Decoding visual working memory in a dynamic context

By Dasja de Leeuw

Artificial Intelligence Master's thesis

Utrecht University

Supervised by Dr. Surya Gayet PhD

25-02-2024

Utrecht University

Abstract

Visual working memory (VWM) is the brain's mechanism for briefly retaining visual information for imminent goal-directed behavior. Existing research has pointed to different brain areas for VWM storage, prompting a discussion. We propose that these conflicting results arise from the static approach with which VWM maintenance has previously been examined. VWM is used to guide behavior in an inherently dynamic visual environment and might therefore be best considered as a dynamic system itself. For instance, task-related goals can change from a moment-to-moment basis, temporarily making some visual stimuli more task-relevant than others. Moreover, brain activity is known to be inherently dynamic, posing the question of whether VWM representations are maintained in a dynamic or stable manner. To address the existing conflict, we have investigated VWM maintenance in two ways: (1) by examining the maintenance of VWM representations over time to adhere to the brain's inherent dynamics and (2) by exploring the differences between VWM representations with different task-relevance states to take the dynamic visual environment into account. Relevancy states were initiated by task-related goals where items could either be currently task-relevant or prospectively task-relevant.

Using a 7-Tesla fMRI dataset and inverted encoding models we have decoded VWM information throughout the human cortex (N=3). We have found that VWM representations are stored in occipital and parietal areas in the cortex, confirming earlier findings. Within these areas, VWM information is maintained in a stable representational format over the course of the 8 seconds retention period. Furthermore, differences in storage location between currently- and prospectively-relevant representations were revealed: currently-relevant representations were present in occipital and parietal areas, while prospectively-relevant representations were present in frontal and parietal areas. Additionally, we found that currently- and prospectively-relevant representations could be stored in similar and even in opposite representational formats in the parietal lobe. These findings show that VWM representations and their storage location can differ between relevancy states, emphasizing the importance of taking dynamic changes into account.

1. Introduction

As humans we heavily rely on visual information to meaningfully interact with our environment. However, as the world around us is constantly changing, visual information can rapidly change or entirely vanish. To avoid the loss of relevant visual information immediately after direct visual input has ceased, our brain uses visual working memory (VWM) (Baddeley, 1998). The cognitive system VWM is able to temporarily maintain visual information in the absence of sensory input to facilitate imminent goal-directed behavior. For instance, imagine you are looking for a friend in a crowded room. As you are scanning the environment an active mental image of your friend's face needs to be maintained. This active maintenance allows the eventual discovery of a found match in the environment, which completes the task of finding your friend. By keeping an active representation in VWM, incoming visual input can be directly influenced by VWM content, thus allowing goal-directed behavior (Gayet et al., 2017). By maintaining crucial visual information over short periods of time, VWM enables the completion of a plethora of tasks in an ever-changing environment.

In the large body of VWM research there exists a lot of debate on which brain areas VWM representations are stored in during maintenance. Based on decoding results, possible storage locations include frontal, parietal and sensory regions (Polanía et al., 2012; Cavanagh et al., 2018; Christophel et al., 2018; Rademaker et al., 2019, Yu et al., 2020). Most studies that attempt to locate VWM storage have been analyzing VWM maintenance as a static system. This is done by averaging retention-related data over the entire retention period (i.e. the onset of memorization until the recall task) and/or assuming all memorized items are stored in the same brain area(s) and represented by the same neural activity pattern.

However, VWM should be considered as a dynamic system to take into account its interaction with the dynamic environment. As the world around us is constantly changing, our goals are constantly changing, making certain memory items more relevant to the task at hand than others. This raises the question of whether VWM representations are always stored in an identical location and representational format in the brain. It could be that VWM storage operates differently for memorized items based on their task-relevance. Additionally, brain activity is known to be inherently dynamic, prompting the question of whether VWM representations are maintained in a stable manner during the entirety of retention (Schneegans et al., 2018; Rule et al., 2019). It could be that in certain brain areas the format in which VWM representations are stored changes over the course of VWM maintenance.

In the present study the focus on the dynamics of VWM maintenance is two-fold. Firstly, we explore the differences in the maintenance of VWM representations with different states of task-relevance. Certain memory items can be made more or less task-relevant when introducing top-down goals. Secondly, we examine the dynamic nature that is inherent to VWM maintenance by investigating how memory representations can change over time during retention. By taking the dynamic nature of VWM into account we aim to contribute to the vast body of VWM research and

possibly provide new insights into the storage locations and representational formats of VWM content.

1.1 Visual working memory maintenance over time

Traditional models propose VWM representations are initiated during perception of the stimulus and maintained through sustained firing of neurons in the prefrontal cortex (Fuster et al., 1971; Funahashi et al., 1989; Miller et al., 1996; Wang, 2001). This was concluded from univariate analyses showing elevated levels of activity in the frontal lobe during retention, while no sustained activity was found in sensory locations such as the visual cortex (Offen et al., 2009). Due to the absence of persistent activity in visual areas, these locations were thought to be uninvolved in VWM maintenance (Super, 2003). More recently, advanced analysis methods such as decoding have been introduced. A decoding model attempts to predict the retained stimulus from neural activation patterns present in the observed brain activity. Unlike univariate analyses, decoding methods adhere to the dynamic nature of brain activity by taking advantage of the multivariate patterns representing stored information. If stimulus predictions made by the decoding model are significantly above chance, it serves as an indication that stimulus-specific information is being stored in the selected location. Locating the storage of memorized representations can not be inferred from elevated univariate responses, thus decoding can offer a unique insight into the storage locations and formats of VWM representations.

Decoding studies have been used to test the sensory recruitment hypothesis, which states that VWM representations are stored in the same neural code as the sensory representations associated with initial stimulus encoding (Serences et al., 2009; Albers et al., 2013; Bettencourt et al., 2016; Rademaker et al., 2019). Storing VWM content in a high-resolution, sensory-like representation would allow straightforward comparison with incoming visual information. This hypothesis challenges the traditional view of VWM maintenance as sustained prefrontal activity and instead proposes VWM representations are stored in sensory-like format in early visual areas. To investigate the role of the visual cortex in VWM maintenance, Harrison and Tong (2009) attempted to decode which of two oriented gratings was retained in memory based on the observed activity patterns in the visual cortex. It was found that visual cortex activity patterns could decode the correct oriented grating during retention with mean accuracy levels reaching 80%. Decoding was even possible when visual cortex activity fell to baseline-activity levels. These results show that the absence of persistent activity in the visual cortex does not mean VWM content is not maintained in this brain area. The finding of successful decoding of remembered oriented gratings throughout the retention period has been replicated many times and found in various occipital, parietal and frontal areas such as the early visual cortex (EVC), intraparietal sulcus (IPS) en frontal eye fields (FEF) (Christophel et al., 2018; Rademaker et al., 2019, Yu et al., 2020).

The presence of VWM information in each individual time point during retention does not reveal whether information is kept at a stable representational format or whether any transformations

occur over the course of retention. Transformations in this context refer to any changes in the neural activation patterns encoding VWM information. A range of studies have found varying results in regard to these possible transformations. Wolff et al. (2020) found stable formats for VWM representations between early and late retention. Alternatively, Oh et al. (2019) found a difference in temporal stability between locations in the brain in an EEG study. In fronto-central areas they found stable formats while in occipito-parietal areas they found dynamic formats, indicating transformations are occurring. In support of this, Yu et al. (2020) also found an instability of representational format in EVC and in IPS, located both in occipito-parietal areas. These results suggest that dynamic changes in the neural code are occurring in occipito-parietal areas throughout VWM retention. Conversely, it has also been found that unstable, dynamic shifts in representational format are only occurring during- and shortly after initial encoding of the stimulus (Stokes et al., 2013; Wolff et al., 2017; Liu et al., 2020). In these studies VWM representations could still be decoded during retention and were found to be kept in a stable format, but they no longer mirrored the representation during initial encoding. This means that transformations are in fact happening during- and shortly after initial encoding, and VWM maintenance does not just rely on a sustained activation of the initial stimulus representation.

Taking all these results together, it seems that VWM maintenance does not solely rely on sustained frontal lobe activity. Decoding studies have helped to reveal that VWM content can be maintained in occipital, parietal and frontal locations, when training and testing on individual timesteps during the retention period. Additionally, with decoding it can be examined whether VWM representations are stored in a stable or dynamic representational format. However, the degree to which representational transformations are occurring is a topic of debate; there is no consensus about the existence and locations of any dynamic transformations of VWM representations.

1.2 Task-relevance in visual working memory

Another way to adhere to the dynamic nature of VWM maintenance is by examining VWM in the context of task-relevance status. In real-life settings, we often have to maintain multiple items in VWM. Moreover, task-related demands can change on a moment-to-moment basis when interacting with a dynamic environment. This means certain memory items can temporarily become more task-relevant than other memory items that are concurrently maintained in VWM. Memory items could be stored in a currently-task relevant state when its representation needs to be used in the current task. Alternatively, a memory item could be stored in a prospectively-relevant state if its representation is not used for the current task, but for an upcoming task in the near future. Important to note is that VWM is a limited system; items compete with each other for resources (Kertzel et al., 2019). Memory items held in different states of task-relevance could influence how these limited resources are divided among the memory items, and thus how and where their respective representations are maintained in VWM.

Research has shown that if top-down goals are initiated by cuing an item to be more task-relevant, a larger amount of memory resources will be allocated to it (Fallon et al., 2016). As a result, other items held in memory will have a smaller amount of resources allocated to them and thus have a weaker memory trace. This is reflected by the finding that decoding strength is higher for currently-relevant representations than for prospectively-relevant representations during retention (Lewis-Peacock et al., 2012; Rose et al., 2016; LaRocque et al., 2017; Sahan et al., 2020). Differences between states of task-relevance is also reflected in the difference in storage location during retention. Christophel et al. (2018) found both the currently- and prospectively-relevant representations of oriented gratings to be present in parietal areas such as IPS and FEF, but only the currently-relevant representation in occipital areas such as EVC. Similarly, Ruijs (yet unpublished results) found the currently-relevant representation to be overrepresented in the occipital lobe compared to the prospectively-relevant representation, whereas parietal regions show no difference in task-relevance.

More recently however, it was found that the prospectively-relevant representation of a memorized orientation can indeed be decoded in EVC when training a decoding model on the robust currently-relevant representation (Yu et al., 2020; Iamshchinina et al., 2021). Iamshchinina et al. (2021) did a reanalysis of the data by Christophel et al. (2018) and could decode the prospectively-relevant representation when training on the currently-relevant representation. In one case training and testing solely on prospectively-relevant representations yielded successful decoding results in the occipital lobe when using highly sensitive 7-Tesla fMRI data (Ruijs, yet unpublished results). These findings demonstrate that the decoding results of VWM representations can be highly dependent on the sensitivity of the analysis method and could be missed entirely by methods that lack sensitivity. Additionally, the results show that currently- and prospectively-relevant representations of orientations can be stored in similar representational formats. If the prospectively-relevant representation can be successfully decoded after training on the currently-relevant representation, it means they are represented by similar neural activation patterns.

Studies using more sensitive analysis methods have also shown that prospectively-relevant representations in VWM can undergo significant transformations. Specifically, it was found that prospectively-relevant items can be represented by neural patterns opposite from currently-relevant items within the same stimulus category (Van Loon et al., 2018; Wan et al., 2020; Yu et al., 2020; Wan et al., 2022). This unexpected finding has been concluded from observing below-chance performance when training on the currently-relevant representation and testing on the prospectively-relevant representation. The possibility that the observed below-chance performance is the result of a blood-oxygen-level-dependent (BOLD)-related undershoot has been ruled out (Van Loon et al., 2018). In the context of oriented gratings, below-chance performance means representations are stored orthogonally to each other (Wan et al., 2020; Yu et al., 2020; Wan et al., 2022). However, opposite representational formats have also been found for memorized objects in the object-selective cortex (Van Loon et al., 2018). It is thought that maintaining the prospectively-relevant representation in an

opposite representational space will protect the representation from distraction while actively maintaining the currently-relevant representation in anticipation of the recall task (Van Loon et al., 2018; Wan et al., 2020; Yu et al., 2020). Transformations to opposite representational space were found for objects in the object-selective cortex, for oriented gratings in EVC and for stimulus locations in IPS (Van Loon et al., 2018; Yu et al., 2020).

Based on the described earlier work, VWM representations with different task-relevance states may be maintained in different locations and in dissimilar representational formats. Earlier studies have also demonstrated that results depend strongly on the sensitivity of the chosen analysis method.

1.3 The current study

In the current study we aim to examine the dynamic nature of VWM maintenance in two ways. Firstly, we study the inherent dynamics of VWM maintenance by examining whether VWM representations are maintained in a stable or dynamic format over the course of retention. Secondly, we study VWM maintenance in a dynamic context by introducing top-down goals and examining the differences in VWM maintenance between currently- and prospectively-relevant memory items. Previous research has shown conflicting results for both of these fields. Decoding studies have shown that by making use of the multivariate nature of brain activity, new discoveries about the location and format of VWM representations can be made. Additionally, studies by Yu et al. (2020) and Iamshchinina et al. (2021) have revealed the chosen analysis method can have a large impact on results and might explain why other studies have found null results. Weaker VWM representations such as prospectively-relevant representations can be missed entirely by methods solely aimed at capturing the robust currently-relevant representation, but can alternatively be picked up by highly sensitive methods.

In an attempt to advance upon previous research and possibly clear up conflicting results we use an ultra-high-field 7-Tesla fMRI dataset (Ruijs, yet unpublished results). When locating VWM representations we employ an assumption-free approach and do not predefine regions of interest. Additionally, we use inverted encoding models to decode memory representations, which are able to provide highly detailed stimulus reconstructions. The existing dataset that is used has implemented a two-state delayed match-to-sample trial setup with orientated gratings as stimuli. Two oriented gratings were successively presented, after which a retro-cue indicated which of the two oriented gratings participants needed to reproduce in an upcoming behavioral task. After a retention period the behavioral task was initiated, in which the orientation of the cued oriented grating needed to be reproduced. Next, a second retro-cue again cues one of the two oriented gratings, followed by another retention period and behavioral task. During the entirety of the first retention period, participants need to concurrently maintain two VWM representations: the cued orientation in a currently-relevant state

and the uncued orientation in a prospectively-relevant state. Therefore, this experimental design allows the manipulation of task-relevance for maintained VWM representations.

1.4 Value of approach

The bulk of VWM research that uses fMRI data has used 3-Tesla fMRI. The usage of ultra-high-field 7-Tesla fMRI data could offer a more sensitive, high-resolution approach to studying VWM in a dynamic context. Compared to 3T fMRI, 7T fMRI can capture changes in the BOLD-signal in a higher spatial resolution and has a higher signal-to-noise ratio (Torrisi et al., 2018; Willems et al., 2021). Therefore, using 7T fMRI data can improve the stimulus reconstruction of a relatively weak neural signal in a decoding context. This is particularly of interest to the current study in which we aim to decode prospectively-relevant VWM content and possibly weaker neural representations over time. With the goal of uncovering weaker memory representations it is beneficial to conduct many trials per participant (Gordon et al., 2017). Therefore, we have opted to use a 7T fMRI dataset (Ruijs, yet unpublished results) with three participants, each having participated in 324 trials. Due to the small sample size, we analyze each participant separately instead of conducting group-level analyses.

Another way to advance upon previous research is by employing an agnostic approach when attempting to locate VWM information in the cortex. Almost all aforementioned fMRI studies have used predefined regions of interest (ROIs) to decode VWM representations. These ROIs are a selection of voxels meant to represent a brain area of interest. Commonly used ROIs in VWM research include the EVC and IPS (Christophel et al., 2018; Rademaker et al., 2019; Yu et al., 2020). ROIs can be defined by anatomical criteria (e.g. using a brain atlas) or functional criteria (e.g. using retinotopic mapping) (Haynes, 2015). Using functional criteria, regions are selected that show specific patterns of activation in response to specific tasks or stimuli. In the current study we refrain from predefining ROIs for a number of reasons. Firstly, as previously mentioned the storage locations of VWM representations are debated. This poses a problem for the pre-selection of ROIs because areas that are wrongly assumed to be uninvolved with VWM maintenance can be discarded from analysis. Because we are using a high-resolution 7T fMRI dataset, it is possible we find unexpected results that have not been previously reported on, or for which very little support exists in earlier work. Secondly, activity within an anatomically defined ROI is not necessarily functionally heterogeneous (Korhonen et al., 2017). Consequently, a predefined ROI could contain a large number of voxels providing irrelevant activity. This could lower the chance of successfully decoding memory information and could lead to the incorrect conclusion that VWM content is not present in the selected ROI. Thirdly, defining a ROI based on functionality within the current dataset could introduce the risk of double-dipping (Kriegeskorte et al., 2009). The term double-dipping here refers to the phenomenon of using the same data for selection and selective testing. For example, for the current study we could compare brain activity between trials where participants successfully maintained visual information and trials where participants failed to do so, based on the behavioral error. This analysis could reveal

certain brain areas that are associated with successful VWM maintenance, which can be selected as ROIs to apply further VWM-related analyses on. Selecting voxels based on their functionality and testing the selected voxels for a similar functionality increases the risk of artificially biasing the results by overfitting on the data. This could result in invalid statistical testing and an overestimation of significant results. Thus, instead of limiting our decoding models to predefined ROIs, we will attempt to decode VWM content throughout the entire cortex to ensure no unexpected results are missed and the mentioned pitfalls are mitigated.

