# Predicting image memorability from evoked feelings

Cheyenne Wakeland-Hart[1] and Mariam Aly[1,2]
[1]Columbia University, Department of Psychology
[2]University of California, Berkeley, Department of Psychology

**Please address correspondence to:**
Cheyenne Wakeland-Hart
cdw2147@columbia.edu

**ORCiD:**
Cheyenne Wakeland-Hart: https://orcid.org/0000-0003-3964-1908
Mariam Aly: https://orcid.org/0000-0003-4033-6134

## Declarations

**Funding:** This work was funded by an NSF GRFP award to C.W.H. and NSF CAREER Award to M.A. (BCS-1844241; BCS-2435322).

**Conflicts of interest/Competing interests:** The authors declare no competing financial or proprietary interests.

**Ethics approval:** This study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Institutional Review Board (IRB) in the Human Research Protection Office of Columbia University.

**Consent to participate:** Informed consent was obtained from all participants included in this study.

**Consent for publication:** N/A

**Availability of data and materials:** The image sets and datasets generated during and/or analyzed during the current study are available in the VAMOS repository, https://osf.io/ufg89/.

**Code Availability**: All code for data analysis associated with the current submission is available at https://osf.io/ufg89/.

**Acknowledgements:** We would like to thank the Alyssano Group for valuable insights on this project. We would also like to thank Benedek Kurdi, Bolei Zhou, Mariann Weierich, Malgorzata Wierzba, and Artur Marchewka for permission to include images from OASIS (B. Kurdi), PLACES (B. Zhou), COMPASS (M. Weierich), and NAPS (M. Wierzba and A. Marchewka) in our new image set. This work was funded by an NSF GRFP award to C.W.H. and NSF CAREER Award to M.A. (BCS-1844241; BCS-2435322).

**CRediT Statement:**

Cheyenne Wakeland-Hart: Conceptualization, Formal analysis, Investigation, Data Curation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing

Mariam Aly: Conceptualization, Methodology, Funding Acquisition, Project Administration, Resources, Supervision, Writing—original draft, Writing—review and editing

## Abstract

While viewing a visual stimulus, we often cannot tell whether it is inherently memorable or forgettable. However, the memorability of a stimulus can be quantified and partially predicted by a collection of conceptual and perceptual factors. Higher-level properties that represent the 'meaningfulness' of a visual stimulus to viewers best predict whether it will be remembered or forgotten across a population. Here, we hypothesize that the feelings evoked by an image, operationalized as the valence and arousal dimensions of affect, significantly contribute to the memorability of scene images. We ran two complementary experiments to investigate the influence of affect on scene memorability, in the process creating a new image set (VAMOS) of hundreds of natural scene images for which we obtained valence, arousal, and memorability scores. From our first experiment, we found memorability to be highly reliable for scene images that span a wide range of evoked arousal and valence. From our second experiment, we found that both valence and arousal are significant but weak predictors of image memorability. Scene images were most memorable if they were slightly negatively valenced and highly arousing. Images that were extremely positive or unarousing were most forgettable. Valence and arousal together accounted for less than 8% of the variance in image memorability. These findings suggest that evoked affect contributes to the overall memorability of a scene image but, like other singular predictors, does not fully explain it. Instead, memorability is best explained by an assemblage of visual features that combine in perhaps unintuitive ways to predict what is likely to stick in our memory.

## Introduction

As our lives move from one moment to the next, our experiences trace lasting impressions in our memory. Yet, not all of what we experience can be easily retrieved. Some experiences will be easily remembered, others forgotten. While individual factors influence what a person will remember, there are also general trends across people in which experiences will be remembered over others (Konkle et al., 2010). Using large-sample data, these trends can be abstracted as a stable stimulus property called *memorability*, which indicates how well an item is remembered across the general population (Bainbridge et al., 2013; Isola et al., 2014; Isola, Parikh, et al., 2011; Isola, Xiao, et al., 2011). Here, we assess how image memorability is affected by the feelings an image evokes, and provide an image set for future studies that incorporates valence and arousal ratings and memorability scores for hundreds of natural scenes (**Figure 1**).

Foundational studies on memorability have found that this property is distinct from, but influenced by, other perceptual and conceptual factors that determine how well an item is remembered (Bylinskii et al., 2015; Dubey et al., 2015; Isola et al., 2014; Khosla et al., 2015). Although much of the variance in memorability remains unexplained, memorability can be influenced by low and high-level perceptual properties – from simple image features to conceptual meaningfulness (Brady et al., 2019; Huebner & Gegenfurtner, 2012; Vokey & Read, 1992). Higher-level properties, such as those that represent the 'meaningfulness' of objects, images, and scenes, are most informative for the memorability of an item (Brady et al., 2019; Dubey et al., 2015; Kramer et al., 2023). For example, images and faces that are unusual are typically highly memorable (Bainbridge et al., 2013; Khosla et al., 2015). Images that contain social information such as faces or human figures are also better remembered (Brady et al., 2019; Kramer et al., 2023).

Images can also be endowed with meaning via the internal feelings and emotional representations that they elicit. The feelings of anger, happiness, and amusement that arise from a visual stimulus color our interpretation of its meaning or relevance. Thus, image memorability could be related to the feelings and emotions evoked by an image. Work relating memorability to our internal feelings about a given image has thus far primarily explored the relationship between discrete emotional states and memorability, finding images that evoke disgust, fear, and amusement to be more memorable than those that evoke awe or contentment (Khosla et al., 2015). However, the relationship between large-sample memorability scores and the continuous affective dimensions that are associated with evoked emotion, such as valence or arousal, has not been directly explored. Part of this limitation arises because most studies of memorability use image sets that do not contain affective information (Isola et al., 2011). Furthermore, the few prior studies examining memorability for emotional scene images have not verified whether memorability scores are consistent across different sets of viewers, specifically for these emotional scene images (Khosla et al., 2015). Thus, it is unclear (i) whether image memorability is reliable for images that span a wide range of evoked arousal and valence and (ii) whether arousal and valence predict memorability.

To answer these questions, we assembled an image set of hundreds of natural scenes varying in affective content, and acquired (i) valence and arousal ratings for each image; and (ii) estimates of how memorable vs. forgettable each image is. We determined the consistency of memorability scores across this large set of scene images that spanned a wide range of valence and arousal. Finally, we created a model that determined the predictive power of valence and arousal on image memorability. Together, these two complementary studies allowed us to assess how robust memorability is to varying affective dimensions, and whether affect explains a substantial amount

of variance in what is memorable or not. We also share our image set VAMOS – Valence, Arousal, and Memorability of Scenes – with the scientific community to inspire future work on interactions between evoked affect and memorability (https://osf.io/ufg89/).

