nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Images with harder-to-reconstruct visual representations leave stronger memory traces
Corresponding author name(s): Qi Lin, John Lafferty, Ilker Yildirim

Reviewer Comments & Decisions:

Decision Letter, initial version:

26th April 2023

Dear Prof Yildirim,

Thank you once again for your manuscript, entitled "From seeing to remembering: Images with harder-to-reconstruct representations leave stronger memory traces", and for your patience during the peer review process.

Your article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

First, we ask that you address the important technical concerns of Reviewer #3 regarding 1) how the reconstruction error is measured (comment 1) and 2) model retraining following the suggestion of Reviewer #3 (comment 2). Next, please respond to all other reviewer concerns, including those by Reviewer #1 and Reviewer #2. We expect you to address Reviewer #2's concerns with additional

discussion rather than additional experimental data.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Giacomo Ariani
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Vision, memorability, deep learning

Reviewer #2: Vision, memory

Reviewer #3: Vision, memorability, deep learning

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:
The paper presents a theoretical viewpoint, back up with a model and experiments, of visual image memorability. To the knowledge of this reviewer, this is the first strong and convincing theory of why some images are more memorable than other. The authors detail a sparse coding model that compresses feature embeddings of images and demonstrates with 3 experiments that the reconstruction errors explains visual image memory accuracy and recognition reaction time. This is the first account that is both theoretically sounded, with experimental data, of visual memorability. The paper is extremely clear, very well written, the experiments, data analyses (partial correlations analyses make a lot of sense as well), statistics are technically sounds and they make a lot of sense (including the distinctiveness-reconstruction error profiles). I read the paper in detail, and I could not find something obviously wrong to report (but I did not connect to the github repository link given in the paper). I like everything about this paper and the results, a very compelling and thorough analyses that I believe will inspire many memory researchers -- it did inspire me.

A couple of remarks, not significant issues: the current work focuses on images dataset of memorability (given the model and the theory, the authors have to start somewhere), it may be worth adding a couple of sentences about how the theory would generalize to video / event memorability, and maybe words/language, although it is possible the language mechanism for word/text memory would be different. The paper is really clear and well written, but could be reduced by 15-20% if space is of concern, as the concept of reconstruction errors is pretty easy to get, and it is repeated a lot in the current manuscript.

Reviewer #2:
Remarks to the Author:
Lin and colleagues address a very important research question by asking how perceptual and memory processes interact. They provide a computational model that combines deep neural nets with sparse coding, aiming to jointly model perception and memory processes. The authors also present experimental data in support of their modelling work. The approach is certainly interesting and I appreciate the effort and dedication that went into combining modelling with experimental work. However, I have fundamental concerns about the conceptual and theoretical framework that limit the overall interpretation and significance of the findings. This does not mean that the results presented here are not meaningful within the limitations of the approach or the work is not interesting. I am, however, not convinced that the findings can be generalized beyond these limitations. I am detailing these concerns below.

-Perception is not a passive process
The authors depict (visual) perception as a passive process. To the contrary, there is strong evidence that perception is active, involving sensory exploration and (motor) sampling (see e.g., Ahissar & Arieli, 2001, Neuron; Schroeder, et al, 2010, Curr Op Neurobiol). In particular, visual perception is a process that cannot be understood without taking eye movements into account. The eyes are actually never still, even when fixating a stimulus. Convergent evidence in the last years also shows that the historical

divide between sensory-only and motion-only brain areas is not accurate (e.g., Niell & Stryker, 2010, Neuron), making it more and more clear that movements interact with and modulate the perception of (visual) stimuli (e.g., Sommer & Wurtz, 2008, Perception). The present work does, however, not include this important relationship.

-Eye movements and memory performance are tightly linked
The authors refer to research on memorability, arguing that some images are systematically more memorable than others across observers. I am, however, not convinced that interpreting these observations is very meaningful without taking eye movements into account. Many studies have demonstrated the tight link between exploration of visual scenes by eye movements and memory performance. Eye movements during encoding are correlated with subsequent memory performance and have been shown to be reinstated during retrieval (e.g., Voss et al., 2017, TiCS; Wynn et al., 2020, PNAS). Moreover, brain activity related to memory processes is also linked to eye movements (e.g., Kragel & Voss, 2022, TiCS), making a strong case for why the relationship between (visual) perception and memory cannot be studied without taking eye movements into account.
The present work does, however, not incorporate the accumulating evidence of a strong relationship between eye movements (being vital for visual perception), memory and neural activity.

