# Unveiling Neural Mechanisms of Memorability: Generative AI Reveals Brain-Behavior Links

**Hyewon Willow Han (hhan228@uwo.ca)**
Western University, London, ON N6A 3K7, Canada
Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada

**Mansoure Jahanian (mjahani@uwo.ca)**
Western University, London, ON N6A 3K7, Canada

**Johann Cardenas (jcarden4@uwo.ca)**
Western University, London, ON N6A 3K7, Canada

**Yalda Mohsenzadeh (ymohsenz@uwo.ca)**
Western University, London, ON N6A 3K7, Canada
Vector Institute for Artificial Intelligence, Toronto, ON M5G 0C6, Canada

# Abstract

**Some images are more memorable than others, yet the underlying neural mechanisms of memorability are not fully understood. In this study, we introduce a novel framework, "MemBrainGen", that connects memorability, a behaviorally defined feature of images, with the human brain responses. The framework utilizes generative deep neural network models to investigate how the human brain processes images based on their memorability across visual and memory regions of the human brain. Using MemBrainGen, we successfully manipulated the memorability of natural images in both increasing and decreasing directions, and observed that predicted activations in early-mid visual regions except for V1 showed no difference in response to memorability changes. However, brain regions associated with face and body categories, and the amygdala exhibited increased predicted activation when image memorability was increased. Most notably, V1 and place-associated regions showed lower predicted activation when images with increased memorability were presented to the model. We confirm our findings by demonstrating that brain activation-maximized images have higher memorability scores compared to their original counterparts in high-level visual and memory regions. Reversely, the memorability scores of these images were decreased in the place-selective regions. We further solidify our tested hypotheses by analyzing an independent fMRI dataset. From the univariate analysis with the independent dataset, we found that the direction of changes in brain activation is consistent with the predictions of our framework. This investigation contributes to our understanding of the cognitive processes involved in visual memory. It demonstrates the potential of integrating generative models with neuroimaging to explore the causal links between brain functions and behaviour, paving the way for the formulation of experimentally testable hypotheses.**

**Keywords:** memorability; visual memory; generative adversarial networks (GANs)

# Introduction

It has been shown that memorability is an intrinsic property of images, consistently observed across different ages and even species (Isola et al., 2011; Jaegle et al., 2019; Almog et al., 2023), and it is a computable feature which can be predicted from models (Khosla et al., 2015; Needell & Bainbridge, 2022; Younesi & Mohsenzadeh, 2024). It is still not fully understood what features make an image more memorable, and the neural mechanisms shaping this behavioral phenomenon. Previous studies using human fMRI have demonstrated that variations in response magnitude within the high-level visual cortex correlate well with the memorability of faces and scene images (Bainbridge et al., 2017; Bainbridge & Rissman, 2018; Lahner et al., 2024). In this work, we aimed to explore how the brain reacts differently to variations in the memorability of images by using a diverse set of images beyond faces and scenes. Inspired by the work of Gu et al. (2022), our research leverages artificial deep neural network (DNN) models as brain encoders and generative adversarial network (GAN) models for cognitive neuroscientific discovery, integrating these with human fMRI data to analyze responses in the face, body, place, and early visual regions of interest (ROI), as well as memory-related ROIs such as the hippocampus and amygdala.

# Framework Overview

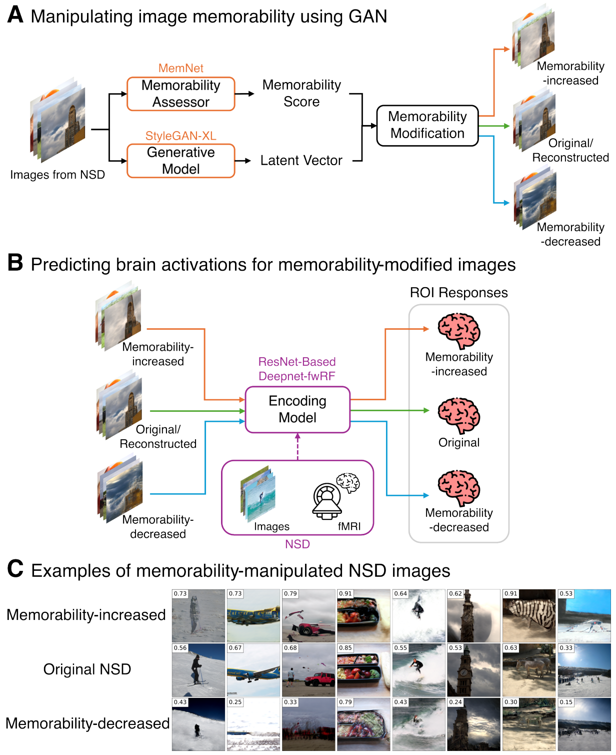

Figure 1: A schematic diagram of MemBrainGen and examples of memorability-modified NSD images.