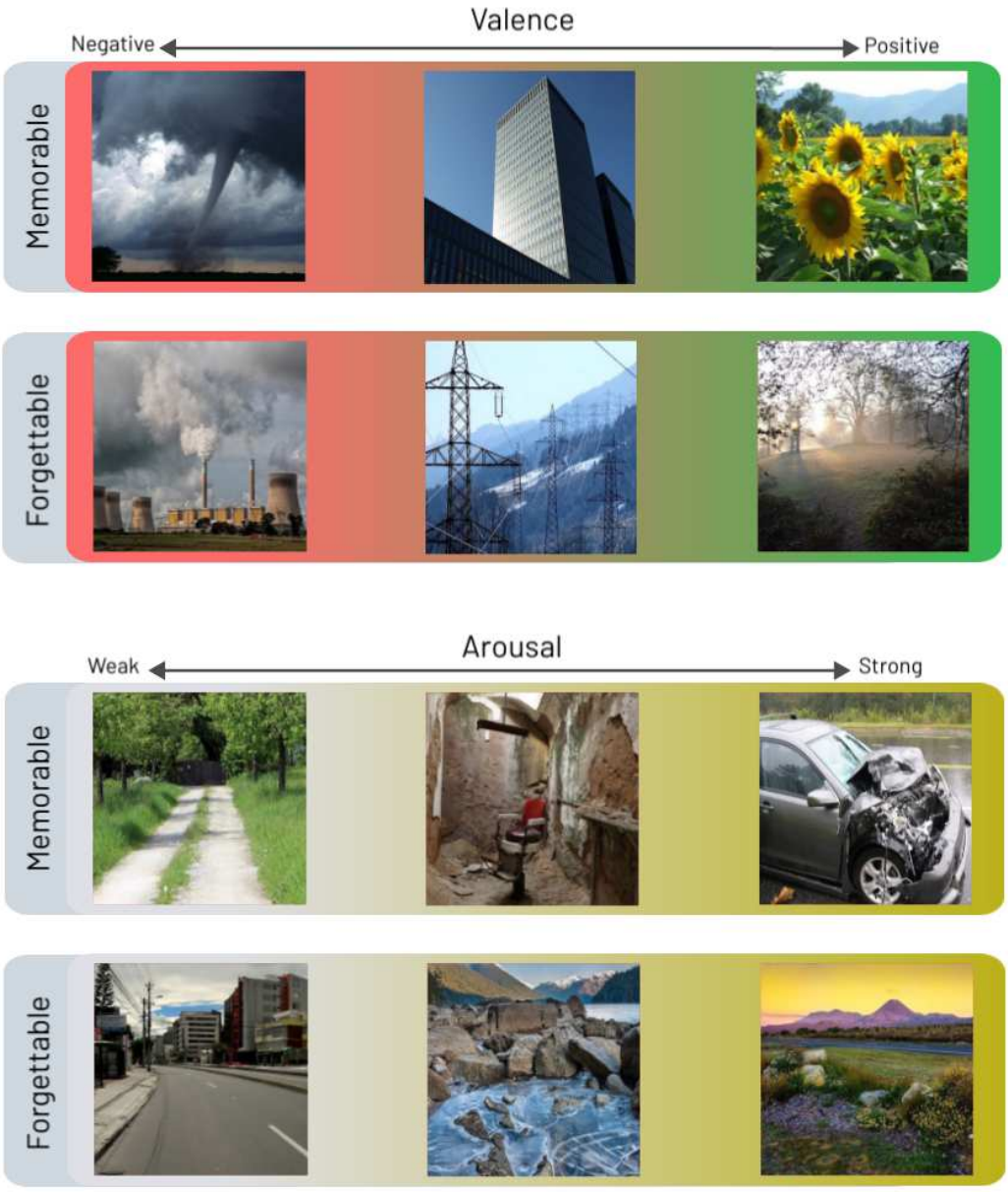


**Figure 1.** Sample images sorted by evoked valence/arousal (from Experiment 1) and memorability (from Experiment 2). Scene images were selected to span a wide range of average valence and arousal scores. The figure above shows examples of memorable and forgettable images from across the spectra of valence and arousal.

## Methods

We conducted two experiments to determine the relationship between the memorability of an image and the feelings that it evokes. The aim of Experiment 1 was to quantify valence and arousal for each image in a set that was selected to evoke a broad affective range. In this experiment, a group of participants provided ratings of the valence and arousal evoked by 906 scene images. These images, spanning a wide range of valence and arousal ratings, were then used in Experiment 2. Experiment 2 determined whether the same images are consistently remembered or consistently forgotten across a population of individuals. Within this experiment, we recorded memory performance as participants completed a continuous recognition task (CRT). After establishing the reliability of memorability scores for emotional images within this experiment, we used the average valence and arousal ratings from Experiment 1 to determine the degree to which memorability is predicted by the feelings that each image provokes.

### Stimuli

Both experiments used 906 real-world color scene images sourced from several existing image sets (Nencki Affective Picture System (NAPS) (Marchewka et al., 2014), Open Affective Standardized Image Set (OASIS) (Kurdi et al., 2017), Complex Affective Scene Set (COMPASS) (Weierich et al., 2019), and Places (Zhou et al., 2018). This constructed image set featured a diverse set of realistic indoor and outdoor scenes with a wide range of associated valence and arousal. By using only scene images, we were able to feature complex real-world stimuli while controlling for the presence of objects that may differentially affect the memorability of scene images with vs. without those objects (e.g. human figures, animals, or legible text, which were all excluded from the stimulus set). All images were cropped to be square and resized to 400 x 400 pixels. This image set, along with valence/arousal ratings and memorability scores for each

scene, can be accessed at the following website: https://osf.io/ufg89/.

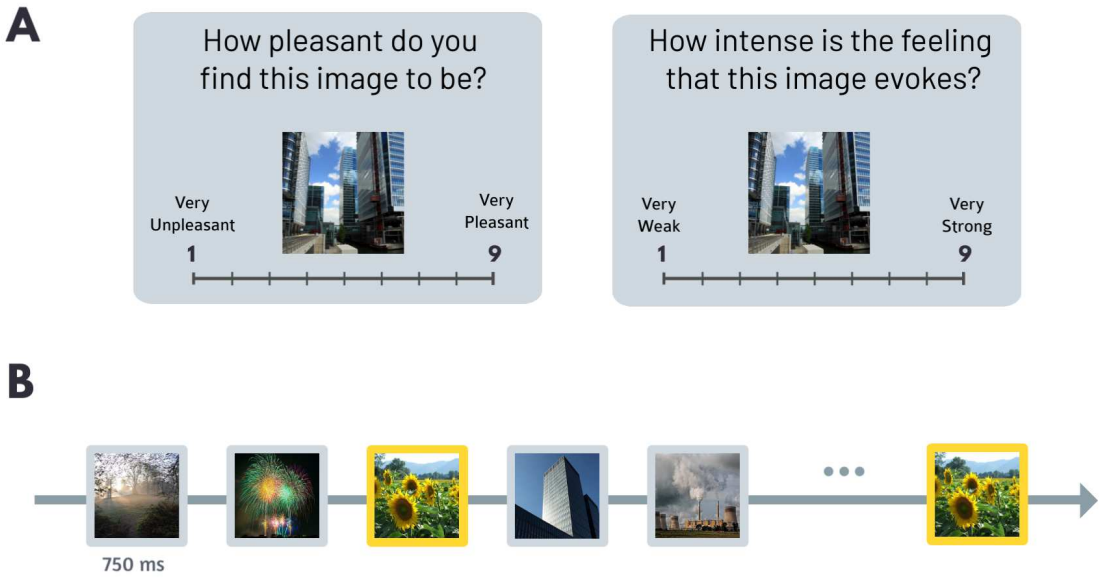


**Figure 2.** Overview of experimental procedures. **A.** The goal of Experiment 1 was to collect standardized ratings of valence and arousal for each of 906 images that were selected to vary in the feelings they evoked. Participants provided valence and arousal ratings using a nine-point Likert scale. On these scales, 1 represents a very negative (valence) or very weak (arousal) judgment and 9 represents a very positive (valence) or very strong (arousal) judgment **B.** A continuous recognition task was used in Experiment 2 to obtain memorability scores for the set of 906 images from Experiment 1. Participants were instructed to respond with a key press when they encountered a previously shown target image (shown in yellow for illustration purposes) within a stream of scene images. The second presentation of this target was separated by at least 30 seconds from the first presentation.

