Running head: Semantic Knowledge & Visual Working Memory

Associating Everything with Everything Else, All at Once: Semantic Associations Facilitate Visual Working Memory Formation for Real-World Objects

Xinchi Yu[1,2], Sanikaa P. Thakurdesai[3], Weizhen Xie[1,3*]

1. Program in Neuroscience and Cognitive Sciences, University of Maryland, College Park
2. Department of Linguistics, University of Maryland, College Park
3. Department of Psychology, University of Maryland, College Park

***Correspondence**: Weizhen Xie, zanexie@umd.edu

Word counts: 5660 (main text)

Author notes: This study is made possible by funding support from the National Institute of Neurological Disorders and Stroke (K99NS126492, PI: W. X.).

Abstract

Integrating prior semantic knowledge with environmental information is essential for everyday cognition, yet how this process affects ongoing perception and memory remains unclear. We investigate this by studying how associative semantic knowledge interacts with perceptual constraints induced by brief encoding times, thereby supporting visual working memory (VWM) for real-world objects. Study 1 reanalyzed data from Quirk et al. (2020), involving 75 participants across 13,750 trials of a VWM task with randomly chosen objects and verbal distraction. We found that objects' semantic associations, estimated by a natural language processing model, predicted trial-level VWM accuracy under brief but not prolonged encoding times (0.2s vs. 1-2s). These results, unaffected by image similarity from computer vision models, were replicated in Study 2 with 50 participants across 11,880 trials. Combined, these findings suggest that semantic associations facilitate effective grouping among VWM items to mitigate perceptual constraints, highlighting the role of semantic knowledge in VWM formation.

Keywords: visual working memory, long-term memory, prior knowledge, semantic association, natural language processing

Public significance

This study demonstrates that human observers frequently use their prior associative semantic knowledge to enhance memory when viewing brief visual displays of random everyday objects. These findings underscore the role of prior knowledge in overcoming perceptual processing limitations, thereby supporting effective everyday visual cognition.

Understanding how prior knowledge integrates with environmental information to sustain ongoing perception and memory remains a critical challenge in comprehending everyday cognition (Chase & Simon, 1973; Collins & Olson, 2014; Potter, 2012; Smith et al., 2022). Inquiry into this challenge has gradually shifted research from well-controlled laboratory stimuli to more naturalistic stimuli, such as images of real-world scenes and objects, highlighting the dynamic interplay between prior knowledge and moment-by-moment cognition (Brady et al., 2019; Xie & Zhang, 2022). For instance, recent studies have demonstrated that everyday objects may be memorized more effectively than simple stimulus features such as colors in visual working memory, VWM (Brady et al., 2016; Chung et al., 2023; but see Li et al., 2020; Quirk et al., 2020) – a capacity-limited mental faculty crucial for higher cognition such as fluid intelligence (Cowan, 2001; Luck & Vogel, 2013). These findings prompt research into computer vision algorithms that capture stimulus-driven attributes in VWM processes (Brady & Störmer, 2023; Liu et al., 2020) and observer-focused research emphasizing prior knowledge – either from past encounters driving visual familiarity (Jackson & Raymond, 2008; Xie & Zhang, 2017c, 2018), from the meaningfulness/integrity of a visual item relative to its visual distortion (Asp et al., 2021; Chung et al., 2024; Sahar et al., 2024; Thibeault et al., 2024), or from the categorical knowledge about an object's class information (Chiou & Lambon Ralph, 2018; Endress & Potter, 2014; Hu & Jacobs, 2021; Markov et al., 2021; Potter, 2012; Wong et al., 2008).

Compared to these progresses, surprisingly little is known concerning the influence of our prior associative semantic knowledge on VWM. Separate from the class information of

object categories, semantic associations can link any arbitrary pair of concepts, reflecting a distributed process of how information is organized in our minds for more efficient computation (Klix, 1978; Kumar, 2021; McClelland & Rogers, 2003; Sipser, 2012). This form of knowledge begins to take root early in life, coinciding with the onset of language acquisition during infancy (Barbir et al., 2023; Lany & Saffran, 2011). As it continues to reinforce through our written and spoken communication, semantic associations exert a pervasive impact on our perception (Xie & Zhang, 2023b) and memory (Tompary & Thompson-Schill, 2021; Xie et al., 2020). For example, semantic associations can effectively group perceptually distinct stimuli within a single display (Green & Hummel, 2006; Nah & Geng, 2022; Roberts & Humphreys, 2011), serving as a chunking mechanism underlying the interaction between pre-existing long-term memory and enhanced VWM storage (Cowan, 2001; Gobet et al., 2001). Yet, despite this intuition, except for some context-specific associations (e.g., using a "*hammer*" to drive a "*nail*" under the action/function context, O'Donnell et al., 2018; Humphreys & Riddoch, 2006; e.g., a "*hair blower*" is not frequently seen in a "*forest*", Võ, 2021), a notable gap persists in understanding how human observers leverage diverse, arbitrary semantic associations to mitigate VWM constraints, potentially due to various known challenges.

First, given the vast number of possible combinations between things in everyday life, it is challenging to identify the strength of associations between any arbitrary pairs of object-related concepts. Some of these arbitrary associations may carry little explicit action or functional meaning, yet they remain semantically related to each other (e.g., "*bunny*" and

"*Easter egg*"). Revealing these arbitrary semantic associations in our everyday language has only become feasible with the advancements in natural language processing (NLP) models, such as the GloVe (Global Vector for Word Representation, Pennington et al., 2014). GloVe is an effective model that processes word-word co-occurrences in large corpora from printed and online texts to extract associations between any concepts in our everyday language (Grand et al., 2022; Pennington et al., 2014). Although features from feed-forward convolutional neural network (CNN) models trained solely on pixel-level information of images can capture the relationship between objects based on their visual similarity within the same semantic category (Simonyan & Zisserman, 2015), NLP word embeddings capture semantic associations between any arbitrary pair of objects across any categories, regardless of their visual resemblance. Using these word embeddings, recent research has suggested that properties of semantic associations can modulate early visual perception (Xie & Zhang, 2023b) and episodic memory retrieval (Xie et al., 2020), even when perceptual contributions to these processes are strictly controlled. Nonetheless, these models or analyses have yet been applied to VWM research, leaving a gap in our understanding concerning the role of semantic associations among everyday objects in VWM and how it may differ from the contribution of image similarity predicted by CNN features (Brady & Störmer, 2023).

Second, the interplay between semantic associations and perceptual processing during VWM formation has remained unclear. Past research has implied that semantic and perceptual factors could exert either a compensatory or addictive impact on VWM. On the one hand, as retaining multiple pieces of information in VWM is inherently effortful (Xie &

Zhang, 2023a), semantic associations – as higher-order information of a visual display with multiple objects – may not become relevant when each individual object is sufficiently encoded. This compensatory relationship would predict more semantic contribution to VWM only when perceptual encoding is limited. On the other hand, it also remains possible that leveraging semantic associations may reflect a deeper level of processing following sufficiency perceptual encoding (Brady & Störmer, 2022; Graham & Golan, 1991). Thus, an alternative account would predict that limited perceptual encoding may hinder the extraction of semantic associations and their subsequent impacts on VWM. Testing these possibilities requires manipulating perceptual processing during VWM formation, for example, by adjusting the time allowed for VWM encoding (e.g., Ye et al., 2024).

