Scaling models of visual working memory to natural images

Christopher J. Bates[1*], George Alvarez[1] and Samuel J. Gershman[1]

[1*]Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, 02138, Massachusetts, USA.

*Corresponding author(s). E-mail(s): cjbates@g.harvard.edu;
Contributing authors: alvarez@wjh.harvard.edu;
gershman@fas.harvard.edu;

Abstract

Over the last few decades, psychologists have developed precise quantitative models of human recall performance in visual working memory (VWM) tasks. However, these models are tailored to a particular class of artificial stimulus displays and simple feature reports from participants (e.g., the color or orientation of a simple object). Our work has two aims. The first is to build models that explain people's memory errors in continuous report tasks with natural images. Here, we use image generation algorithms to generate continuously varying response alternatives that differ from the stimulus image in natural and complex ways, in order to capture the richness of people's stored representations. The second aim is to determine whether models that do a good job of explaining memory errors with natural images also explain errors in the more heavily studied domain of artificial displays with simple items. We find that: (i) features taken from state-of-the-art deep encoders explain coarse-grained and some fine-grained aspects of trial-by-trial difficulty in natural images, while several reasonable baselines do not; and (ii) deep visual encoders could reproduce set-size effects but overall offered a poorer explanation of human data in the artificial domain. Together, our results suggest that people may rely on distinct cognitive systems or brain areas in artificial versus natural task domains. Moving forward, our approach offers a scalable way to build a more generalized understanding of VWM representations by combining recent advances in both AI and cognitive modeling.

Keywords: Visual working memory, deep neural networks, psychophysics

1 Introduction

When viewing an image, what details do we store in memory over the short term? What is the nature of the cognitive bottleneck that restricts how much information we can retain and recall? These and related questions have been pursued for the last several decades, leading to the discovery of a number of striking behavioral phenomena, including set-size, attraction, repulsion, and inter-item interaction effects [1, 2]. Mathematical models offer compelling and principled explanations for many of these phenomena [3–8]. However, while these models are able to test competing theories about the nature of people's memory representations and capacity limits, they lack generality. Critically, they cannot predict what people will recall about *natural images*. While challenging, it is crucial to study memory for more ecological stimuli, since findings are likely to reveal important cognitive design principles that cannot be discovered by studying more simplified and artificial settings alone [9]. Moreover, given that our visual systems are optimized primarily to operate on natural images, it is reasonable to ask whether many of the phenomena we have identified in artificial domains are related to this adaptation.

The effort to study visual memory in more ecological settings is hindered in part by the same kinds of technological challenges facing much of vision science. Due to the visual system's complexity, we have long lacked precise models of the computations carried out in the visual stream. A simultaneous challenge lies in stimulus design. In order to probe the richness of our representations in the domain of natural images, we need methods to continuously vary stimuli in ways that appear natural to participants. Deep learning is beginning to offer effective tools to solve both of these problems.

In order to build a general computational account of VWM for natural images, we need a theory of where visual features come from. We argue that the most parsimonious hypothesis is that VWM is primarily built on top of feature detectors residing in the visual stream, and that our memory systems select subsets of these features and store noisy or compressed versions of them. Arguably, the most precise models to-date for computations carried out along our visual streams come from certain classes of deep neural networks (DNNs) [10, 11]. Thus, a reasonable starting place would be to select features from these networks as candidates for the features that feed into VWM.

In order to predict behavior, we next need to combine the selected deep neural network features with a noise model. Here, we adopt the Target Confusability Competition (TCC) model [7]. This model is a generalization of a standard signal detection model, which assumes two response options, to tasks with arbitrary numbers of choices. Critical to our purposes, it can generate predictions for any feature space, including the kinds of complex, high-dimensional

feature vectors that are likely needed to capture human visual representations, such as those derived from DNNs. The TCC model is flexible in this way because it relies only on pairwise similarity scores between the target stimulus and each response alternative. Thus, the stimuli can be represented in any hypothesized feature space, as long as a valid similarity metric can be applied. Incorrect responses are assumed to result from a noise process that corrupts the similarity scores (specifically, additive Gaussian noise).