## Experiment 1

Although the image sets from which we sourced our scenes sometimes included valence and arousal ratings for each image, these ratings were neither standardized across image sets nor uniformly obtained. We therefore obtained participant ratings of valence and arousal for the 906 scene images using an online experiment. We then used a linear regression model to determine whether an image's memorability (obtained from Experiment 2) was strongly and significantly predicted by its evoked valence and arousal.

*Participants.* We recruited 528 adults (43% women, mean age = 33.4 years old, mean education = 14.6 years, race: White = 69%, Black = 7%, Asian = 14%, Native Hawaiian/Pacific Islander = 2%, more than 1 race/Other = 8%) via the online experimental platform Prolific and the Columbia University SONA participant recruitment system. Participants were compensated either at a rate of $6.50/hour or with 1 course credit for participation in the approximately 30-minute experiment. Participants were required to have an IP address in the United States and speak fluent English. Participants provided consent in compliance with procedures approved by the Columbia University Institutional Review Board (IRB).

*Procedure.* Participants viewed a self-paced stream of images randomly selected from the full 906 image set. Images were presented centrally on a white background above a question that asked participants to judge the image along a scale. Participants were asked to first rate the valence of an image ("How pleasant do you find this image to be?") using a nine-point Likert scale (1 = very unpleasant, 9 = very pleasant) (**Figure 2A**). Immediately following this judgment, participants were asked to rate the arousal of that same image ("How intense is the feeling that this image evokes?") using a nine-point Likert scale (1 = very weak, 9 = very strong). Participants used a mouse or trackpad to move the slider along the Likert scale to complete their ratings.

412 participants (78% of our sample) viewed and rated approximately 100 images each. The remaining participants viewed and rated 35-76 images each, largely due to experimental bugs that led to image presentation errors or errors collecting ratings. To make up for that data loss, we ensured that each image was seen and rated by at least 20 participants in our full sample; and the vast majority of images (98%) were seen and rated by at least 40 participants. Further, we repeated our analyses with data only from the 412 participants who saw and rated ~100 images each, and obtained the same pattern of results.

The valence and arousal of each image was calculated using the average of all participant responses for that image. These ratings were entered into a GLM to predict the memorability of an image (obtained in Experiment 2, below) given its valence and arousal.

## Experiment 2

We calculated memorability scores for the 906 scene images used in Experiment 1 with an online experiment and determined the consistency of image memorability across groups of participants.

*Participants.* We recruited 640 adults from the United States via the online experimental platform Amazon Mechanical Turk (MTurk). Participants were compensated at a rate of $6.50/hour for participation in the approximately 10-minute experiment. To be eligible, participants must have completed at least 500 tasks with a 98% approval rate on MTurk (Wakeland-Hart et al., 2022). They were also required to have an IP address in the United States and speak fluent English (as indicated by self-report and comprehension questions). Participants provided consent in compliance with procedures approved by the Columbia University Institutional Review Board (IRB). After participants provided informed consent, demographic information was collected. Participants then completed three basic English comprehension questions which were designed to ensure participants were proficient in English and could fully understand the task instructions. 565 participants successfully completed the task and were included in the final sample (42% women, mean age = 37.9 years old, mean education = 15.8 years, race: White = 84%, Black = 6%, Asian = 7%, American Indian/Native American = 1%, More than 1 race = 1%). The remaining 75 participants were excluded from the final sample because they did not attempt the task, answered the English screening questions incorrectly, or failed at least 30 vigilance checks during the main task.

*Procedure*. Participants completed a continuous recognition task (CRT) adapted from the Memorability Experiment Maker (Bainbridge, 2017) (**Figure 2B**). We used this task because it is commonly used in studies of image memorability, including early demonstrations of this phenomenon (Isola et al., 2011; Isola et al., 2014). The continuous recognition task featured a continuous stream of 240-260 images which appeared for 750 ms each, presented centrally against a white background. This variation in the number of images viewed was due to experimental bugs that led to image presentation errors. To make up for that data loss, we ensured that each image was seen by at least 40 participants. During the 1000 ms interstimulus interval, a central fixation cross was presented against a white background (Bainbridge et al., 2017; Khosla et al., 2015). A subset of images (52-60 images) from this stream were selected to be 'target images' and repeated once during the task, at least 30 seconds (17 images) apart. Separating target images in this way ensures that estimates of image memorability are due to long-term memory representations as opposed to working memory. Another subset of non-target 'vigilance images' repeated once during the stream at a shorter interval (1-5 images apart). There were 59-64 of these vigilance images; the variation in the number of vigilance images across participants is due to experimental bugs. These closer repeats served to maintain task vigilance and were designed to be easily detected by participants. In addition to the target and vigilance images, non-target 'filler images' were shown in between repeated images. Target, vigilance, and filler images were randomly sampled from the stimulus set; each image therefore appeared as a target, vigilance, and filler image across participants. Analyses of image memorability considered across-participant performance on target and filler images.

During the continuous recognition task, participants were instructed to press the 'r' key when they saw an image that was shown previously in the stream. Responses to each image were

recorded. Correct identification of a repeated image was considered as a ‘hit’ and failure to identify a repeated image was considered as a ‘miss’. Misidentification of the first (or only) presentation of an image as a repeat was classified as a ‘false alarm’. As noted above, participants who ‘missed’ over 30 vigilance checks (roughly 50%) were excluded from the data analyses.

We then assigned memorability scores to each of the 906 images in the set using the average performance of participants for each image, including responses to the image in both the filler and target conditions. In sum, memorability scores for each image were represented by a corrected recognition rate (CR). This CR score was calculated by subtracting the false alarm rate for that image (the proportion of times that image was falsely identified as a repeat when it was a first or only presentation) from the hit rate for that image (the proportion of times that image was successfully identified as a repeat) across participants. Finally, a split-half consistency analysis was conducted to determine the stability of the CR of each image (see Data Analysis).

## Data analysis

*Consistency analysis.* To test whether image memorability was reliable across individuals, we conducted a consistency analysis in which the CR for each image was derived separately using two random halves of the participant data and then the Spearman rank correlation was obtained across the two halves (Isola, et al., 2011). We conducted 1,000 random split-halves, obtaining the Spearman correlation each time, and then averaged the Spearman correlations across these iterations, resulting in an average across-participant consistency score.