The current study therefore aims to investigate how semantic associations among everyday objects – quantified as NLP word embedding similarities across object labels – interacts with perceptual encoding constraints induced by different encoding times to affect VWM. Relevant data meeting these criteria for testing the proposed hypotheses are available from Quirk and colleagues (2020), where participants attempted to encode a multi-object display in a VWM task within a limited (0.2s) or longer (>=1s) encoding time. As Quirk and colleagues have primarily focused on contrasting participants' VWM for these real-world objects with that for the simpler stimulus feature of colors, they have not tested the role of semantic associations among objects during VWM formation. We therefore first analyzed Quirk and colleague's data with a fresh perspective. This is followed by a pre-registered study (https://osf.io/t3nf5) with a similar design to further test our hypotheses. Combined, as

we gather evidence based on existing data with a replication, our findings underscore structured semantic knowledge – operationalized as semantic associations in NLP models – as an important avenue to enrich our understanding of everyday VWM function in real-world contexts.

Methods

Participants

Study 1. We used the dataset from a previous study (Quirk et al., 2020; https://osf.io/vq37u/), where 75 participants completed a VWM task involving real-world objects and colored squares in 3 behavioral experiments (i.e., Experiments 1a-c in the original paper). These participants were reported to be aged 18-31 with self-reported normal or corrected-to-normal visual acuity and normal color vision and completed the study for monetary compensation ($10/h). Data from these participants were included for these following reasons. First, all participants had completed a VWM task with real-world objects randomly chosen and presented in a single display. Second, the authors have provided trial-level data with information regarding which objects had been presented during the study. Third, these experiments also included a within-subject manipulation of encoding time, containing a relatively limited and a longer encoding time (e.g., 0.2s vs. >1s) within a conventional VWM task. Fourth, the research sample were tested within an English-speaking environment, considering that the current NLP model was trained and tested within an English-speaking context.

Apart from Quirk et al. (2020, Experiments 1a-c), as of the time in preparation of the manuscript, there have been a few other studies that have tested VWM for multiple real-world objects in a single visual display (Brady et al., 2016; Brady & Störmer, 2022; Chung et al., 2023; Li et al., 2020; Thibeault et al., 2024; Experiment 2 in Quirk et al., 2020). These additional studies were not included for the current research due to at least one of these following reasons: (1) raw data were not available (Brady et al., 2016); (2) participants were not English speakers (Li et al., 2020); or (3) the experiment only involved one encoding time condition, making the within-subject contrast between encoding time conditions implausible (Brady & Störmer, 2022; Chung et al., 2023; Thibeault et al., 2024; Experiment 2 in Quirk et al., 2020).

Study 2. We recruited 50 college students (19.96 ± 1.56 [mean ± SD] years old; 34 females, 16 males) from the University of Maryland, College Park, who reported normal or corrected-to-normal vision and being native English speakers. They participated in the study for course credits, following a protocol approved by the local IRB. This study was pre-registered (https://osf.io/t3nf5) with the goal of replicating the findings of Study 1. All data and analytical codes are available online via the Open Science Framework: https://osf.io/v3ckh/.

Stimuli

Following prior research (Brady et al., 2016; Quirk et al., 2020), images of real-world objects from 200 distinct categories in our everyday visual environment (e.g., couch, cup, bag, etc.) were used for the current study. Each of the 200 category-specific image sets

contains 15 perceptually distinct exemplars, yielding a total of 3000 images of objects in the entire stimulus set (http://bradylab.ucsd.edu/stimuli.html). Each trial contained a study and a test display of these everyday objects. For the study display, 6 categorically distinct objects were randomly chosen from the 200 category-specific image sets. For the test display, the target object would repeat from the study display, whereas the foil object was randomly selected from any of the categories not included in the study array. Stimuli were displayed on an invisible ring in fixed, equidistant positions. Stimuli were generated and displayed on a white background using MATLAB and the Psychtoolbox (Brainard, 1997) in both Study 1 and 2.

Procedures

All participants completed a VWM task with concurrent verbal distraction with a similar task structure as outlined in **Figure 1A**. On each trial, participants first saw two digits for 1s, for which they would need to rehearse explicitly or silently in mind throughout the trial. Afterwards, 6 placeholder dots around the center of the screen appeared for 1s, followed by six real-world objects randomly sampled from 200 distinct semantic categories (15 unique objects/category) at the corresponding locations. Within each experiment, these objects were presented for either 0.2s, 1s or 2s (see **Table 1** for task design across different experiments), followed by blank screen of 6 dots for 0.8s. To prepare the participants for response, one of the placeholders enlarged for 0.5s, cuing the location that would be later tested (Quirk et al., 2020), making the delay period a total of 1.3s. Upon the test, two objects were presented above and below the cued location, one matching with the object previously shown at that

location, and the other being randomly sampled from another unshown object category to act as the foil. Participants were instructed to press a set of buttons (e.g., UP or DOWN arrow keys) to indicate which item appeared in the study display at that location. Afterwards, participants tried to make a response based on the digits that they were rehearsing by either typing them out (Study 1) or recognizing whether a newly presented digit pair was the same or different as compared with the previous ones at the beginning of the trial (Study 2). In the latter case, half of the trials contained the same digit pair whereas another half of trials contained a different pair of digits, and participants used another set of buttons to make a match-or-not judgment (e.g., LEFT or RIGHT arrow key for match and not-match, respectively). In Study 1, participants also performed a similar VWM task using colored squares instead of real-world objects (Quirk et al., 2020). These different stimulus types were blocked and randomly intermixed throughout a testing session. For the current study, we only retained the data using the real-world objects for subsequent analyses.

Study 1 also employed multiple experimental designs to test the generalization of experimental findings across different conditions (Quirk et al., 2020). These designs varied by employing different encoding times in the VWM task (0.2, 1, 2s in Experiment 1a; 0.2 and 2s in Experiments 1b-c), by either blocking or randomly intermixing different encoding times within a block (Experiments 1a-b used blocking, while Experiment 1c used random intermixing), and by implementing explicit or silent verbal rehearsal for the concurrent digit task (Experiment 1b). We coded these design variables as potential predictors of participants' trial-level performance, such as encoding time as either short (0.2s) or long (1-2s). Since

explicit or silent verbal rehearsal did not affect task performance (Quirk et al., 2020), we combined trials with explicit and silent verbal rehearsal tasks. Other experimental factors, including the location of the presented objects, the categories and identities of individual objects, and the tested locations, were randomly chosen across trials. Participants completed a minimum of 100 trials per encoding time condition to allow reliable estimates of VWM task performance, yielding a total of 13,750 trials across participants (see **Table 1** for details).

Study 2 used the same experimental design as in Study 1, with the following exceptions. First, to focus on VWM for real-world objects, we only retained the object condition in Study 1. Second, to ensure consistency with Experiments 1b-c in Study 1, we only included two encoding time conditions, presenting objects for either 0.2s or 2s in a blocked design. The order of these different conditions/blocks was randomized across participants. Third, we employed a 5s time limit for the VWM response for the interest of time. Fourth, participants completed blocks of 60 trials for 4-6 blocks depending on the availability of testing time within a 1h experimental session, yielding 120 or 180 trials per encoding time condition and a total of 11,880 trials across participants (see **Table 1** for details).