We note that an alternative to TCC would be to add noise directly to the DNN representations, rather than to pairwise similarity scores. For instance, one could add Gaussian noise to each dimension of each DNN representation, then compute similarity scores using the noise-corrupted vectors, and finally take the maximum score as the response. This would lead to a model that is mathematically similar, but raises the complication that the model's behavior then depends on nuisance factors, such as the dimensionality of the visual representations and statistical moments of the activation values. Here, we are most interested in whether the *representational geometry* of people's VWM representations is similar to that of a candidate DNN layer [12]. That is, does higher pairwise similarity in the DNN layer's representational space predict higher confusability in people's memories?

Our TCC-based models build on the original work in important ways. First, while Schurgin et al. [7] refit the model's single noise parameter (d') for each set-size, here we show that feature spaces from select DNN layers can reproduce set-size effects without fitting separate noise parameters. Second, Schurgin et al. derived a psychological similarity function from perceptual similarity judgments, without identifying the origin of this similarity function. We show how DNNs can be used to derive similarity functions that are predictive of VWM for natural images. This also yields a practical benefit by obviating the need to collect pairwise similarity judgments, which is impractical for very large stimulus spaces.

We apply our modeling framework to VWM for both natural images and artificial stimuli (color and orientation), comparing several different DNN-based feature representations. To evaluate the models, we used a combination of quantitative metrics (correlation, likelihood) and qualitative checks (summary statistics derived from from the models and data). We show that our framework can capture many aspects of both natural and artificial stimuli, but that there is a sharp divergence in performance between the two stimulus sets, with our DNNs fairing worse in the artificial domain.

2 Results

2.1 Continuous report with natural images

To study VWM for natural images, we analyzed data collected by [13]. Stimuli were generated using StyleGAN [14] (a generative adversarial network) trained to produce novel, naturalistic indoor scenes (Fig. 1). We will refer to this as

Fig. 1 Evenly-spaced samples from one wheel in the Scene Wheels experiment (radius=8).

the "Scene Wheels dataset". On each trial, participants performed a continuous report task, where the stimulus and all response alternatives were evenly sampled from the circumference of a circle, which was drawn in a randomly-sampled 2D plane in the GAN's high-dimensional latent space. Trial difficulty was controlled at a coarse level by changing the radius of the circle in latent space. Larger radii resulted in more distinct response alternatives, since they were further away from each other in code-space. The dataset includes 25 total "wheels" (circles in latent space), with five unique center points and five different radii around each center point.

Model zoo. We compare TCC models constructed based on a wide range of feature spaces, including layers from deep vision models and simpler baseline models. Our two simplest baselines are the raw pixel vectors (length $3 \times 256 \times 256$) and the RGB channel averages (length 3). We also include the latent representation from a β-Variational Autoencoder (β-VAE) [15] as a more sophisticated baseline. Deep autoencoder models have been explored as tools to learn better image and video compression algorithms for technological applications [16, 17], as well as to model human visual memory [18–20]. In addition to baseline models, we consider networks trained on the ILSVRC ImageNet classification challenge (both the 1,000-way and 22,000-way versions) and networks trained on the Contrastive Language-Image Pre-training (CLIP) objective [21]. The CLIP objective is conceptually related to classification, but it encourages networks to learn semantically richer outputs that capture all the information contained in a typical image caption rather than a single class label. We selected a subset of pre-trained models provided by OpenAI, including models based on the ResNet-50 backbone (and larger variants of the same architecture), which is a convolutional network, and Vision Transformer, which is non-convolutional but also shown to be human-like [22]. For the ImageNet classifiers, we took several classic, pre-trained networks from the Torchvision repository. We also took pre-trained ConvNext models [23] (a recent convolutional competitor to Vision Transformers) from Facebook's Huggingface repository. Finally, we took a "harmonized" version of ResNet-50 from the repository provided by [24], which is optimized to encourage classification decisions to depend on the same areas in the image that humans rely on when making the same decisions.