*GLMs.* A general linear regression model (GLM) in R with quadratic and linear predictors was

used to predict the memorability score of an image (from Experiment 2) from its average valence and arousal ratings (from Experiment 1). We included quadratic predictors to test whether the most extreme images in terms of valence were most memorable (i.e. both the most positive and the most negative images). We did not expect arousal to have a quadratic relationship with memorability but we nevertheless included the quadratic term for completeness. Within this model, both valence and arousal were fixed effects and both variables were z-scored. The model formula was as follows:

Model #1:

*image memorability ~ image valence+ image valence$^2$ + image arousal +image arousal$^2$*

A model including interaction terms for both linear and quadratic effects was also run. There were no significant interactions and thus that model is not reported here.

We also constructed a logistic regression model to predict the memory of individuals (i.e., whether a given image was remembered or not for a given participant, from Experiment 2; image memory, 0 = miss; 1 = hit) from group-averaged valence and arousal ratings for each image (from Experiment 1). Valence and arousal ratings were z-scored, and both linear and quadratic terms were included, as follows:

Model #2:

*image memory ~ image valence + image valence$^2$ + image arousal + image arousal$^2$ + (1 | participant)*

That is, for each participant in Experiment 2, we predicted their individual image memory from the group-averaged valence/arousal ratings for each image from Experiment 1.

Finally, we constructed a model to predict image memorability (estimated from the across-participant data in Experiment 2) from individual-participant arousal and valence ratings (rather than group-averaged ratings, from Experiment 1) for each image. Valence and arousal ratings were z-scored separately within each individual, and both linear and quadratic terms were included, as follows:

Model #3:

$$image\ memorability \sim image\ valence + image\ valence^2 + image\ arousal + image\ arousal^2 + (1 \mid participant)$$

That is, for each participant in Experiment 1, we used their own valence and arousal ratings for each image to predict image memorability from the across-participant data in Experiment 2.

## Results

We used the data collected in Experiments 1 and 2 to characterize the relationship between the memorability of each image and the affect associated with it. We first separately calculated and validated measures of valence and arousal (Experiment 1) and memorability (Experiment 2). We then predicted image memorability from image valence and arousal using linear and quadratic regression models.

## Validating affect ratings and image memorability

*Affect ratings validation.* We calculated a measure of the affect associated with each image, operationalized as the mean valence and mean arousal (across participants) for each of the 906 images (Experiment 1). Consistent with prior work (Kurdi et al., 2017; Marchewka et al., 2014; Weierich et al., 2019), the image set demonstrated a wide range of both valence and arousal (mean valence = 4.65, standard deviation = 1.49, range = (1.33, 8.42); mean arousal = 4.46, standard deviation = 0.83, range = (2.55, 7.31)). These measures of valence and arousal were significantly, but modestly, consistent across participants (Valence Spearman's rank correlation: $\rho = 0.10$, $p < 0.001$; Arousal Spearman's rank correlation:: $\rho = 0.11$, $p < 0.001$). Valence and arousal were modestly correlated with each other (Pearson's $r = 0.16$, $t(904) = 5.01$, $p < .001$).

If the most positive and negative images evoke the highest arousal (Kuppens et al., 2012), we might additionally observe a quadratic relationship between valence and arousal. We therefore used a regression model to predict the arousal of each image using linear and quadratic valence predictors. We found a significant quadratic component, indicating that both strongly negative and strongly positive images (vs. relatively neutral images) are related to higher arousal (Quadratic component: $\beta = 0.23$, $SE = 0.0082$, $t(904) = 27.63$, $p < .001$; Linear component: $\beta = -1.98$, $SE = 0.076$, $t(904) = -25.99$, $p < 0.001$; **Figure 3**). Together, these findings validate our measures of valence and arousal by replicating prior work showing their linear and quadratic relationships (Kuppens et al., 2012).

Prior work has shown that the affect associated with an image is predictive of whether an individual goes on to remember that image (Kensinger, 2007). Thus, we used a logistic regression model (0 = miss; 1 = hit) to predict the memory of individuals in Experiment 2 using average

valence and arousal ratings from Experiment 1 as predictors, with both linear and quadratic terms for each (Model #2, see Methods/Data analysis). We found that both valence and arousal were significant predictors of participants' image memory. Images that were more strongly negative were better remembered by individuals (Valence, linear effect: $\beta = -0.08$, SE = 0.01, $z(564) = -7.43$, $p < 0.001$). There was also a significant quadratic effect of valence, in which moderately negative images were better remembered than extremely negative images or positive images (Valence, quadratic effect: $\beta = -0.07$, SE = 0.01, $z(564) = -5.12$, $p < 0.001$). Images that were more strongly arousing were also better remembered by individuals (Arousal, linear effect: $\beta = 0.15$, SE = 0.01, $z(564) = 10.26$, $p < 0.001$). There was also a quadratic effect of arousal, reflecting superior memory for highly arousing images but similar memory for low- to moderate-arousal images (Arousal, quadratic effect : $\beta = 0.04$, SE = 0.008, $z(564) = 4.21$, $p < 0.001$). Together, these results served as a validity check by verifying prior work linking higher arousal and more negative valence to superior memory (Mather & Nesmith, 2008; Kensinger, 2009; Mather, 2007). However, these affective dimensions represented only a small portion of individual memory determinants: very little variance in individual memory was explained by valence and arousal (mcFadden's $R^2 = 0.003$). Note, however, that this analysis considered only what a *given individual* remembers or forgets, and not the consistency with which a given image was remembered vs. forgotten across individuals (i.e., its memorability); we examine memorability below.

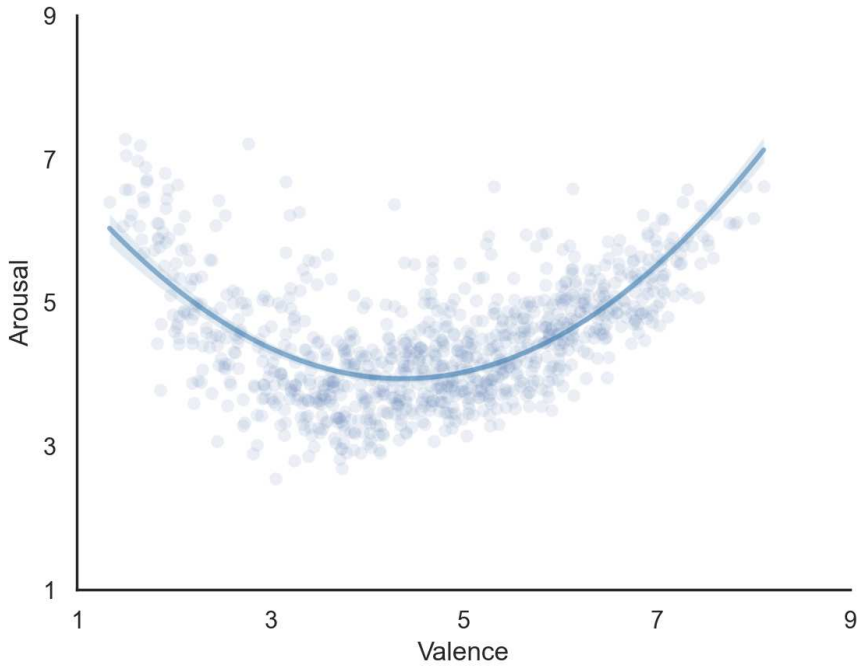


**Figure 3.** Average valence and arousal scores for each image (Experiment 1) were compared, revealing a positive quadratic relationship between valence and arousal. Both extremely negative (low valence) and extremely positive (high valence) images were generally more highly arousing while neutral images were less arousing. The blue line indicates model predictions with 95% CIs; the dots indicate mean valence and arousal for each image