Modeling Trial-level Features

To identify how the semantic associations among presented real-world objects affect VWM formation, we first quantified the strength of semantic association between any pair of presented objects in each trial based on the GloVe word embeddings of the objects' labels. Additionally, to account for other trial-level factors, for example, encoding time, the image-

based visual resemblance of the presented objects, and the image or semantic similarity between the presented target and foil items, we extracted these trial-level measures based on the methods outlined below. These trial-level predictors were then included in a general linear mixed model as outlined in *Statistical Analysis* to predict participants' trial-by-trial VWM recognition accuracy.

Semantic Associations Among Study Objects

On each trial, we defined and calculated the semantic association between any non-redundant object pairs as the cosine similarity of the GloVe word embeddings of each object's categorical labels. These categorical labels are available in the original stimulus set, and we also have inspected these labels and edited them when necessary to ensure that they accurately reflect the presented real-world objects. As some object categories may be better described using more than one word (e.g., "cooking pan"), we retained the word embeddings for each individual word (e.g., "cooking" and "pan"). With a visual display of 6 randomly selected objects, there may be more than 6 words to describe these items. For each pair words *i* and *j*, namely W_i and W_j, we calculated their cosine similarity to capture their semantic association, as follows,

$$S(W_i, W_j) = \frac{W_i \cdot W_j}{||W_i||\,||W_i||}$$

We first obtained an estimate of semantic association between a pair of objects by averaging $S(W_i, W_j)$ over all non-redundant word pairs between the two categories. We then computed the average of these cosine similarity measures across all non-redundant pairs of objects as a measure of the overall semantic association across all objects in a visual display.

By chance, some visual display would contain objects that are mostly semantically related with one another (e.g., "*bunny*" and "*Easter egg*", "*hat*" and "*shoe*" on the *right* side of **Figure 1B**), whereas other is less semantically related (e.g., "*chessboard*", "*compass*", and "*microscope*" on the *left* side of **Figure 1B**). There is, therefore, a continuum of semantic association strength among objects within a given multi-object display, suggesting a source of variance for trial-by-trial VWM recognition accuracy.

Image Similarity Among Study Objects

Based on a similar approach, we also attempted to quantify the visual resemblance among the presented objects based on a pre-trained feedforward CNN model, namely VGG-16 (Simonyan & Zisserman, 2015). VGG-16 is trained based on the pixel-level information of images containing real-world objects, making it an appropriate model in the current context. Although layers of VGG-16 features have been used to infer perceptual or conceptual information underlying object recognition and visual memory (Rust & Mehrpour, 2020; Simonyan & Zisserman, 2015), they primarily capture the image-level information available for visual categorization, considering that objects from the same category tend to be visually similar. This factor is in sharp contrast with the GloVe word embeddings that contain little information regarding the visual resemblance of concepts; instead, the semantic association estimate from GloVe completely abstracts away from any visual inputs. To distinguish these factors in predicting trial-level VWM success, we therefore quantified each object image's VGG-16 features of the last max pooling layer (Brady & Störmer, 2023) and calculated the average cosine similarity of these VGG-16 features between non-redundant

object pair as a measure of visual similarity of the object images within a given display. Considering that this image-driven metric may capture primarily categorical information instead of semantic associations among objects, it serves as a good control variable in the current study to parse out image-related variances in participants' VWM task performance.

Semantic Association and Image Similarity between Target and Foil Objects at Test

To factor out the variance in trial-level VWM recognition performance driven by how similar or dissimilar the target and foil items were to each other (Brady & Störmer, 2023; Xie & Zhang, 2017b), we quantified both the semantic association and image similarity between the target and foil objects at test for each trial. These metrics can be directly calculated as the cosine similarity between the target and foil items, based on their GloVe word embeddings or VGG-16 features respectively, serving as additional covariates.

Statistical Analyses

We used a generalized linear mixed model to fit trial-level data to predict participants' trial-by-trial VWM recognition performance, which was coded as 1 for correct and 0 for incorrect responses. This approach allows us to evaluate the variances accounted for by trial-level predictors, while partialling out the variances introduced by different participants or experimental designs. In this model, participants' overall likelihood of recognition success ($P_{success}$) across trials can therefore be directly modeled based on a linear combination of trial-level predictors with a logistic link,

$$\ln\left(\frac{P_{success}}{1-P_{success}}\right) \sim \beta_0 + \beta_{enc\_time} + \beta_{VGG\_Study} + \beta_{GloVe\_Study} + \beta_{VGG\_Test} + \beta_{GloVe\_Test} + \beta_{VGG\_Study\times enc\_time} + \beta_{GloVe\_Study\times enc\_time} + \beta_{VGG\_Test\times enc\_time} + \beta_{GloVe\_Test\times enc\_time} + 1|(\text{Subject: Experiment})$$

, where β_{enc_time} captures the variance accounted for by encoding time (coded as shorter: 0.2s vs. longer: >=1s); β_{GloVe_Study} and β_{VGG_Study} the average semantic association and image similarity among presented study objects, respectively; β_{GloVe_Test} and β_{VGG_Test} the semantic association and image similarity between the target and foil objects at test, respectively; and $\beta_{X\times enc\_time}$ the interaction effects between encoding time and the respective variables *X*. All continuous variables (e.g., semantic association or image similarity among all study objects or between target and foil objects at test) were standardized before regression. Model fitting was implemented via the *fitglme* function in MATLAB. To achieve a focal test on the planned effect related to the semantic factors, we compared the models without and with the two key semantic predictors, namely β_{GloVe_Study} and $\beta_{GloVe\_Study\times enc\_time}$ to evaluate the extent to which adding these terms is necessary to better account for participants' trial-by-trial VWM task performance. In this approach, the potential influence of other predictors on participants' VWM task performance is less consequential for our interpretation, as the variance related to these covariates has been factored out. Additionally, to enhance interpretation, we conducted separate regression analyses for data from each encoding time condition. Statistical significance was evaluated using Wald's Z test, and all reported p-values are two-tailed.

Results

Study 1: Quirk et al., 2020

We first re-analyzed the data from 3 experiments across 75 participants and 13,750 trials from Quirk et al. (2020). To ensure that participants actively retained the presented objects in VWM under concurrent verbal distraction, we restricted our analysis only to the trials where participants had successfully recalled the presented digits in each trial. Furthermore, we also excluded occasional trials where participants failed to respond using the instructed keys for the VWM task (< 0.1% of all trials). This reduced our trials count to 12,939 trials (94.10% of all 13,750 trials). As a sanity check, we first replicated the original study and others (Brady et al., 2016; Li et al., 2020; Quirk et al., 2020) and observed a clear enhancement effect of encoding time on VWM task performance (β = 0.59, SE = 0.045, Z = 12.98, $p = 2.94\times10^{-38}$). That is, participants' VWM recognition success is much higher under a longer encoding time of 1-2s (mean = 83.51%) relative to a shorter one of 0.2s (mean = 73.77%). Complementary analysis based on all 13,750 trials yield highly comparable findings (see **Supplementary Figure S1**).

