Direct perception of affective valence from vision

Saeedeh Sadeghi[1*], Zijin Gu[2], Eve De Rosa[1], Amy Kuceyeski[3, 4], Adam Anderson[1*]

1) Department of Psychology, Cornell university, NY, US

2) School of Electrical and Computer Engineering, Cornell University, NY, US

3) Department of Radiology, Weill Cornell Medicine, NY, US

4) Department of Statistics and Data Science, Cornell University, NY, US

Affect has been posited as the basic subjective feeling component underlying all experience, arising directly from internal states of the body and the closely related proximal senses [1,2]. By contrast, in the distal senses such as vision, affect is thought to be indirect, largely mediated by higher level processes[3]. Evidence from machine learning suggests affective features may be embedded in the ecological statistics of the external environment [4,5] but their causal role, if any, remains unclear. Here we provide evidence from machine and human observers that affective valence can be decoded directly from visual features. A visual valence (VV) model of low-level image statistics trained to predict Normative Valence (NV) in 8000 emotionally charged images transferred even more robustly to predicting valence of abstract paintings without conceptual content. Manipulating VV and availability of conceptual analysis, enhanced the contribution of VV to valence experience. In the brain, in contrast with NV, VV resided exclusively in early and mid-level visual areas. Employing a deep generative network [6], brain activity in these regions synthesized new images containing predicted positive versus negative VV. There are distinct competing modes determining valence experience, one indirectly from meaning, and the other derived from ecological visual statistics, affording direct perception of valence as an apparent objective property of the external world.

Main

Subjective feelings are thought to be supported by the representation of internal interoceptive bodily states [1,7–9], which construct emotions when integrated with the higher order "meaning" of external cues and retrieved memories [1]. However, the term affect, as coined by the nineteenth-century psychologist W. Wundt, more broadly reflects the basic feeling ingredient underlying all

sensations [10]. Affect, as originally conceptualized by Wundt, is a fundamental component of all experience, extracted as if a feature from the external world [11]. While bears and sharks may reflexively “look" dangerous or babies are approachable [10], this is thought to be mediated by an evoked aversion or fondness at a deeper level of abstraction [12,13], intermixing with afferent bodily states to affect us [1,7,9]. It remains an open question whether valence, the positive and negative dimension of affect and the primary dimension of meaning [14], is a core component of visual experience, derived directly from the ecological regularities of the environment [15].

Valenced feelings may be correlated with the visual statistics of our environment [4] and with activity patterns in the visual system [16–20]. Here we examined the extent and variation of these correlations and their causal status in the brain and for behavior. Deep neural networks may afford decoding of affective dimensions and emotional categories from visual inputs [18] by deriving high-level internal representations related to specific objects and associated conceptual meanings. We took a different approach by hypothesizing and estimating superficial valence, i.e. a perceptual gist, which we call Visual Valence (VV). VV derivation is hypothesized to originate from basic compositional visual features of valenced entities in the environment, affording ecological visual statistics reflecting valence. The visual system might exploit such ecological visual correlates of valence for efficient representation. If so, one might expect VV to have a causal status with distinct behavioral and neural markers from the conceptual analysis of valence, dominant in normative valence acquisition methods [21]. The expected competing hypothesis is that VV reflects aliased Normative Valence (NV), having no independent causal role, or being causal but originating from conceptual associations outside the visual system. As such, in each study we examined whether VV and NV have competing or shared origins in brain, behavior, or computation. These

investigations would shed light on a potential notion of affect directly embedded in the visual experience [11], expanding core affect to the visual modality [1,2].

1- A Visual Valence Model (VVM)

We compiled a dataset of nearly 8000 photos from nine publically available affective image datasets to examine whether low-level visual features predict valence rating norms. Rather than feeding all image pixels into a deep model, which would lead to high performance in predicting emotions relying on conceptually rich representations, we selected a set of basic global features, demonstrating abstract composition, i.e. perceptual gist, independent of specific objects. We selected 142 total features, including variances from the first convolutional neural network layer of AlexNet [22], average color, proportion of different colors, spatial frequencies, symmetry, structural diversity; as well as a control for presence of human faces to track potential contributions from higher-level features.

Prior to any model training, images were transformed from the large pixel space into this substantially smaller space of global visual features. The Visual Valence Model (VVM, Supplementary **Figure S1**), consisted of a trained random forest model that used these low-level features to predict normative valence (NV), i.e. the average valence rating of each image among individuals on a 1-9 (extremely negative to extremely positive) scale. The out-of-sample predictions were correlated with ground-truth NV ratings, indicating a small to moderate effect size ($r(7982)=0.33$, $p<0.001$). Presence of faces did not appreciably change this relationship ($\Delta r=0.016$, $z=1.19$, $p=0.23$), nor was the relationship sensitive to the choice of regression algorithm (see Supplementary Material).

We expected better than chance, albeit not too large an association between visual features and NV [4,23], since human NV ratings also reflect object recognition and conceptual valence processing of scene content. This was also reflected in how the VVM predictions were much more restricted in range than human observers' NV with the distributions differing significantly in variance (**Figure 1.a**; $F(7983/7983)=14.2$, $p<0.001$). VVM predictions support evidence for the visual valence (VV) hypothesis, that visual features correlate to some degree with conceptual valence content and valence experience.

2- Transfer of VV to Abstract Paintings

VVM predictions may be merely a correlate of conceptual content access. By contrast, if visual features have a distinct functional role in the experience of valence, then removing conceptual content from images should not influence the relationship between VVM predictions and NV. We next performed a transfer test, examining whether the VVM predicted valence ratings of art, using abstract paintings as the test data. Even though VVM was trained on real photos of emotional scenes, its predictions not only remained correlated with human NV ratings of abstract paintings (N=500, each rated by 20 individuals [24]), but also increased in predictive association ($r(498)=0.47$, $p<0.001$; **Figure 1.c**).

Comparing the distributions of valence ratings to VVM predictions also revealed a more restricted range similar to abstract art (**Figure 1.b**), which differed from realistic photos with conceptual content (**Figure 1.a;** $F(7983/499)=3.8$, $p<0.001$). VV now explained 20.1% of variance of NV in abstract paintings, nearly twice the amount of photos with conceptual content. These results are consistent with extremes in NV reflecting access to conceptual content [21]. Without this access, machine and human observers demonstrated greater similarity in range and correspondence

of valence experience. This provides evidence that VV has some causal role in valence experience independent of conceptual analysis. While abstract art may not contain explicit conceptual content, VV may be regulated by the processing of weak conceptual associations.

3- Manipulating conceptual access

As a stronger test of the causal role of VV and its direct perception, Study 3 experimentally manipulated the presence of VV and availability of conceptual processing for stimulus processing. Direct perception was operationally defined as independence from access to higher level-conceptual representations for VV expression. We selected realistic photos (N=280) with weak NV, i.e., varying marginally from neutral (5 ± 0.63; range width=0.6 sd) on 1-9 scale) but with strong bi-modal VV, splitting the images into positive (VV+) and negative (VV-) set based on VVM predictions (VV- range=4 to 4.8 (width=1.5 sd); VV+ range=5.5 to 6.6 (width=1.9 sd)) (**Figure 2a**). By design, VV had little relation to NV in this subset of images ($r(278)=0.04$, $p>0.1$) (**Figure 2b)**.