*Memorability validation*. Many studies of memorability have used neutral images, making it unclear whether across-participant reliability in what is remembered vs. forgotten extends to images that are selected to span a wide range of evoked feelings (Wakeland-Hart et al., 2022). To evaluate whether images within our affective image set were consistently remembered across people, we performed a split-half consistency analysis. We calculated the memorability scores for each image using data from half of the participants who completed the continuous recognition task (Experiment 2). We compared these scores to the memorability scores for each image derived from the left-out participant data. We found that image memorability was highly reliable

(mean Spearman's rank correlation across 1000 split halves: $\rho$ = 0.32, $p < 0.001$; mean memorability = 0.66, standard deviation = 0.08; **Figure 4**). In other words, in our sample, individuals tended to remember or forget the same images.

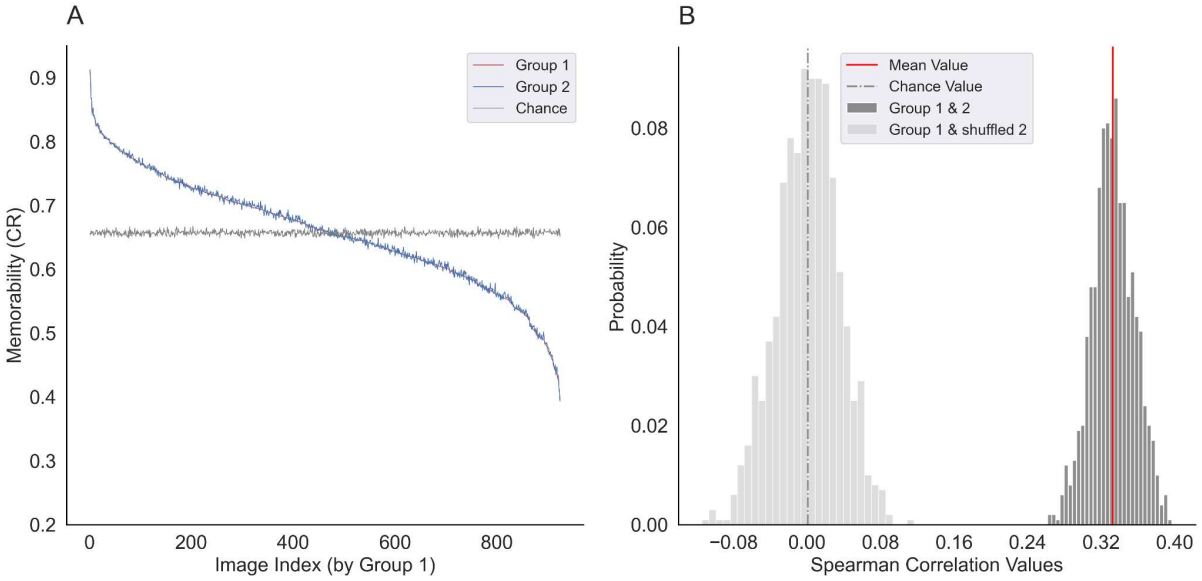


**Figure 4**. **(A)** Assessment of image memorability. Participants were randomly divided into two groups 1000 times. For each iteration, memorability scores for each image were obtained for Group 1 and sorted from highest to lowest. Memorability scores for each image were then obtained for Group 2 and sorted by the same image index as Group 1. For visualization purposes, the figure above shows the memorability of each image averaged across 1000 iterations for Group 1 (red) and Group 2 (blue); note that the Spearman correlation reported in the text was conducted for each iteration separately before being averaged. The gray line shows memorability scores averaged over 1000 iterations in which image identities were randomly shuffled each time. If there is no consistency in what is remembered vs. forgotten across groups, the line for Group 2 should approximate this permuted average. As expected, however, image memorability was highly reliable, as indexed by a statistically significant Spearman correlation between Group 1 and Group 2 (see text for details). **(B)** Distribution of split-half correlations. Each of the Spearman correlations between Group 1 and Group 2 for all 1000 iterations formed a distribution (dark gray) that could be compared to the null distribution of correlations between Group 1 and shuffled Group 2 image identities (light gray). The distribution of Group 1 vs. Group 2 correlations is centered around the mean ($\rho$ = 0.33) and demonstrates no overlap with the Group 1 vs. shuffled Group 2 distribution. This further demonstrates the reliability of the memorability ratings, as there are no instances of the Group 1 vs. shuffled Group 2 correlation (null distribution) reaching the mean value of the Group 1 vs. Group 2 Spearman correlation in 1000 iterations of the test.

One limitation of the above analysis is that our image set included images that were not strongly negative or positive. Could weakly valenced images be driving the overall reliability of memorability for our image set? To test this, we removed neutral images from our image set and examined memorability for only those images that were strongly negative (valence less than or equal to 3) and strongly positive (valence greater than or equal to 7). Memorability for these most negative and positive images was still reliable (mean Spearman's rank correlation across 1000 split halves: $\rho = 0.306$, $p < 0.001$).

We next repeated this approach for those images in the upper half of arousal ratings based on a median split, to determine whether memorability was reliable for the most arousing images. Here, too, we found that memorability was still reliable when only considering these relatively arousing images (mean Spearman's rank correlation across 1000 split halves, $\rho = 0.310$, $p < 0.001$).

Together, these findings show that memorability is robust: the same images are consistently remembered vs. forgotten even in an image set that spans negative to positive affect and low to high arousal. Further, image memorability remains robust even when considering only non-neutral images and only the most arousing images.

## Predicting memorability from image valence and arousal

Memorability has been conceptualized as an image property that arises from a combination of visual features (Rust & Mehrpour, 2020). Valence and arousal may be features that contribute to memorability. However, the direct relationship between these three image properties has not been characterized. How predictive of memorability are valence and arousal? To understand the

influence of these affective measures on memorability, we created a fixed-effects model including both average valence and average arousal as predictors (Model #1, see Methods/Data analysis).

The results from this model revealed that, controlling for the other variable, both valence and arousal were weak but statistically significant predictors of memorability. For valence, the linear predictor indicated that, overall, negative images were better remembered than positive images ($\beta$ = -0.017, SE = 0.0030, t(904) = -5.81, $p < 0.001$). Valence was also negatively quadratically related to memorability ($\beta$ = -0.014, SE = 0.0035, t(904) = -3.93, $p < 0.001$), with both extremely positive and extremely negative images being slightly less memorable than images that were moderately negative. These results, in sum, show a modulated negative relationship between valence and memorability when controlling for arousal: weakly negative images are most memorable and strongly positive images are least memorable (**Figure 5A**). For arousal, only the linear term significantly predicted memorability (Linear: $\beta$ = 0.027, SE = 0.0039, t(904) = 6.73, $p < 0.001$; quadratic: $\beta$ = 0.004 , SE = 0.0022, t(904) = 1.64, p = 0.10), indicating that more arousing images are generally more memorable (**Figure 5B**).