TCC model. We construct a separate TCC model for each layer in each architecture, as well as each baseline (see Fig. 2 for a schematic). For each

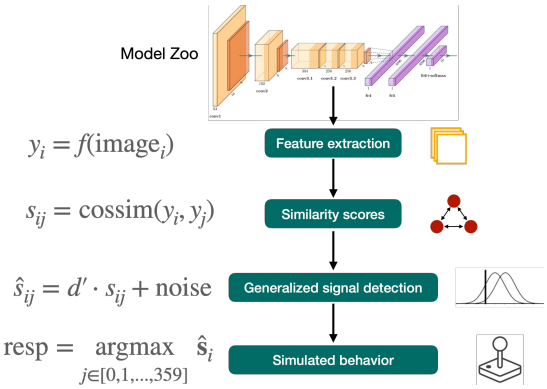

Fig. 2 Schematic overview of our modeling pipeline using DNN features and the TCC model. For a given DNN model and layer within that model, we take the (flattened) activations from that layer after feeding in a stimulus image and each response alternative from the scene wheel, in turn. There were always 360 evenly-spaced response options. For each option j, we computed cosine similarity between that option's activation vector (y_j) and the stimulus's (y_i). After scaling by a constant factor d', we added independent Gaussian noise with unit variance to each of the 360 similarity scores to produce corrupted similarity scores. Finally, we assumed responses were the argmax of these noisy scores.

trial, we compute all pairwise similarities between the target stimulus and each of the 360 options along the response wheel. We then multiply these 360 similarity scores by a scaler, d', which corresponds to the memory strength for an exact match (similarity = 1), and therefore controls response accuracy. Finally, we add independent Gaussian noise with unit variance to each of the scores and take the option with the max score as the model's response on that simulated trial. (Note that it would be mathematically equivalent to scale the variance of the noise, rather than the similarity scores.) We simulated 8000 model responses for each trial in the dataset.

Trial difficulty rank-order analysis. For each architecture considered, we searched for the layer that best matched human data. For each layer, we fit our only model parameter, d', according to model likelihood. We conducted a grid search over d' values and used a histogram approximation to the model likelihoods. We then estimated the Spearman correlation coefficient between the human and model mean absolute error per trial. Because there was a large number of unique stimuli compared to the number of responses collected, it was necessary to bin trials. (Note that nearby stimuli on a given response wheel tended to be highly similar.) We divided each scene wheel into 12 evenly-sized bins, and for each bin, we averaged errors across all trials for which the target stimulus fell within that bin.

Results of the Spearman analysis are presented in Fig. 3 (top panels). The left-most panel (all trials aggregated) demonstrates that features taken from our selected CLIP and ImageNet classifier models capture trial-by-trial difficulty better than baselines. Within radius, our best models still explain some fine-grained variance in the rank-order of difficulty, but the amount explained differs by radius, whereas baseline models do not. As expected, baselines also