To manipulate access to promote perceptual gist over conceptual processing, one group (Full-viewing, N = 80 ratings per image) had unlimited exposure time and response window; while in another group (Limited-viewing condition, N=80 ratings per image), images were inverted and and viewing time restricted to 100 ms and requiring speeded judgments. Participants in both conditions were asked to report subjective valence with a 2-alternative forced choice response (positive or negative). Under these conditions, observers reported *de novo* valence experience to images that lacked strongly positive or negative NV (**Figure 2c, 2d**). Under full-viewing conditions prior NV norms associated with the presence of weak conceptual valence content explained valence responses better than the VVM's predictions of VV ($z=5.7$, $p<0.001$; mixed-

model fit comparison, ΔAIC=92.7). By contrast, in the limited-viewing condition, VV better explained responses than NV (ΔAIC=286.5; **Figure 2e**).

A significant interaction between VV and viewing condition revealed that VV had a larger impact on valence responses during limited-viewing compared to full-viewing($z=-5.5$, $p<0.001$); whereas an interaction in the opposite direction between NV and viewing condition, revealed that NV having a greater impact during full-viewing ($z=5.7$, $p<0.001$). Overall, the image by image correspondence between machine and human observers' decoding of VV increased under limited and decreased under full viewing (**Figure 2e**).

While limited viewing conditions would reduce access to high level object or mid-level categories, we further accounted for the possible role of salient complex categories (i.e., presence of humans, animals, artifact vs natural, and indoor vs outdoor). The greater contributions of VV in limited viewing ($z=-5.4$, $p<0.001$) and effect of NV in full-viewing ($z=5.8$, $p<0.001$) held after statistically controlling for these salient content related effects. This is consistent with VV as residing in complex lower-order features rather than higher level objects or scene inferences.

The salience of VV was not only limited to brief stimulus presentations but it also predominated as observers made faster assessments. As expected, there were faster responses in the Limited-viewing compared to the Full-viewing condition (beta=0.32, $t=6.50$, $p<0.001$). In general, faster valence judgments led to reports more driven by VV ($z=-3.7$ $p<0.001$; full-viewing: $z=-2.3$, $p<0.05$, in limited-viewing: $z=-2.65$, $p<0.01$; **Figure 2f).** These effects were again persistent when controlling for higher level content ($z=-3.9$, $p<0.001$; full-viewing: $z=-2.1$, $p<0.05$; limited-viewing: $z=-2.7$, $p<0.01$).

The double dissociation between valence type (visual versus conceptual) and processing mode (brief vs extended exposures and fast vs slow processing) demonstrates that although

correlated in normal experience (Study 1 and 2), VV and NV are distinct competing valence signals, with NV dependent on deeper conceptual content and VV directly perceived from visual features, processed distinctly from mid-level scene types and conceptual categories.

4- Neural bases of Visual Valence

If VVM's predictions reflect a valid representation of valence embedded in visual feature space, then visual valence should reside in the visual system and differ from regions traditionally associated with amodal valence responses [25,26]. During fMRI BOLD scanning [25], participants (N=20) viewed 128 images varying between extreme positive and negative emotional valence, rating valence experience under conditions of long exposure, favoring conceptual valence access. A whole-brain GLM univariate analysis of BOLD data identified brain regions whose activation, either positively or negatively, correlated with participant valence ratings for each image on a bipolar scale of most negative to most positive, i.e. the average positivity minus negativity rating would reflect NV. **Figure 3a** shows the significant clusters (false positive rate; FPR=0.05) associated with NV from the human observers and VV obtained from the VVM (**Supplementary Tables S1 and S2** for full list of anatomical regions).

Although the stimulus selection was independent of the VVM, the model's predictions transferred to this smaller image data set, correlating with average subjective valence ratings (NV) across the images ($r(126)=0.46$, $p<0.001$). Despite this correlation, VV overlapped with only 4% of voxels associated with NV (**Supplementary Table S3**). Specifically, NV (red in **Figure 3a**) correlated with higher levels of the cortical hierarchy centered on the default mode network, including regions traditionally associated with the experience of valence across modalities. In addition, there were activations in regions associated with high-level visual scene content

processing [27], including the lingual, fusiform and parahippocampal cortices. By contrast, VV (**Figure 3a** yellow; **Figure 3b** red and dark blue) correlated with lower-level modality-specific visual regions, including the occipital pole, intracalcarine cortex, and mid-level regions including the lateral occipital cortices.

To establish the reliability of these findings, we employed the open access 7T fMRI Natural Scenes Dataset [28] of 8 participants viewing 8-10k images across 30-40 sessions (a total of 73k image stimuli). Images were selected from various categories but were largely neutral in NV. A group-level analysis in the entire dataset employing VV as predictors confirmed that VV resided primarily in the visual system (**Figure 3b; Supplementary table S.4).** The larger number of stimuli allowed for the examination of regions that supported negative (VV-) and positive (VV+) valence with greater reliability, not confounded by conceptual biases that potentially exist within a limited number of stimuli. Across data sets, we found that VV- was differentially associated with more posterior intracalcarine cortices and VV+ with more anterior fusiform and lateral occipitotemporal regions (**Supplementary table S.4)**.

Beyond providing evidence for the distinct biological bases of VV and NV, it is evident that VV originates from the visual system. While the visual system is capable of representing high-level content [29], the current evidence suggests that the neural basis of VV relies more on decoding the external appearance of visual stimuli rather than internal neural representations of their high-level categories.

5- Synthesizing images from visual system

While VV was largely correlated with activity in modality-specific visual regions, it is unclear if it is the generator of VV. In contrast to Study 4 that measured a brain response to stimuli,

in Study 5 we reversed this approach, generating synthetic images to afford a visualization of images that maximize or minimize activation in visual regions most correlated with VV [6]. This also allowed for visualization of whether VV in the visual system did associate with any mid or high-level conceptual categories more than others, even though it did not result from a direct conceptual analysis. To do so, we employed NeuroGen [6], a brain-to-stimulus generative model, trained on the Natural Scenes Dataset. A small subset of images (N=100) were rated on valence. When examining the transfer of VVM predictions to these images, we found a small but reliable correlation, ($r(98)=0.22$, $p<0.05$).

Within the visual regions identified from the NSD, we considered voxels that maximally responded to positive (VV+) and minimally responded to negative (VV-) features and vice versa (**Figure 3b** yellow and light blue). Synthesized images took place at the image class level (between classes, e.g., bear, kite), and specific visual features (within classes, e.g. a bear with positive and negative VV). To determine whether these regions are generators of VV, the resulting synthesized images **(Figure 4a**) were then fed into the VVM. As a machine observer, the VVM discriminated image valence at both the image class levels ($t=3.0$, $df=198$, $p<0.01$; between sample, between positive and negative classes; **Figure 4c**) and visual feature optimized within classes of the same content (paired t-test, $t=2.7$, $df=149$, $p<0.01$, within-sample; **Figure 4b**). To further examine the potential categorical correlates of VV we performed an analysis of image categories based on WordNet tree hierarchy across the entire 1000-items list of classes from which images were generated (horizontal bars in **Figure 5**). The most informative high-level category was “Artifact” (52.2%) and “Living thing” (40.7%) with VV- and VV+ classes following the same division ($\chi^2 =88.5$, $df=1$, $p<0.001$)(supplementary **table S.5** for the full list of positive and negative classes).