Similar arguments can be applied to the author's appraisal of the levels-of-processing account by Craik & Lockhart, proposing that a memory trace is a by-product of the perceptual analysis of the incoming sensory signals and that a 'deeper' analysis is associated with better retention in memory. The authors argue that previous accounts are 'completely missing the automatic nature of perceptual processing.' As outlined above, I am not convinced that perception is automatic in nature. Rather, it is an active process that involves movement of the sensors. In the case of the levels-of-processing account, the authors propose that 'a memory trace is a by-product of the perceptual analysis of the incoming sensory signals and that a 'deeper' analysis is associated with better retention in memory.' Previous research has, however, shown that the relationship between eye movements and memory performance is particularly strong for complex visual stimuli (e.g., Voss et al., 2017), which could, at least to some extent, explain different 'levels of processing'. On the basis of these arguments, I also do not agree that the presented model can be a good candidate for exploring the neurocomputational basis of the subsequent memory effect.

Together, I am not convinced that a novel model will provide fundamental insight into how perception and memory interact, if the model does not take the fundamental relationships between visual sampling via eye movements and perceptual processes into account.

Reviewer #3:
Remarks to the Author:
The authors investigated the interplay between perception and memory via memorability of images in an interesting and innovative way. They fit a sparse coding model to activation patterns of a DCNN model trained for image classification, and found that memorable images show larger reconstruction error.

I have the following main concerns for the approaches used in this study:

1- The reconstruction error is measured as the Euclidean Distance between the original activation vector and the reconstructed activation by the SPC model. Previous studies (Jaegle et al. 2019) have shown

that memorable images evoke pattern activities with larger magnitude in DCNNs, and authors also mentioned this in the paper. Then if the reconstruction error is not normalized by the magnitude of activation patterns, it is a misleading measure. If the authors use the normalized MSE as the reconstruction error, they may find an opposite result, i.e., memorable images will have smaller reconstruction error or maybe even no significant difference between memorable and forgettable images.

2- Previous studies have shown that memorable images are more object-centric images. This means that specific units in the layers of model will be more responsive to these images and others not. Therefore, the way the authors sampled 1000 elements of activation patterns from the high dimensional activations of the DCNN to train their models seem suboptimal for memorable images compared to forgettable images (which are more texture- like images). To have a fair training, the authors should first use a dimensionality reduction technique with an independent set of images to reduce the dimension of activation patterns in each layer to 1000 and then use that on their training with the Isola et al. dataset.

3- Previous work by Boer and Potter 2018 has show that memorable images show enhanced recognition in ultra RSVP paradigm, this means images that are easier to grasp quickly will be remembered and recognized better. How the authors compare this with their finding in their RSVP experiment?

Author Rebuttal to Initial comments

Response to Reviewer #1

The paper presents a theoretical viewpoint, back up with a model and experiments, of visual image memorability. To the knowledge of this reviewer, this is the first strong and convincing theory of why some images are more memorable than other. The authors detail a sparse coding model that compresses feature embeddings of images and demonstrates with 3 experiments that the reconstruction errors explains visual image memory accuracy and recognition reaction time. This is the first account that is both theoretically sounded, with experimental data, of visual memorability. The paper is extremely clear, very well written, the experiments, data analyses (partial correlations analyses make a lot of sense as well), statistics are technically sounds and they make a lot of sense (including the distinctiveness-reconstruction error profiles). I read the paper in detail, and I could not find something obviously wrong to report (but I did not connect to the github repository link given in the paper). I like everything about this paper and the results, a very compelling and thorough analyses that I believe will inspire many memory researchers -- it did inspire me.

We thank the Reviewer for the accurate summary of our paper and we are grateful for their positive assessment!

A couple of remarks, not significant issues: the current work focuses on images dataset of memorability (given the model and the theory, the authors have to start somewhere), it may be worth adding a couple of sentences about how the theory would generalize to video / event memorability, and maybe words/language, although it is possible the language mechanism for word/text memory would be different.

This is a great suggestion! It did not previously occur to us to think about domain-general mechanisms of memorability but this makes perfect sense. We have now added a new discussion paragraph about how our framework could generalize to other types of stimuli (added text in purple on Page 12 under the *Beyond image memorability* section; also pasted below for reader's convenience).

> "Beyond what we studied here (i.e., static images), recent work has begun to establish that memorability is a more domain-general phenomenon than we studied it here (focusing on the memorability of static images). This includes the stimulus-driven memorability of videos (Cohendet et al., 2019, *ICCV*), events (Xu et al., 2021, *NeurIPs*), and even words (Lau et al., 2018, *Quarterly Journal of Experimental Psychology*). We believe that the presented framework -- compression based reconstruction error -- should be a good starting point for characterizing what may be a shared, domain-general mechanism of memorability. Future work should build on our work to explore models that integrate compression based reconstruction error with different forms of encoding sequential inputs, including using recent advances in neural network models for encoding of videos and events (Majumdar et al., 2023, *arXiv*), and text-based natural language input (Radford et al., 2019, *OpenAI*)."