Next, of primary interest, we then focus on the effect of semantic associations among objects in the study array, after controlling for various covariates including image similarity among study objects, semantic association between target and foil objects, and the image similarity between target and foil objects (see **Statistical Analyses** for the full general linear mixed model). We identified a significant main effect of the study objects' overall semantic association on VWM task performance (β = 0.087, SE = 0.031, Z = 2.78, p = 0.0054), in that

the higher average semantic association of a visual display, the more likely an observer would correctly recognize the target item after a short delay. Notably, there was a significant interaction effect between the study objects' overall semantic association and encoding time (β = -0.095, SE = 0.046, Z = -2.08, p = 0.038; see **Table 2**). As demonstrated in the *left* panel in **Figure 2A**, these effects were primarily driven by the greater improvement of VWM task performance driven by higher semantic associations among objects in the study display when it was presented with a shorter encoding time of 0.2s relative to a longer encoding time (1-2s). This was confirmed by separately analyzing the short and long encoding time trials (see **Table 3**). That is, for short encoding time trials, semantic association was a significant predictor for VWM accuracy (β = 0.087, SE = 0.031, Z = 2.81, p = 0.0050), improving VWM accuracy from below 70% to near 80% from a semantically unrelated visual display to a semantically related display. In contrast, this clear increase was absent for long encoding time trials, in that semantic associations among objects within the study display did not have a significant contribution to VWM recognition accuracy (β = -0.0066, SE = 0.034, Z = -0.20, p = 0.84).

Critically, the significant interaction effect between semantic associations among objects in the study display and encoding time could not be accounted for by other trial-level covariates. Based on a comparison between models without and with semantic association and its interaction with encoding time as additional predictors, we found that the latter provided a significantly better fit to the data (ΔAIC = -4, log-likelihood ratio test: $\chi^2_{(2)}$ = 7.76, p=0.021; see **Table 2**), suggesting unique semantic contributions to VWM formation

under perceptual processing constraints. Alternatively, can these accuracy effects be accounted for by response bias? We found this unlikely, considering that the mapping between the target location and the correct response was randomized across trials. Furthermore, additional analysis replacing participants' response accuracy in each trial with key responses corresponding to the top versus bottom objects (coded as 1 and 0 respectively) did not reveal any significant findings from the predictors, yielding chance-level predictions (see the *right* panel in **Figure 2A**). Hence, the current effect of semantic associations among study objects on VWM accuracy should not be confounded by response bias.

Study 2: Pre-registered Replication

Building upon Study 1, we next aimed to replicate the key interaction effect between the average semantic association among study objects of a visual display and encoding time on VWM formation. As we simplified the original task design by removing the color condition, we were able to test more trials per participant. Across 50 participants, we analyzed a total of 11,880 trials, where 10,822 trials contained correct responses for the concurrent digit task and proper key presses within a 5s response time window for the VWM task, accounting for 91.09% of all trials. Again, using these trials as an indicator for participants' active retention of the task content within a given trial, we replicated the widely observed enhancement effect of prolonged encoding time on VWM for real-world objects (β = 0.89, SE = 0.052, Z = 17.14, $p = 5.72\times10^{-65}$). Consistent with Study 1, participants' VWM recognition success was much higher under a longer encoding time of 2s (mean = 86.07%)

relative to a shorter one of 0.2s (mean = 73.00%). Data including all trials also provide similar findings (see **Supplementary Figure S1**).

Next, we examined the extent to which findings from Study 1 could be replicated in Study 2. Using the same analytical approach, we identified a significant main effect of semantic associations among study objects of the study display on VWM task performance (β = 0.14, SE = 0.032, Z = 4.38, $p = 1.18\times10^{-5}$). Critically, there was a significant interaction effect between semantic associations among study objects and encoding time (β = -0.12, SE = 0.053, Z = -2.34, p = 0.020; see **Table 2**). Again, these effects were primarily driven by a greater improvement in VWM recognition accuracy as a function of semantic associations among study objects under the short encoding time condition (β = 0.14, SE = 0.032, Z = 4.31, $p = 1.65\times10^{-5}$), relative to the long encoding time condition (β = 0.011, SE = 0.043, Z = 0.26, p = 0.79), as demonstrated in separate regression analyses (see **Table 3**). Consistent with the previous results, VWM recognition accuracy can increase by up to ~10% as the semantic association among objects within a visual display increase under a brief encoding time of 0.2s, whereas such an effect would disappear when the encoding time extends to 10 times longer (see the *left* panel in **Figure 1B**).

Furthermore, as these effects were estimated after controlling for covariates, factors like image similarity among study objects, semantic similarity between target and foil objects, and the image similarity between target and foil objects could not account for these results. Similarly, a comparison between models without and with semantic association and its interaction with encoding time as additional predictors suggested that adding trial-level

predictors associated with semantic associations among study objects could significantly better account for the data (ΔAIC = -15, log-likelihood ratio test: $\chi^2_{(2)} = 19.32$, $p = 6.38\times10^{-5}$, see **Table 2**), replicating the findings in Study 1. Furthermore, additional analysis replacing participants' response accuracy in each trial with key responses corresponding to the top versus bottom objects did not reveal any significant findings from the predictors (see the *right* panel in **Figure 2B**).

While these findings were on par with the results in Study 1, we also identified some nuances. For example, while data from Study 1 did not reveal an effect of the image similarity between target and foil on VWM recognition performance[1], some recent work suggest otherwise (Brady & Störmer, 2023). Results from Study 2 appears to be in line with these recent findings, in that visually similar foil can lead to worse VWM recognition, potentially due to the interference at the retrieval phase (β = -0.059, SE = 0.031, Z = -1.90, p = 0.058), despite with attenuated statistical evidence relative to the original study (Brady & Störmer, 2023). In contrast, semantically similar foil, however, did not lead to a similar detrimental effect for VWM accuracy in either Study 1 (β = -0.019, SE = 0.031, p = 0.55) or Study 2 (β = 0.017, SE = 0.032, Z = 0.55, p = 0.59; see **Table 2**). These results suggest a potential distinction between the relevance of semantic similarity and image similarity between target and foil items during VWM retrieval.

[1] In Study 1, this effect was numerically in the same direction (β = -0.023, SE = 0.031, Z = -0.73, p = 0.47) as that in Study 2 and in Brady & Störmer (2023).

Discussion

Human observers rarely encode real-world stimuli as mere data transferred onto a tabula rasa. A fundamental question in visual cognition within naturalistic settings is, therefore, how prior knowledge facilitates visual processing, particularly through associative semantic knowledge grounded in everyday language experiences. Building on a prior study (Quirk et al., 2020) with a replication (Study 2), our trial-by-trial analysis reveals previously undocumented yet robust influences of associative semantic knowledge on VWM formation across two separate research samples with a total of 125 participants across over 25,000 trials in 4 experiments. These results cannot be explained by image similarity captured by a feedforward computer vision model that categorizes images based solely on pixel-level information, idiosyncratic trial-level covariates such as the similarity between target and foil items at test, nor participants' response biases. As these findings emphasize the impact of semantic associations among real-world objects beyond mere class information on VWM formation, they hold significant implications for scaling up our understanding of visual cognition in the real world.