To decode VWM information from fMRI voxel responses inverted encoding models (IEMs) are used. These decoding models have been previously used in the context of decoding VWM representations from fMRI- and EEG data (Oh, et al., 2019; Rademaker et al., 2019; Wan et al., 2020; Yu et al., 2020). In the context of fMRI, IEMs are trained to encode voxel responses from presented stimuli. After training, the resulting weights matrix can be inverted to decode the presented stimulus from unseen voxel responses. The stimulus reconstruction of an IEM is built from a predefined number of so-called channel responses. Because of the multiple channel responses, the stimulus reconstruction mirrors a tuning function in feature space, showing higher responses for similar stimulus features and lower responses for dissimilar stimulus features (Scotti et al., 2021). Multiple channel responses also allow IEMs to decode information on a continuous scale, allowing more detailed stimulus reconstructions than those of binary classification models such as support vector machines. A continuous response also allows comparison with the actual stimulus or with behavioral responses. Another advantage of the continuous output scale is that transformations to opposite neural formats can be easily interpreted from IEM output. Because the orientation of oriented gratings is used as the memory item, an opposite format would be revealed by a 90° shift in the IEM reconstruction. In the context of orientations, this would indicate a transformation to an orthogonal representational format. This transformation could be present for the prospectively-relevant memory item, which makes an IEM a useful decoding model for this study (Yu et al., 2020).

1.5 Research questions and expectations

In the following section the research questions and their respective expectations based on earlier research will be listed. A table overview can be found in Figure 1.

1.5.1.a Where in the cortex is currently-relevant VWM content stored?

We can expect to find currently-relevant VWM representations in frontal, parietal and occipital regions. This is based on successful decoding in the EVC, IPS, FEF and the prefrontal cortex (PFC) (Polanía et al., 2012; Cavanagh et al., 2018; Christophel et al., 2018; Rademaker et al., 2019, Yu et al., 2020).

1.5.1.b Where in the cortex is prospectively-relevant VWM content stored?

The storage location of prospectively-relevant VWM representations is less agreed on. Most studies report a presence in IPS and FEF, while studies employing more sensitive methods point to the EVC (Christophel et al., 2018; Yu et al., 2020; Iamshchinina et al., 2021). It is also proposed that prospectively-relevant representations are stored in a 'hidden' activity-silent format, which means information is maintained as a pattern of synaptic weights (Stokes, 2015; Christophel et al., 2017; Wolff et al., 2017). When information is represented by an activity-silent format, it can not be decoded. Nonetheless, we expect to decode prospectively-relevant VWM content in frontal and parietal regions and perhaps even in occipital regions, as we are using highly sensitive methods.

1.5.2.a Where in the cortex is currently- and prospectively-relevant VWM content stored in similar representational formats?

Similar representational formats have previously been found in EVC and in IPS (Van Loon et al., 2018; Yu et al., 2020; Iamshchinina et al., 2021). Therefore, we can expect to find similar formats for currently- and prospectively-relevant VWM representations in the occipital and parietal lobe.

1.5.2.b Where in the cortex is currently- and prospectively-relevant VWM content stored in orthogonal representational formats?

In addition to examining whether currently- and prospectively-relevant VWM representations are stored in similar formats, we will examine whether they are stored in formats orthogonal to each other. Earlier studies finding orthogonal storage for prospectively-relevant representations compared to currently-relevant representations have pointed to the EVC (Yu et al., 2020). Thus, we can expect to find orthogonal representational formats in the occipital lobe.

1.5.3.a Where in the cortex is VWM content maintained in stable representational formats?

Stable VWM representations could be found in various locations throughout the cortex. Multiple studies have shown stable representations during the retention period, after initial encoding of the cue (Stokes et al., 2013; Wolff et al., 2017; Liu et al., 2020). Oh et al. (2019) located stable representations in fronto-central areas. However, these results are all based on EEG data or found in monkey PFCs, which makes it difficult to anticipate the locations of stable representations. Based on these earlier findings, it is possible we find some stable representations, though it is not entirely clear in what locations.

1.5.3.b Where in the cortex is VWM content maintained in changing representational formats?

Based on unstable representational formats in EVC and IPS, we can expect to find changing representational formats in occipito-parietal areas (Oh et al., 2019; Yu et al., 2020).

Research question		**Analysis**	**Expectations**	**Results**
1. Where in the cortex is VWM content stored?	a. Where in the cortex is currently-relevant VWM content stored?	CMI → CMI	Occipital, parietal and frontal lobe	Occipital and parietal lobe
	b. Where in the cortex is prospectively -relevant VWM content stored?	UMI → UMI	Occipital, parietal and frontal lobe	Parietal and frontal lobe
2. Where in the cortex are currently- and prospectively-relevant VWM content stored in similar representational formats?	a. Where in the cortex is currently- and prospectively -relevant VWM content stored in similar representational formats?	CMI → UMI, UMI → CMI	Occipital and parietal lobe	Right-parietal lobe
	b. Where in the cortex is currently- and prospectively-relevant VWM content stored in orthogonal representational formats?	CMI → orthogonal UMI, UMI → orthogonal CMI	Occipital lobe	Left-parietal lobe
3. Where in the cortex is currently-relevant VWM content maintained in stable vs. changing representational formats?	a. Where in the cortex is VWM content maintained in stable representational formats?	(early → late + late → early)/2	Frontal lobe	Occipital and parietal lobe
	b. Where in the cortex is VWM content maintained in changing representational formats?	(early → early + late → late)/2 - (early → late + late → early)/2	Occipital and parietal lobe	-

Figure 1. *Overview of research questions, expectations, analyses and results.*

2. Methods

2.1 Dataset description

The dataset was collected by Ruijs (yet unpublished results) as part of a Utrecht University Master's thesis (Neuroscience and Cognition Master's program). The dataset includes 7T fMRI data as well as behavioral data (N=3). For the current study only the 7T fMRI data from the experiment will be used for analysis.

2.1.1 Participants

Three right-handed participants were recruited (one female, two male), aged 23-40 years. Each participant had prior experience with being in an fMRI scanner. One participant received monetary compensation for their participation. All fMRI scanning sessions were performed in the Spinoza Centre for Neuroimaging in Amsterdam.

2.1.2 Procedure

The experiment consists of 27 runs of 12 trials each, conducted across several sessions (and across several days). For each trial, the participant is expected to perform two successive delayed match-to-sample tasks. The onset of a trial is initiated by a fixation screen that is shown for 1000ms. Next, the first stimulus is presented (800ms), followed by a fixation screen (800ms), followed by the second stimulus (800ms). The presented stimuli are both donut-shaped Gabor gratings with a faded border. The orientations of the two Gabor gratings are always at least 6° apart. After the second grating, a donut-shaped mask with a retro-cue (1 or 2) in the center is shown (800ms), cueing one of the two previously shown gratings. Participants are required to reproduce the orientation of the cued grating in a future behavioral task. After the retro-cue a retention period with a fixation screen follows (8000ms). Next, the behavioral task is initiated by the appearance of a randomly orientated line. The participant is required to match the orientation of the line with the orientation of the cued Gabor grating, using a rotary dial. The participant has 3000ms to complete this task. Then a second retro-cue is shown, again cueing one of the previously shown orientations with a probability of 50% for each orientation (800ms). The retro-cue is again followed by a retention period (8000ms) and the behavioral task (3000ms), after which the trial ends. An overview of the sequence of events during a trial can be found in Figure 2.

Because there are two different orientations and two retro-cues in each trial, there are four possible trial conditions:

1. Orientation 1 is cued by both retro-cues (repeat-trial).
2. Orientation 1 is cued by the first retro-cue and orientation 2 is cued by the second retro-cue (switch-trial).

3. Orientation 2 is cued by both retro-cues (repeat-trial).
4. Orientation 2 is cued by the first retro-cue and orientation 1 is cued by the second retro-cue (switch-trial).

During the first retention period, the cued memory item (CMI) will likely be stored in a currently-relevant state in anticipation for the behavioral response. Simultaneously, the uncued memory item (UMI) can not be dropped from memory entirely, given the possibility that this orientation might be cued by the second retro-cue in switch-trials. Therefore, the UMI will likely be stored in a prospectively-relevant state during the first retention period. This experimental design allows us to examine both the CMI and UMI during the first retention period.

To test whether the experimental setup works and the UMI is actually retained in VWM, we examined the behavioral response in the second retention period in switch-trials for all three participants separately. Note that during the second retention period in switch-trials, the orientation cued by the second retro-cue has been maintained in a UMI status during the first retention period. We tested the mean absolute error of the behavioral response in the second retention period in switch-trials against a null distribution, defined by 10,000 permutations. To create a single permutation, the orientation labels of all trials were shuffled randomly. For each trial the absolute error was computed by subtracting the observed behavioral response from the randomly shuffled orientation. This resulted in a null distribution of 10,000 mean absolute error values. The observed mean absolute error was compared against this null distribution. For all three participants we found the observed mean absolute error during the second delay in switch-trials is significantly lower than the null distribution, i.e. lower than chance level (p<.001). This shows that participants could successfully recall the UMI during the second retention period, meaning the UMI is also maintained during the first retention period. With this experimental design we can therefore adequately analyze VWM content during the first retention period in different states of task-relevance: currently task-relevant or prospectively task-relevant.

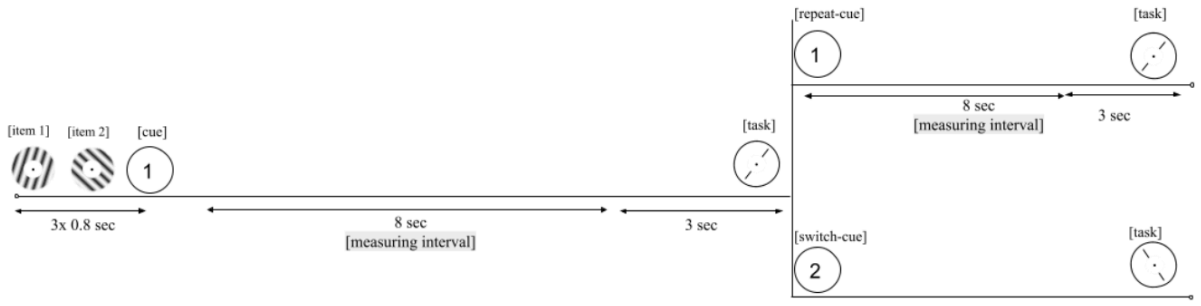

Figure 2. *Overview of the trial setup, from Ruijs (yet unpublished results).*

2.1.3 Stimuli

All stimuli and tasks were created and presented using MATLAB and Psychtoolbox3 (version 3.0.17). Stimuli were displayed on a 698x393 mm LCD- screen (BOLDscreen 32, Cambridge Research Systems, Rochester, UK) with a resolution of 1920x1080 pixels, RGB color, a contrast ratio of 1400:1 and a frame rate of 120 Hz. The screen was positioned 1850 mm from the participant's eyes and was visible through a mirror placed on the participant's head coil.

All stimuli are specified in dva (degrees visual angle) and converted into pixels based on pixel size and distance from the screen. Stimuli were projected onto a gray background (rgb=3x128). A bullet point fixation dot was continuously visible in the center of the screen throughout the experiment (black outer circle (rgb=3x0, dva=0.2) and white inner circle (rgb=3x255, dva=0.1)). The fixation dot was only absent during presentation of a numerical retro-cue (font=Calibri, font-size=60 pixels). The sinusoidal Gabors gratings were created using a 0.6 grating contrast, a target size of 1.2 dva (10 cycles per dva). The whole Gabor grating was enlarged to 5 dva. The masks that are presented during the retro-cues were created by 100x100 pixel brown noise, also enlarged to 5 dva. Gabor gratings and masks were presented in a donut shape centered around the fixation dot (dva=5-by-5) with a cosine blur around all edges of 0.5 dva.

The previously mentioned four trial conditions are balanced to be equally present in each run, meaning each condition was presented (12/4 =) three times each run. Stimulus orientations are computed by selecting (3*27=) 81 equally spaced orientations from the 180° stimulus space. These stimulus orientations are counterbalanced with the four trial conditions, as well as the two possible presentation timings (the first or second orientation shown in the sequence) across the whole experiment. The two stimulus orientations in each trial are constrained to be at least 6° apart, to ensure they are perceived as two different orientations by human viewers.

2.2 fMRI data acquisition and -processing

2.1.1 fMRI data acquisition

fMRI data are acquired by a 7T Philips Achieve scanner. T2-weighted functional images were taken using a 32-channel head coil with a resolution of 1.688x1.688x1.7 mm with 57 slices, each containing 128x128 voxels, which are acquired for each repetition time (TR). A TR of 1500ms is used, and a time to echo (TE) of 22.49ms. A flip angle of 70° is applied. A gradient echo sequence with a SENSE acceleration factor of 2.4, multiband factor 3 and anterior-posterior encoding is used. A second order B0 based shim of the functional scan's field of view is used. This captures most of the brain except the cerebellum and parts of the anterior temporal lobes. Each run contains 244 TRs (where a single TR lasts 1500ms) resulting in a total scan time of 366 seconds for each run. At the start of each run, a delay of 16 seconds is included before the onset of the first trial with the goal of returning the BOLD-signal to baseline. After the end of the last trial a delay of 8 seconds is added to capture the remaining BOLD-signal from the last trial. In addition, for each session two supplementary functional

scans are acquired, which are recorded using the opposite phase-encoding direction from the functional runs to correct for image distortion. A high resolution T1-weighted scan is acquired for two participants (the third participant's scan was already available). This scan is automatically segmented using Freesurfer, after which segments were manually edited to minimize errors using ITK-SNAP. The resulting gray-white segmentations can be used to construct a gray-matter mask. Several anatomical T1-weighted scans in the same resolution as the functional scans are also collected, which aid in aligning each session's functional data to each participant's anatomical scans. The described fMRI acquisition procedure was also implemented by Van Ackooij et al. (yet unpublished results).

2.2.2 Preprocessing fMRI data

Several preprocessing steps have been performed to prepare the 7T fMRI data for analysis (Ruijs, yet unpublished results). The following procedure has also been implemented by Van Ackooij et al. (yet unpublished results). The preprocessing steps include motion correction and determining (1) the distortion transformation, (2) the transformation in brain position between and within functional scans and (3) the transformation co-registering the functional data to the same space T1. The product of these transformations is a single transformation matrix used to transform the functional data to the high-resolution T1 anatomy. No further spatial or temporal smoothing has been applied.

2.2.3 fMRI data selection

To decode VWM representations from retention-related activity, the correct TRs need to be selected from the data. For each trial, the onset of the retention period is identified as the exact moment in time when the retro-cue disappears from the screen. At this timestamp, 4.5 seconds is added to account for the hemodynamic response associated with the BOLD-signal, thus selecting the maximum amount of retention-related activity. The first TR that follows this timestamp is chosen as the first TR containing retention-related activity. The five subsequent TRs are additionally selected as containing retention-related activity, resulting in six TRs total per trial. The six TRs together span a timeframe of 7.5 seconds and partly overlap with the retention period of 8 seconds (with 4.5 seconds added). The amount of overlap depends on the amount of time between the onset of the retention period and the first TR that follows this timestamp. For each trial, the six selected TRs are averaged to obtain the average retention-related activity per trial. To test cross-temporal decoding, TRs were averaged in pairs, i.e. 1-2 (early) and 5-6 (late). This resulted in two retention parts per trial. We chose to omit TRs 3 and 4 from cross-temporal analyses to ensure the early and late retention parts have minimal overlap in the BOLD-signal.

For the behavioral task during the experiment, participants were required to reproduce the cued orientation using a rotary dial. Trials where the error in degrees for the behavioral task was too large are omitted from analysis. This is because we assume that for these trials, participants were retaining an orientation too disparate from the cued orientation and therefore decoding VWM content

for these trials would not be informative. Outliers in error were identified using the error distribution for each participant separately. If the absolute error of a trial was larger than 3 standard deviations from the mean, this trial was omitted from analysis. Ultimately, participant 1 contributed 316 trials (324-8 = 316), participant 2 contributed 313 trials (324-11 = 313) and participant 3 contributed 320 trials (324-4 = 320).