The Natural Scenes Dataset (NSD) contains 7T fMRI responses from 8 subjects to natural images (Allen et al., 2022). We trained the encoding model for each subject, using responses from 1,000 images that were shared across all subjects as the validation set.

We manipulated the memorability of images in both increasing and decreasing directions (Figure 1A). As NSD lacked memorability scores, we used MemNet (Khosla et al., 2015) to predict. We utilized a GAN inversion approach to convert images into latent vectors for the memorability modification (Younesi & Mohsenzadeh, 2022). StyleGAN-XL (Sauer et al., 2022) pre-trained on ImageNet was used for the inversion.

We first excluded images not shown to all subjects, screened out overly disrupted images, and excluded those altered beyond recognition after memorability control. The screening process resulted in 102 images from the shared

1,000 in NSD used for the experiment. The feature-weighted receptive field (fwRF) model with ResNet-18 (He et al., 2016) was used as the brain encoder (St-Yves & Naselaris, 2018) (Figure 1B).

## Results

### Effects of image memorability modification on brain activation

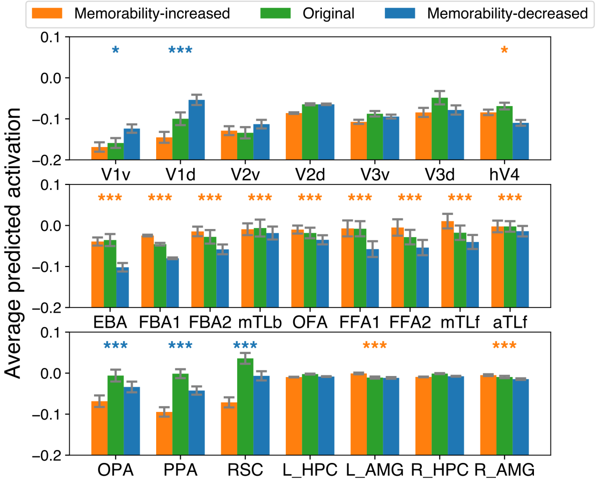

Figure 2: Average predicted brain activation per ROI based on memorability alteration.

To examine the effects of memorability modification of image stimulus on predicted brain activations, 102 images from NSD were reconstructed with increased and decreased memorability scores. Figure 2 depicts the magnitude of predicted brain activations for each ROI analyzed. Predicted activations of V4, face and body ROIs and amygdala increased with increased memorability. More notably, V1 and none of the place ROIs showed any significant changes in our anticipated direction but exhibited the reverse, showing less active responses when images with increased memorability were presented.

### Memorability of brain activation-maximized images

We posited that images capable of eliciting higher brain activity within the higher-level ROIs would yield higher memorability scores. Using NeuroGen (Gu et al., 2022), we generated 320 pairs of initial and activation-maximized images for each ROI. Figure 3 shows changes in memorability between those image pairs per ROI. No significant difference was found in memorability scores between activation-maximized and initial images for most of the early-mid visual ROIs, supporting our hypothesis that early-mid visual regions do not exert a significant influence on memorability. However, the memorability of activation-maximized images in the remaining ROIs exhibited a significant increase, except for the place ROIs. For the place ROIs, the memorability score decreased significantly for the activation-maximized images, which also supports the reverse pattern of those ROIs predicted using MemBrainGen.

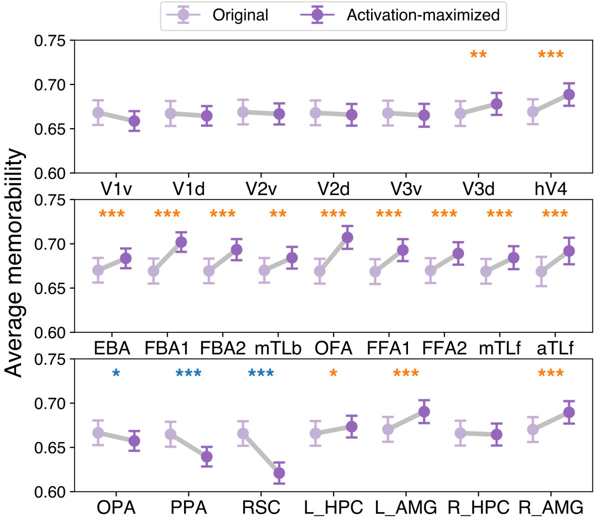

Figure 3: Changes in memorability when the activation of the target brain region was maximized.