Phenomenologically, our results suggest that human observers leverage prior semantic knowledge cumulated from everyday language experiences to group or chunk multiple arbitrary visual objects to facilitate ongoing VWM processes. This influence of long-term associative semantic knowledge on VWM processing is separate from other forms of long-term memory influences, such as the effects of visual familiarity of individual items (Jackson & Raymond, 2008; Xie & Zhang, 2017b, 2017d), the integrity of an individual item

relative to visual distortion (Asp et al., 2021; Chung et al., 2024; Sahar et al., 2024; Thibeault et al., 2024), or categorical knowledge about an object's class information (Chiou & Lambon Ralph, 2018; Endress & Potter, 2014; Hu & Jacobs, 2021; Markov et al., 2021; Potter, 2012; Wong et al., 2008). Instead, these findings align more with the rapid extraction of configural information or pattern goodness across multiple visual objects during VWM formation (Brady & Alvarez, 2011; Howe & Brandau, 1983; Xie & Zhang, 2017a). Our findings extend these prior results from the perceptual domain to the conceptual domain, demonstrating that the contribution of configural information to VWM formation can occur rapidly within 0.2s at the semantic level.

Theoretically, why would such higher-order information be rapidly extracted during VWM formation? Conceptually, prior associative semantic knowledge may provide a cognitive map to scaffold the effective grouping or chunking of multiple perceptually distinct items in VWM (Gobet et al., 2001), thereby stabilizing these temporary memory information over time (Peer et al., 2021). If we assume that information in VWM is retained as a bound representation (Peterson et al., 2015; Thyer et al., 2022; Yu & Lau, 2023), remembering semantically associated items may help reduce the need to remember multiple unique VWM representations, thereby reducing VWM load. Despite this intuition, however, the role of semantic associations among *arbitrary* pairs of everyday objects in this VWM grouping process has remained largely underspecified until now. Except for various context-dependent associations typically seen in the context of action/function (Humphreys & Riddoch, 2006; O'Donnell et al., 2018) or specific scenes (Võ, 2021), in prior VWM research using real-

world objects, the role of diverse arbitrary associative semantic knowledge in VWM encoding is often overlooked, especially under the typical randomization procedures of object selection and concurrent verbal distraction. Our findings refine this understanding by revealing the significant effect of long-term semantic memory on VWM, delineating and replicating conditions under which this effect may emerge.

First, the contribution of semantic associations to VWM formation is more pronounced when perceptual processing is constrained by brief encoding times (e.g., 0.2s). Intuitively, prolonged encoding times may prompt deeper processing beneficial to integrating an ongoing memory item into prior associative knowledge (Graham & Golan, 1991). However, our findings favor an alternative account: longer encoding times may not provide ideal conditions for semantic contributions to VWM formation. Instead, semantic associations appear to be rapidly extracted within 0.2s of encoding time, contributing more to VWM formation when perceptual processing is interrupted, highlighting a compensatory relationship between semantic and perceptual information during visual memory formation (Naspi et al., 2023). Yet, our current study only tested a fixed set size and a limited range of perceptual encoding times. Future experiments with psychophysics methods can determine the necessary encoding times and set sizes for semantic effects to emerge or diminish.

Second, our findings indicate that the overall effects of semantic associations on VWM formation persist across variations in experimental conditions, for example, regardless of whether a concurrent digit task was explicitly or silently implemented as verbal

distraction, as found in Study 1. As semantic information is rapidly extracted to interact with perceptual processing, the digit task may not fully block verbal encoding. Instead, this task may serve as an active interference, urging participants to retain information in working memory before further interruption. Future research may further investigate the extent to which other experimental procedures, such as perceptual masking (Enns & Di Lollo, 2000) and other forms of dual task (e.g., addition of a concurrent physical load; Xie & Zhang, 2023a), may modulate this interplay between semantic and perceptual processing during VWM formation.

Collectively, our current findings illuminate how human observers optimize the encoding of novel visual information by balancing perceptual inputs and semantic knowledge in a coordinated manner. Articulating how the human brain supports this coordination remains a key issue for future research – an important step toward elucidating how human visual cognition operates in the real world.

References

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater Visual Working Memory Capacity for Visually Matched Stimuli When They Are Perceived as Meaningful. *Journal of Cognitive Neuroscience*, 1–17. https://doi.org/10.1162/jocn_a_01693

Barbir, M., Babineau, M. J., Fiévet, A.-C., & Christophe, A. (2023). Rapid infant learning of syntactic–semantic links. *Proceedings of the National Academy of Sciences*, *120*(1), e2209153119. https://doi.org/10.1073/pnas.2209153119

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 942–958. https://doi.org/10.1037/xlm0001014

Brady, T. F., & Störmer, V. S. (2023). Comparing memory capacity across stimuli requires maximally dissimilar foils: Using deep convolutional neural networks to understand visual working memory capacity for real-world objects. *Memory and Cognition*. https://doi.org/10.3758/s13421-023-01485-5

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7459–7464.

Brady, T. F., Störmer, V. S., Shafer-Skelton, A., Williams, J. R., Chapman, A. F., & Schill, H. M. (2019). Scaling up visual attention and visual working memory to the real world. *Psychology of Learning and Motivation - Advances in Research and Theory*, *70*, 29–69. https://doi.org/10.1016/bs.plm.2019.03.001

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

Chiou, R., & Lambon Ralph, M. A. (2018). The anterior-ventrolateral temporal lobe contributes to boosting visual working memory capacity for items carrying semantic information. *NeuroImage*, *169*(December 2017), 453–461. https://doi.org/10.1016/j.neuroimage.2017.12.085

Chung, Y. H., Brady, T. F., & Störmer, V. S. (2023). No Fixed Limit for Storing Simple Visual Features: Realistic Objects Provide an Efficient Scaffold for Holding Features in Mind. *Psychological Science*, *34*(7), 784–793. https://doi.org/10.1177/09567976231171339

Chung, Y. H., Tam, J., Wyble, B., & Störmer, V. S. (2024). Conceptual information of meaningful objects is stored incidentally. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001339

Collins, J. A., & Olson, I. R. (2014). Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin & Review*, *21*(4), 843–860.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General*, *143*(2), 548–565.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences*, *4*(9), 345–352.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Graham, S., & Golan, S. (1991). Motivational influences on cognition: Task involvement, ego involvement, and depth of information processing. *Journal of Educational Psychology*, *83*(2), 187–194.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01316-8

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1107–1119. https://doi.org/10.1037/0096-1523.32.5.1107

Howe, E. S., & Brandau, C. J. (1983). The temporal course of visual pattern encoding: Effects of pattern Goodness. *The Quarterly Journal of Experimental Psychology*, *35A*, 607–633.