All fMRI data were z-scored to ensure voxel responses between experimental runs are within the same range and each voxel contributes equally to the overall signal. Z-scoring is performed over time, for each voxel individually. Each voxel at each timepoint is z-scored by subtracting the mean voxel response per run and dividing by the standard deviation of the voxel response per session. This ensures that for each voxel over time the mean is 0 and the standard deviation is 1. The standard deviation per session was used instead of the standard deviation per run, because we observed large differences in voxel responses between sessions but not between runs within the same session. Using the standard deviation per session instead of per run ensures a more precise estimation of the variation of voxel responses.

2.3 Decoding VWM content

In order to answer the research questions we aimed to decode VWM representations from the selected 7T fMRI voxel responses. In other words, we construct a model that attempts to predict the orientation of the retained stimulus from the activation patterns present in the voxel responses. To decode VWM representations we used inverted encoding models (IEMs) within a searchlight analysis. We used a leave-one-run-out cross-validation approach where we iteratively take each run as a test set, and use the remaining data as a training set. This was chosen to ensure each trial could be decoded, while avoiding training and testing on the same data. All analyses were performed in MATLAB and using the toolbox *CoSMoMVPA*.

For our analyses we opted for an assumption-free approach by using a searchlight analysis instead of predefining ROIs (Etzel et al. 2013). The motivation behind this is because storage locations of VWM representations are debated in existing work. To execute the searchlight analyses, the function *cosmo_searchlight* was used. In a searchlight analysis, the chosen decoding model is applied to each gray matter voxel, together with its spherical neighborhood with a fixed radius. The results of the decoding model applied to the voxel neighborhood are projected onto the center voxel, allowing whole-brain multivariate analysis. This approach has several advantages. Firstly, it is a multivariate pattern analysis (MVPA) approach which means it takes advantage of the existing neural activation patterns that allow discrimination between experimental conditions. Secondly, it is a whole-brain approach that requires no a priori region specification. This way no unexpected activity can be missed due to only considering a subset of voxels for analysis. For our analyses, a spherical voxel neighborhood is defined with a radius of three, resulting in an average neighborhood of 86 voxels.

2.3.1 Inverted encoding models

To decode VWM representations from voxel responses, inverted encoding models (IEMs) were used. The architecture of the IEM that is used in this study is formulated by Rademaker and colleagues (2019). The IEM analysis consists of two steps: (1) estimation of an encoding model using the training set and (2) inverting the model to obtain stimulus reconstructions for the test set. An overview of an IEM architecture can be found in Figure 3.

Firstly, an encoding model is estimated using the training set, with the goal of predicting voxel responses from the observed stimulus orientations. The training set consists of a matrix of voxel responses A (trials x voxels) and a matrix of associated stimulus orientations B (trials x orientation degrees). A single row of B contains 1 at the index of the actual orientation and 0 at all the other orientations. Additionally, the encoding model requires nine predefined basis functions. The basis functions are serving as sensitivity functions the model uses to assign sensitivity values to various stimulus orientations. These basis functions are modeled as cosine functions in 180° space and are also referred to as channels. The nine predefined channels are combined with B to create a new matrix C (trials x channels). A single row of C consists of the pointwise product of the corresponding row in B with each of the nine channels. The voxel responses matrix A is then related to the channel responses matrix C by a weight matrix W (voxels x channels) (see Equation 1). W essentially quantifies the sensitivity of each channel for each voxel. W is estimated by least-squares linear regression and can be used to predict the voxel responses in C using the weights of the channels as predictors (i.e. as regressors in a linear regression).

$$A = WC$$

Equation 1.

After estimation of the encoding model, W can be used for decoding the unseen test set samples. Specifically, W will be applied to the voxel responses of the test set with the goal of obtaining a stimulus reconstruction for each test set sample. This is achieved by using the inverse of W using the Moore-Penrose pseudoinverse. The inverted weights matrix W can be applied to the voxel responses of the unseen test set B_{test} (trial x voxel), resulting in the stimulus reconstructions C_{test} (trial x channel) (see Equation 2). W is multivariate in nature, meaning it uses the sensitivity profiles across all voxels to jointly build the stimulus reconstructions. Each trial in the test set will be given an estimated response for each of the nine predefined channels.

$$C_{test} = (W^T W)^{-1} W^T B_{test}$$

Equation 2.

The process of estimating *W* and building the stimulus reconstructions is repeated for a total of 20 iterations. In each iteration, the centers of each of the nine channels shift one degree. For each iteration, the matrix *C* is computed anew with stimulus orientations *B* and the nine shifted channels. Also *W* is estimated anew and inverted to obtain the stimulus reconstructions, built using the shifted channels. After each iteration, the resulting stimulus reconstructions are stored. Thus, after 20 iterations the entire 180° stimulus space has been taken into account. When omitting this iterative shifting step, stimulus reconstructions of the test set would be biased as they would rely on the arbitrary placement of the nine channels (Scotti et al., 2021). After 20 iterations, the stimulus reconstructions will have the shape (trials x 180 channels) and the stimulus reconstructions are predicted at 1 degree steps. Thus, the resulting stimulus reconstruction for a single trial in the test set consists of 180 channel responses.

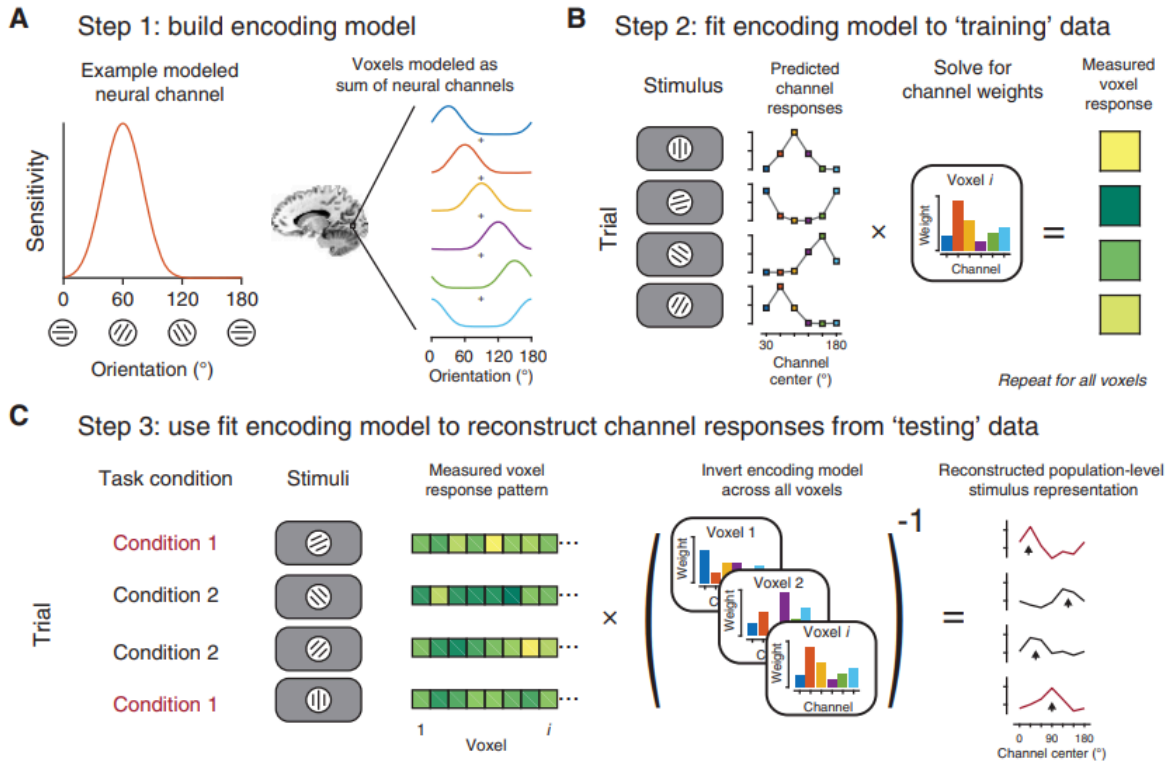

Figure 3. *IEM architecture, from Sprague et al. (2018).*

2.3.2 Evaluation of IEM output

How to reliably evaluate the channel responses of an IEM is a topic of debate. Standard approaches involve shifting the channel responses to center around 0 degrees, and averaging the shifted responses over all trials. Metrics that evaluate these average responses include amplitude, slope, bandwidth and fidelity (fitting of a cosine function). However, the align-and-average procedure tends to lose lots of important decoding information and can be prone to heavy outlier bias (Scotti et al., 2021). Moreover, decoding results based on these metrics can be difficult to interpret. They rest on the incorrect

assumption that a monotonic relationship exists between these metrics and the amount of stimulus-specific information in the brain signal. Important to understand is that an IEM stimulus reconstruction is formed by the predefined basis functions. A perfect stimulus reconstruction would therefore be the exact shape of the basis function centered on the correct orientation. The aforementioned metrics do not take the characteristics of the basis channels into account. These metrics incorrectly assume the shape of the IEM channel responses is associated with neural tuning in the brain, while the shape of the channel responses is actually associated with the shape of the predefined basis functions.

In an attempt to mitigate these pitfalls we have opted for an evaluation method of the channel responses that takes the shape of the predefined basis functions into account. Our approach uses a correlation table to evaluate IEM channel responses trial by trial (Scotti et al., 2021). This entails the computation of correlation coefficients between the channel responses of a single trial and each of the 180 predefined basis functions, each centered on a distinct orientation in 180° stimulus space. This results in 180 correlation coefficients per trial, where a relatively high correlation coefficient for the basis function centered on orientation x indicates a good fit between the IEM stimulus reconstruction and orientation x. To further quantify the evaluation of the stimulus reconstruction, these 180 correlation coefficients are ranked in descending order. The rank index of the actual stimulus orientation is chosen as the metric to evaluate decoding, where a lower rank index corresponds to a more accurate decoding. Another option would have been to simply select the orientation corresponding to the highest correlation coefficient in the ranking and use the absolute distance in degrees between this orientation and the actual stimulus orientation as an error measure. The problem with this approach is that in the presence of two orientation stimuli, the orientation with a stronger presence in the voxel responses will always be chosen over the other orientation. Because in our experimental design participants are required to retain two stimuli in VWM, the presence of two representations in voxel responses is a possibility. Therefore, we have opted for choosing the rank index of the correct stimulus orientation as a decoding metric. This way weaker representations will not be missed due to the presence of stronger representations in voxel responses. While the correlation table method has been used before in the context of evaluating IEM channel responses, the addition of the ranking method is an entirely new approach.

After decoding is complete for a given voxel's searchlight neighborhood, each trial is associated with a correlation table rank. The average rank index over trials is computed, normalized to a range of [-1, 1] and stored in the center voxel in the output map of the searchlight function.

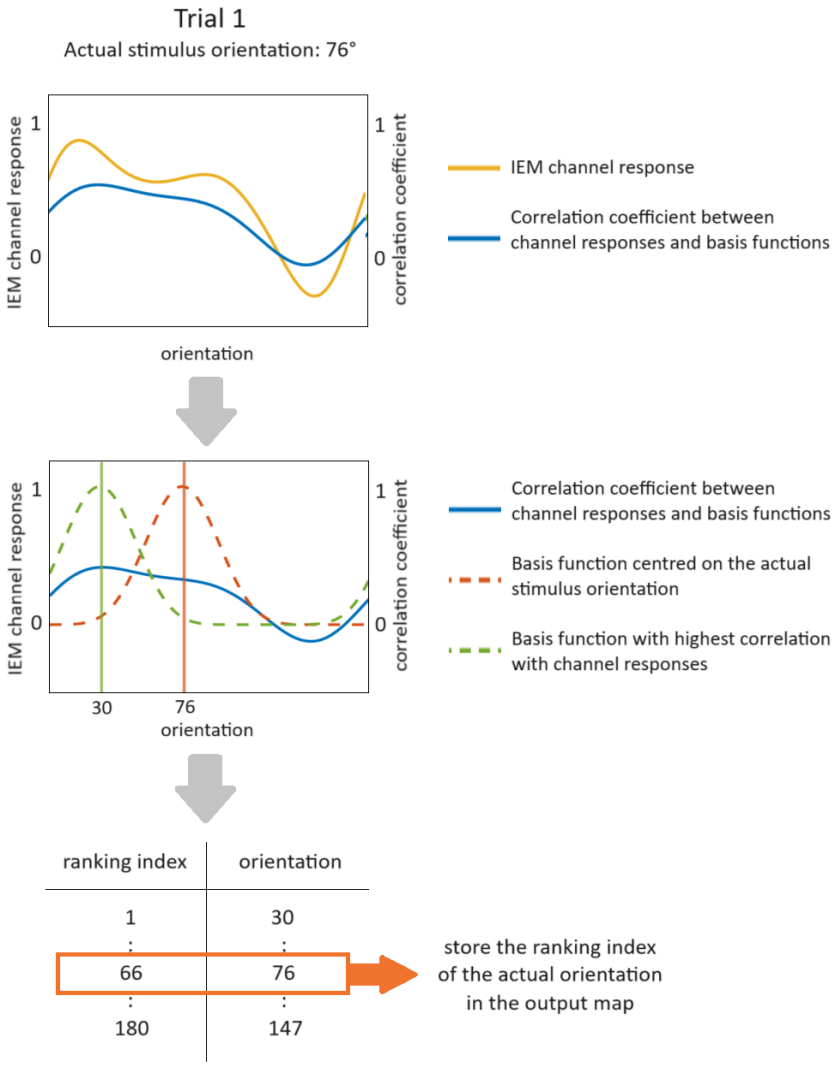

Figure 4. *IEM evaluation approach: correlation table and ranking method.*

2.3.3 Significance testing with permutations

To test whether the observed decoding results differ significantly from chance level, a null distribution was defined by executing 1,000 permutations. Rademaker et al. (2019) also used 1,000 permutations for their significance testing of IEM decoding results. Permutation testing is a non-parametric method with minimal assumptions and can be used in classification contexts to estimate the dependence between class labels and observations (Stelzer et al. 2013). Typically, the null hypothesis is defined as the independence between class labels and observations. To estimate the null distribution empirically a large number of permutations are applied to the class labels, after which corresponding decoding results are obtained. The significance level for rejecting the null hypothesis is computed by comparing the observed decoding results against the empirically estimated null distribution. Because our analysis pipeline uses a novel ranking method as a decoding metric, permutation tests offer a robust and

straightforward way to test our results for significance without having to adhere to the assumptions associated with parametric testing.

To construct the null distribution using permutations, the trials and their associated orientations were shuffled before the searchlight analysis. During the searchlight analysis, each voxel neighborhood used the same fixed 1,000 sets of shuffled orientation labels. A single permutation iteration entails selecting the correlation table rank index associated with the orientation of a randomly shuffled trial instead of the rank of the correct orientation. For each permutation iteration, the average, normalized rank was stored in the center voxel in the output map of the searchlight function.

Another way of defining a null distribution with permutations is by iteratively flipping the signs of the observed decoding results at random. This method essentially tests the results against zero. The problem with this method is that the null distribution is not subjected to any existing bias in the analysis pipeline. This would lead to testing biased observed decoding results to an unbiased null distribution, which leads to unreliable significance testing. Additionally, significance testing by testing against zero is less precise because the standard deviation of the null distribution can be overestimated when simply flipping the signs of decoding results. To avoid these pitfalls we use 1,000 permutation maps that have been made by shuffling the orientation labels randomly and evaluating the IEM channel responses with these randomly shuffled labels. Because the permutation maps are constructed within the analysis pipeline, every permutation is based on exactly the same data and processed by exactly the same analysis pipeline. The only thing that has changed is the mapping between the decoding stimulus reconstructions and the actual stimulus orientations. Thus, if the observed decoding results are significantly higher than the permutation decoding results, it serves as a strong indicator that orientation information is present in the selected voxel responses.

2.3.4 Threshold-free cluster enhancement

After a searchlight analysis, we obtain an output map where each voxel is associated with a decoding result. To test these decoding results for significance, the 1,000 permutation maps that have been constructed can be used. However, it is not desirable to execute voxel-wise significance testing as this could not identify any meaningful, spatially-extended signal. Instead, we want to take the decoding results of neighboring voxels into account to differentiate between spatially-extended signals and noise. Because their respective neighborhoods partly overlap, neighboring voxels are assumed to contain similar information and thus should show similar decoding results. A voxel with a relatively high decoding result is more likely part of a meaningful signal if its neighboring voxels show similarly high decoding results. Alternatively, an isolated high decoding result likely reflects noise. Implementing a clustering algorithm facilitates the discrimination between noise and a spatially-extended signal.