Together, these results confirm that neural activity in visual regions are not only correlated with VV, but are sufficient for synthesizing VV back as visual inputs, with distinct regions in the visual system differentially supporting positive and negative features.

Discussion

We provide support for a distinct form of valence that arises from visual perception. This visual valence is related to but functionally distinct from high level inferences about the content of what is perceived [3,18,30–32]. This view is compatible with an emerging decentralized view of affect, with sensory modality-specific valence as a fundamental component of perceiving [2,10]. The tradition of ecological vision [15] argues visual perception develops as a result of navigating the external environment and discovering its inherent structure, more than forming and relying on internal higher-level inferences. Just as texture gradients provide visual cues, or affordances, for perception of near or far distance, low or mid-level visual features can provide cues to positive or negative feelings. Events "look" valenced, particularly when observers have limited access to deep contents on which traditional valence depends.

In recent theorizations, salience is the main recognized scheme of the entanglement between the visual system and affective processing [1,33], supporting on online reentrant feedback [34] and rapid scene analysis [35,36]. Beyond salience, top down input into visual processing regions may support representations of valence [16–20]. Across two fMRI datasets, and thousands of images, we demonstrate a computational metric of valence correlated with activity within modality-specific visual regions, largely devoid of higher order inputs strongly engaged by normative valence. Visual valence may then represent affective priors embedded in the visual system that depend upon intrinsic activity and reentrant feedback within visual regions.

Interoceptive modalities, including those that derive their stimulation from the proximal external senses, such as pain or caress [37], sweet and bitter taste (Wang et al., 2018) and maybe even fruity or rancid smells [38], have sensory receptors that are evolutionarily tuned to information broadly consistent with valence [2,39]. By contrast in the exteroceptive senses, such as vision, there are no receptors for valence. They require discovering predictors of valence from the process of perception, if such associations exist in the environment [12,23,40]. This ecological perspective aligns with the Bayesian understanding of brain function [41] where observers learn the transitional probabilities from seeing to feeling.

Visual valence only accounted for 5-20% of variance in valence. VV fails to predict valence of deep content at the individual item level much more often than not. Rather than serving as a valence predictor for specific events, VV would arise from an aggregate of experiences. As such VV operates as a distinct mode, competing with traditional valence, in which observers decode valence gist [42]. The basis for such valence gist learning may originate from affective-motivational circuitry. Reward or punishment adaptively tunes the visual system, toward representation of predictive physical properties [43]. This modified response is maintained even after the stimulus is no longer motivationally relevant [44,45], supporting long-term rewiring of the visual system. Until now it has been thought that such learning alters the salience of associated physical features, rather than representing valence itself.

The function of visual valence aligns with an implicit versus explicit distinction in memory [46] and dual system theory [47], suggesting a fast, emotional heuristic system I, in contrast to a slow, deliberate system II, with the former supporting implicit affective judgements [48] and the latter associated with explicit emotional thoughts and memories [1]. Implicit preferences need not require explicit conceptual inferences [48,49]. Visual valence embedded in the global perceptual features can

acquire an affective charge of its own. Playing a role even when conceptual content may indicate to feel differently, e.g., as in a positively valenced insect or spider (see **Figure 4**). VV would guide feelings and behavior in a more implicit way, similar to how facial features influence judgements of attraction or trustworthiness without access to knowledge about traits [50,51]. Unlike faces, the the present studies unbind visual features from specific localized objects, to highly abstract universal features of the environment. Valence features may be represented in each perceptual system [2,39] and may even have multiple representations within one perceptual system, such as for faces or body shapes.

Affect can arise from different levels of abstraction, ranging from conceptual processes rooted in identifying concrete objects and their relationships within scenes to experiences more closely tied to abstract lower-level visual patterns. This hierarchical perspective aligns with theories on aesthetic appreciation of artworks [52], or visual appeal of nature [53]. We show that VV is not isomorphic with this distinction, but yet resides in regions that distinguish the natural from artificial and as such relates to biophilia, i.e. the preference for nature [54]. That valence is derived from visual features across highly distinct domains, from explicitly emotionally evocative events, and abstract art, to scenes from the natural and the built world, points to a fundamental convergence to which our brains are attuned. Independently of visual features, valenced feeling states regulate visual system function. Positive states result in more global and negative states more focal visual processing [55] which may align how we experience the natural and built world [35]. More than the contents of visual perception, visual valence may align with a way of seeing.

Methods

1-Developing VVM

Realistic image database:

A database of real-world naturalistic photographs was compiled by combining the following datasets: EmoMadrid [56], Emotional Picture System (EmoPics [57]), Geneva Affective Picture Database (GAPED [58]), International Affective Picture System (IAPS [59]), Nencki Affective Picture System (NAPS [60]), Erotic subset for (NAPS ERO [61]), Open Affective Standardized Image Set (OASIS [62]), The Set of Fear Inducing Pictures (SFIP [63]), Socio-Moral Image Database (SMID [64]). Duplicate images were detected and excluded. Grayscale images were also excluded. The result was a set of 7984 images. Average valence and arousal ratings of images from each dataset were rescaled from their original range to the 1-to-9 scale, to be comparable across different datasets (1 indicating most negative and 9 indicating most positive).

Visual feature extraction:

The following visual features were extracted for each image as the model input.

Average color: colored images in the RGB space were transformed into HSV color map, and the average hue, average saturation, and average brightness across all pixels were estimated (3 features).

Color names: the proportion of pixels that are labeled as each basic color name were estimated based on a previous study [65], that provides a mapping from each point in the RGB space into one of 11 basic colors as perceived and labeled by humans (black, blue, brown, green, gray,

orange, pink, purple, red, white, yellow). The result was 11 features indicating the amount of each color in an image.

Frequency: Spatial frequency amplitudes are employed by the visual system for early rapid visual processing to obtain an initial sketch of scenes [5,66,67]. These amplitudes reflect the texture of the image in different frequencies corresponding to different orientations. We followed the method used in [5] for extracting frequency amplitudes based on the discrete fourier transform. Briefly, images were first converted to gray scale. Then the 2-D space was divided into 4 orientations by 4 lines having 0, pi/4, pi/2, or 3pi/4 angles from the horizon. Each orientation line was then divided into 20 equally spaced points on a logarithmic scale. The amplitude in each point was interpolated, resulting in 20*4=80 total features.