To visualize the effects of valence and arousal simultaneously, we made a 3-dimensional plot of valence, arousal, and memorability (**Figure 5C**). This plot shows that the most memorable images are generally those that are more negatively valenced and more arousing. However, the peak of the memorability plane indicates that moderately negative, rather than extremely negative, images are most memorable.

Of note, this model (Model #1) predicted memorability better than a model with only linear valence and arousal terms ($AIC_{quadratic}$= -1911.2, $AIC_{linear}$= -1899.3). However, while valence and

arousal predicted image memorability, the relationships were relatively weak (i.e., small beta values); and, indeed, only a modest amount of variance in memorability was explained by evoked affect ($R^2$ = 0.078). This pattern of results remained consistent when the interaction between valence and arousal was included as a predictor in the model, with no significant effect of the interaction itself.

Given the relatively weak relationships observed between valence, arousal, and memorability, we conducted an exploratory analysis to assess whether category-level scene information could provide additional insights or enhance memorability prediction. We labeled each image as being either natural or manmade; for manmade images, we also labeled them as being either indoor or outdoor. We then examined the relationship between these category labels and valence ratings, arousal ratings, and memorability scores. Natural images were generally more positive and arousing compared to manmade images (Valence: $\beta = 0.62$, SE = 0.08, $t = 7.97$, $p < 0.001$; Arousal: $\beta = 0.24$, SE = 0.04, $t = 5.32$, $p < 0.001$). However, natural images were not more memorable than manmade images ($\beta = 0.003$, SE = 0.005, $t = 0.8$, $p = 0.42$). We obtained the same pattern of results when comparing natural images only to manmade images in outdoor settings, given that natural images were always outdoors. Within the manmade category, outdoor scenes evoked more positive valence and higher arousal than indoor scenes (Valence: $\beta = 0.18$, SE = 0.05, $t = 3.43$, $p < 0.001$; Arousal: $\beta = 0.10$, SE = 0.03, $t = 3.49$, $p < 0.001$). However, these scene categorizations did not significantly predict memorability scores ($\beta = 0.0004$, SE = 0.003, $t = 0.13$, $p = 0.9$).

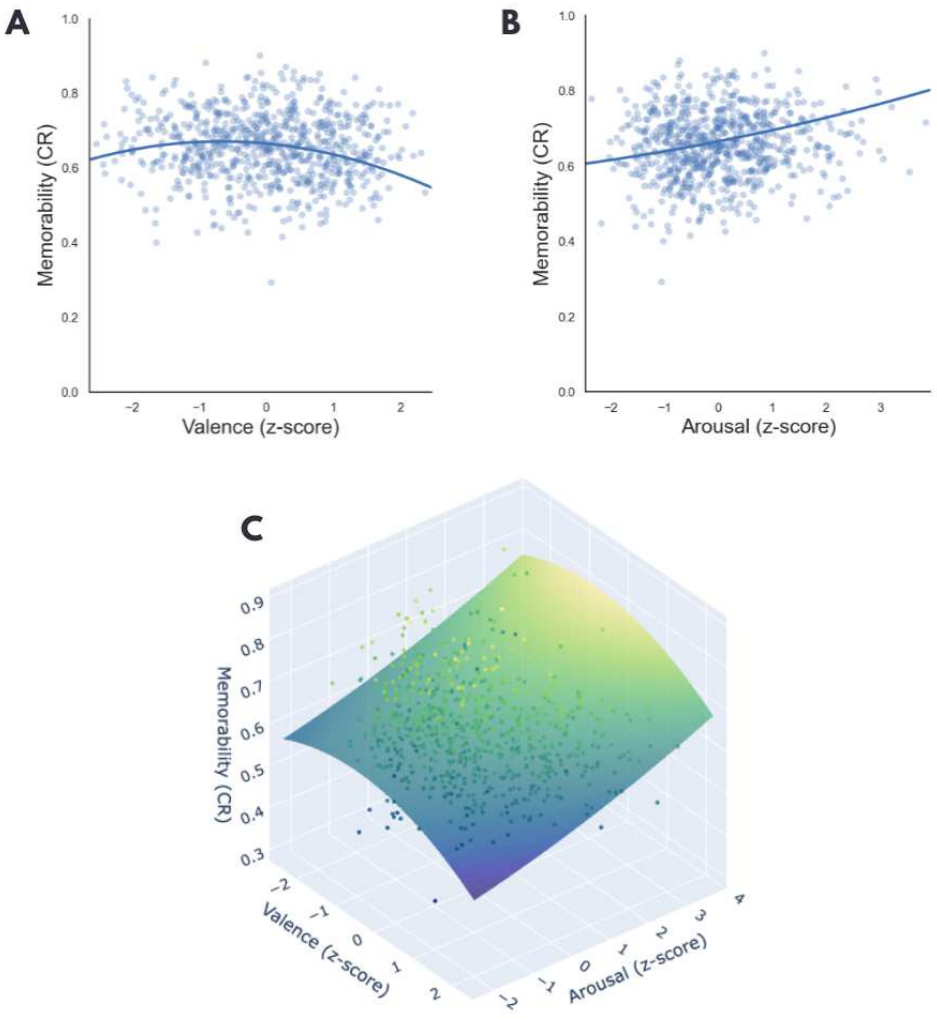


**Figure 5.** Relationship between image memorability and evoked affect. **A.** Valence predicts memorability (operationalized as corrected recognition) with both negative linear and negative quadratic relationships. The blue line depicts model predictions; the dots indicate mean valence and memorability for each image **B.** Arousal is positively and linearly related to memorability. The blue line depicts model predictions; the dots indicate mean arousal and memorability for each image **C.** The simultaneous influence of valence and arousal on memorability is depicted using multidimensional topography. Darker hues indicate lower memorability and lighter hues indicate higher memorability. Each point represents an image within the valence-arousal-memorability space.

*Investigating the relationship between image memorability and individual-participant ratings of evoked affect.* While both across-participant valence ratings and across-participant arousal ratings were predictive of memorability, the relationships were relatively weak. Additionally,

across-participant ratings of arousal and valence were only modestly reliable, raising the possibility that these ratings do not do a good job of predicting memorability because they do not adequately capture individual emotion responses. For example, a given image may have an average valence rating of 5 on a 1-9 scale, but this rating may mischaracterize an underlying distribution in which some participants view it very negatively and others very positively. Thus, average ratings may do a poor job at predicting memorable images if memorable images lead to polarized valence ratings.

We therefore examined the relationship between memorability (estimated from across-participant data in Experiment 2) and measures of image-evoked affect (from Experiment 1), but using individual-participant ratings from Experiment 1 rather than group-averaged ratings. That is, for each participant in Experiment 1, we used their own valence and arousal ratings for each image to predict image memorability from the across-participant data in Experiment 2 (Model #3, see Methods/Data analysis).

