, Alink A, Charest I, Kriegeskorte N, Anderson MC. 2015. Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nat Neurosci* **18**:582–589. doi:10.1038/nn.3973

Woolrich MW, Ripley BD, Brady M, Smith SM. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* **14**:1370–1386.

Xue G. 2018. The neural representations underlying human episodic memory. *Trends Cogn Sci* **22**:544–561.

Acknowledgements

We thank Joyce van Arendonk, Joost Verchuren for the assistance of data acquisition and memory performance evaluation, Sjoerd Meijer for the memory performance evaluation, and Yingjie Shi for commenting on previous versions of the manuscript. This work was supported by a PhD fellowship from the Chinese Scholarship Council (201606990020) awarded to W.L.

Author contributions

WL, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; NK, Conception and design, Analysis and interpretation of data, Drafting or revising the article; GF, Conception and design, Analysis and interpretation of data, Drafting or revising the article.

Competing interests

The authors declare no competing interests.