Symmetry: Symmetry was estimated using the method proposed in [22], based on filters from a Convolutional Neural Networks (CNN) that take colors, edges, textures, shapes and objects into account, approximating the function of the early visual processing in humans. We included the left-right, up-down, and left-right-up-down symmetry (3 features)

Diversity: We used the method in (Brachmann et al., 2017) to calculate variances of the first convolutional layer responses of a well-established CNN trained to recognize objects. The first layer of this network consists of 96 filters each detecting luminance edges at certain frequencies and orientations or different color gradients, resembling the brain's early visual processing. Each image was divided into n*n equally sized subregions with n taking 15 possible values of 2, 4, 6, …, 30. Three types of variances were estimated for each n resulting in 45 total features. The three variance types were: 1) richness: total variance over all filters and subregions (many features respond at many regions), 2) median variability across different regions: variance of each filter across all subregions and then the median of the result across filters. 3) variability:

variances of all filter responses across each subregion and then median of the result across all subregions. Together, these features index how uniform or varied are the subregions of the image and how varied individual subregions are in terms of low-level visual features, independent of the location of subregions.

Faces: Three features were extracted related to faces in an image: number of faces, size of the largest face relative to the image size, total size of all faces relative to image size. We used the MTCNN toolbox in Matlab to automatically detect faces and compute the value of these three features for each image.

Random Forest Regression:

We used the random-forest algorithm with 10-fold cross validation to test predictability of valence or arousal from each of the feature sets in the real-world image database. A random forest is an ensemble of decision trees each trained on a subset of features. The random forest's prediction for a set of input features is then the average prediction of all trees. While individual trees tend to overfit, this averaging prevents overfitting to the training data. Random forest was implemented using Matlab's regression tree ensembles with 500 ensemble trees. We defined prediction accuracy of a random forest or linear regression model as the Pearson correlation coefficient between the out-of-sample predictions and ground-truth values.

2- Abstract paintings

Abstract artwork consisted of a database of 500 paintings from the Museum of Modern and Contemporary Art of Trento and Rovereto's (MART) collection [24]. Each painting in the MART database had been rated by 20 individuals in terms of emotional valence on a 1 (highly negative)

to 7 (highly positive) scale. Raters did not see any information about the creator of the artwork or its title. We estimated the average valence rating of each painting as the groundtruth. Valence ratings were rescaled to the 1-9 range to match the real photo database rating scale.

3- Full- and Limited-viewing

Stimuli:

We selected the most neutral images from the realistic image dataset which was used to train VVM. The first criterion to select images was to have an arousal rating smaller than 3.5 (on a 1-9 scale). Among these low-arousing images, we selected N=600 images with the most neutral valence rating (valence in a symmetrical range centered at 5). The valence range around the rating of 5 that included 600 images was between 4.37 and 5.63. Out of these 600 images, we then selected the N=140 with the highest VV, based on VVM's predicted valence (VV+ set) and N=140 with the lowest VV (VV- set). The two image sets were further controlled for not having significantly different subjective valence and arousal based on the average human ratings in the original dataset (positive set valence: mean=5.16, sd=0.35; negative set valence: mean=5.11, sd=0.34; positive set arousal: mean=2.42, sd=0.68, negative set arousal: mean=2.32, sd=0.65). The positive valence set had an average visual valence of 5.78, significantly higher than the negative image set which had an average visual valence of 4.56 ($p<0.001$). This procedure resulted in two sets of images that were indistinguishable in terms of human affective ratings, with one set having positive VV and the other having negative VV, merely based on VVM's predictions (**Figure 2a**). The entire stimulus selection procedure was conducted with a Matlab script, with no active involvement of researchers, that could potentially bias the selection.

Participants:

A total of 180 individuals participated in the study. Participants included paid Amazon Mechanical Turk workers (N= 25), and Cornell University undergraduate students participating for course credit (N=155). Amazon Mechanical Turk workers were paid 1$ for rating each 56 images. Inclusion criteria for Mechanical Turk workers were having done at least 100 tasks on the Mechanical Turk website with at least a 95% approval rate.

Task and procedure:

There were two valence rating conditions: Full-viewing and Limited-viewing. Each participant was randomly assigned to either condition (between-subject design). In each trial of both conditions, an image was displayed and the participant had to determine whether the image induced more positive versus more negative feeling (2-alternative valence repsonse). In the Full-viewing condition, in each trial, the image was displayed until a response key was pressed with no time limit. In the Limited-viewing condition, the image was displayed upside-down and only for 100 ms, followed by a 3500 ms response time limit. Participants were instructed to enter their responses with 'A' and 'L' keys on the keyboard corresponding to positive and negative valence. All images were resized to have a width of 400 pixels, while maintaining the original aspect ratio. The 280 image stimuli were randomly divided into 5 blocks of 56 images (28 from the negative and 28 from the positive set). Each participant performed between 1 to 5 blocks. Each stimulus was in total rated by 80 individuals in the Full-viewing and 80 individuals in the Limited-viewing condition. All conditions and instructions were implemented in Qualtrics online platform (with embedded HTML/JavaScript). Participants performed the task in a remote online mode on a computer browser.

Analysis: Mixed-effect regression was used to analyze the results to account for within and between individual variability in responses or response times. In order to take mid and high level conceptual categories into account, the set of 280 image stimuli were rated in terms of presence of humans (binary), animacy (binary), naturalness (1=very artificial, 7=very natural), and representing outdoors (1=completely indoors, 7=completely outdoors). Two raters independently performed the ratings. Internal consistency of the two ratings was confirmed (Chronbach's alpha> 0.9 for all categories), and the average was used in the analysis.

Among the entire data, in the limited viewing condition 1.6% of trials were timed out and dropped from further analysis.

4- fMRI experiment

Data Collection and Participants:

Previously collected data from a published study from our lab consisting of 16 participants was included (the visual experiment of [25]. Additional data was collected for 4 individuals following the same procedure, resulting in a total of 20 participants included in the analysis.

Task and procedure:

We analyzed data from the visual experiment of a previous study [25]. Stimuli were a set of 128 colored images from the IAPS database, consisting of an equal number from each category of fear, disgust, neutral, or pleasant. In each trial of the task, one image was displayed for 3 seconds, followed by a blank screen for 4 seconds. Participants then rated the positivity and negativity of the image each for 3 seconds (on a scale of 1-7). There was an inter-trial interval of 4 seconds

displaying a blank screen. Data was collected using a 3.0 Tesla fMRI machine with a repetition time (TR) of 2000 ms. Voxel size for the T1-weighted anatomical images was 1mm and for the functional images was 3.5 mm in each dimension (see [25] for more details about data collection).

Pre-processing:

Regular preprocessing steps were conducted in SPM followed by normalization of each individual's functional images into the MNI space (interpolating data to 2mm voxel dimensions)[25]. An activation map was estimated on the un-smoothed data in response to the stimulus in each trial (using the canonical function in SPM). This resulted in one whole-brain t-map for each trial for each individual (20*128 total beta-maps). t-maps were normalized by subtracting the mean.