Hu, R., & Jacobs, R. A. (2021). Semantic influence on visual working memory of object identity and location. *Cognition*, *217*, 104891. https://doi.org/10.1016/j.cognition.2021.104891

Humphreys, G. W., & Riddoch, M. J. (2006). Features, objects, action: The cognitive neuropsychology of visual object processing, 1984–2004. *Cognitive Neuropsychology*, *23*(1), 156–183. https://doi.org/10.1080/02643290542000030

Jackson, M. C., & Raymond, J. E. (2008). Familiarity Enhances Visual Working Memory for Faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556–568. https://doi.org/10.1037/0096-1523.34.3.556

Klix, F. (1978). On the presentation of semantic information in human long-term memory. In H. Ebbinghaus & A. König (Eds.), *Band 186, Heft 1 1978* (pp. 26–38). De Gruyter.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development: Interactions between statistical and semantic information. *Developmental Science*, *14*(5), 1207–1219. https://doi.org/10.1111/j.1467-7687.2011.01073.x

Li, X., Xiong, Z., Theeuwes, J., & Wang, B. (2020). Visual memory benefits from prolonged encoding time regardless of stimulus type. *Journal of Experimental Psychology:*

Learning Memory and Cognition, *46*(10), 1998–2005. https://doi.org/10.1037/xlm0000847

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 of the United States of America*, *117*(51), 32329–32339. https://doi.org/10.1073/pnas.2006752117

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

Markov, Y. A., Utochkin, I. S., & Brady, T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision*, *21*(3), 18. https://doi.org/10.1167/jov.21.3.18

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

Nah, J. C., & Geng, J. J. (2022). Thematic object pairs produce stronger and faster grouping than taxonomic pairs. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(12), 1325–1335. https://doi.org/10.1037/xhp0001031

Naspi, L., Stensholt, C., Karlsson, A. E., Monge, Z. A., & Cabeza, R. (2023). Effects of Aging on Successful Object Encoding: Enhanced Semantic Representations Compensate for Impaired Visual Representations. *The Journal of Neuroscience*, *43*(44), 7337–7350. https://doi.org/10.1523/JNEUROSCI.2265-22.2023

O'Donnell, R. E., Clement, A., & Brockmole, J. R. (2018). Semantic and functional relationships among objects increase the capacity of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1151–1158. https://doi.org/10.1037/xlm0000508

Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring Knowledge with Cognitive Maps and Cognitive Graphs. *Trends in Cognitive Sciences*, *25*(1), 37–54. https://doi.org/10.1016/j.tics.2020.10.004

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peterson, D. J., Gözenman, F., Arciniega, H., & Berryhill, M. E. (2015). Contralateral delay activity tracks the influence of Gestalt grouping principles on active visual working memory representations. *Attention, Perception, & Psychophysics*, *77*(7), 2270–2283. https://doi.org/10.3758/s13414-015-0929-y

Potter, M. C. (2012). Conceptual Short Term Memory in Perception and Thought. *Frontiers in Psychology*, *3*(113), 1–11. https://doi.org/10.3389/fpsyg.2012.00113

Quirk, C., Adam, K. C. S., & Vogel, E. K. (2020). No evidence for an object working memory capacity benefit with extended viewing time. *eNeuro*, *7*(5), 1–13. https://doi.org/10.1523/ENEURO.0150-20.2020

Roberts, K. L., & Humphreys, G. W. (2011). Action-related objects influence the distribution of Visuospatial attention. *Quarterly Journal of Experimental Psychology*, *64*(4), 669–688. https://doi.org/10.1080/17470218.2010.520086

Rust, N. C., & Mehrpour, V. (2020). Understanding Image Memorability. *Trends in Cognitive Sciences*, *24*(7), 557–568. https://doi.org/10.1016/j.tics.2020.04.001

Sahar, T., Gronau, N., & Makovski, T. (2024). Semantic Meaning Enhances Feature-Binding but not Quantity or Precision of Locations in Visual Working Memory. *Memory and Cognition*.

Simonyan, K., & Zisserman, A. (2015, April 10). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of the 3rd International Conference on Learning Representations.* ICLR, San Diego, CA. http://arxiv.org/abs/1409.1556

Sipser, M. (2012). *Introduction to the Theory of Computation* (3rd ed.). Cengage Learning.

Smith, M. E., Pitts, B. L., Newberry, K. M., Elbishari, Y., & Bailey, H. R. (2022). Prior knowledge shapes older adults' perception and memory for everyday events. In *Psychology of Learning and Motivation* (Vol. 77, pp. 233–262). Elsevier. https://doi.org/10.1016/bs.plm.2022.07.005

Thibeault, A. M. L., Stojanoski, B., & Emrich, S. M. (2024). Investigating the effects of perceptual complexity versus conceptual meaning on the object benefit in visual working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *24*(3), 453–468. https://doi.org/10.3758/s13415-024-01158-z

Thyer, W., Adam, K., Diaz, G., Vogel, E. K., Sánchez, I. N. V., & Awh, E. (2022). Storage in visual working memory recruits a content-independent pointer system. *Psychological Science*. https://osf.io/uhbx5/

Tompary, A., & Thompson-Schill, S. L. (2021). Semantic influences on episodic memory distortions. *Journal of Experimental Psychology: General*, *150*(9), 1800–1824. https://doi.org/10.1037/xge0001017

Võ, M. L. H. (2021). The meaning and structure of scenes. *Vision Research*, *181*(January), 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Wong, J. H., Peterson, M. S., & Thompson, J. C. (2008). Visual working memory capacity for objects from different categories: A face-specific maintenance effect. *Cognition*, *108*(3), 719–731. https://doi.org/10.1016/j.cognition.2008.06.006

Xie, W., Bainbridge, W. A., Inati, S. K., Baker, C. I., & Zaghloul, K. A. (2020). Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nature Human Behaviour*, *4*(9), 937–948. https://doi.org/10.1038/s41562-020-0901-2

Xie, W., & Zhang, W. (2017a). Discrete item-based and continuous configural representations in visual short-term memory. *Visual Cognition*, *25*(1–3), 21–33. https://doi.org/10.1080/13506285.2017.1339157

Xie, W., & Zhang, W. (2017b). Dissociations of the number and precision of visual short-term memory representations in change detection. *Memory and Cognition*, *45*(8), 1423–1437. https://doi.org/10.3758/s13421-017-0739-7

Xie, W., & Zhang, W. (2017c). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory and Cognition*, *45*(4), 677–689. https://doi.org/10.3758/s13421-016-0679-7

Xie, W., & Zhang, W. (2017d). Familiarity speeds up visual short-term memory consolidation. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1207–1221. https://doi.org/10.1037/xhp0000355

Xie, W., & Zhang, W. (2018). Familiarity speeds up visual short-term memory consolidation: Electrophysiological evidence from contralateral delay activities. *Journal of Cognitive Neuroscience*, *30*(1), 1–13. https://doi.org/10.1162/jocn_a_01188

Xie, W., & Zhang, W. (2022). Pre-existing long-term memory facilitates the formation of visual short-term memory. In T. Brady & W. A. Bainbridge (Eds.), *Visual Memory* (pp. 84–104). https://doi.org/10.4324/9781003158134-6

Xie, W., & Zhang, W. (2023a). Effortfulness of visual working memory: Gauged by physical exertion. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001391

Xie, W., & Zhang, W. (2023b). Pupillary evidence reveals the influence of conceptual association on brightness perception. *Psychonomic Bulletin & Review*, *22*(14), 4228. https://doi.org/10.3758/s13423-023-02258-6

Ye, C., Guo, L., Wang, N., Qiang, L., & Xie, W. (2024). Perceptual encoding benefit of visual memorability on visual memory formation. *Cognition*, *248*, 105810. https://doi.org/10.1016/j.cognition.2024.105810

Yu, X., & Lau, E. (2023). The Binding Problem 2.0: Beyond Perceptual Features. *Cognitive Science*, *47*(2), e13244. https://doi.org/10.1111/cogs.13244