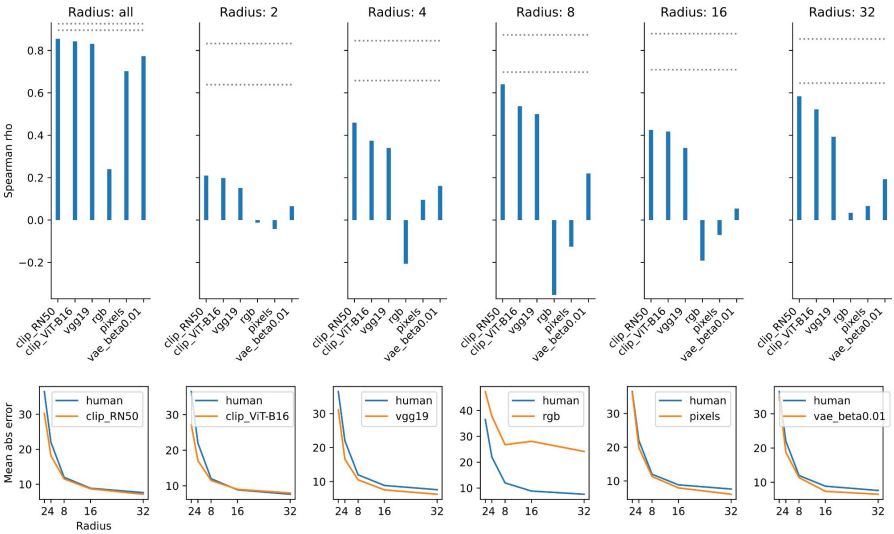

Fig. 3 Rank-order difficulty results for Scene Wheels dataset. Top panels: Spearman rank correlations for trial difficulty between best layer in selected DNN architectures and humans. Dotted lines are an indication of the noise ceiling. Specifically, we took bootstrap resamples of human responses within each radius, and for each resample we computed the Spearman correlation coefficient between it and the original data. The lines are the fifth and 95th percentile. Blue bars indicate p-values less than 0.05. Bottom panels: Comparison of human and model mean errors within each wheel radius.

had lower likelihoods (Table 1). As another way to compare models, we also plotted mean error per radius for humans and models (Fig. 3, bottom panels). Interestingly, both our VAE model and raw pixels capture the relationship between error and radius just as well as our best models, even while failing to capture more fine-grained variance within each radius.

Table 1 Comparison of models and baselines on Scene Wheels dataset

Model	Log-likelihood	Spearman
RGB channel means	-25722	0.24 ($p < 0.001$)
Pixels	-23077	0.70 ($p < 0.001$)
VAE	-22935	0.77 ($p < 0.001$)
CLIP-RN50	-22628	0.85 ($p < 0.001$)
CLIP-ViT-B16	-22647	0.84 ($p < 0.001$)
VGG-19	-22606	0.83 ($p < 0.001$)

We also conducted a comparison across DNN architectures to examine what factors might lead an architecture to explain more variance in this experiment (Fig. 4). We considered several dimensions, including number of images seen during training, type of architecture, and number of trainable parameters. Since we are unable to do an exhaustive search over these factors (and various

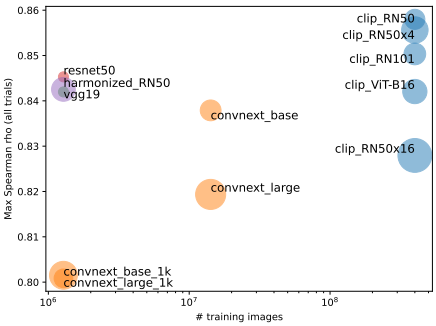

Fig. 4 Extended cross-model comparison on Scene Wheels dataset. Marker radius is proportional to number of trainable parameters. ConvNext models labeled '1k' were trained on the 1,000-way ImageNet classification dataset and the others were trained on the 22,000-way version.

confounds may exist), we present qualitative results, which may be suggestive for future work.

Overall, we find that architecture, number of trainable parameters, and number of training images may all be important factors. For each architecture, we selected the best layer according to its Spearman correlation in the rank-order difficulty analysis. We find that the highest correlations are achieved by the CLIP pre-trained networks, which also saw the most images during training. At the same time, we see that within the class of ConvNext models, increasing the number of training images increases correlation. Although number of training images may be confounded with objective, since the better performing ConvNext models were trained on the 22,000-way classification task as opposed to 1,000-way. Another possibility is that training objectives that encourage richer semantic information at the output layer lead to higher correlations. Keeping objective and training set fixed, we also see that some architectures outperform others. Within CLIP-trained models, the Vision Transformer does worse than several convolutional architectures. Within models trained on ImageNet 1,000-way classification, VGG-19 and ResNet50 outperform ConvNext. Finally, we see evidence that continuing to increase model size after a certain point leads to lower correlations. Within both CLIP and ConvNext architectures, the largest models fare worse than smaller ones, at least when the number of training images is large. If this effect is indeed mediated by training set size, it could be related to the finding that very large classifiers tend to find more ways to "cheat" the dataset in ways that humans do not [25, 26].