Univariate analysis:

Whole-brain univariate analysis was conducted to compare neural correlates of subjective valence and VV. Subjective valence for each image was defined as the participant's rating for positivity minus negativity. Note that VV for each stimulus was the same across all subjects (VVM's prediction based on low-level features), while subjective valence was distinct for each individual based on the participant's ratings. The Pearson correlation between each voxel's beta activation for each individual and visual valence across trials was estimated. The correlation coefficient was then converted to a z-score using Fisher transformation, resulting in one z-map for each subject. The z-maps were then tested across individuals using a two-tail t-test to determine whether each voxel's value was significantly different from zero. Cluster thresholding was done using the 'ETAC' option (Equitable Thresholding And Clustering) in AFNI's 3dttest++ function, with a smoothing kernel size of FWHM=6mm. ETAC option for multiple comparison correction

uses a voxelwise false positive rate (FPR) of 5% at several different commonly used significance thresholds to find robustly significant clusters, not relying on arbitrary user parameters [68]. Both positive and negative clusters (voxels whose activity increased with either positivity or negativity) were considered as being sensitive to valence, while clusters were either positive or negative. A similar procedure was conducted to obtain whole-brain correlates of subjective valence. Structures associated with significant clusters were named based on probability maps from the Harvard-Oxford Cortical Atlas and the Harvard-Oxford subcortical atlas, using 'autoaq' script in FSL.

NSD univariate analysis: We also performed a second group-level univariate analysis across the 8 subjects NSD [28]; see next section). We first estimated the z-score associated with the correlation between VV estimated for each stimulus and the beta activation in response to it. We then performed a one-sample t-test across the 8 subjects' z-maps. The voxel-level significance threshold was set at $p<0.001$ and the minimum cluster size was estimated using AFNI's 3dclustsim routine to correct for multiple comparisons at $p<0.05$.

5- Synthesizing images

We used a previously developed generative framework NeuroGen [6] to synthesize images. NeuroGen works by concatenating an image generator with neural encoding models by feeding the generator output image to the encoding model and obtaining the predicted activation. During the optimization, the gradient flows from the region's response back to the synthetic image and then to the noise vector. Thus images that can achieve targeted brain responses (predicted) can be then iteratively synthesized via optimization of the noise vector, which is the input to the generator. NeuroGen was previously shown effective in modulating macro-scale brain responses [69]. Here we

follow previous work and use the generator from BigGAN-deep [70], a conditional generative adversarial network (GAN) that synthesizes images conditioned on ImageNet classes. There is a truncation parameter during noise vector sampling that controls the balance between image fidelity and variety. Our encoding models were trained with eight individuals' brain responses in the NSD. Different from previous work that uses NeuroGen at a regional level [71], here the plugged encoding model into NeuroGen was a whole-brain voxel-wise model. The model had VGG19 [72] as a backbone and a spatial and feature factorized linear readout [73].

Our objective function was:

$$L = \sum_{i=1}^{8}(predicted\_activation * visual\_mask * valence\_mask).$$

The visual mask was a combination of 24 early and higher-order visual regions in NSD, identified using floc and retinotopic mapping, and the valence mask was obtained by estimating correlation coefficient between VV and trial-by-trial beta map, followed by an uncorrected voxel-wise threshold of $p<0.001$ for each individual's map and a minimum cluster size of 10 voxels.

During optimal image generation, we first identified the top 50 image classes for 3 different conditions: 1) maximize the objective (positive classes), 2) minimize the objective (negative classes), 3) achieve maximal variance of the objective (middle classes). The class selection was done by ordering the indices of the 1000 ImageNet classes based on the average or variance of predicted activation of class-representative synthetic images generated from 100 random initializations, with the truncation parameter set to 1. After that, we performed two optimizations for each identified class (total 50x3=150 classes), one maximizing the objective (VV+) and the other minimizing the objective (VV-). For each optimization, the class information was embedded as a one-hot code into the class vector and fixed, and the noise vector was initialized from a truncated normal distribution and optimized. The truncation parameter here was set to 0.4 to have

a good balance of image fidelity and variety. Overall, this procedure resulted in 3(class conditions) x 50(classes within each condition) x 2 (VV+ or VV-) = 300 images. The generated images were then fed into VVM to estimate VV for comparing between and within class conditions. We also used the *ntlk* package in Python interfacing with WordNet lexical database in order to analyze the hierarchical tree of categories associated with the top positive and negative class labels.

Data Availability

The following repository contains the VVM trained random forest model, a sample script for extracting visual features and predicting VV, and the features extracted for the data used in this project: https://github.com/saeedeh/Visual-Valence-Model

The stimuli used in study 3, an online demo of the full and limited-viewing condition tasks, and the behavioral results is available at: https://github.com/saeedeh/sensoryValence-onlineStudy

Figures

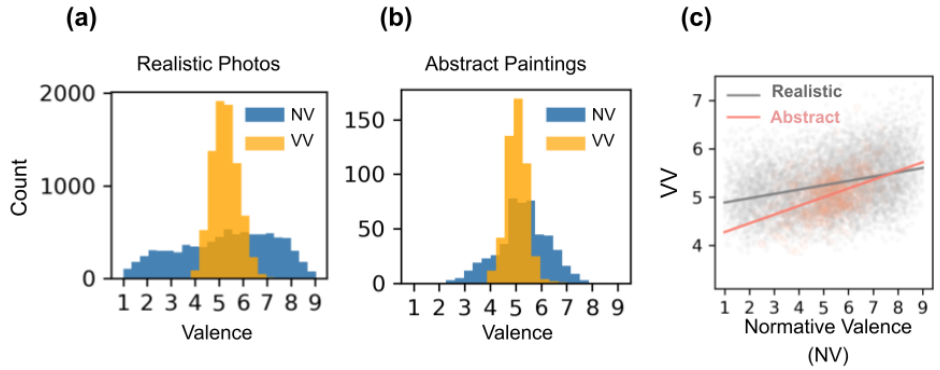

Figure 1. a,b) *Frequency distribution of normative (NV) and visual (VV) valence across realistic photos in study 1 (**a**), and abstract paintings in study 2 (**b**). **c**) Relationship between VV (Y-axis) and NV (x-axis) for realistic photos and abstract paintings. Each scattered point represents one image. The best-fitting lines are presented.*