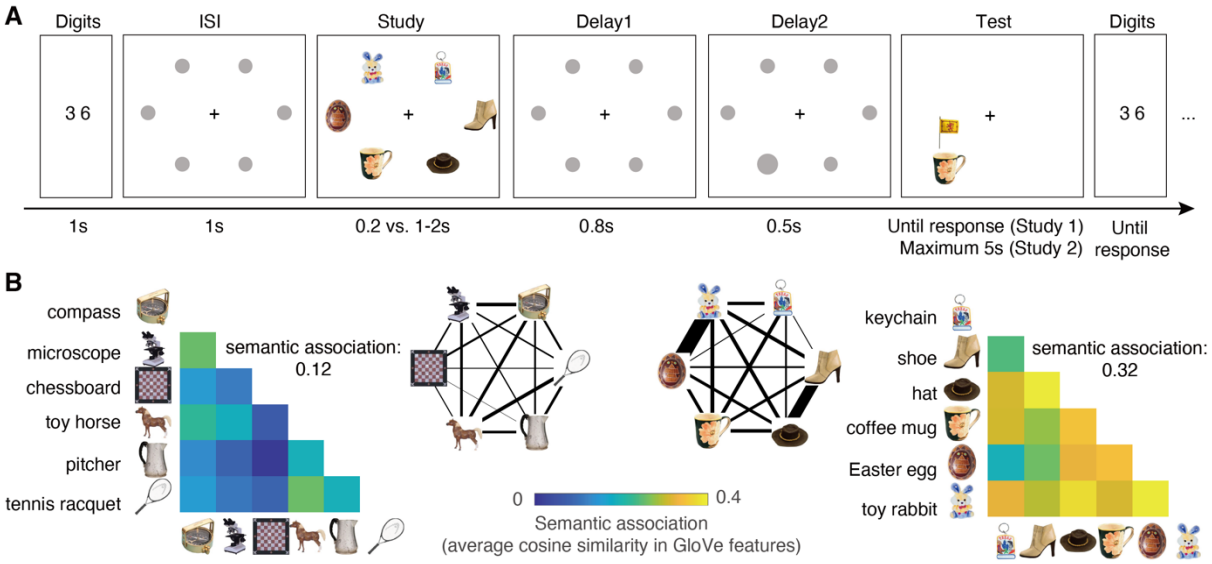

Figure 1. **Task design and analysis of semantic associations among objects within a visual display**. (A) An example trial of the VWM task using real-world objects from different categories with concurrent verbal distraction. Study 1 required participants to directly recall the digits presented at the end of a trial, while Study 2 asked participants to recognize whether the digits presented at the end were the same or different compared to the digit shown at the beginning of the trial. The other parameters and procedures were consistent across both studies. (B) Illustrations of visual displays with objects depicting lower semantic association (*left*) and higher semantic association (*right*). The overall average semantic association among objects in a visual display is estimated as the mean association strength across all non-redundant pairs of object labels. The width of the lines connecting the objects represents the pair-wise semantic association strength.

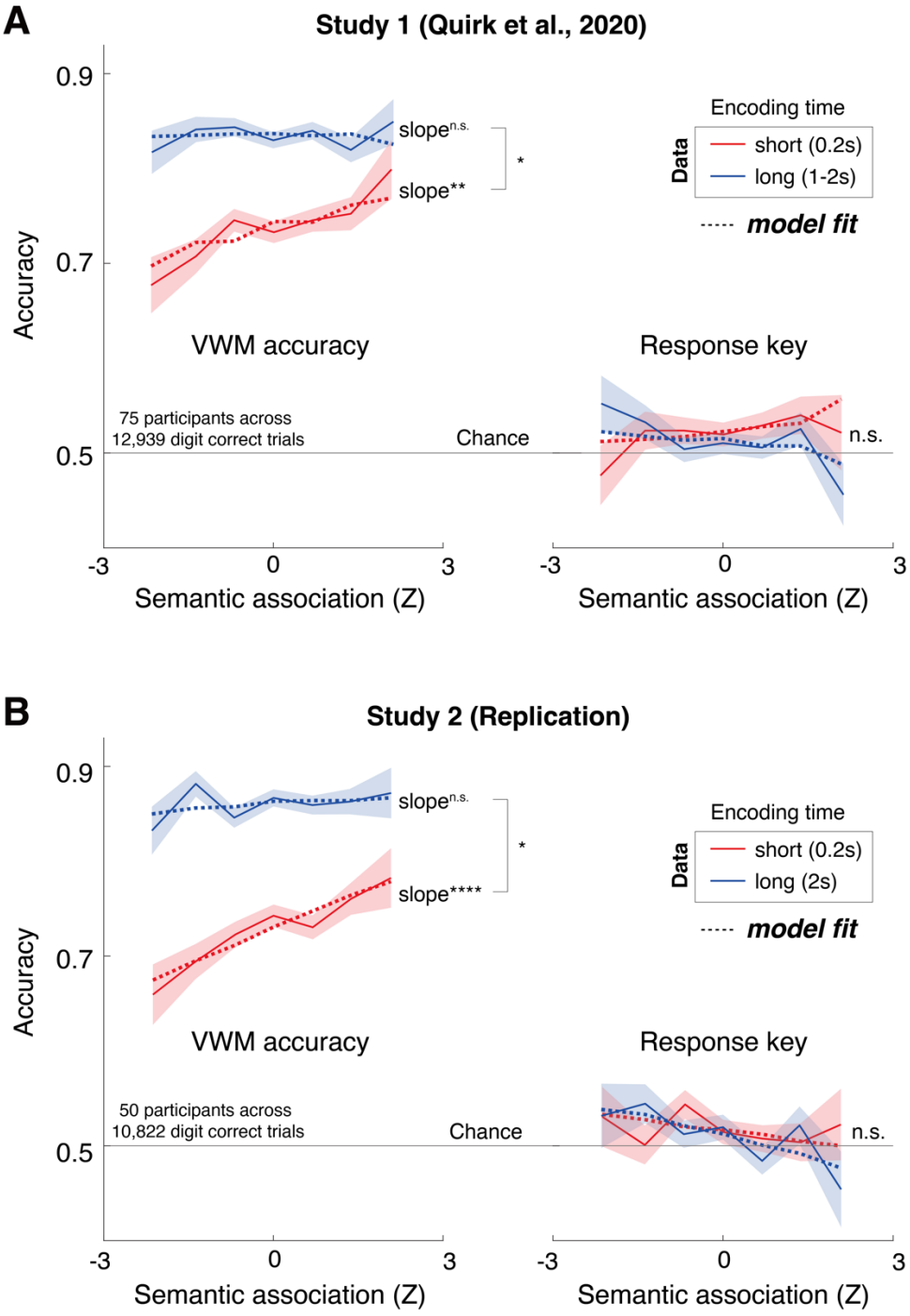

Figure 2*. Semantic associations among objects within the study display predict VWM accuracy but not response key press, under a short encoding time of 0.2s across Study 1 (A) and Study 2 (B).** Solid lines represent the mean estimates across trials and participants, dashed lines represent the best fit prediction of the generalized linear mixed model. The shaded areas represent bootstrapped 95% confidence intervals. These results are based on trials with correct digit responses, excluding occasional non-response trials, which account for more than 90% of the data. Results based on all trials without any data exclusion remain consistent with these findings and are summarized in ***Supplementary Figure S1.

Table 1. Overview of the study design for the data involved in our current analyses.