The paper is really clear and well written, but could be reduced by 15-20% if space is of concern, as the concept of reconstruction errors is pretty easy to get, and it is repeated a lot in the current manuscript.

We greatly appreciate the Reviewer's continued positive assessment of our work.

Per the Reviewer's suggestion, we have gone through the paper (in particular, the results section) and tried to shorten the wording wherever possible (but we were not able to shorten it to the extent the Reviewer suggested). Below we show one example where we tried to reduce the redundancy (first paragraph of Study 3, Page 8):

> "Studies 1 and 2 revealed reconstruction error as a novel source of image memorability and retrieval efficiency using a publicly available dataset. These results suggest that reconstruction error might be modulating the process of encoding itself, in terms of how deeply an image should be encoded~~, and thus yielding these measurable relationships~~. To further establish that reconstruction error is associated with a functionally distinct set of computations in the mind

and brain, beyond the vision-only computations implemented in DCNN, ~~the non-linear feature hierarchies trained for image classification captured in distinctiveness,~~ we turn to a prediction made by our modeling framework and test it in a 'model-driven psychophysics' experiment in Study 3."

We additionally would like to note that we believe that the length of the paper is not a concern.

Response to Reviewer #2

Lin and colleagues address a very important research question by asking how perceptual and memory processes interact. They provide a computational model that combines deep neural nets with sparse coding, aiming to jointly model perception and memory processes. The authors also present experimental data in support of their modelling work. The approach is certainly interesting and I appreciate the effort and dedication that went into combining modelling with experimental work.

We thank the Reviewer for this accurate summary of our paper. We are also grateful that they found our work interesting and appreciated our approach. These are also some of the reasons why we felt compelled to submit this manuscript to *Nature Human Behaviour* in the first place.

However, I have fundamental concerns about the conceptual and theoretical framework that limit the overall interpretation and significance of the findings. This does not mean that the results presented here are not meaningful within the limitations of the approach or the work is not interesting. I am, however, not convinced that the findings can be generalized beyond these limitations. I am detailing these concerns below.

Below we provide point-by-point responses, including the revisions we made in the manuscript for addressing each issue.

-Perception is not a passive process

The authors depict (visual) perception as a passive process. To the contrary, there is strong evidence that perception is active, involving sensory exploration and (motor) sampling (see e.g., Ahissar & Arieli, 2001, Neuron; Schroeder, et al, 2010, Curr Op Neurobiol). In particular, visual perception is a process that cannot be understood

without taking eye movements into account. The eyes are actually never still, even when fixating a stimulus. Convergent evidence in the last years also shows that the historical divide between sensory-only and motion-only brain areas is not accurate (e.g., Niell & Stryker, 2010, Neuron), making it more and more clear that movements interact with and modulate the perception of (visual) stimuli (e.g., Sommer & Wurtz, 2008, Perception). The present work does, however, not include this important relationship.

We thank the Reviewer for this perspective. We agree with the Reviewer on the importance of eye movements for visual processing, and the fact that we are not addressing them. We now remedy this issue with a new discussion paragraph that discusses the role of eye movements and how it can be integrated within our framework. We discuss the details of this new discussion paragraph in our response to the Reviewer's next point. We use the rest of the current response for making two clarifications.

First, we wish to state that visual perception is a multi-faceted, complex process. Eye movements are an important aspect of this process, but at the same time, we believe that it is evident that many questions about vision can be productively pursued without making particular commitments about eye movements. A classic example is the work of David Marr and his colleagues (summarized and expanded in his book Vision), which rarely, if ever, mentions eye movements, and yet has been incredibly helpful helping many students of vision make sense of visual perception. A more modern example is some of the recent progress in vision research based on using DCNNs as a model family (with static input). The list goes on, but our point is that certain aspects of visual perception can be productively studied without making strong commitments about eye movements.

Second, we wish to clarify our perspective of perception: we don't think vision is passive (in fact, we never make any such statement in the manuscript); instead, we think vision is active in the sense that it involves inference. Inference can be formalized as an active process (e.g., belief updating, evidence accumulation) and it has been productively integrated with frameworks that involve actual action components (as in Karl Friston's active learning framework and John Tsotsos's work in attention).

Once again, we thank the Reviewer for pointing us to this literature, which we now include in a new Discussion paragraph as we detail in the next response.

-Eye movements and memory performance are tightly linked

The authors refer to research on memorability, arguing that some images are systematically more memorable than others across observers. I am, however, not convinced that interpreting these observations is very meaningful without taking eye movements into account. Many studies have demonstrated the tight link between exploration of visual scenes by eye movements and memory performance. Eye movements during encoding are correlated with subsequent memory performance and have been shown to be reinstated during retrieval (e.g., Voss et al., 2017, TiCS; Wynn et al., 2020, PNAS). Moreover, brain activity related to memory processes is also linked to eye movements (e.g., Kragel & Voss, 2022, TiCS), making a strong case for why the relationship between (visual) perception and memory cannot be studied without taking eye movements into account.