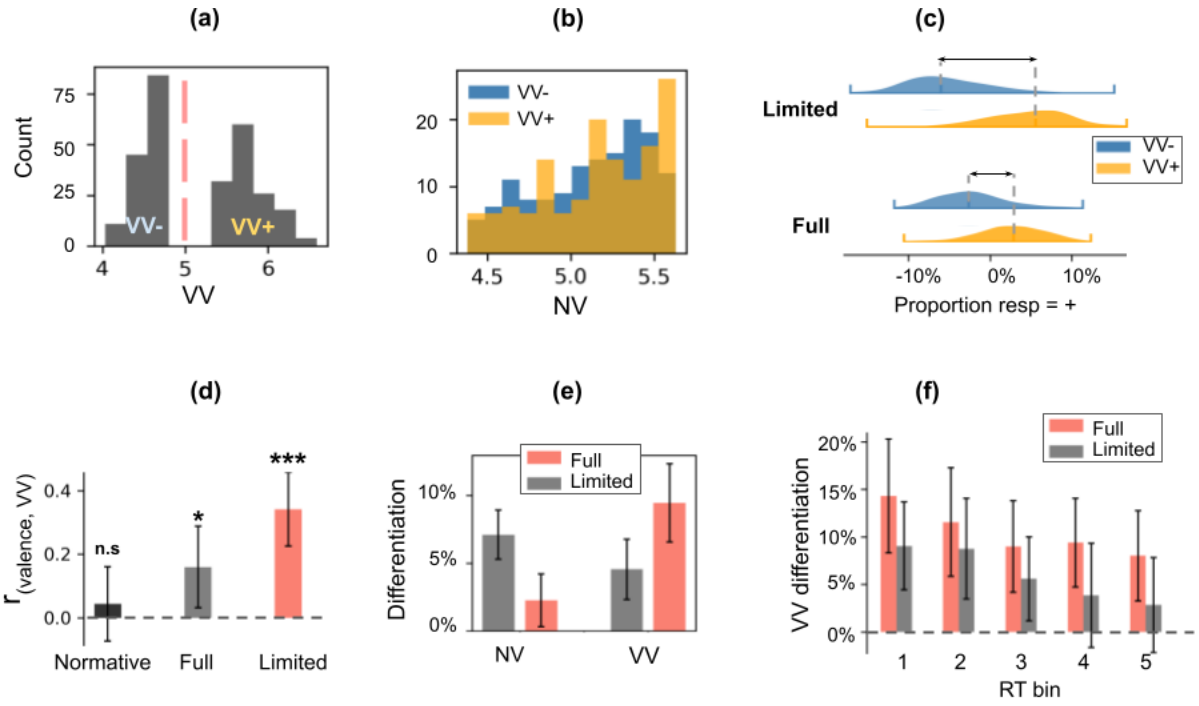

Figure 2. Stimulus characteristics and results of Study 3. a) *Visual valence (VV) distribution for stimulus set.* ***b)*** *Normative Valence (NV) distribution for the VV+ and VV- splits of stimuli are entirely overlapping.* ***c)*** *The relative proportion of times an image was rated as positive for each image set (VV- blue, VV+ orange). Percentages are subtracted from each participant's baseline (the average tendency of pressing "positive"), and distribution across participants is presented (dashed lines represent median).* ***d)*** *Spearman correlation coefficient between VVM's predicted valence and subjective valence ratings obtained from the original naturalistic image dataset, 2-alternative responses in full-viewing condition, and limited viewing condition (n.s.: not significant; *: p<0.05; ***; p<0.001).* ***e)*** *Average differentiation of images by NV or VV for full versus limited viewing conditions.* ***f)*** *Average differentiation by VV associated with the Response Time (RT) quantile-based bins. VV+ = visual valence positive +; VV- = visual valence negative. Error bars represent 95% confidence intervals after Bonferroni multiple comparison adjustment.*

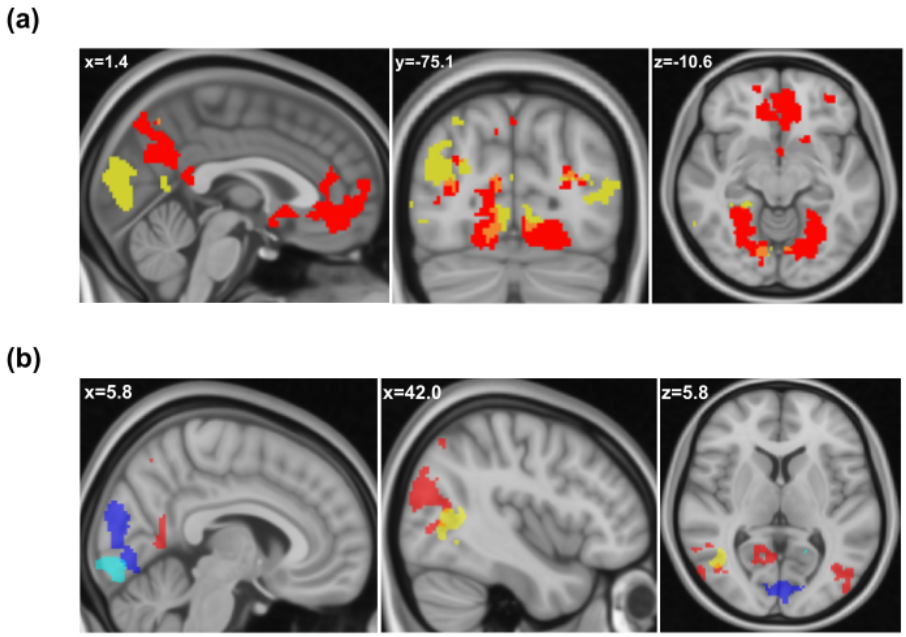

Figure 3. *Neural correlates of normative and visual valence (study 4).* ***a)*** *A computational metric of image visual valence (VV) corresponded with largely modality-specific posterior visual regions (yellow). Normative valence (NV) derived from observer ratings of valence experience revealed recruitment of higher-order multimodal regions traditionally associated with valence experience (red). Regions were defined by valence selectivity, increasing to either positive or negative valence; whole-brain, corrected at false positive rate of 0.05.* ***b)*** *color coded significant positive and negative visual valence clusters associated with Chikazoe data [25] (128 stimuli, 20 participants) and Natural Scene Dataset (NSD); whole-brain, corrected at false positive rate of 0.05; Chikazoe+: red; NSD+: yellow; Chikazoe-: dark blue; NSD-: light blue.*

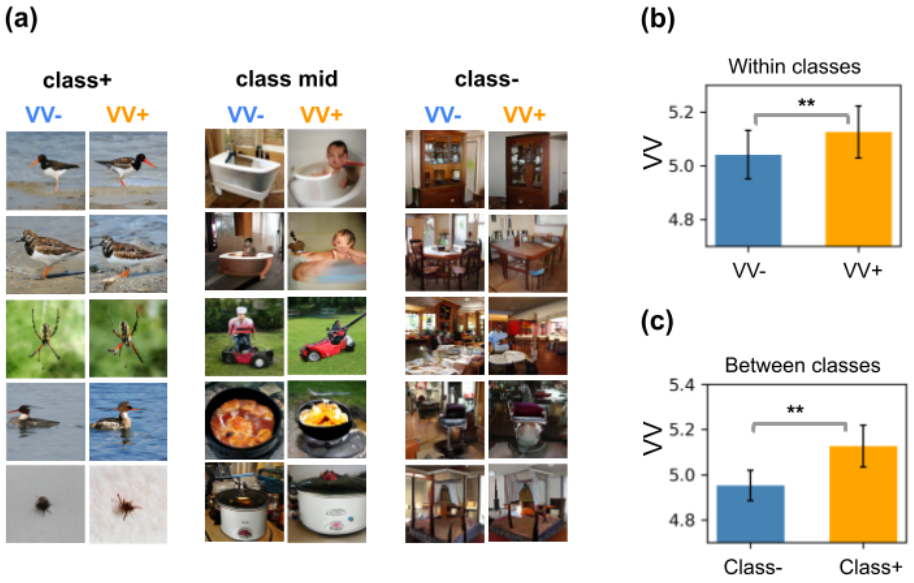

***Figure 4**. **a)** Sample generated images for the five most positive (class+), most negative (class-) and middle classes. Within each class, once an image is generated to maximize the sum of individuals' weighted mask (VV+) and once to minimize it (VV-). **b)** Within-class comparison of the VVM's estimate of VV for the generated VV- and VV+ images (each N=150). **c)** between-class comparison of VV for the generated images of positive classes (class+; N=100) with negative classes (class-; N=100). Bars indicate mean and error bars indicate 95% confidence interval. Confidence intervals in c control for the between-class variability. * p<0.05 **p<0.01 ***p<0.001*