Valence and arousal remained poor predictors of memorability, but were *weaker* predictors when using individual-participant ratings as opposed to group-averaged ratings. The strongest predictors of memorability were the linear valence and arousal terms (Valence, linear effect: $\beta$ = -0.007, SE = 0.0002, $z(527) = -32.67$, $p < 0.001$; Arousal, linear effect: $\beta = 0.007$, SE = 0.0002, $z(527) = 29.96$, $p < 0.001$), reflecting higher memorability of images that were more negative and more arousing – as we observed in our model with group-averaged ratings of valence and arousal. When using individual-participant ratings of valence, however, we did not observe a quadratic effect of valence (Valence, quadratic effect: $\beta = 0.0003$, SE = 0.0002, $z(527) = 1.63$, $p = 0.10$). Finally, we observed a quadratic effect of arousal (Arousal, quadratic effect: $\beta = 0.002$, SE = 0.0002, $z(527) = 13.06$, $p < 0.001$), in which low- to moderate levels of arousal were associated

with similar memorability, but highly arousing images were more memorable. Altogether, however, the model with individual-participant valence and arousal ratings explained even less variance than our primary model with group-averaged valence and arousal ratings: only 1% of the variance in memorability was explained ($R^2$ = 0.013), as opposed to 8% in our primary model .

These findings are inconsistent with the possibility that group-averaged affective ratings only modestly predict memorability because individuals disagree on the affect evoked by images. Instead, group-averaged ratings were better predictors of memorability than individual ratings, as would be expected if idiosyncratic emotional responses have little bearing on what is consistently remembered across individuals.

Together, our analyses converge in showing that images that are more arousing are associated with better memory, both when predicting memory of individuals and when examining the consistency of memory across individuals (i.e., memorability). Further, images that are rated as *moderately* negative at the group level are better remembered, both when predicting memory of individuals and when examining the consistency of memory across individuals. This quadratic effect of valence – in which moderately rather than extremely negative images are better remembered – failed to reach significance when individual, rather than group-averaged, ratings of valence were used to predict image memorability across individuals; but we do not over-interpret this null effect given that *individual* ratings of affect generally did a poor job of explaining *across-individual* memorability (1% explained variance), as noted above.

## Discussion

We investigated how the affective dimensions of valence and arousal influence the likelihood that

a given image was more consistently remembered over others. Specifically, we tested the relationship between valence, arousal, and memorability by 1) constructing one of the largest sets of real-world scenes to contain normative valence and arousal ratings and memorability scores and 2) building a statistical model that determined the predictive power of valence and arousal on image memorability. Generally, we found that while evoked affect is significantly related to memorability, the variance in memorability explained by arousal and valence is extremely small. Instead, memorable images can span the continuum from low to high arousal and negative to positive valence.

Our results extend prior work exploring the contribution of valence and arousal to individuals' memories (Kensinger, 2007) and work showing that the general memorability of an image is influenced by the internal feelings that an image evokes, such as feelings of disgust, amusement, and fear (Khosla et al., 2015). We contribute to this line of work by determining the nature and strength of the relationships between valence, arousal, and memorability. We found negative images to be, on the whole, more consistently remembered across individuals than neutral or positive images. Intriguingly, *moderately* negative images were more memorable than highly negative images, a point we discuss in more detail further below. More arousing images were also more memorable than less arousing images. Further, valence and arousal made independent, albeit modest, contributions to image memorability. Together, our results show that both the degree of stimulation and the degree of positivity or negativity evoked by an image can influence whether it persists in our memories.

Our findings complement prior work demonstrating that, at a group level, the magnitude of the memory benefit to emotional over neutral images is reliable over delays from a day to several

years (Schümann et al., 2020), even when the size of the emotional memory boost for individual participants was variable over time (Schümann et al., 2020). Together, our results and those of Schümann et al. suggest that there is across-participant stability in the superiority of memory for emotional images, perhaps due to the intrinsic properties of the negative images, such as visual distinctiveness and semantic content, which may evoke similar cognitive operations across individuals and contribute to consistent memory outcomes. In other words, memory of negative images is reliably enhanced despite idiosyncratic thoughts or feelings within individuals that may make individual-level memories more variable over time.

These results provide evidence that valence and arousal do not solely, or even strongly, determine memorability. Instead, they join a network of features that imbue an image with meaning and influence its memorability. For example, prior work has shown that images containing social information (Isola et al., 2014), atypical content (Saleh et al., 2013), and food (Kramer et al., 2023) are more memorable. Additionally, images that evoke emotions such as disgust or fear are more memorable (Khosla et al., 2015). Overall, it seems that images with salient novelty, utility, or behavioral relevance are inherently more likely to be consistently remembered across individuals. These attributes, related to the high-level semantic meaning and conceptual features of an image (Isola et. al., 2014; Hovhannisyan et. al., 2021), along with, to a lesser extent, perceptual features like visual complexity (Kyle-Davidson et. al., 2023; Brook et. al., 2024; Kramer et. al., 2023), collectively contribute to how consistently an image is remembered across people. Further work can determine how these image attributes combine to influence memorability. For example, it would be informative to determine whether moderately negative images are better remembered than extremely negative images because the former are more distinctive or novel. It could be that extremely negative images – such as burning buildings – are

more frequently seen in the media than moderately negative images – like cars stuck in unusual places; if so, image valence may have its effects on memorability partly via distinctiveness or novelty. However, this possibility is complicated by the fact that prior work has found inconsistent results with respect to whether more distinctive or more typical items are more memorable. Some accounts of memorability posit that there is a simple linear relationship between the typicality of an item and its memorability (Bainbridge et. al., 2013; Xie et. al., 2020). Others, however, suggest that the typicality vs. distinctiveness of an image contribute to memorability in a complex fashion and may be related to similarity or discriminability at different levels of the hierarchy of visual features (Kramer et. al., 2023; Koch et. al., 2020). Further work could determine the separate and interacting effects of visual and emotional features on image memorability, and investigate whether the contributions of valence to memorability are related to image distinctiveness, typicality, or neither.

This study is among the first to relate image memorability to the dimensions of affect that signify our internal feelings related to each image (Davis & Bainbridge, 2023; Khosla et al., 2015). However, it is important to acknowledge the ways in which future research can expand upon the findings within this study. For example, future studies can investigate the relationship between affect and memorability for images with social content. Image sets that include photos of people may have a larger capacity to trigger extremes in emotion and lead to larger relationships between emotion and memorability than what can be observed from scenes. Additionally, the relationship between affect and memory can be further studied in work that accounts for broader affective states that extend over large periods of time (Mather & Knight, 2009; Tambini et al., 2017). Further, although commonly used in memorability research (Isola et al., 2011; Isola et al., 2014; Li et al., 2022; Wakeland-Hart et al., 2022; Davis et al., 2022), the continuous recognition

task provides a measure of recognition memory over a relatively short timescale (30 seconds to several minutes) and with no distinct consolidation period between encoding and retrieval. Future work could explore the relationship between image-evoked affect and memorability after a period of consolidation lasting days to weeks, by utilizing a delayed memory task rather than the continuous recognition test. This is particularly important because the effects of memorability, and the interaction between memorability and emotion in predicting individuals' memory, can change with consolidation (Morales-Calva & Leal, 2024).