The present work does, however, not incorporate the accumulating evidence of a strong relationship between eye movements (being vital for visual perception), memory and neural activity.

We thank the Reviewer for raising this important and fascinating point. We would like to clarify that the main prediction of our model is that images that are harder to reconstruct receive deeper processing during encoding. Our model in its current form, does not make specific predictions for or against the role of visual sampling via eye movement during this process.

That being said, it is possible that this deeper processing could be realized (at least, in part) via visual sampling behavior (e.g., more fixations on images with harder to reconstruction representations). Consistent with this idea, prior work on memorability and eye movements has found that both a larger number of fixations and higher consistency in the fixation maps across individuals predict higher memorability scores (Lyu et al., 2020, *JoV*; Bylinskii et al., 2015, *Visual Cognition*). In response to this excellent point (and the relevant point raised below on the relationship between depth of processing and eye movements), we have now added in discussion about this potential relationship (added text in purple on Page 12 under the *A candidate for exploring the computational basis of the subsequent memory effect and active visual exploration* Section; also pasted below).

> "Previous work has linked active visual exploration (e.g., eye movements) with successful memory formation and brain activity (for reviews, see Voss et al., 2017; Ryan et al., 2020; Kragel and Voss, 2022). Moreover, previous work demonstrated that images with larger numbers of fixations and greater "fixation-map" consistency across individuals are more memorable (Bylinskii et al., 2015; Lyu et al., 2020). Future work should explore the possibility that reconstruction error, in the service of effectively modulating
>
> depth-of-encoding of an image, might drive spontaneous visual exploration behavior and modulate brain activity (e.g., potential regions of exploration include hippocampus, frontoparietal regions and medial temporal lobe)."

However, at an empirical level, the results we currently present in our manuscript readily suggest that eye movement is unlikely to be the *only* mechanism linking our computational account and memory performance. As shown in our Study 3 (Figure 5B), the memory benefit of large reconstruction error can be observed at stimulus presentation times as short as 84 and 167 milliseconds, which are durations not long enough for more than one fixation (~200 ms on average according to a recent large dataset [Wilming et al., 2017, *Sci Data*]). Therefore, we suspect that the deeper processing includes not only the overt shift of attention through eye movements but also other forms of selective attention, such as attention gain control that allows the refinement and sharpening of the sensory representation. Future work can explore the relative degree to which visual sampling via eye movements and more covert forms of attention contribute to the link between reconstruction error and memory. However, this topic is out of the scope for the current study.

Similar arguments can be applied to the author's appraisal of the levels-of-processing account by Craik & Lockhart, proposing that a memory trace is a by-product of the perceptual analysis of the incoming sensory signals and that a 'deeper' analysis is associated with better retention in memory. The authors argue that previous accounts are 'completely missing the automatic nature of perceptual processing.' As outlined above, I am not convinced that perception is automatic in nature. Rather, it is an active process that involves movement of the sensors.

We appreciate the Reviewer's emphasis on the active aspect of perception but we would like to emphasize that 'automatic' is not equivalent to being 'passive'.

By automatic, we mean the spontaneous and stimulus-driven nature of visual perception. Thus, rather than drawing a contrast between being passive or active, we are primarily concerned with the contrast between being spontaneous vs. being intentional. In the current study, we are interested in how the selective retention of certain information in memory, in addition to being influenced by the intention of the observer (e.g., being instructed to pay attention to faces rather than scenes), could also be driven by perceptual processing that is more stimulus-dependent (e.g., devoting more resources to analyze what's in a scene) and happens without the observer explicitly exerting deliberate control. We believe eye movements can occur automatically or deliberately; so we would consider those eye movements that occur automatically within the scope of automatic/spontaneous processing.

In the revised manuscript, to clarify our view of perception and prevent miscommunication to other readers, we made the following modification when we first bring up the automaticity of perception (see Page 2,; also pasted below, added text in

purple):

> "To date, all empirical demonstrations of the level-of-processing effect rely on the use of an orienting task (as reviewed in Cermak and Craik, 2014), completely missing the automatic (i.e., spontaneous and stimulus-driven) nature of perceptual processing."

In the case of the levels-of-processing account, the authors propose that 'a memory trace is a by-product of the perceptual analysis of the incoming sensory signals and that a 'deeper' analysis is associated with better retention in memory.' Previous research has, however, shown that the relationship between eye movements and memory performance is particularly strong for complex visual stimuli (e.g., Voss et al., 2017), which could, at least to some extent, explain different 'levels of processing'. On the basis of these arguments, I also do not agree that the presented model can be a good candidate for exploring the neurocomputational basis of the subsequent memory effect.