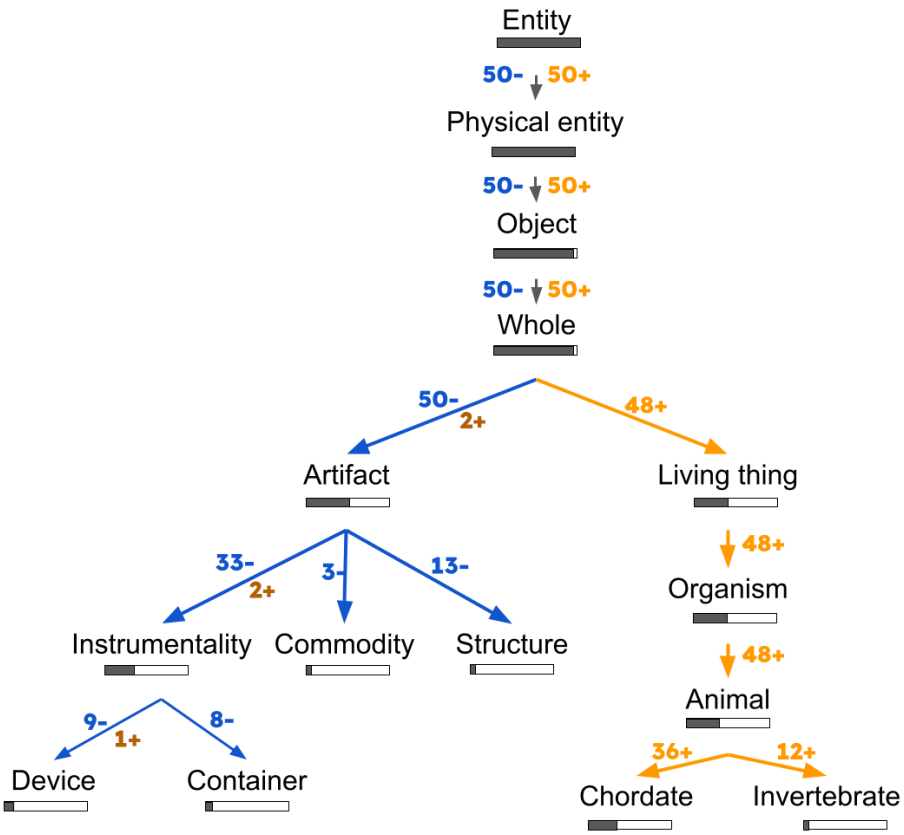

Figure 5. *Hierarchical tree structure of class labels based on the WordNet database. Horizontal gray bars represent the proportion of classes across the ImageNet's 1000 class list. Only nodes containing more than 5% of the list (>20 classes) are included in the graph. The number of classes out of the top 50 most positive classes (blue) and most negative classes (orange) are shown on arrows.*

References

1. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2017).
2. Kryklywy, J. H., Ehlers, M. R., Anderson, A. K. & Todd, R. M. From architecture to evolution: Multisensory evidence of decentralized emotion. *Trends Cogn. Sci.* **24**, 916–929 (2020).
3. Vuilleumier, P. & Driver, J. Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. *Philos. Trans. R. Soc. B Biol. Sci.* **362**, 837–855 (2007).
4. Redies, C., Grebenkina, M., Mohseni, M., Kaduhm, A. & Dobel, C. Global image properties predict ratings of affective pictures. *Front. Psychol.* **11**, 953 (2020).

5. Rhodes, L. J. *et al.* The role of low-level image features in the affective categorization of rapidly presented scenes. *PLOS ONE* **14**, e0215975 (2019).
6. Gu, Z. *et al.* NeuroGen: activation optimized image synthesis for discovery neuroscience. *NeuroImage* **247**, 118812 (2022).
7. Critchley, H. D. & Garfinkel, S. N. Interoception and emotion. *Curr. Opin. Psychol.* **17**, 7–14 (2017).
8. Damasio, A. & Carvalho, G. B. The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **14**, 143–152 (2013).
9. Seth, A. K. Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* **17**, 565–573 (2013).
10. Wundt, W. M. & Judd, C. H. *Outlines of psychology*. (W. Engelmann, 1902).
11. Anderson, A. K. Toward an objective neural measurement of subjective feeling states. (2015).
12. Damiano, C., Walther, D. B. & Cunningham, W. A. Contour features predict valence and threat judgements in scenes. *Sci. Rep.* **11**, 19405 (2021).
13. Goetschalckx, L., Andonian, A., Oliva, A. & Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. in *Proceedings of the ieee/cvf international conference on computer vision* 5744–5753 (2019).
14. Osgood, C. E., May, W. H. & Miron, M. S. *Cross-cultural universals of affective meaning*. vol. 1 (University of Illinois Press, 1975).
15. Gibson, J. J. The Ecological Approach to Visual Perception: Classic Edition. (1979).
16. Čeko, M., Kragel, P. A., Woo, C.-W., López-Solà, M. & Wager, T. D. Common and stimulus-type-specific brain representations of negative affect. *Nat. Neurosci.* **25**, 760–770 (2022).
17. Gao, C. & Shinkareva, S. V. Modality-general and modality-specific audiovisual valence processing. *Cortex* **138**, 127–137 (2021).
18. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).
19. Miskovic, V. & Anderson, A. Modality general and modality specific coding of hedonic valence.

Curr. Opin. Behav. Sci. **19**, 91–97 (2018).

20. Shinkareva, S. V. *et al.* Representations of modality-specific affective processing for visual and auditory stimuli derived from functional magnetic resonance imaging data. *Hum. Brain Mapp.* **35**, 3558–3568 (2014).
21. Itkes, O., Kimchi, R., Haj-Ali, H., Shapiro, A. & Kron, A. Dissociating affective and semantic valence. *J. Exp. Psychol. Gen.* **146**, 924 (2017).
22. Brachmann, A. & Redies, C. Using convolutional neural network filters to measure left-right mirror symmetry in images. *Symmetry* **8**, 144 (2016).
23. Lakens, D., Fockenberg, D. A., Lemmens, K. P. H., Ham, J. & Midden, C. J. H. Brightness differences influence the evaluation of affective pictures. *Cogn. Emot.* **27**, 1225–1246 (2013).
24. Yanulevskaya, V. *et al.* In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. in *Proceedings of the 20th ACM international conference on Multimedia* 349–358 (2012).
25. Chikazoe, J., Lee, D. H., Kriegeskorte, N. & Anderson, A. K. Population coding of affect across stimuli, modalities and individuals. *Nat. Neurosci.* **17**, 1114 (2014).
26. Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. The brain basis of emotion: a meta-analytic review. *Behav. Brain Sci.* **35**, 121–143 (2012).
27. Price, C. J., Devlin, J. T., Moore, C. J., Morton, C. & Laird, A. R. Meta-analyses of object naming: Effect of baseline. *Hum. Brain Mapp.* **25**, 70–82 (2005).
28. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
29. Logothetis, N. K. & Sheinberg, D. L. Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
30. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).
31. Pourtois, G., Dan, E. S., Grandjean, D., Sander, D. & Vuilleumier, P. Enhanced extrastriate visual

response to bandpass spatial frequency filtered fearful faces: Time course and topographic evoked-potentials mapping. *Hum. Brain Mapp.* **26**, 65–79 (2005).