For the current study threshold-free cluster enhancement (TFCE) is used (Smith et al., 2009). TFCE is a clustering algorithm that tests the existence of an effect by looking at the amount of support

from neighboring voxels. Firstly, TFCE is applied to the observed decoding results. To perform significance tests, TFCE is also applied to the 1,000 null-decoding results. Afterwards, each voxel's TFCE score is compared against the voxel's null distribution of TFCE scores to obtain significance results, expressed in z-scores. Cluster-based significance testing is found to be more sensitive to finding an existing effect than voxel-wise significance testing (Smith et al., 2009). These steps were executed using the function *cosmo_montecarlo_cluster_stat*, which automatically corrects for multiple comparisons.

When TFCE is applied to the observed decoding results, it assigns each voxel a value indicating the amount of spatially-extended support. If a center voxel shows high decoding results and is surrounded by voxels with similarly high decoding results, TFCE will assign a relatively high value to the center voxel. Alternatively, if high decoding results are surrounded by low decoding results, TFCE will assign a low value to the center voxel. Thus, the assigned TFCE value indicates the level of spatial support of a voxel. In other clustering approaches, the resulting voxels' values are compared against a predefined cluster-forming threshold, which means only voxels with a value higher than this threshold are labeled as part of a significant cluster. However, TFCE uses a threshold-free approach which means a cluster-forming threshold does not need to be defined beforehand. Defining such a threshold beforehand is often arbitrary while its choice could have large effects on the cluster results. Instead, TFCE iterates through a number of different cluster-forming thresholds and computes the amount of spatial support under each of these thresholds by multiplying the height (decoding result) with the extent (amount of neighboring voxels with the same result). The resulting TFCE value of a voxel is the integral of all supporting threshold outcomes, i.e. the area under the curve.

The described TFCE procedure is repeated for all 1,000 null-decoding results, which results in 1,000 TFCE output maps. To test for significance, the observed TFCE values can be compared against this null distribution. TFCE and the significance tests were executed using the function *cosmo_montecarlo_cluster_stat*, which performs TFCE on the observed data, tests the results for significance using the given null distribution and automatically corrects for multiple comparisons. The function returns an output map where each voxel is provided with a z-score indicating how the observed TFCE score for that voxel relates to its null distribution, whole-brain corrected. The resulting z-scores indicate how many standard deviations the observed TFCE value is from the mean in the null data. A threshold of $z>1.64$ can be used for one-sided significance testing and a threshold of $z>1.96$ and $z<-1.96$ can be used for two-sided significance testing. The used function requires two parameters to be specified: the mean value of the input data under the null hypothesis h_0 and the threshold step *dh* with which the cluster-forming threshold is incremented iteratively by the TFCE algorithm. The mean under the null hypothesis is set at $h_0=0$, since decoding results have been normalized to range [-1, 1]. It is recommended to set *dh* such that *dh**100 roughly corresponds to the range of values of the input map, to ensure a reasonable tradeoff between speed and accuracy. The values in the input map are ranged from -1 to 1. The function *cosmo_montecarlo_cluster_stat* applies

TFCE for negative and positive data separately. Therefore, the threshold step is set at *dh*=0.01 to cover the range [-1, 0] and [0, 1].

An important remark of the TFCE procedure is that the observation of a significantly high z-score for a voxel does not serve as indication that information is maintained in the location of this particular voxel. A significantly high z-score merely indicates that the voxel is part of at least one significant cluster of voxels in which information is maintained.

2.4 Analyses

To answer each individual research question the appropriate data to train and test the decoding model on were selected. The following section specifies which data were selected for each respective research question. A table overview can be found in Figure 1.

2.4.1.a Where in the cortex is currently-relevant VWM content stored?

To answer this research question we attempt to decode currently-relevant VWM representations by training and testing on the CMI. The first and the second retention period are analyzed separately. Results from the first retention period can be used to adequately identify differences between currently- and prospectively-relevant VWM representations and their storage locations. Results from the second retention period can be used as additional results as to where currently-relevant information might be stored when there is no longer a prospectively-relevant item to be maintained. Note that for the second retention period we will analyze the representation of the orientation that is cued by the second retro-cue and that both repeat-trials and switch-trials are included in this analysis. Significant decoding results are defined as voxels that show a z-score of 1.64 or higher, after TFCE.

2.4.1.b Where in the cortex is prospectively-relevant VWM content stored?

For this research question we attempt to decode prospectively-relevant VWM representations by training and testing on the UMI. Only the first retention period is analyzed, as this is the only retention period during which a representation can be labeled as prospectively-relevant. Significant results are defined as z-scores of 1.64 or higher.

2.4.2.a Are currently- and prospectively VWM representations stored in similar representational formats?

For this research question we only analyze the first retention period. We will train the decoding model on the CMI and test on the UMI (CMI → UMI), and train on the UMI and test on the CMI (UMI → CMI). This is also referred to as cross-decoding. Z-scores of 1.96 or higher show significant positive cross-decoding and imply currently- and prospectively-relevant representations are retained in similar representational formats. Alternatively, z-scores of -1.96 or lower show significant negative cross-decoding and imply currently- and prospectively-relevant representations are retained in similar,

but orthogonal representational formats. In the case that we find successful cross-decoding (either positive or negative) it implies that both the CMI and the UMI are present concurrently in the selected voxel neighborhood. Thus, any significant cross-decoding results can be additionally used to answer the previous research questions concerning where currently- and prospectively-relevant representations are stored.

2.4.2.b. Are currently- and prospectively-relevant VWM representations stored in orthogonal representational formats?

In addition to similar representational formats, we will test whether currently- and prospectively-relevant representations are stored in formats orthogonal to each other. This is achieved by training on the CMI and testing on the orientation orthogonal to the UMI and vice versa. Z-scores of 1.96 or higher imply currently- and prospectively-relevant representations are stored in formats orthogonal to each other. Z-scores of -1.96 or lower imply they are stored in similar formats (not orthogonal to each other).

2.4.3.a Where in the cortex is VWM content maintained in stable representational formats?

For the temporal analyses we will be using early- and late retention-related data from the first retention period, as previously mentioned. In order to analyze where in the cortex stable representations are maintained, we firstly need to determine where in the cortex representations are present during both early and late retention. Therefore, we have taken the average of early-within (training and testing on early delay; early → early) and late-within (training and testing on late delay; late → late). After applying TFCE, the significant results ($z>1.64$) were identified. Voxels with significant results were selected for a mask, identifying locations where representations are present during both early and late retention.

Within this selected mask, we apply cross-temporal analyses. Specifically, we train on early retention and test on late retention (early → late), and vice versa (late → early). These two cross-temporal directions are averaged and TFCE is applied to the results. Z-scores of 1.64 or higher imply VWM representations are stored in stable representational formats during retention.

2.4.3.b Where in the cortex is VWM content maintained in changing representational formats?

To answer this research question the aforementioned mask was used again. The decoding results of cross-temporal decoding (early → late and late → early) were subtracted from within-temporal decoding (early → early and late → late). TFCE was applied to the results. Z-scores of 1.64 or higher would imply VWM representations are stored in changing representational formats during retention. This would be inferred from observing significantly worse cross-temporal decoding results compared to the within-temporal decoding results.

3. Results

In the following section the results are reported per research question. A table overview can be found in Figure 1.

3.1.a Where in the cortex is currently-relevant VWM content stored?

For all three participants, we were able to decode currently-relevant VWM content by training and testing on the cued representation during the first retention period (z_{tfce}>1.64) (see Figures 5, 7, 9). Significant decoding was found for participant 1 and 3 in the occipital lobe and for all three participants in the parietal lobe. Additionally, we can see that for all participants significant decoding seems to be more left lateralized, especially in parietal regions.

Furthermore, we found some unexpected negative significant results for participant 2 (z_{tfce}<-1.64). This observation indicates that decoding was significantly below chance in these locations. Negative decoding results were found in the frontal lobe and the clusters were relatively small.

We also attempted to decode the CMI during the second retention period, where the CMI is the orientation cued by the second retro-cue. This was done as an additional analysis to examine where in the cortex currently-relevant representations are stored, however, in the absence of a prospectively-relevant representation. Again, we were able to successfully decode the CMI during the second retention period for all three participants (see Figures 6, 8, 10). For participant 1 and 2 we found significant CMI decoding in locations throughout the occipital, parietal and frontal lobe. Participant 3 only showed significant decoding in occipital regions.

3.1.b Where in the cortex is prospectively-relevant VWM content stored?

To decode prospectively-relevant VWM representations we trained and tested on the UMI. The UMI could be decoded successfully for participant 2 and 3, and could not be decoded for participant 1 (see Figures 11, 12). For participant 2, a small cluster was found in the right-frontal lobe (z_{tfce}>1.64). For participant 3, a small cluster was found in the left-frontal lobe (z_{tfce}>1.64). No significant within-decoding of the UMI was found in occipital or parietal regions.

3.2.a Are currently- and prospectively-relevant VWM representations stored in similar representational formats?

To examine the similarity of currently- and prospectively-relevant representations we cross-decoded in two directions: CMI → UMI and UMI → CMI. For participant 1 we found significant negative cross-decoding for the UMI → CMI direction in the left-parietal lobe (z_{tfce}<-1.96) (see Figure 13). Negative cross-decoding suggests CMI and UMI are stored in orthogonal representational formats in these locations. No positive cross-decoding was found for participant 1.

For participant 2 we found positive cross-decoding for UMI → AMI in the parietal lobe (z_{tfce}>1.96) (see Figure 14). This implies the CMI and UMI are stored in similar representational formats. No negative cross-decoding was found for participant 2.

For participant 3 we could find no significant cross-decoding in any direction.

3.2.b Where in the cortex is attended and unattended VWM content stored in orthogonal representational formats?

To test the presence of orthogonal storage for the CMI and UMI, we performed cross-decoding by training on the UMI and testing on the orientation orthogonal to the CMI. For participant 1 significant cross-decoding was found in the left-parietal lobe, confirming the negative cross-decoding described in the previous paragraph (z_{tfce}>1.96) (see Figure 15). This serves as an additional indication that prospectively-relevant information can be stored in a representational format orthogonal to the CMI. However, we have not been able to find evidence of orthogonal representations in the other two participants.

3.3.a Where is VWM content maintained in stable representational formats?

To answer this research question we restricted the search area for TFCE to locations where currently-relevant VWM content is stored during both early and late delay. Within this restricted area we tested cross-temporal decoding between early and late delay. For all three participants we found significant cross-temporal decoding for almost the entire area (z_{tfce}>1.64) (see Figures 16, 17, 18). This suggests that in areas that contain VWM content during early and late delay, representational formats during early and late delay are highly similar. In other words, VWM representations are maintained in a stable fashion within the occipital and parietal lobe.

3.3.b Where is VWM content maintained in changing representational formats?

For all three participants we did not find a significant difference in decoding strength between within-temporal decoding and cross-temporal decoding. In other words, within areas that contain VWM content during both early and late delay, cross-temporal decoding is not significantly worse than within-temporal decoding. This suggests that the representational format of VWM content is maintained in a stable fashion when comparing early to late delay. This finding confirms the results of the previous research question: VWM content is kept in a stable format instead of a changing format in occipital and parietal areas.

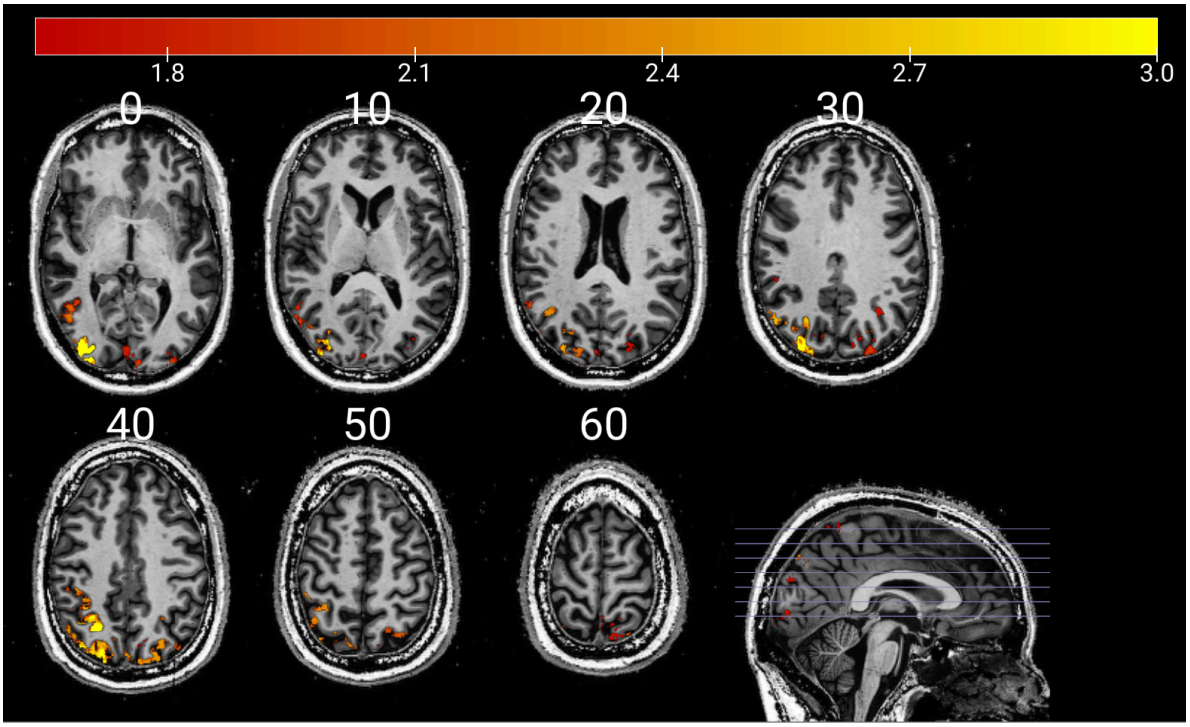

Figure 5. *CMI decoding during the first retention period for participant 1 (z>1.64). The significant decoding results are colored red-yellow, where a red voxel indicates a z-score of 1.64 or 1.96 (depending on the analysis) and a yellow voxel indicates a z-score of 3.0 or higher. Thus, any colored voxels indicate whole-brain corrected above-chance decoding. A color scale is included in the figure. The significant decoding results are projected onto the anatomical scan of the participant for interpretability. Each presented slice is an axial (horizontal) slice and is accompanied by a number indicating the index of the slice. Each axial slice corresponds to one of the lines in the sagittal cross-slice in the bottom-right of the figure, where the first top-left axial slice corresponds to the bottom line in the sagittal cross-slice.*

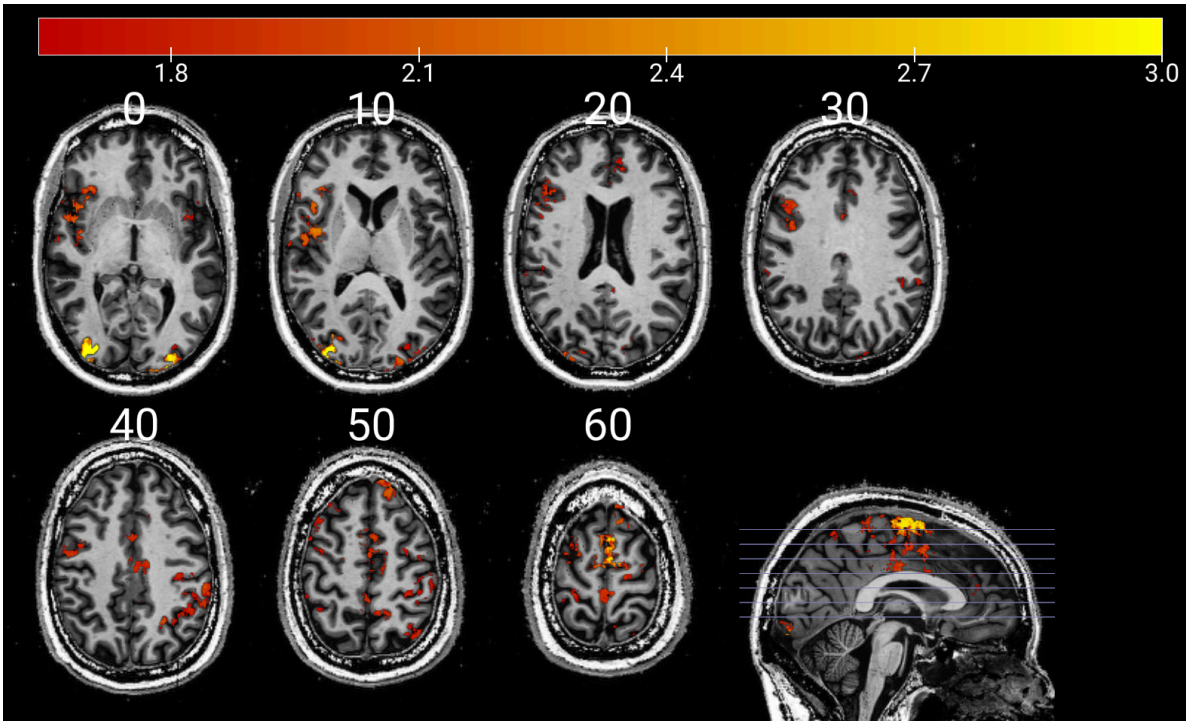

Figure 6. *CMI decoding during the second retention period for participant 1 (z>1.64). See the caption of Figure 5 for an explanation of the figure.*