We thank the Reviewer for proposing this possibility, which is also discussed in the corresponding new discussion paragraph mentioned above.

However, our perspective in that paragraph comes from a slightly different angle. In particular, we think the present model can still provide important insights into the neurocomputational basis of the subsequent memory effect, despite not explicitly incorporating eye movements in its current form. Empirical work has demonstrated compellingly the importance of hippocampal activity (both in terms of overall BOLD activation measured in fMRI [e.g., Voss et al., 2011, *PNAS*] and theta oscillations measured in intracranial EEG [e.g., Kragel et al., 2021, *Sci. Adv.*]) in guiding visual exploration during encoding and thus enhancing subsequent memory performance. However, the critical 'why' question still remains: why are heightened hippocampal theta oscillations and thus more revisitation eye movements observed in some trials but not others? We think that our model provides a potential answer to this critical question: That reconstruction error might drive both active visual exploration behavior and modulate neural activity, and thus lead to the observed patterns of memorability. More specifically, we argue that the reconstruction error resulting from the compression of sensory signals could (partially) inform this sampling process.

Finally, in an additional new discussion paragraph, we discuss how the reconstruction error in the model can be refined to support richer patterns of visual active exploration behavior. (see Page 12, purple text; also pasted below):

> "Our model can also be extended to represent the reconstruction error with

spatially-varying gradients, and thus generate predictions of how the depth-of-processing can be modulated at finer-levels of granularity (covertly, or via a combination of covert processes and eye movements)."

Together, I am not convinced that a novel model will provide fundamental insight into how perception and memory interact, if the model does not take the fundamental relationships between visual sampling via eye movements and perceptual processes into account.

We thank the Reviewer for these insightful and inspiring ideas! We hope that the foregoing response helps to clarify our view of perception, the theoretical contribution of our work to understanding the interface between perception and memory and how it relates to the existing literature on visual sampling and memory.

Response to Reviewer #3

The authors investigated the interplay between perception and memory via memorability of images in an interesting and innovative way. They fit a sparse coding model to activation patterns of a DCNN model trained for image classification, and found that memorable images show larger reconstruction error.

We are delighted by the Reviewer's positive assessment of our work.

I have the following main concerns for the approaches used in this study:

1- The reconstruction error is measured as the Euclidean Distance between the original activation vector and the reconstructed activation by the SPC model. Previous studies (Jaegle et al. 2019) have shown that memorable images evoke pattern activities with larger magnitude in DCNNs, and authors also mentioned this in the paper. Then if the reconstruction error is not normalized by the magnitude of activation patterns, it is a misleading measure. If the authors use the normalized MSE as the reconstruction error, they may find an opposite result, i.e., memorable images will have smaller reconstruction error or maybe even no significant difference between memorable and forgettable images.

We thank the Reviewer for raising this concern. First, we would like to clarify that in our

original manuscript, we did consider L2-Norm as a relevant measure extracted from the DCNNs, but ultimately chose distinctiveness (e.g., euclidean distance to nearest neighbor) because in our dataset distinctiveness predicted memorability better than L2-Norm (see Figure S2). In addition, we also found that L2-Norm is highly correlated with distinctiveness (r = 0.95, see Figure S3). (Both Figures S2 and S3 were presented in the original manuscript.)

To more directly address the Reviewer's concern, we repeated all of our analyses using the normalized MSE measure (as suggested by the Reviewer) and presented the results below as Figures R1 and R2. Specifically, we normalized the reconstruction errors with the L2-norm of the corresponding DCNN activation. As seen below, with this new measure of reconstruction error, although there is quantitative changes in terms of the contribution of specific layers, the pattern of results remains qualitatively similar, that is, large reconstruction error still correlates with higher memorability and shorter response times during retrieval (especially for the reconstruction errors from Layer 7, which we focus on in our RSVP experiment in Study 3). These results further support the conclusion that the observed relationship between reconstruction error and memory strengths cannot be explained away by measures derived from vision-only models and highlight the importance of compression-based reconstruction error in capturing the transformation of sensory signals into memory traces.