32. Schupp, H. T. *et al.* Selective visual attention to emotion. *J. Neurosci.* **27**, 1082–1089 (2007).
33. Miller, M. & Clark, A. Happily entangled: prediction, emotion, and the embodied mind. *Synthese* **195**, 2559–2575 (2018).
34. Edelman, G. M. & Gally, J. A. Reentry: a key mechanism for integration of brain function. *Front. Integr. Neurosci.* 63 (2013).
35. Oliva, A. & Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.* **155**, 23–36 (2006).
36. VanRullen, R. & Thorpe, S. J. The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.* **13**, 454–461 (2001).
37. Löken, L. S., Wessberg, J., Morrison, I., McGlone, F. & Olausson, H. Coding of pleasant touch by unmyelinated afferents in humans. *Nat. Neurosci.* **12**, 547–548 (2009).
38. Lapid, H. *et al.* Neural activity at the human olfactory epithelium reflects olfactory perception. *Nat. Neurosci.* **14**, 1455–1461 (2011).
39. Todd, R. M., Miskovic, V., Chikazoe, J. & Anderson, A. K. Emotional Objectivity: Neural Representations of Emotions and Their Interaction with Cognition. *Annu. Rev. Psychol.* **71**, null (2020).
40. Palmer, S. E. & Schloss, K. B. An ecological valence theory of human color preference. *Proc. Natl. Acad. Sci.* **107**, 8877–8882 (2010).
41. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
42. Bookbinder, S. H. & Brainerd, C. J. Emotionally negative pictures enhance gist memory. *Emotion* **17**, 102 (2017).
43. Hickey, C. & Peelen, M. V. Neural mechanisms of incentive salience in naturalistic human vision. *Neuron* **85**, 512–518 (2015).

44. Miskovic, V. & Keil, A. Escape from harm: linking affective vision and motor responses during active avoidance. *Soc. Cogn. Affect. Neurosci.* **9**, 1993–2000 (2014).
45. Rhodes, L. J., Ruiz, A., Ríos, M., Nguyen, T. & Miskovic, V. Differential aversive learning enhances orientation discrimination. *Cogn. Emot.* **32**, 885–891 (2018).
46. Graf, P. & Schacter, D. L. Implicit and explicit memory for new associations in normal and amnesic subjects. *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 501 (1985).
47. Kahneman, D. *Thinking, fast and slow*. (macmillan, 2011).
48. Zajonc, R. B. Feeling and thinking: Preferences need no inferences. *Am. Psychol.* **35**, 151 (1980).
49. Zajonc, R. B. On the primacy of affect. (1984).
50. Stirrat, M. & Perrett, D. I. Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychol. Sci.* **21**, 349–354 (2010).
51. Thornhill, R. & Gangestad, S. W. Facial attractiveness. *Trends Cogn. Sci.* **3**, 452–460 (1999).
52. Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K. & O'Doherty, J. P. Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nat. Hum. Behav.* **5**, 743–755 (2021).
53. Kardan, O. *et al.* Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature? *Front. Psychol.* **6**, 471 (2015).
54. Kellert, S. R. & Wilson, E. O. *The biophilia hypothesis*. (Island press, 1995).
55. Schmitz, T. W., De Rosa, E. & Anderson, A. K. Opposing influences of affective state valence on visual cortical encoding. *J. Neurosci.* **29**, 7199–7207 (2009).
56. Carretié, L., Tapia, M., López-Martín, S. & Albert, J. EmoMadrid: An emotional pictures database for affect research. *Motiv. Emot.* **43**, 929–939 (2019).
57. Wessa, M. *et al.* EmoPics: Subjektive und psychophysiologische Evaluation neuen Bildmaterials für die klinisch-bio-psychologische Forschung. *Z. Für Klin. Psychol. Psychother.* **39**, 77 (2010).
58. Dan-Glauser, E. S. & Scherer, K. R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **43**, 468–477 (2011).

59. Lang, P. J., Bradley, M. M. & Cuthbert, B. N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent. Study Emot. Atten.* **1**, 39–58 (1997).
60. Marchewka, A., Żurawski, Ł., Jednoróg, K. & Grabowska, A. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav. Res. Methods* **46**, 596–610 (2014).
61. Wierzba, M. *et al.* Erotic subset for the Nencki Affective Picture System (NAPS ERO): cross-sexual comparison study. *Front. Psychol.* **6**, 1336 (2015).
62. Kurdi, B., Lozano, S. & Banaji, M. R. Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* **49**, 457–470 (2017).
63. Michałowski, J. M. *et al.* The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and nonfearful individuals. *Behav. Res. Methods* **49**, 1407–1419 (2017).
64. Crone, D. L., Bode, S., Murawski, C. & Laham, S. M. The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PloS One* **13**, e0190954 (2018).
65. Van De Weijer, J., Schmid, C. & Verbeek, J. Learning color names from real-world images. in *2007 IEEE Conference on Computer Vision and Pattern Recognition* 1–8 (IEEE, 2007).
66. Crouzet, S. M. & Thorpe, S. J. Low-level cues and ultra-fast face detection. *Front. Psychol.* **2**, 342 (2011).
67. Gaspar, C. M. & Rousselet, G. A. How do amplitude spectra influence rapid animal detection? *Vision Res.* **49**, 3001–3012 (2009).
68. Cox, R. W. Equitable thresholding and clustering: a novel method for functional magnetic resonance imaging clustering in AFNI. *Brain Connect.* **9**, 529–538 (2019).
69. Gu, Z., Jamison, K., Sabuncu, M. R. & Kuceyeski, A. Modulating human brain responses via optimal natural image selection and synthetic image generation. *ArXiv Prepr. ArXiv230409225* (2023).
70. Brock, A., Donahue, J. & Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *ArXiv Prepr. ArXiv180911096* (2018).

71. Gu, Z., Jamison, K., Sabuncu, M. & Kuceyeski, A. Personalized visual encoding model construction with small data. *Commun. Biol.* **5**, 1382 (2022).
72. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv Prepr. ArXiv14091556* (2014).
73. Klindt, D., Ecker, A. S., Euler, T. & Bethge, M. Neural system identification for large populations separating "what" and "where". *Adv. Neural Inf. Process. Syst.* **30**, (2017).

Author contributions

S.S. and A.K.A. conceptualized the project and performed data analysis and modeling. S.S., A.K.A, and E.D. designed the study 3 experiment. Z.G. and A.K. performed the image synthesis in study 5. S.S. and A.K.A. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.