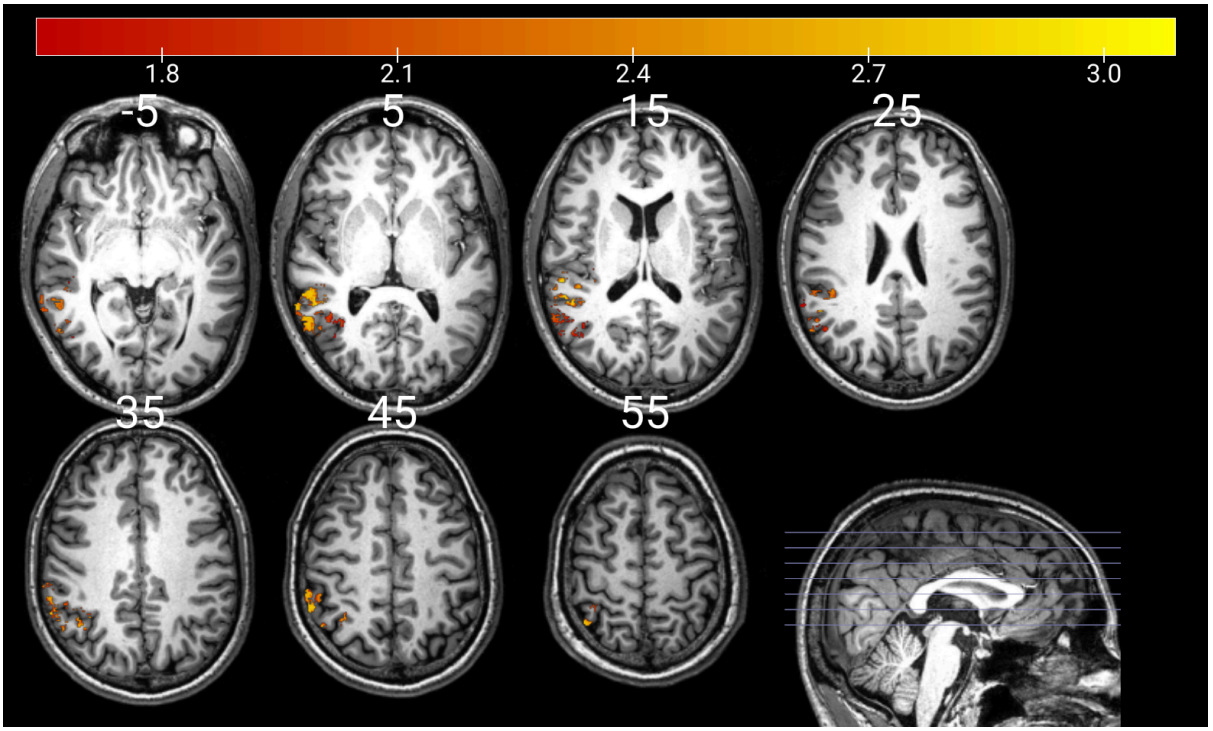

Figure 7. *CMI decoding during the first retention period for participant 2 (z>1.64). See the caption of Figure 5 for an explanation of the figure.*

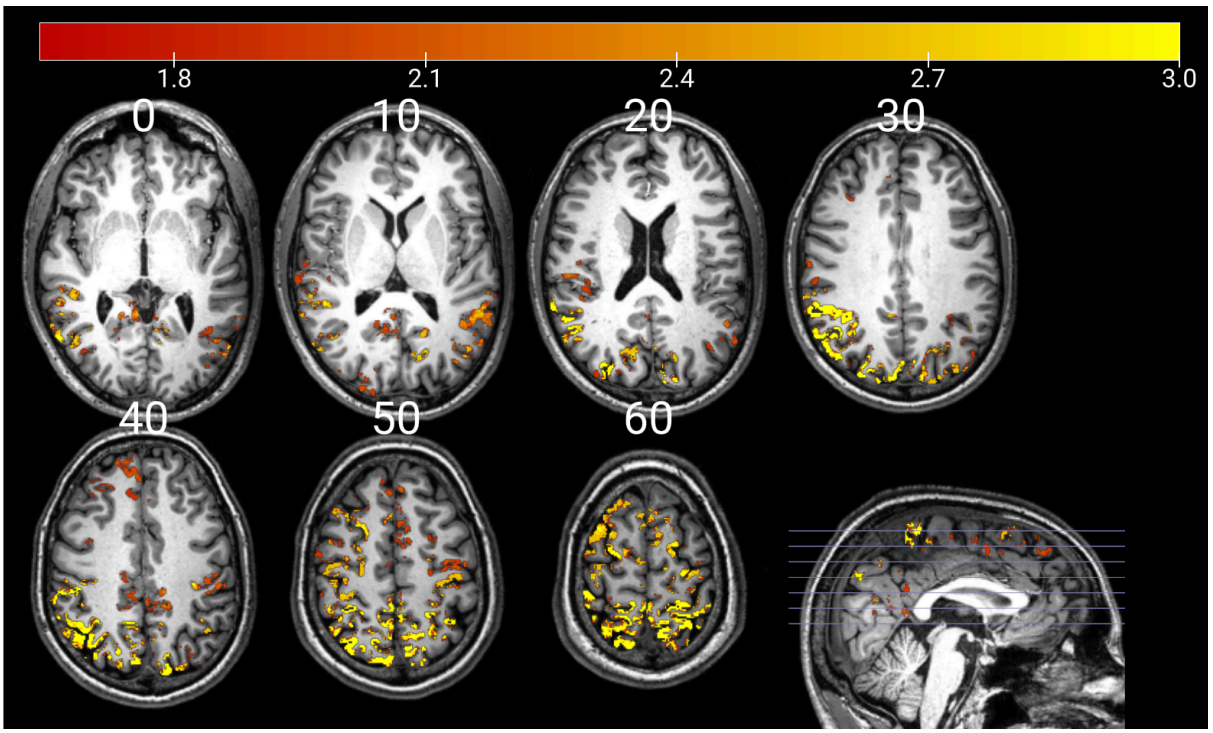

Figure 8. *CMI decoding during the second retention period for participant 2 (z>1.64). See the caption of Figure 5 for an explanation of the figure.*

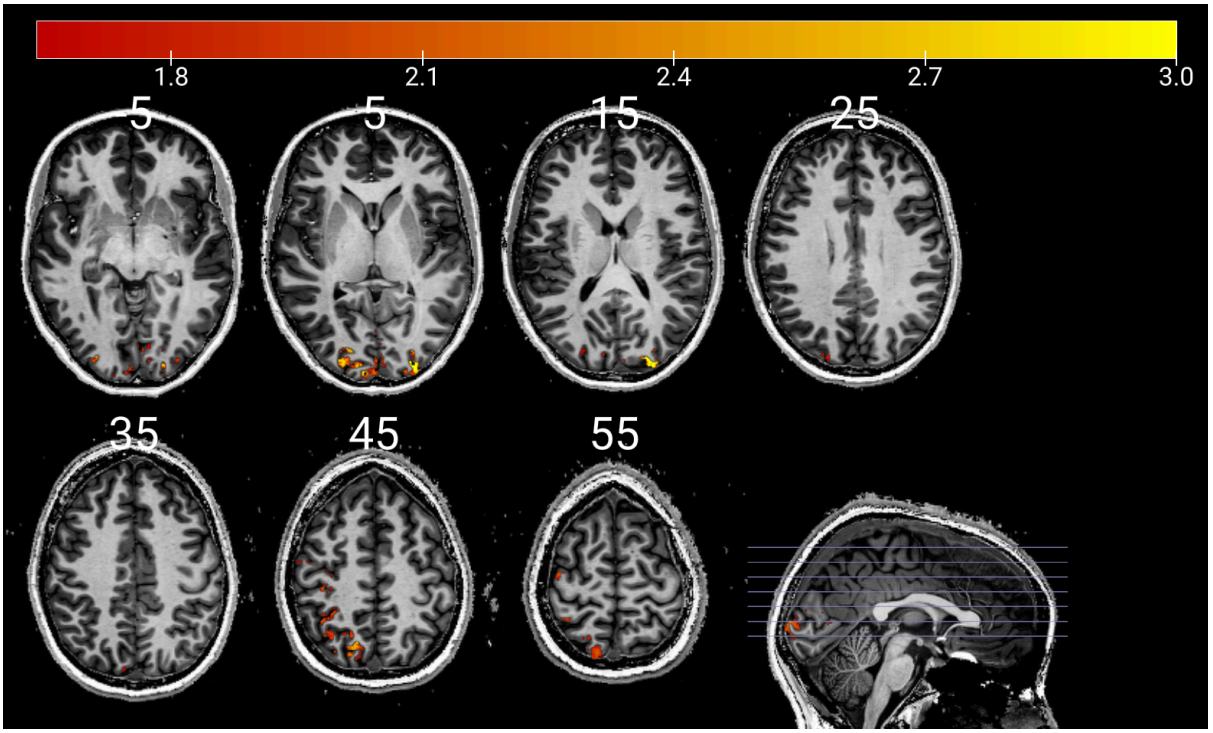

Figure 9. *CMI decoding during the first retention period for participant 3 (z>1.64). See the caption of Figure 5 for an explanation of the figure.*

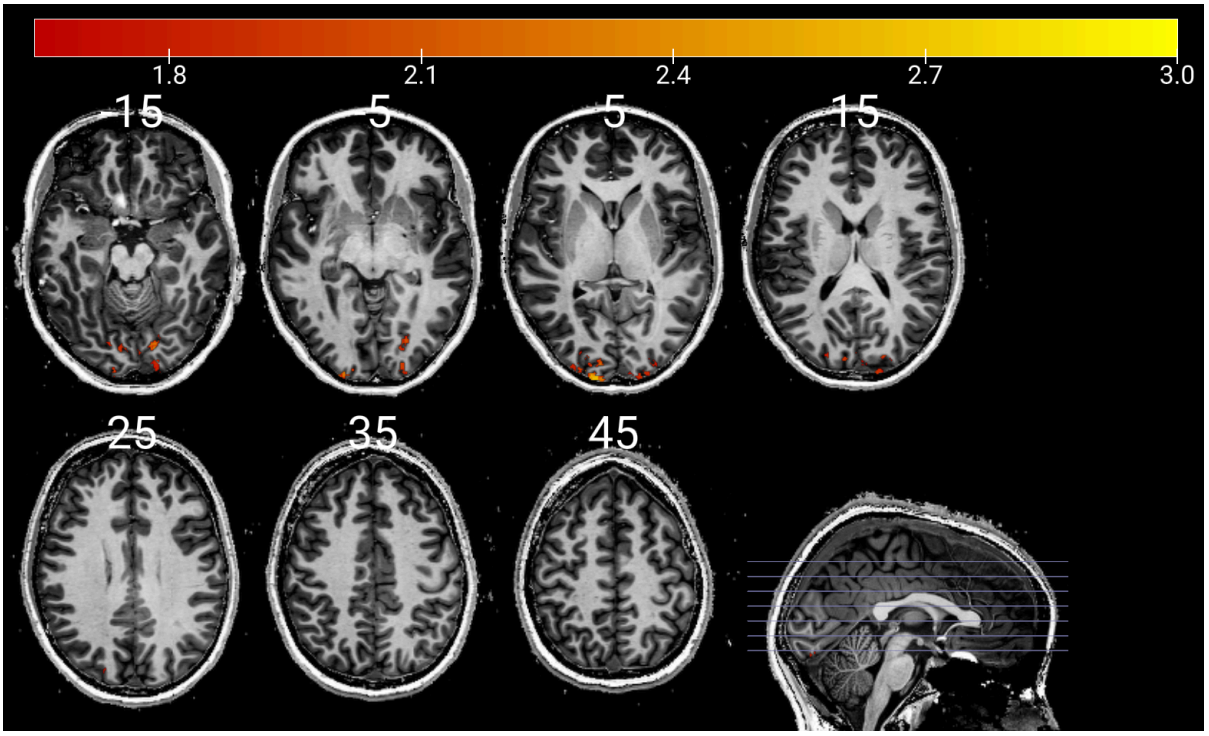

Figure 10. *CMI decoding during the second retention period for participant 3 (z>1.64). See the caption of Figure 5 for an explanation of the figure.*

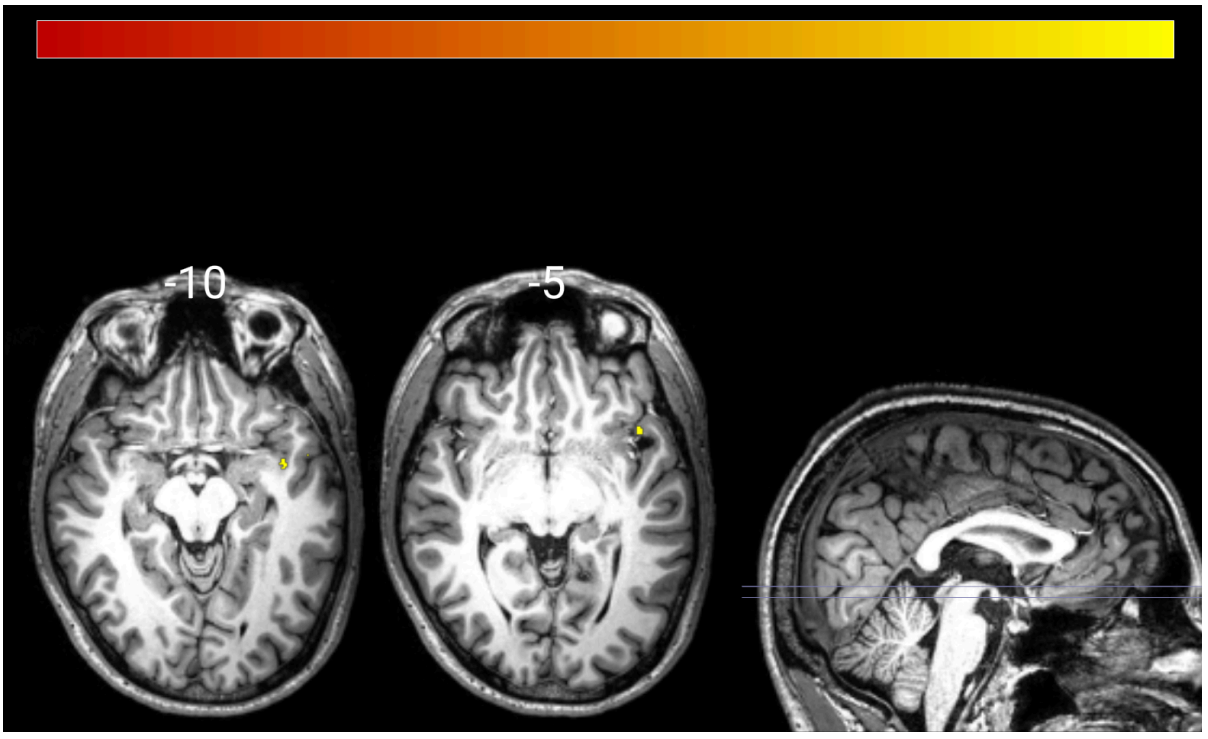

Figure 11. *UMI decoding for participant 2 (z>1.64). All voxels are made yellow for visibility. See the caption of Figure 5 for an explanation of the figure.*