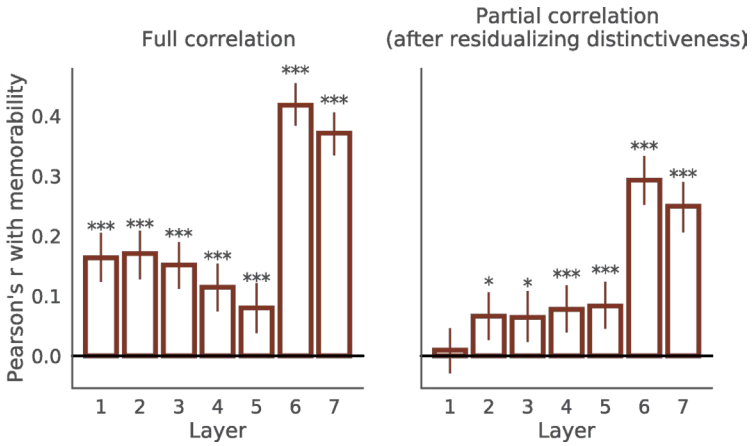

*Figure R1. Pearson's r between memorability and reconstruction error (normalized with L2 Norm of the activation vector) and partial Pearson's r after accounting for distinctiveness. Error bars represent 95% confidence intervals from 1000 bootstrapping iterations. ***: p < .001, **: p < .01, *: p <.05.*

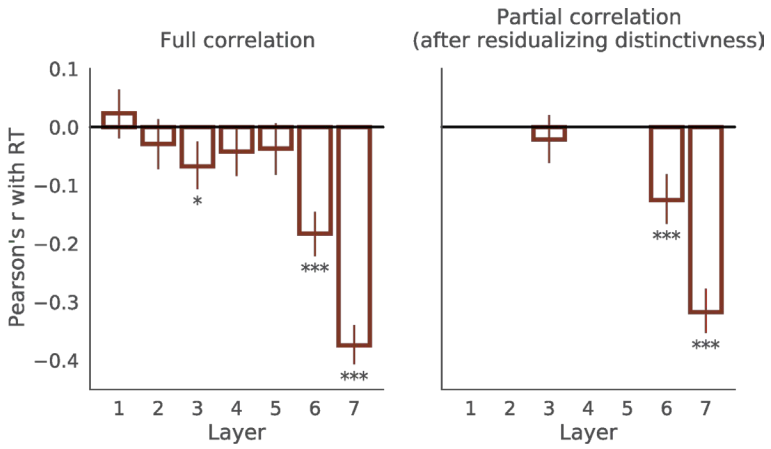

Figure R2. Pearson's r between response times during retrieval and reconstruction error (normalized with L2 Norm of the activation vector) and partial Pearson's r after accounting for distinctiveness. Error bars represent 95% confidence intervals from 1000 bootstrapping iterations.

****: p < .001, **: p < .01, *: p <.05. Note that partial correlation was only conducted for layers where the full correlation was statistically significant.*

We include these new results in the revised manuscript as Figs. S4 and S5, and refer to them within the main text (see Page 4, footnote 4).

2- *Previous studies have shown that memorable images are more object-centric images. This means that specific units in the layers of model will be more responsive to these images and others not. Therefore, the way the authors sampled 1000 elements of activation patterns from the high dimensional activations of the DCNN to train their models seem suboptimal for memorable images compared to forgettable images (which are more texture- like images). To have a fair training, the authors should first use a dimensionality reduction technique with an independent set of images to reduce the dimension of activation patterns in each layer to 1000 and then use that on their training with the Isola et al. dataset.*

We thank the Reviewer for raising this important point. To address it, we re-trained the SPC model exactly as suggested by the Reviewer and repeated the analysis with memorability and response times. Specifically, for each of the 7 layers in the DCNN, in a set of 8000 images randomly sampled from the Microsoft COCO dataset (Lin et al., 2014, *ECCV*), we learned the transformation of top 1000 principal components. We then applied the transformation to the activations evoked by the images in the Isola dataset

and thus reduced the dimensionality of each image's activation to 1000. The 1000 PCs (instead of 1000 randomly sampled units) were then used for training the sparse coding models and obtaining the reconstruction errors. We refer to this version of the model as the PCA-SPC model.

On this model, we conducted the same set of analysis as we did in Studies 1 and 2, including the full correlation and partial correlation analyses to account for the effect of distinctiveness. Typically, PCA is seen as just another preprocessing step in common data analysis and modeling pipelines, but in our case, because PCA is essentially a compression method, it is theoretically consequential to incorporate it into the model. Thus, for the sake of comprehensiveness and this theoretical concern, we also wished to demonstrate that the observed relationship between the reconstruction errors of the PCA-SPC model and memorability is not confounded by "the reconstruction error of PCA" alone. Thus, in an additional partial correlation analysis, we account for this PCA loss, which we quantify as the euclidean distance between the original DCNN activation vector (with all units in a given layer) and the reconstructed vector by inverting the transformation matrix and applying it to the reduced 1000-PC for each image.

As shown below in Figures R3 and R4 (also shown as Supplemental Figures 6 and 7), although there is a change in terms of the contribution of specific layers, we were delighted that all the results remained qualitatively similar (even with slightly stronger effect sizes in some cases), that is, large reconstruction error still correlates with higher memorability and shorter response times during retrieval. These correlations also survive residualizing PCA loss.