Study	Study 1[a] N = 13,750 trials			Study 2 N = 11,880 trials
	Exp. 1a	Exp. 1b	Exp. 1c	Replication experiment
Encoding time conditions	0.2, 1, 2s	0.2, 2s	0.2, 2s	0.2, 2s
Number of participants	25	25	25	50
Encoding time design[b]	within-subject, blocked	within-subject, blocked	within-subject intermixed	within-subject, blocked
Concurrent digit task	silent	half silent, half explicit	silent	silent

Note: (a) Study 1 contains data from Quirk et al. (2020). (b) The design here refers to whether the encoding time conditions are blocked or intermixed across blocks within each participant. Exp. = Experiment.

Table 2. Predicting VWM accuracy across encoding time condition based on generalized linear mixed models using trial-level features

Predictors	Study 1 (Quirk et al., 2020)						Study 2 (Replication)					
	Model 1[a]			Model 2[b]			Model 1			Model 2		
	β	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
enc_time	0.59	0.045	**<0.0001**	0.59	0.045	**<0.0001**	0.89	0.052	**<0.0001**	0.89	0.052	**<0.0001**
VGG_Study	0.011	0.031	0.71	0.0075	0.031	0.81	0.028	0.032	0.39	0.020	0.032	0.53
VGG_Test	-0.024	0.031	0.45	-0.023	0.031	0.47	-0.061	0.031	0.053	-0.059	0.031	0.058
GloVe_Test	-0.0012	0.031	0.97	-0.019	0.031	0.55	0.047	0.031	0.13	0.017	0.032	0.59
VGG_Study × *enc_time*	0.026	0.045	0.57	0.030	0.045	0.51	0.030	0.052	0.57	0.036	0.052	0.49
VGG_Test × *enc_time*	-0.017	0.045	0.71	-0.018	0.045	0.69	-0.061	0.051	0.23	-0.062	0.051	0.22
GloVe_Test × *enc_time*	0.038	0.045	0.40	0.057	0.046	0.21	-0.043	0.052	0.41	-0.017	0.053	0.75
GloVe_Study				0.087	0.031	**0.0054**				0.14	0.032	**<0.0001**
GloVe_Study × *enc_time*				-0.095	0.046	**0.038**				-0.12	0.053	**0.020**
Model fit metric (AIC)	12640			12636			10138			10123		
Comparison (Model 2 - Model 1)	ΔAIC = -4, log-likelihood ratio test: $\chi^2_{(2)} = 7.76$, $p = 0.021$						ΔAIC = -15, log-likelihood ratio test: $\chi^2_{(2)} = 19.32$, $p < 0.0001$					

Note: (a) Model 1 removes the term related to the semantic association among objects within a study display and its interaction effect with encoding time. (b) Model 2 contains all terms. The comparison between these two models therefore informs us about the necessity of adding these semantic predictors in accounting for participants' trial-by-trial VWM task performance. Statistically significant effects with $p < 0.05$ are bolded in the table.

***Table 3*.** Predicting VWM accuracy for each encoding time condition based on generalized linear mixed models using trial-level features

Predictors	Study 1 (Quirk et al., 2020)						Study 2 (Replication)					
	short encoding time (0.2s)			long encoding time (1-2s)			short encoding time (0.2s)			long encoding time (2s)		
	β	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*	*β*	*SE*	*p*
VGG_Study	0.0087	0.031	0.78	0.036	0.034	0.28	0.019	0.032	0.55	0.054	0.042	0.20
VGG_Test	-0.027	0.031	0.39	-0.041	0.033	0.22	-0.055	0.031	0.074	-0.12	0.041	**0.0028**
GloVe_Study	0.087	0.031	**0.0050**	-0.0066	0.034	0.84	0.14	0.032	**<0.0001**	0.011	0.043	0.79
GloVe_Test	-0.020	0.031	0.51	0.033	0.034	0.34	0.018	0.031	0.57	0.00038	0.044	0.99

Note. Statistically significant effects with $p < 0.05$ are bolded in the table.

Supplementary Materials

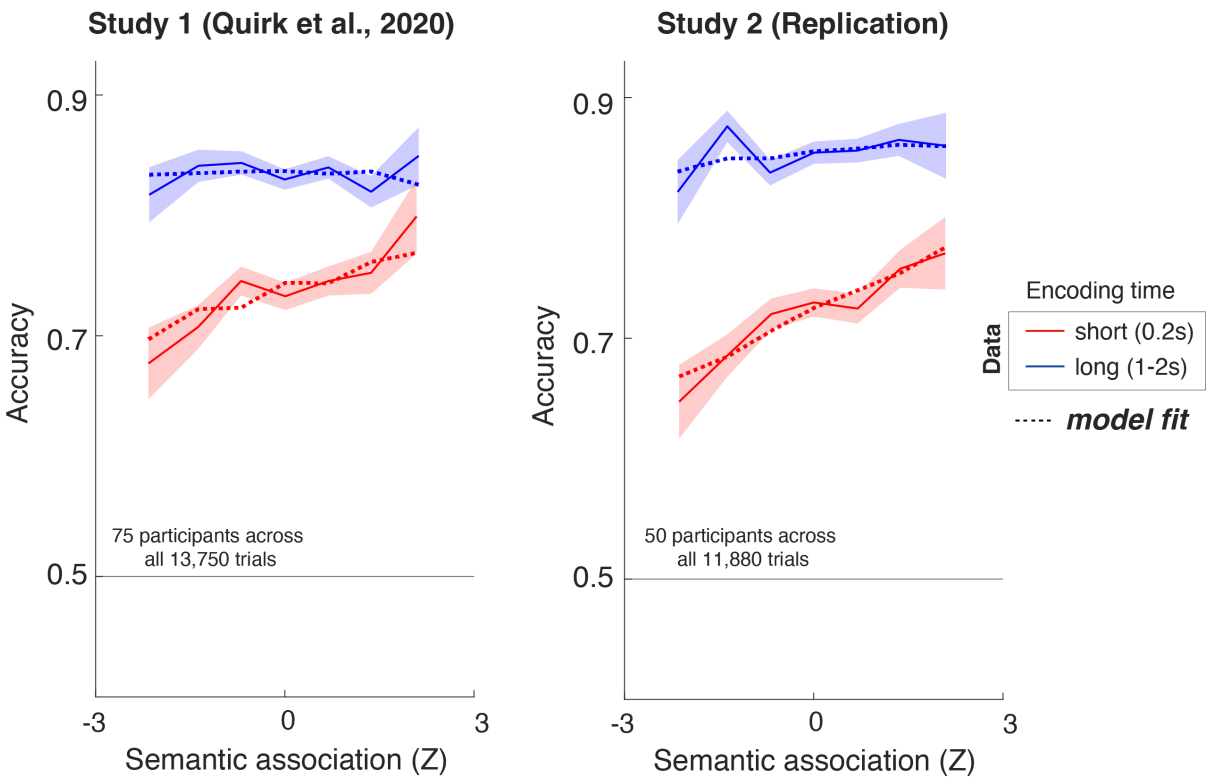

***Figure S1*. Semantic associations among objects within the study display predict VWM accuracy under a short encoding time of 0.2s across Study 1 and Study 2 using all trials.** Solid lines represent the mean estimates across trials and participants, dashed lines represent the best fit prediction of the generalized linear mixed model. The shaded areas represent bootstrapped 95% confidence intervals.