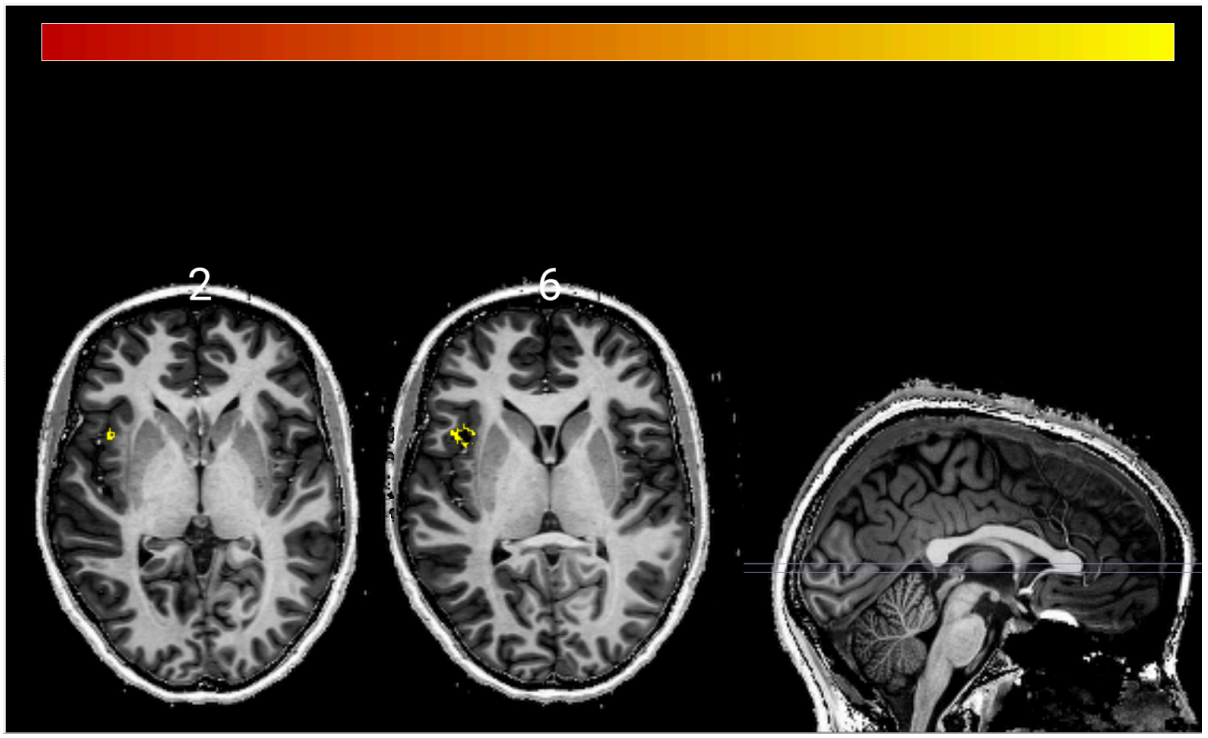

Figure 12. *UMI decoding for participant 3 (z>1.64). All voxels are made yellow for visibility. See the caption of Figure 5 for an explanation of the figure.*

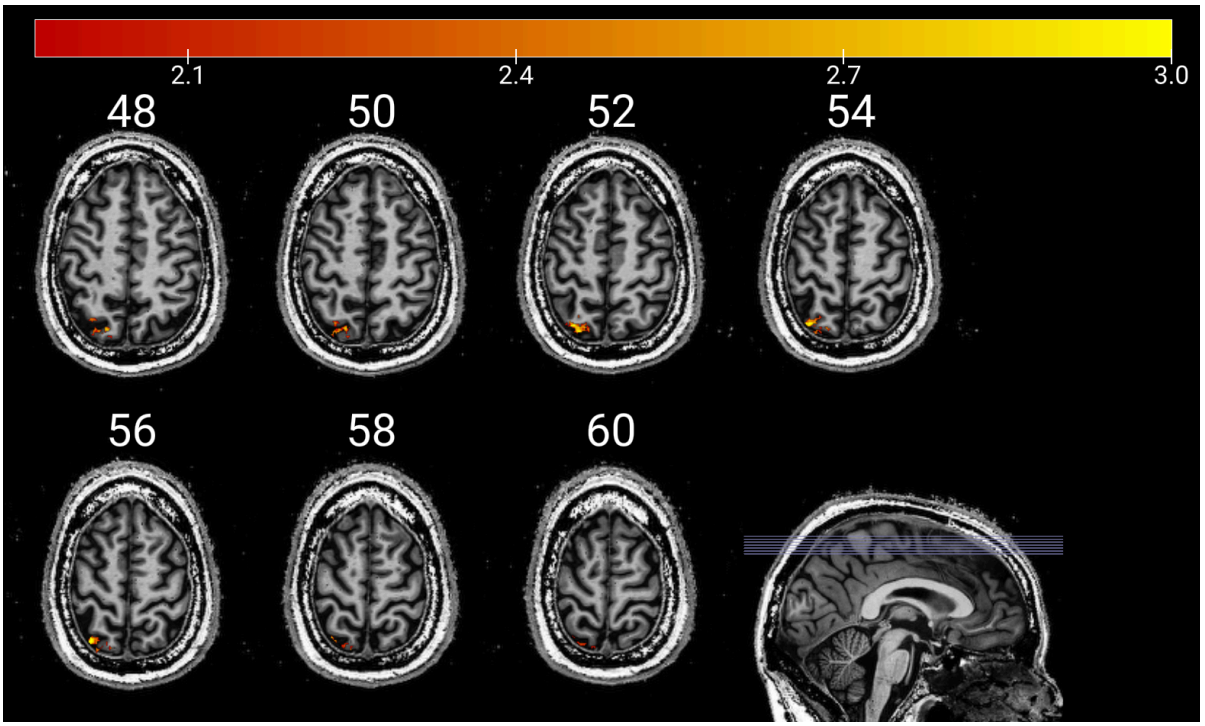

Figure 13. *UMI → CMI negative cross-decoding for participant 1 (z<-1.96). Z-scores have been sign-flipped for visibility. See the caption of Figure 5 for an explanation of the figure.*

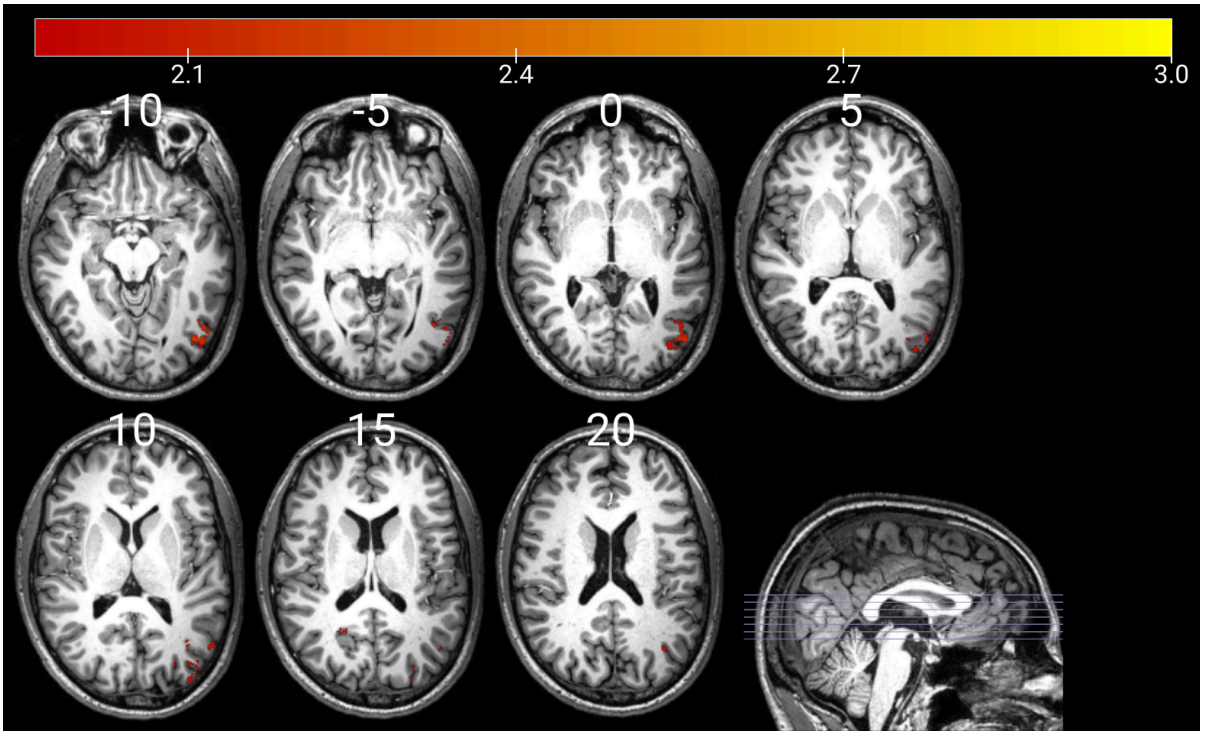

Figure 14. *UMI → CMI cross-decoding for participant 2 (z>1.96). See the caption of Figure 5 for an explanation of the figure.*

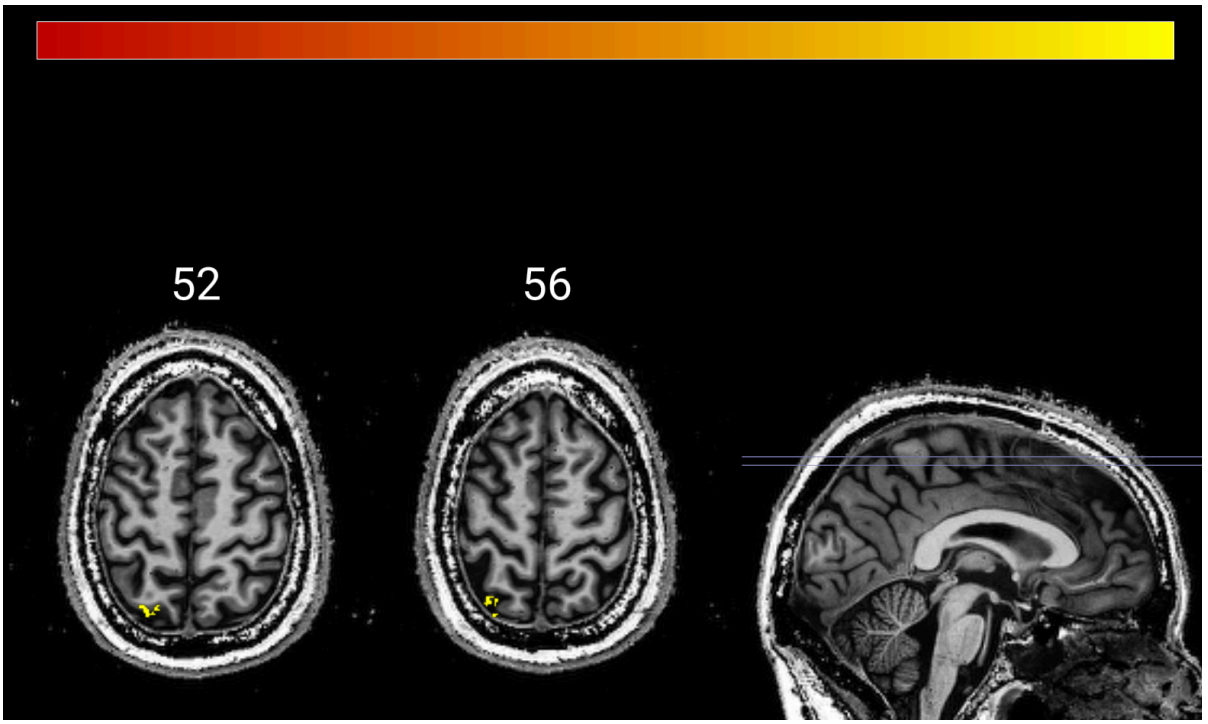

Figure 15. *UMI → orthogonal CMI cross-decoding for participant 1 (z>1.96). All voxels are made yellow for visibility. See the caption of Figure 5 for an explanation of the figure.*

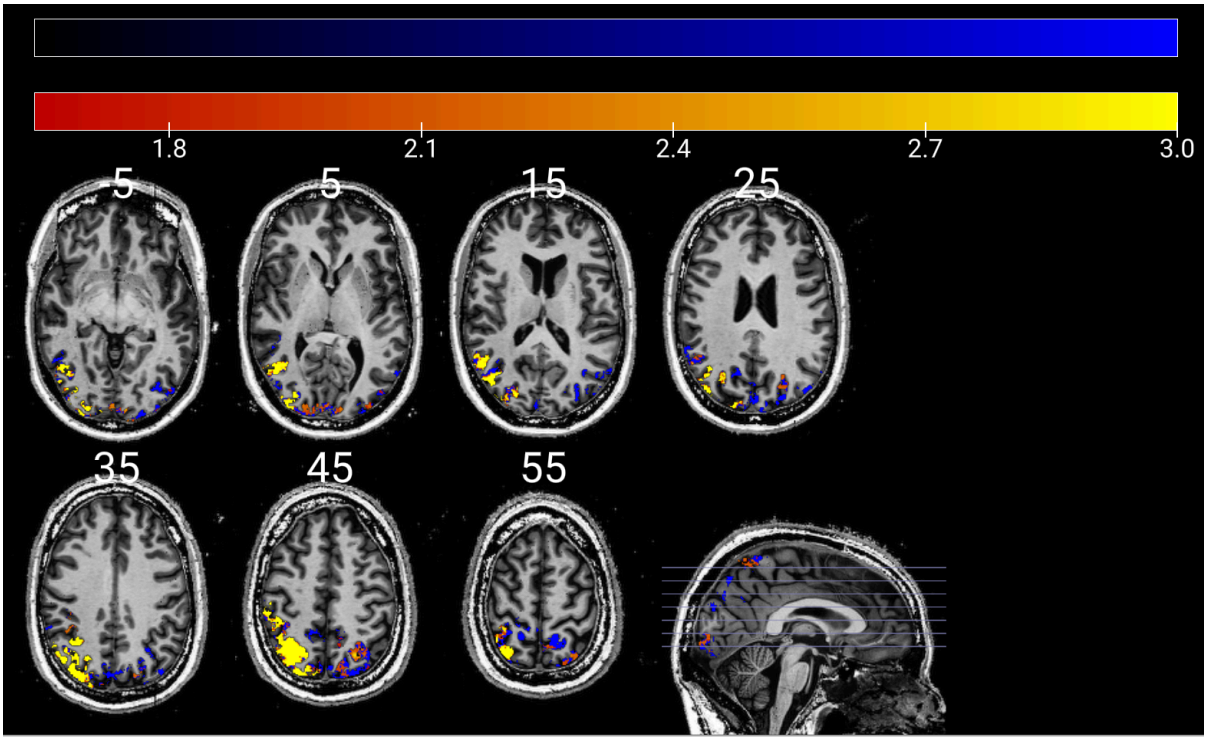

Figure 16. *Cross-temporal decoding (early → late and late → early) for participant 1 (z>1.64). In blue: areas where VWM content is stored during early and late delay, in red-yellow: significant cross-temporal decoding. See the caption of Figure 5 for an explanation of the figure.*

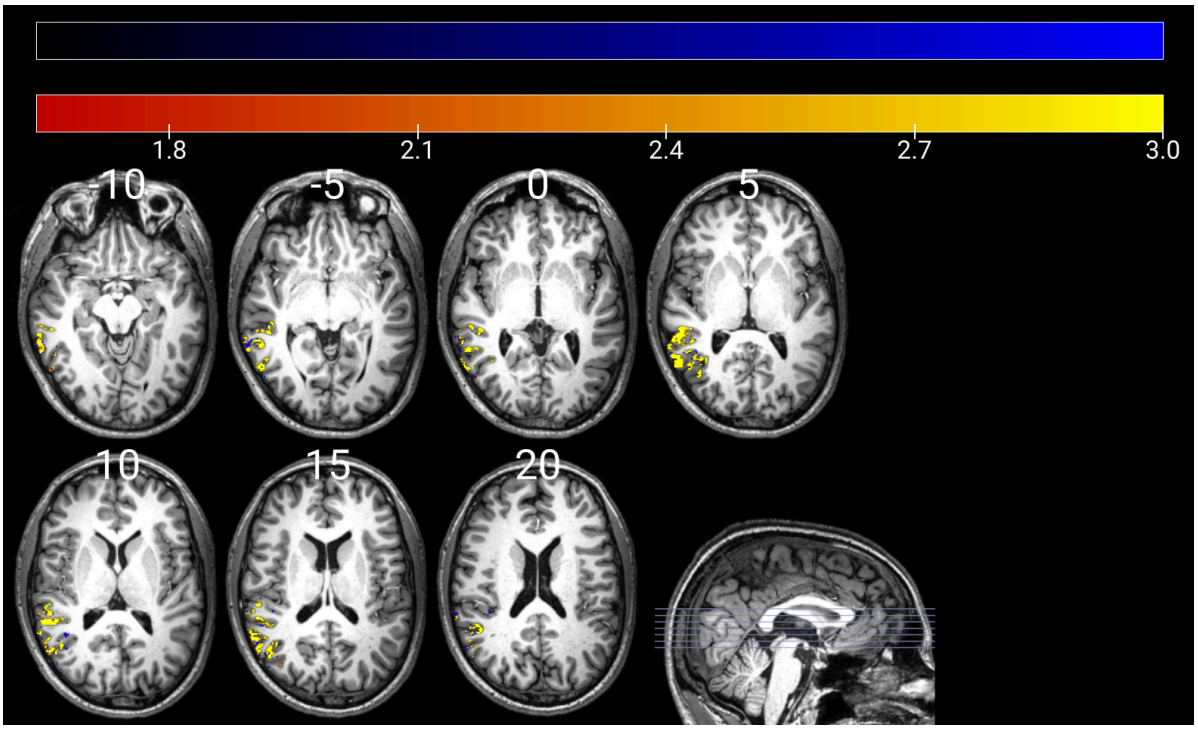

Figure 17. *Cross-temporal decoding (early → late and late → early) for participant 2 (z>1.64). In blue: areas where VWM content is stored during early and late delay, in red-yellow: significant cross-temporal decoding. See the caption of Figure 5 for an explanation of the figure.*