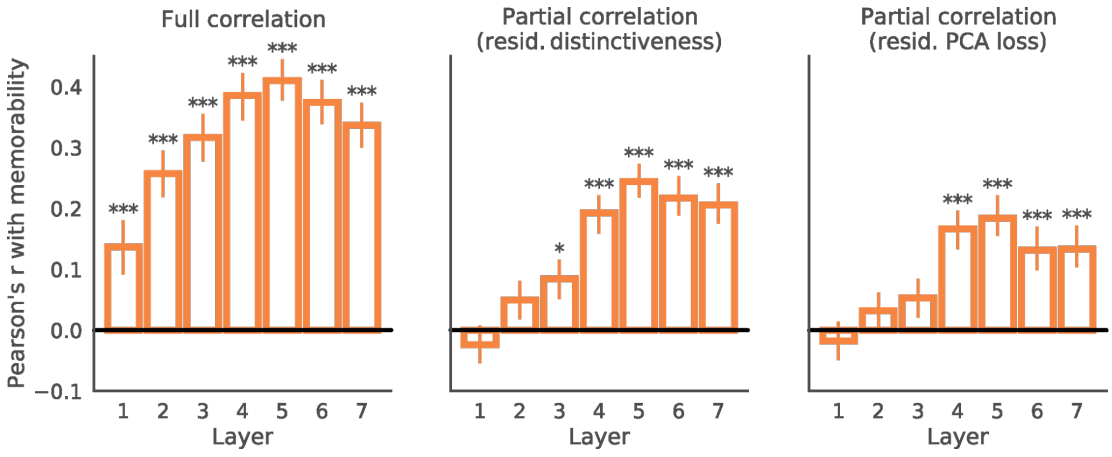

*Figure R3. Pearson's r between memorability and reconstruction error of the PCA-SPC model, partial Pearson's r after accounting for distinctiveness and partial Pearson's r after accounting for PCA loss. Error bars represent 95% confidence intervals from 1000 bootstrapping iterations. ***: p < .001, *: p <.05.*

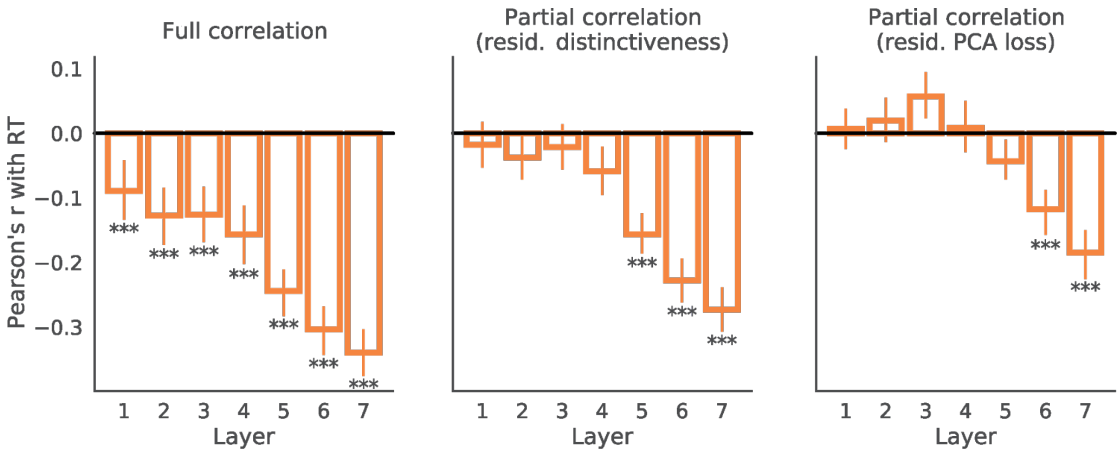

*Figure R4. Pearson's r between response times during retrieval and reconstruction error of the PCA-SPC model, partial Pearson's r after accounting for distinctiveness and partial Pearson's r after accounting for PCA loss. Error bars represent 95% confidence intervals from 1000 bootstrapping iterations. ***: p < .001.*

It is important to emphasize that the main point of novelty of our paper is to demonstrate the *theoretical* importance of the reconstruction error resulting from the compression of sensory signals: we show that this reconstruction error goes beyond vision-only computations implemented in a DCNN by providing a quantitative signal for characterizing the interface between perception and memory. However, the SPC model presented in the original manuscript is one instantiation of this theory, which we did not attempt to optimize in terms of its architecture and certain hyperparameters that go into its training – rather, it was a conscious decision to keep it as simple as possible (we did not want our theory to depend on the specific choices we make).