Finally, future work can examine how the affective determinants of memorability may change over the lifespan. Although not focused on this question, our samples included a broad age range – which lends both strengths and potential limitations. The wide age range increases the generalizability of our findings, ensuring that the pattern of effects obtained was not limited to relatively young college students. However, it is important to rule out that the wide age range could have biased our results, for example due to cognitive and emotional changes that occur over the lifespan (Carstensen, et. al., 2000; Craik & Bialystok, 2006; Kensinger, 2009). We believe this is unlikely for a few reasons. First, the samples in both of our Experiments (which we used to obtain valence/arousal ratings and memorability scores) included a wide age range; that is, the valence and arousal ratings should be representative of a broad age group and so too should the memorability scores. In that way, the relatively modest correlation between valence/arousal and memorability cannot be due to drastically different age ranges that were tested. Second, when we eliminated the small but statistically significant difference in mean age across samples by removing data from the oldest participants, we still obtained the same pattern of results. Finally, reanalyzing the data with only participants 18-40 years old led to the same pattern of results as well: in all cases, valence and arousal were only modestly related to memorability, and the significance (or lack thereof) of particular effects was unchanged. Together, this suggests that the

broad age range we tested can be considered a strength for generalizability rather than a limitation.

In sum, the findings of this study contribute to our understanding of memorability as being aided, in small part, by how negatively we view scenes and how intensely they make us feel. Nevertheless, we show that valence and arousal neither singularly nor jointly account for a large amount of variance in image memorability. Instead, they join the assemblage of conceptual and perceptual visual features that contribute to an image's overall memorability.

## References


Bainbridge, W. A. (2017). The memorability of people: Intrinsic memorability across transformations of a person's face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 706–716. https://doi.org/10.1037/xlm0000339

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology. General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Brady, T. F., Alvarez, G. A., & Störmer, V. S. (2019). The Role of Meaning in Visual Memory: Face-Selective Brain Activity Predicts Memory for Ambiguous Face Stimuli. *Journal of Neuroscience*, *39*(6), 1100–1108. https://doi.org/10.1523/JNEUROSCI.1693-18.2018

Brook, L., Kreichman, O., Masarwa, S., & Gilaie-Dotan, S. (2024). Higher-contrast images are better remembered during naturalistic encoding. Scientific reports, 14(1), 13445. https://doi.org/10.1038/s41598-024-63953-5

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178. https://doi.org/10.1016/j.visres.2015.03.005

Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of personality and social psychology*, 79(4), 644–655.

Craik, F.I.M, Bialystok, E. (2006). Cognition through the lifespan: mechanisms of change. 10(3), 131-138.https://doi.org/10.1016/j.tics.2006.01.007

Davis, T. M., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences*, *120*(28), e2302389120. https://doi.org/10.1073/pnas.2302389120

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What Makes an Object Memorable? *2015 IEEE International Conference on Computer Vision (ICCV)*, 1089–1097. https://doi.org/10.1109/ICCV.2015.130

Goetschalckx, L., Adonian, A., Oliva, A., & Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties. 2019 *IEEE International Conference on Computer Vision (ICCV).* https://doi.org/10.48550/arXiv.1906.10112

Hovhannisyan, M., Clarke, A., Geib, B. R., Cicchinelli, R., Monge, Z., Worth, T., Szymanski, A., Cabeza, R., & Davis, S. W. (2021). The visual and semantic features that predict object memory: Concept property norms for 1,000 object images. Memory & cognition, 49(4), 712–731. https://doi.org/10.3758/s13421-020-01130-5

Huebner, G. M., & Gegenfurtner, K. R. (2012). Conceptual and Visual Features Contribute to Visual Memory for Natural Images. *PLOS ONE*, *7*(6), e37575. https://doi.org/10.1371/journal.pone.0037575

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). *Understanding the Intrinsic Memorability of Images*.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). *What makes an image memorable?*

Kensinger, E. A. (2007). Negative Emotion Enhances Memory Accuracy: Behavioral and Neuroimaging Evidence. *Current Directions in Psychological Science*, *16*(4), 213–218. https://doi.org/10.1111/j.1467-8721.2007.00506.x

Kensinger E. A. (2009). Remembering the Details: Effects of Emotion. *Emotion review : journal of the International Society for Research on Emotion*, 1(2), 99–113.

https://doi.org/10.1177/1754073908100432

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and Predicting Image Memorability at a Large Scale. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2390–2398. https://doi.org/10.1109/ICCV.2015.275

Koch, G. E., Akpan, E., & Coutanche, M. N. (2020). Image memorability is predicted by discriminability and similarity in different stages of a convolutional neural network. *Learning & memory (Cold Spring Harbor, N.Y.)*, *27*(12), 503–509. https://doi.org/10.1101/lm.051649.120

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects. *Journal of Experimental Psychology. General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981. https://doi.org/10.1126/sciadv.add2981

Kuppens, P., Tuerlinckx, F., Russell, J., & Barrett, L. (2012). The Relation Between Valence and Arousal in Subjective Experience. *Psychological Bulletin*, *139*. https://doi.org/10.1037/a0030811

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Kyle-Davidson, C., Zhou, E. Y., Walther, D. B., Bors, A. G., & Evans, K. K. (2023). Characterising and dissecting human perception of scene complexity. Cognition, 231, 105319. https://doi.org/10.1016/j.cognition.2022.105319

Li, X., Bainbridge, W.A. & Bakkour, A. (2022) Item memorability has no influence on value-based

decisions. Sci Rep 12, 22056. https://doi.org/10.1038/s41598-022-26333-5

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, *46*(2), 596–610. https://doi.org/10.3758/s13428-013-0379-1

Mather, M., & Nesmith, K. (2008). Arousal-Enhanced Location Memory for Pictures. *Journal of memory and language*, 58(2), 449–464. https://doi.org/10.1016/j.jml.2007.01.004

Mather, M., & Sutherland, M. R. (2009). Disentangling the effects of arousal and valence on memory for intrinsic details. *Emotion Review : Journal of the International Society for Research on Emotion*, *1*(2), 118–119. https://doi.org/10.1177/1754073908100435

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

Saleh B., Farhadi, A., Elgammal, A. (2013) Object-Centric Anomaly Detection by Attribute-Based Reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 787-794.

Schümann, D. & Joue, G. & Jordan, P. & Bayer, J. & Sommer, T. (2019). Test-retest reliability of the emotional enhancement of memory. *Memory*. 28. 1-11. 10.1080/09658211.2019.1679837.

Tambini, A., Rimmele, U., Phelps, E. A., & Davachi, L. (2017). Emotional brain states carry over and enhance future memory formation. *Nature neuroscience*, 20(2), 271–278. https://doi.org/10.1038/nn.4468

Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, *20*(3), 291–302. https://doi.org/10.3758/BF03199666

Wakeland-Hart, C. D., Cao, S. A., deBettencourt, M. T., Bainbridge, W. A., & Rosenberg, M. D. (2022). Predicting visual memory across images and within individuals. *Cognition*, *227*,

105201. https://doi.org/10.1016/j.cognition.2022.105201

Weierich, M. R., Kleshchova, O., Rieder, J. K., & Reilly, D. M. (2019). The Complex Affective Scene Set (COMPASS): Solving the Social Content Problem in Affective Visual Stimulus Sets. *Collabra: Psychology*, *5*(1), 53. https://doi.org/10.1525/collabra.256

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

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009