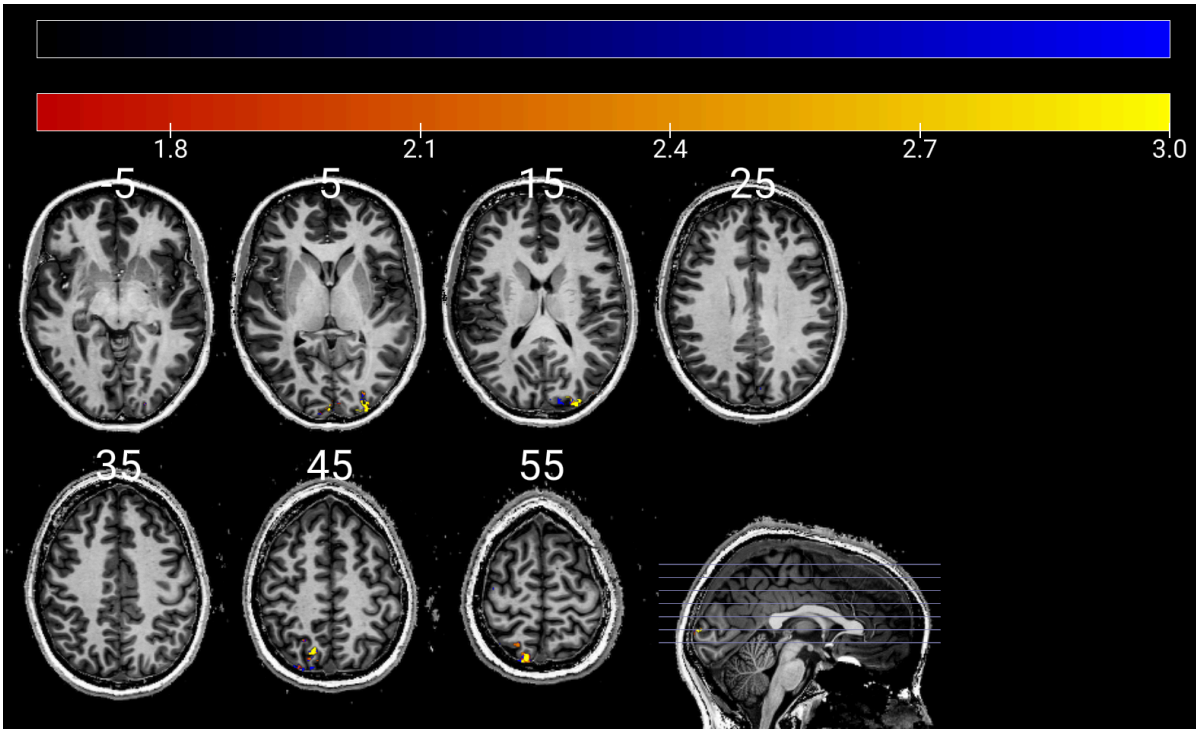

Figure 18. *Cross-temporal decoding (early → late and late → early) for participant 3 (z>1.64). In blue: areas where VWM content is stored during early and late delay, in red-yellow: significant cross-temporal decoding. See the caption of Figure 5 for an explanation of the figure.*

4. Discussion

In the current study visual working memory (VWM) is analyzed in various dynamic contexts. Firstly, the storage locations and representational formats of currently- and prospectively-relevant VWM representations are examined. Secondly, the maintenance of currently-relevant VWM representations are examined over the course of the retention period. The goal behind this approach is to challenge the static view of VWM when analyzing storage locations and formats of VWM representations. Instead, we adopted a dynamic view of VWM to analyze VWM representations by considering relevance status and maintenance over time.

In an attempt to advance upon previous research we have used a 7-Tesla fMRI dataset and employed various analysis techniques that are known for their sensitivity, i.e. inverted encoding models and threshold-free cluster enhancement. Moreover, searchlight analyses were used to take on an assumption-free approach when locating VWM representations. Additionally, we have created a novel approach for IEM output evaluation to ensure both strong and weak representations can be decoded when both are present in the same location in the brain.

4.1 Where in the cortex is VWM content stored?

Currently-relevant VWM representations were decoded in the occipital and parietal lobe, and could additionally be decoded in the frontal lobe during the second retention period. These findings are in agreement with earlier work decoding orientations from the EVC, IPS, FEF and PFC (Polanía et al., 2012; Cavanagh et al., 2018; Christophel et al., 2018; Rademaker et al., 2019, Yu et al., 2020).

Prospectively-relevant VWM representations could be decoded in relatively small areas in the frontal lobe, though in different hemispheres. Moreover, the cross-decoding results show the presence of prospectively-relevant representations in the parietal lobe for two participants, again in different hemispheres. This is inferred from the (either positive or negative) significant cross-decoding between the currently- and prospectively-relevant representation, which indicates the presence of both representations (Yu et al., 2020; Iamshchinina et al., 2021). The presence of prospectively-relevant representations in the frontal and parietal lobe is supported by earlier work pointing to IPS and FEF (Christophel et al., 2018; Yu et al., 2020; Iamshchinina et al., 2021). Prospectively-relevant representations were previously also decoded in the occipital lobe, which we have not been able to replicate (Yu et al., 2020; Iamshchinina et al., 2021; Ruijs, yet unpublished results).

Based on these results, it seems that certain brain areas have a preference for either the currently- or the prospectively-relevant representation. In the occipital lobe, only the currently-relevant representation could be decoded. This is in line with previous findings suggesting an overrepresentation of currently-relevant information compared to prospectively-relevant information in the EVC (Ruijs, yet unpublished results; Christophel et al., 2018). This is also in support of the sensory recruitment hypothesis, stating currently-relevant information is stored in

sensory-like representations in the EVC (Rademaker et al., 2019). Storing currently-relevant information in sensory-like, high-resolution representations could allow straightforward comparison with incoming visual input. Alternatively, prospectively-relevant information does not need to be stored in a sensory-like format as its representation is not currently task-relevant. Storing the prospectively-relevant representation in a sensory-like format would thus only serve as a distracting factor in the maintenance of currently-relevant information and its comparison with incoming visual input.

In the frontal lobe, only the prospectively-relevant representation could be decoded during the first retention period. This is in accordance with the finding that prospectively-relevant representations are represented more strongly in the frontal lobe than currently-relevant representations, specifically in FEF (Christophel et al., 2018). It is suggested that VWM representations in frontal regions are stored in a lower-resolution format, abstracted away from sensory-like representations. In this view, it makes sense the prospectively-relevant representation is stored in the frontal lobe, as it is not necessary to actively maintain this representation in a high-resolution, sensory-like format like the currently-relevant representation. Furthermore, storage in the frontal lobe would avoid disruption of the active maintenance of the currently-relevant representation in sensory areas. Several frontal areas have also been shown to store VWM information while simultaneously coordinating how and where this information is being represented (Ester et al., 2015). Thus, it is possible that the found decoding results in the frontal lobe reflect this top-down control in regard to the storage of prospectively-relevant representations. Prospectively-relevant representations could be stored in the frontal lobe while simultaneously being subjected to top-down coordination processes. In the same line of thought, the active maintenance of currently-relevant representations in sensory areas would not have to be coordinated by this top-down coordination in the frontal lobe.

The level of consistency across participants is also worth mentioning. While the decoding results of the currently-relevant representation is fairly stable across participants, the prospectively-relevant representation is not. Significant decoding of the prospectively-relevant representation was found in different areas (frontal and parietal) and in different hemispheres (right and left) when comparing between participants. It has previously been shown the prospectively-relevant representation is relatively weak, causing decoding difficulty in certain brain areas (Christophel et al., 2018; Ruijs, yet unpublished results). The weaker signal in combination with the conservative whole-brain correction could have led to low decoding performance in certain brain areas, for some participants. Inconsistency across participants could also be due to any shortcomings of our analysis pipeline. It is possible (one of) the chosen techniques fail to adequately capture the weaker prospectively-relevant representation. On the other hand, the robust currently-relevant representation seems to be effectively captured by the chosen techniques.

To conclude, we can find currently-relevant representations in the occipital and parietal lobe and prospectively-relevant representations in the frontal and/or parietal lobe. This shows that VWM representations can be stored in different brain areas based on their task-relevance state. Currently-relevant representations might be overrepresented in sensory-related areas for active, high-resolution maintenance and in preparation for comparison with incoming visual input. Prospectively-relevant representations could be overrepresented in frontal areas to reduce disruption of the active maintenance of currently-relevant representations. These representations are possibly stored in frontal areas in a lower-resolution format, abstracted away from their sensory-like representation and subjected to top-down processes to coordinate their maintenance.

4.2 Where in the cortex are currently- and prospectively-relevant VWM content stored in similar representational formats?

The cross-decoding results show similar representational formats for currently- and prospectively-relevant representations in the right-parietal lobe of one participant. This is in accordance with studies finding similar representational formats in IPS (Van Loon et al., 2018; Yu et al., 2020; Iamshchinina et al., 2021). For another participant, orthogonal representations were present in the left-parietal lobe. Even though this is found in only one participant, two separate analyses have confirmed the existence of orthogonal storage. Orthogonal storage of prospectively-relevant stimulus orientations in parietal areas is a novel finding. Previous work has located orthogonal representations of orientations in EVC, i.e. the occipital lobe (Yu et al., 2020). Opposite representational formats for objects were found in the posterior fusiform cortex, also known as the object-selective cortex, located in occipito-temporal areas (Van Loon et al., 2018). Yu et al. (2020) did find storage of stimulus locations in opposite representational formats in IPS, i.e. in the parietal lobe. However, our results show opposite storage for stimulus orientation, not of stimulus location, in parietal areas.
As previously mentioned, we have been able to find orthogonal storage in the parietal lobe with two distinct analyses. Thus, it is possible a new location has been revealed for the orthogonal storage of remembered orientations.

Interestingly, we only found significant cross-decoding results in the UMI → CMI direction, not in the CMI → UMI direction. It has been shown that if two datasets have differing signal-to-noise ratios, cross-decoding is the strongest when training on the low signal-to-noise ratio dataset and testing on the high signal-to-noise ratio dataset (Van den Hurk et al., 2019). In this case, this is confirmed by significant cross-decoding when training on the weaker UMI representation and testing on the stronger CMI representation and not vice versa. While we do not find cross-decoding results in both directions, we can base our interpretation of the similarity between the CMI and UMI on the best performing direction, i.e. UMI → CMI (Van den Hurk et al., 2019).

As with the within-decoding results of the prospectively-relevant representations, the cross-decoding results are showing inconsistencies between participants. Because the cross-decoding

results partly rely on the decoding of the prospectively-relevant representation, this could be due to the inadequacy of our analysis pipeline in the context of decoding this representation. Alternatively, the found inconsistencies could signify individual differences in maintaining multiple VWM representations, one in a currently-relevant state and one in a prospectively-relevant state. It has even been proposed that the debate on the storage location(s) of VWM representations is due to such individual differences (Pearson et al., 2019). Individual differences have been found in VWM capacity and in the cognitive control of attention in the presence of distraction (Luck et al., 2013; Gulbinaite et al., 2014). It is possible that the amount of cognitive control explains some found inconsistencies. For instance, a high amount of cognitive control could be reflected by orthogonal representational formats, where the goal is to actively protect the currently-relevant representation from interference in anticipation for the recall task. Alternatively, a low amount of cognitive control could be reflected by similar representations, in which case the brain wastes no additional resources to differentiate between currently- and prospectively-relevant representations. These differences in cognitive control are not necessarily reflected by differences in behavior, i.e. recall error or reaction time (Gulbinaite et al., 2014). Another possible explanation for inconsistent results in the current study is that the similarity of the currently- and prospectively-relevant representations changes over the course of the retention period (Van Loon et al., 2018). Similar representations could be found at the start of retention, then drop to baseline, before dropping below baseline to orthogonal representations right before the recall task. This could very well reflect the differences we found between participants, considering the retention period was averaged over time. For one participant similar representations were found, for another orthogonal representations and for another no similarity at all.

The cross-decoding results show that when currently- and prospectively-relevant representations are concurrently stored in the same brain area, they can either be stored in similar or in orthogonal representational formats. These two different storage methods could be explained by individual differences in the concurrent maintenance of currently- and prospectively-relevant VWM representations or could both be employed, but at different timepoints during the retention period.

4.3 Where in the cortex is currently-relevant VWM content maintained in stable vs. changing representational formats?

Cross-temporal analyses show stable VWM representations over the course of the retention period in occipital and parietal areas. Consistent results across participants confirms the notion that our analysis pipeline is able to adequately capture the robust currently-relevant representation. However, our results contradict some previous findings. Unstable, changing representational formats were found in the occipital and parietal lobe based on unsuccessful cross-temporal generalization in these areas (Oh et al., 2019; Yu et al., 2020). Alternatively, stable representational formats are associated with the frontal lobe (Oh et al., 2019). Because we found no significant decoding of currently-relevant

representations in the frontal lobe during the first retention period, our analyses do not reveal whether representations are also kept in stable format in the frontal lobe.

The observed cross-temporal generalization in the current study could also have been the result of the lag in BOLD-response, which might have led to a significant overlap in signal between early and late retention. Faster measuring techniques or longer retention periods could reveal unsuccessful cross-temporal generalization and thus dynamic changes in the representational format of VWM representations.

It is also possible that slightly weaker VWM representations were previously overlooked by less sensitive analysis techniques, which resulted in the inability to generalize across time in occipital and parietal areas. Using a highly sensitive analysis pipeline, our results suggest VWM representations in the occipital and parietal lobe are maintained in a stable representational format during the retention period. Storing currently-relevant memory items in stable representational formats likely reflects actively sustained attention (Chun, 2011). Bringing the currently-relevant memory items into the focus of attention could facilitate active maintenance in anticipation of the behavioral task with the goal of minimizing distraction and memory decay.

4.4 Limitations

Several limitations to the current study should be acknowledged. As previously mentioned, we were not able to decode the prospectively-relevant representation in all expected areas, most notably the occipital lobe. This could have been caused by a number of reasons.

Firstly, it could be due to our choice of analysis methods, such as the novel ranking-based approach. In order to improve our chances of decoding both the currently- and prospectively-relevant representations from the fMRI voxel responses, we created a new method to evaluate IEM output. Using a ranking-based method instead of a winner-takes-all approach would result in a more sensitive decoding metric. This would lead to a higher chance of decoding weaker representations such as prospectively-relevant representations, without being overshadowed by the robust currently-relevant representations. Conversely, the ranking-based method is a non-parametric method and thus lacks some sensitivity compared to parametric approaches. The alignment between the found decoding results and previous results showcases that our evaluation method seems to work adequately for decoding currently-relevant VWM representations. In addition, the cross-temporal decoding results show highly consistent results across participants, confirming the effectiveness of the method. However, the evaluation method seems to lack some sensitivity when it comes to decoding prospectively-relevant VWM representations. Both within-decoding results and cross-decoding results of prospectively-relevant representations were inconsistent across participants. Furthermore, we were not able to decode the prospectively-relevant representations in the occipital lobe, even when using cross-decoding. Perhaps the currently-relevant representation was so strongly present in the occipital lobe that, despite the sensitive ranking-based approach, the prospectively-relevant representation

could not be decoded. As previous studies that use highly sensitive analysis techniques have been able to decode prospectively-relevant information in occipital areas, it seems that our evaluation method was not able to adequately capture this representation in each brain area. Therefore, it might be possible that we missed other existing findings due to the chosen analysis techniques.

Secondly, it could be that we needed more data to train the decoding model on. The prospectively-relevant representation is associated with a relatively weak signal in brain activity, which means a decoding model needs a larger amount of training data to successfully decode its representation compared to the robust currently-relevant representation.