Table 1: Univariate analysis results on the fMRI dataset by Lahner et al. (2024).

| ROI | p-value | Cohen's d | Effect direction |
|---|---|---|---|
| V1 | 0.9648 | -0.0116 | LM >HM |
| V2 | 0.8381 | -0.0538 | LM >HM |
| V3 | 0.3753 | 0.2364 | HM >LM |
| V4 | 0.0092$^{*}$ | 0.7798 | HM >LM |
| RSC | 0.1484 | -0.3950 | LM >HM |
| FFC | 0.0000$^{***}$ | 1.6306 | HM >LM |
| LO1 | 0.0012$^{**}$ | 1.0454 | HM >LM |
| LO2 | 0.0011$^{**}$ | 1.0516 | HM >LM |
| LO3 | 0.0491 | 0.5562 | HM >LM |
| L-HPC | 0.7497 | 0.0840 | HM >LM |
| R-HPC | 0.8291 | -0.0568 | LM >HM |
| L-AMG | 0.9512 | -0.0167 | LM >HM |
| R-AMG | 0.9558 | -0.0151 | LM >HM |

### Testing generated hypotheses on an independent fMRI dataset

We conducted an univariate analysis using an independent fMRI dataset (Lahner et al., 2024). The dataset consists of 78 pairs of semantically similar images from 15 subjects. The analysis revealed that V4 ($p < 0.05$), FFC ($p < 0.001$), LO1, and LO2 ($p < 0.01$) exhibit higher beta values when high memorable images are shown, supporting our findings that responses in higher visual perceptual areas increase when high memorable images are presented. We also observed that R-HPC, L-AMG, R-AMG, and RSC showed the opposite response (Table 1), with higher activation when low memorable images were shown; however, this difference did not remain significant after correcting for multiple comparisons at $\alpha$ = 0.05.

# Acknowledgments

This study was generously supported by the Canada First Research Excellence Fund (CFREF) through a BrainsCAN grant to YM, a NSERC Discovery Grant to YM, and a Vector Institute Research Grant to YM. The authors would also like to thank Vector Institute for providing computing resources. This work was also supported by a Brain Canada Platform Support Grant.

# References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Almog, G., Alavi Naeini, S., Hu, Y., Duerden, E. G., & Mohsenzadeh, Y. (2023). Memoir study: Investigating image memorability across developmental stages. *Plos one*, *18*(12), e0295940.

Bainbridge, W. A., Dilks, D. D., & Oliva, A. (2017). Memorability: A stimulus-driven perceptual neural signature distinctive from memory. *NeuroImage*, *149*, 141–152.

Bainbridge, W. A., & Rissman, J. (2018). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(1), 8679.

Gu, Z., Jamison, K. W., Khosla, M., Allen, E. J., Wu, Y., St-Yves, G., ... Kuceyeski, A. (2022). Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, *247*, 118812.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *Cvpr 2011* (pp. 145–152).

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife*, *8*, e47596.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the ieee international conference on computer vision* (pp. 2390–2398).

Lahner, B., Mohsenzadeh, Y., Mullin, C., & Oliva, A. (2024). Visual perception of highly memorable images is mediated by a distributed network of ventral visual regions that enable a late memorability response. *Plos Biology*, *22*(4), e3002564.

Needell, C. D., & Bainbridge, W. A. (2022). Embracing new techniques in deep learning for estimating image memorability. *Computational Brain & Behavior*, *5*(2), 168–184.

Sauer, A., Schwarz, K., & Geiger, A. (2022). Stylegan-xl: Scaling stylegan to large diverse datasets. In *Acm siggraph 2022 conference proceedings* (pp. 1–10).

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202.

Younesi, M., & Mohsenzadeh, Y. (2022). Controlling memorability of face images. *arXiv preprint arXiv:2202.11896*.

Younesi, M., & Mohsenzadeh, Y. (2024). Predicting memorability of face photographs with deep neural networks. *Scientific Reports*, *14*(1), 1246.