From this perspective, we see the PCA-SPC model as an extension of this core theory, and future work should more systematically explore optimal motifs for how the compression-based reconstruction errors might be computed in the mind and brain. This future work can focus on exploring the optimal architecture and sets of hyperparameters, including the PCA-SPC model, variational autoencoders, architectures with dynamic computational graphs, and spatially-varying reconstruction errors (in response to R2's insightful comments on eye-movements), among others.

We now include the above results in a pair of new supplementary figures (Supplementary figures 6 and 7) and refer to them in a new discussion paragraph on Page 12 (text in purple; also pasted below):

"The present work explores only one way of computing compression-based reconstruction error

--- using SPC to compress 1000 randomly sampled DCNN units in a layer. As a result, what is compressed in the model only partially captures what is in the DCNN activations. Refining this architecture, simply by replacing the random sampling strategy with an additional compression block (via principal component analysis) appears readily effective."

3- Previous work by Boer and Potter 2018 has show that memorable images show enhanced recognition in ultra RSVP paradigm, this means images that are easier to grasp quickly will be remembered and recognized better. How the authors compare this with their finding in their RSVP experiment?

We thank the Reviewer for bringing up this highly relevant study. In fact, we believe our results are perfectly consistent with the findings from Broers, Potter and Nieuwenstein (2018, *Psychon Bull Rev*). In our RSVP experiment (Figures 5B and 5D), we also observe that the memory performance for images with large versus small distinctiveness differ very early on – at the shortest presentation time we tested, 34 ms. This suggests that the effect of distinctiveness emerges very early on, consistent with Broers et al. study.

Accordingly, we would anticipate that the divergence between memorable and forgettable images at even shorter timescales (e.g., 13 ms and 27 ms in Broers et al.) to be primarily driven by distinctiveness. However, the critical contribution of our study is that as these likely primarily feedforward processes unfold, soon after, still rapidly and just as spontaneously, there seems to be additional memory encoding guided by the reconstruction error. This is synergistic with Broers et al., but suggests the involvement of fundamentally different kinds of computations than they suspected, certainly by 84 ms for images with large distinctiveness (and by 167 ms for images with small distinctiveness), and may be even sooner.

We reflect this foregoing discussion in the revised manuscript (see purple text on Page 11; also pasted below).

"We believe that the adaptive deployment of these perceptual mechanisms, alongside with purely feedforward computations (as implemented in standard DCNNs as we tested here), can be temporally mapped out with increasingly greater resolution. The RSVP results we presented in Study 3 are broadly

consistent with a behavioral report finding that the difference between memorable and forgettable images emerges at presentation times as brief as 13 ms (Broers et al., 2018). In our study, the effect of distinctiveness also emerges at the shortest duration (34 ms) we tested on whereas the effect of reconstruction error is only detectable at 84 ms and not before 34 ms, suggesting a temporal window of a few 10s of milliseconds after encountering a stimulus in which additional perceptual mechanisms, beyond feedforward processing, may be initially recruited. Future work should use behavioral paradigms -- ultra-fast RSVP (Broers et al., 2018)} and others -- and computational implementations of these adaptive perceptual mechanisms to uncover a precise temporal evolution of the encoding processes that ultimately lead to traces of remembered or forgotten memories."

Decision Letter, first revision:

3rd January 2024

Dear Dr. Yildirim,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Human Behaviour manuscript, "From seeing to remembering: Images with harder-to-reconstruct representations leave stronger memory traces" (NATHUMBEHAV-23020563A). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Please also address the additional marked-up edits we have proposed within the reporting summary. Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would hope to receive your revised paper, with all of the requested files and forms within two-three weeks. Please get in contact with us if you anticipate delays.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

Nature Human Behaviour offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "From seeing to remembering: Images with harder-to-reconstruct representations leave stronger memory traces". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Cover suggestions

We welcome submissions of artwork for consideration for our cover. For more information, please see our guide for cover artwork.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

ORCID

Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

For information regarding our different publishing models please see our Transformative

Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

Best regards,
Alex McKay
Editorial Assistant
Nature Human Behaviour

On behalf of

Giacomo Ariani
Editor
Nature Human Behaviour

Reviewer #1:
None

Reviewer #2:
None

Reviewer #3:
Remarks to the Author:
The authors have addressed my previously raised concerns; the revised manuscript has improved; I have no further comments.

Author Rebuttal, first revision:

Response to reviewers

Reviewer 3:

The authors have addressed my previously raised concerns; the revised manuscript has improved; I have no further comments.

We thank the Reviewer for their constructive and insightful feedback and we are glad

that the reviewer found our paper to have been improved.

Final Decision Letter:

Dear Prof Yildirim,

We are pleased to inform you that your Article "Images with harder-to-reconstruct visual representations leave stronger memory traces", has now been accepted for publication in Nature Human Behaviour.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words)

related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,

Giacomo Ariani
Editor
Nature Human Behaviour

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**