Thirdly, participants were analyzed separately, which means our results are difficult to compare with similar studies conducting group-level analyses (Christophel et al., 2018; Rademaker et al., 2019). It is possible that the weaker prospectively-relevant representation shows more successful decoding results throughout the cortex if the data of more participants were collected and analyzed at the group-level. Nonetheless, seeing we have been able to decode this weaker representation on the individual level in two of the three participants by simply training and testing on the prospectively-relevant representation highlights the sensitivity of our approach.

Fourthly, for significance testing whole-brain correction was applied, making it highly conservative. Because of this conservativeness, relatively weak decoding results in certain locations might have been deemed as insignificant, such as the prospectively-relevant representation in occipital areas.

Moreover, because a searchlight analysis was used we have only been able to analyze information retained in local clusters of voxels. It could be that prospectively-relevant items are represented by larger, more distributed neural activation patterns. In this case, the searchlight analysis would not be able to decode the prospectively-relevant representation, as it is unfit to capture information encoded in a more distributed fashion across multiple brain regions (Haynes, 2015).

Apart from the inability to decode prospectively-relevant representations in all expected locations, there are a few limitations to be pointed out. Firstly, since we only analyzed three participants it is not possible to compute and report group-level statistics. This makes it difficult to generalize any found results to the population. Secondly, we have analyzed the maintenance of stimulus features, specifically stimulus orientations. It could be the case that different stimulus features are maintained differently in VWM. Previous research has already shown different findings in VWM maintenance for stimulus location for instance (Yu et al., 2020). Thirdly, the retention period did not include anything but a fixation point and was kept at a stable length of 8 seconds across all trials. This is not representative for real-world tasks, where distractions and uncertainty about retention time can be present. For instance, it could be the case that we found VWM representations to be maintained in a stable representational format because no distraction was present and retention time was stable. Alternatively, when interacting with a dynamic and unpredictable environment it could be that the strength of the memory representation fluctuates over the course of a retention period

of an unknown length. In the presence of distracting visual information, dynamic transformations of VWM representations could be happening to mitigate interference and thus memory decay. Thus, incorporating distraction and varying retention times into an experimental design could offer a more comprehensive understanding of VWM maintenance.

4.5 Future research

In the following section several directions for future research are discussed, based on the found results of the current study. Firstly, the presence of orthogonal storage as shown by cross-decoding does not reveal which of the two representations has undergone a transformation to an orthogonal representation. Earlier cross-temporal analyses have suggested it is the prospectively-relevant representation that undergoes these transformations during the retention period (Wan et al., 2020). This is likely a strategy to mitigate interference of the prospectively-relevant representation with maintenance of the currently-relevant representation in anticipation for the recall task. Furthermore, our found results suggest stable maintenance during the retention period of currently-relevant representations. This serves as an indication that it is indeed not the currently-relevant, but the prospectively-relevant representation that is transformed during maintenance. It would be interesting to apply the cross-temporal decoding procedure to the prospectively-relevant representation. This might confirm the transformation to an orthogonal representational format, as seen by our cross-decoding results. It could also be checked if the representation transforms back to its initial representational format once it becomes currently-relevant, similarly found by Van Loon et al. (2018). Additionally, it would be interesting to look into whether opposite representations in neural activation patterns are also present with other stimulus features (e.g. color, shape) or even other modalities (e.g. auditory working memory). While for stimulus orientations opposite representational formats are realized by a 90° rotation, it is unclear how other features and/or modalities could be transformed to opposite representational space and whether this is occuring.

Moreover, for the current study we have analyzed the similarity between currently- and prospectively-relevant representational formats. A more thorough look into their respective representational formats could reveal more about VWM maintenance and the differences between task-relevance states. Specifically, it could be examined to what extent currently- and prospectively-relevant representations mirror the sensory representation during initial encoding of the stimulus. This could add to the discussion surrounding the sensory recruitment hypothesis, which states VWM representations share the same neural code as their corresponding sensory representations facilitating comparison with incoming visual input (Serences et al., 2009; Albers et al., 2013; Bettencourt et al., 2016; Rademaker et al., 2019). The sensory recruitment hypothesis is assumed to not apply to prospectively-relevant representations, as these representations are not currently task-relevant and thus there is no need to compare with incoming visual input. This notion is also in line with the results of the current study, where the currently-relevant representation is present

in the sensory-related areas and the prospectively-relevant representation is absent in these areas. However, previous studies have been successful in decoding the prospectively-relevant representation in EVC (or more generally in the occipital lobe), posing the question of whether these representations are in fact stored in sensory-like code in these visual areas (Yu et al., 2020; Iamshchinina et al., 2021; Ruijs, yet unpublished results). It could be examined to what extent VWM representations generalize to their sensory representations using a cross-decoding setup in which the decoding model is trained on the VWM representation and tested on the sensory representation. With the current dataset this cross-decoding setup is difficult to implement because the stimuli (i.e. the Gabor gratings) were presented both for only 0.8 seconds and in succession of each other. We were unable to adequately differentiate between the two sensory representations in analysis. Obtaining separate measurements of stimulus presentations (without a memory task) during longer intervals could ensure a better capture of the sensory representations during initial encoding, which could be used in the cross-decoding setup. Furthermore, it can be examined whether the similarity to sensory-like code changes over the course of retention and in which brain areas this can be found. This could reveal more about the possible intricacies of the sensory recruitment hypothesis and its relation to prospectively-relevant representations.

Lastly, we have seen a difference in storage location of currently-relevant content between the first and second retention period. Currently-relevant information was only found to be present in the frontal lobe during the second retention period, not during the first retention period. This could mean that currently-relevant information in the absence of prospectively-relevant information is stored differently than in the presence of prospectively-relevant information. In other words, it could be that the brain handles the storage of a single memory item differently than the storage of a currently-relevant item in the presence of a prospectively-relevant item. With the current dataset, this hypothesis is difficult to test because the second retention period could be biased by the previous events of the trial. In a repeat-trial, the already robust currently-relevant representation could be made even stronger after the second retro-cue, while in a switch-trial the currently-relevant representation had to be reinstated from a prospectively-relevant state. Thus, an experimental setup that only includes the maintenance of a single memory item can facilitate a proper comparison with multi-item storage. This could give an additional insight into the dynamics of VWM maintenance.

4.6 Concluding remarks

Using a combination of sensitive analysis techniques and an assumption-free searchlight approach we have decoded VWM representations while taking its dynamic nature into account. To allow meaningful interaction with the environment and adequately prepare for imminent goal-directed behavior, the human brain can handle the storage of memory items differently. Specifically, our results show the brain discriminates between memory items based on their task-relevance status by storing these in different locations in the brain. Only currently-relevant memory items are stored in

sensory-associated areas, likely to ensure a high-resolution representation in anticipation for the upcoming task as formulated by the sensory recruitment hypothesis. Alternatively, prospectively-relevant memory items are stored in parietal and frontal areas, likely to abstract away from their high-resolution format and to avoid interference with the currently-relevant memory item. To further reduce disruption, currently- and prospectively-relevant memory items can even be stored in orthogonal representational formats when they are concurrently stored in the same brain area. The fact that orthogonal storage has not been found in each participant could reflect individual differences in cognitive control of attention related to VWM maintenance. Furthermore, to facilitate active VWM maintenance in a dynamic environment and avoid memory decay, currently-relevant memory items are kept in a stable representational format and do not undergo dynamic transformations. The active, stable memory trace likely reflects the focus of attention to anticipate the usage of the memorized information for upcoming goal-directed behavior.

To effectively adapt to changes in the environment and thus to changes in self-initiated goals, the human brain can discriminate between memory items with a different task-relevance status by their storage location in the brain and their representational format in relation to each other. Moreover, in the ever-changing environment the brain retains currently task-relevant visual information in a stable representational format to adequately complete the upcoming task. These results emphasize the importance of considering VWM as a dynamic system, allowing meaningful interaction with the dynamic world around us.

References

van Ackooij, M., Paul, J., van Helden, J., Hendrikx, E., Gayet, S., van der Stoep, N., & Harvey, B. (2023). Tuned responses to visual short-term memory load in a frontoparietal topographic map hierarchy. https://doi.org/10.21203/rs.3.rs-2560452/v1

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & De Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427-1431. https://doi.org/10.1016/j.cub.2013.05.065

Baddeley, A. (1998). Working memory. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie, 321(*2-3), 167-173. https://doi.org/10.1016/S0764-4469(97)89817-4

Bettencourt, K. C., & Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature neuroscience*, *19*(1), 150-157. https://doi.org/10.1038/nn.4174

Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T., & Kennerley, S. W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature communications, 9*(1), 3498. https://doi.org/10.1038/s41467-018-05873-3

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*, *21*(2), 111-124. https://doi.org/10.1016/j.tics.2016.12.007

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J. D. (2018). Cortical specialization for attended versus unattended working memory. *Nature neuroscience*, *21*(4), 494-496. https://doi.org/10.1038/s41593-018-0094-4

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia, 49*(6), 1407-1409. https://doi.org/10.1016/j.neuropsychologia.2011.01.029

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron, 87*(4), 893-905. https://doi.org/10.1016/j.neuron.2015.07.013

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: promise, pitfalls, and potential. *Neuroimage, 78*, 261-269. https://doi.org/10.1016/j.neuroimage.2013.03.041

Fallon, S. J., Zokaei, N., & Husain, M. (2016). Causes and consequences of limitations in visual working memory. *Annals of the New York Academy of Sciences, 1369*(1), 40-54. https://doi.org/10.1111/nyas.12992

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology, 61*(2), 331-349.

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science, 173*(3997), 652-654.

Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J. D., Paffen, C. L., Van der Stigchel, S., & Sterzer, P. (2017). Visual working memory enhances the neural response to matching visual input. *Journal of Neuroscience*, *37*(28), 6638-6647. https://doi.org/10.1523/JNEUROSCI.3418-16.2017

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., … & Dosenbach, N. U. (2017). Precision functional mapping of individual human brains. *Neuron, 95*(4), 791-807. https://doi.org/10.1016/j.neuron.2017.07.011

Gulbinaite, R., Johnson, A., de Jong, R., Morey, C. C., & van Rijn, H. (2014). Dissociable mechanisms underlying individual differences in visual working memory capacity. *Neuroimage, 99*, 197-206. https://doi.org/10.1016/j.neuroimage.2014.05.060

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*, 632-635. https://doi.org/10.1038/nature07832

Haynes, J. D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron, 87*(2), 257-270. https://doi.org/10.1016/j.neuron.2015.05.025

van den Hurk, J., & Op de Beeck, H. P. (2019). Generalization asymmetry in multivariate cross-classification: When representation A generalizes better to representation B than B to A. BioRxiv, 592410. https://doi.org/10.1101/592410

Iamshchinina, P., Christophel, T. B., Gayet, S., & Rademaker, R. L. (2021). Essential considerations for exploring visual working memory storage in the human brain. *Visual Cognition*, *29*(7), 425-436. https://doi.org/10.1080/13506285.2021.1915902

Kerzel, D., & Witzel, C. (2019). The allocation of resources in visual working memory and multiple attentional templates. *Journal of Experimental Psychology: Human Perception and Performance, 45*(5), 645. https://doi.org/10.1037/xhp0000637

Korhonen, O., Saarimäki, H., Glerean, E., Sams, M., & Saramäki, J. (2017). Consistency of regions of interest as nodes of fMRI functional brain networks. *Network Neuroscience, 1*(3), 254-274. https://doi.org/10.1162/NETN_a_00013

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience, 12*(5), 535-540. https://doi.org/10.1038/nn.2303

LaRocque, J. J., Riggall, A. C., Emrich, S. M., & Postle, B. R. (2017). Within-category decoding of information in different attentional states in short-term memory. *Cerebral Cortex*, *27*(10), 4881-4890. https://doi.org/10.1093/cercor/bhw283

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of cognitive neuroscience*, *24*(1), 61-79. https://doi.org/10.1162/jocn_a_00140

Liu, J., Zhang, H., Yu, T., Ni, D., Ren, L., Yang, Q., Lu, B., Wang, D., Heinen, R., Axmacher, N. & Xue, G. (2020). Stable maintenance of multiple representational formats in human visual

short-term memory. *Proceedings of the National Academy of Sciences*, *117*(51), 32329-32339. https://doi.org/10.1073/pnas.2006752117

van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *ELife*, *7*, e38677. https://doi.org/10.7554/eLife.38677

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in cognitive sciences, 17(8), 391-400. https://doi.org/10.1016/j.tics.2013.06.006

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of neuroscience, 16*(16), 5154-5167. https://doi.org/10.1523/JNEUROSCI.16-16-05154

Offen, S., Schluppeck, D., & Heeger, D. J. (2009). The role of early visual cortex in visual short-term memory and visual attention. *Vision research, 49*(10), 1352-1362. https://doi.org/10.1016/j.visres.2007.12.022

Oh, B. I., Kim, Y. J., & Kang, M. S. (2019). Ensemble representations reveal distinct neural coding of visual working memory. *Nature communications*, *10*(1), 5665. https://doi.org/10.1038/s41467-019-13592-6

Polanía, R., Paulus, W., & Nitsche, M. A. (2012). Noninvasively decoding the contents of visual working memory in the human prefrontal cortex within high-gamma oscillatory patterns. *Journal of Cognitive Neuroscience, 24*(2), 304-314. https://doi.org/10.1162/jocn_a_00151

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature neuroscience*, *22*(8), 1336-1344. https://doi.org/10.1038/s41593-019-0428-x

Pearson, J., & Keogh, R. (2019). Redefining visual working memory: A cognitive-strategy, brain-region approach. *Current Directions in Psychological Science, 28*(3), 266-273. https://doi.org/10.1177/0963721419835210

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136-1139. https://doi.org/10.1126/science.aah7011

Ruijs, F. R. A. M., (yet unpublished results). *Decoding attended and unattended visual information in visual working memory*. Utrecht University. https://studenttheses.uu.nl/handle/20.500.12932/43773

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current opinion in neurobiology, 58*, 141-147. https://doi.org/10.1016/j.conb.2019.08.005

Sahan, M. I., Sheldon, A. D., & Postle, B. R. (2020). The neural consequences of attentional prioritization of internal representations in visual working memory. *Journal of Cognitive Neuroscience, 32*(5), 917-944. https://doi.org/10.1162/jocn_a_01517

Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience, 38*(21), 4859-4869. https://doi.org/10.1523/JNEUROSCI.3440-17.2018

Scotti, P. S., Chen, J., & Golomb, J. D. (2021). An enhanced inverted encoding model for neural reconstructions. bioRxiv, 2021-05. https://doi.org/10.1101/2021.05.22.445245

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological science*, *20*(2), 207-214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage, 44*(1), 83-98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Sprague, T. C., Adam, K. C., Foster, J. J., Rahmati, M., Sutterer, D. W., & Vo, V. A. (2018). Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *Eneuro, 5*(3). http://dx.doi.org/10.1523/ENEURO.0098-18.2018

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage, 65*, 69-82. https://doi.org/10.1016/j.neuroimage.2012.09.063

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron, 78*(2), 364-375. https://doi.org/10.1016/j.neuron.2013.01.039

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in cognitive sciences*, *19*(7), 394-405. https://doi.org/10.1016/j.tics.2015.05.004

Super, H. (2003). Working memory in the primary visual cortex. *Archives of neurology, 60*(6), 809-812. https://doi.org/10.1001/archneur.60.6.809

Torrisi, S., Chen, G., Glen, D., Bandettini, P. A., Baker, C. I., Reynolds, R., Yen-Ting Liu, J., Leshin, J., Balderston, N., Grillon, C., & Ernst, M. (2018). Statistical power comparisons at 3t and 7t with a GO/NOGO task. *NeuroImage, 175*, 100–110. https://doi.org/10.1016/j.neuroimage.2018.03.071

Wan, Q., Cai, Y., Samaha, J., & Postle, B. R. (2020). Tracking stimulus representation across a 2-back visual working memory task. *Royal Society open science*, *7*(8), 190228. https://doi.org/10.1098/rsos.190228

Wan, Q., Menendez, J. A., & Postle, B. R. (2022). Priority-based transformations of stimulus representation in visual working memory. *PLOS Computational Biology*, *18*(6), e1009062. https://doi.org/10.1371/journal.pcbi.1009062

Wang, X. J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in neurosciences, 24*(8), 455-463. https://doi.org/10.1016/S0166-2236(00)01868-3

Willems, T., & Henke, K. (2021). Imaging human engrams using 7 Tesla magnetic resonance imaging. *Hippocampus, 31*(12), 1257-1270. https://doi.org/10.1002/hipo.23391

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature neuroscience*, *20*(6), 864-871. https://doi.org/10.1038/nn.4546

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. *PLoS biology*, *18*(3), e3000625. https://doi.org/10.1371/journal.pbio.3000625

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS biology*, *18*(6), e3000769. https://doi.org/10.1371/journal.pbio.3000769