Sources of trial-by-trial difficulty. A successful model of memory should be able to explain why average difficulty varies across trials, even within set size. A straightforward way to make a trial difficult for a participant is to provide response options that are very similar to the target stimulus (e.g., by choosing a small scene wheel radius). However, it is also possible that stimulus-specific factors modulate difficulty, independent of the response options. For

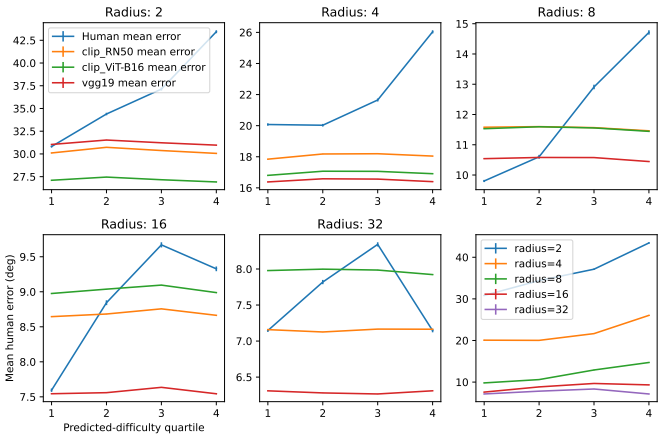

Fig. 5 Results of trial-difficulty regression analysis. We trained regression models using recursive feature elimination to predict human mean error on each stimulus. All input features summarized a stimulus with a single value. Each panel considers trials within a single radius, since radius modulated difficulty but was not a stimulus-intrinsic factor (see Main text). The blue, monotonic increasing lines in radii 2, 4, and 8 indicate that the regression models capture fine-grained aspects of stimulus-specific difficulty. The limits in predictive accuracy of the model can be observed in the larger radii, which are no longer monotonic. Larger radii are harder to explain, because errors are smaller in magnitude and less variable. Note the lines in each panel corresponding to the TCC models are flat, which indicates they could not explain this stimulus-specific variability. Error bars are bootstrapped 90% confidence intervals using 1000 random train-test splits. The bottom-right panel re-plots the regression model predictions on a single plot for easier comparison.

instance, set-size effects in classic multi-item displays are usually explained in terms of visual load [1, 27]. These accounts posit that a limited resource is spread in some manner across items (or features) in the display, and thus average precision for each item necessarily decreases as a function of the quantity of items stored. It is possible, for example, that an analogue to the set-size effect exists in natural images, although a challenge lies in identifying what counts as an item.

To test for stimulus-specific factors of difficulty, we searched for summary statistics (see Methods) that increased monotonically with human mean error, where each statistic summarized a stimulus image with a single number. We found no individual statistic that exhibited a monotonic relationship. However, we conducted a regression analysis to examine whether linear combinations of the same features could produce a monotonic function (Fig. 5). We found this to be the case, suggesting there exist factors that drive mnemonic difficulty which are independent of the set of response options in the experiment. We then asked whether our TCC models could explain this relationship and found that they could not (see lines in plot corresponding to models, which are mostly flat).

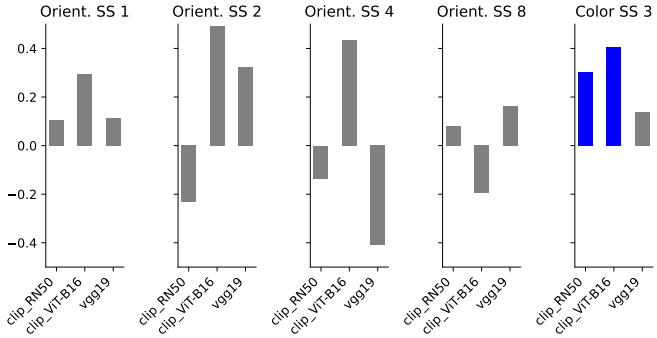

Fig. 6 Results of the Spearman correlation analysis after refitting our models to color and orientation stimuli.

2.2 Continuous report with color and orientation

In our experiments with artificial images, we analyzed previously collected data from experiments studying color [28] and orientation [6] working memory. Both experiments we analyzed used continuous report tasks. In the color memory experiment, every item in each display was probed. In the orientation experiment, one item was probed at random. In addition to examining rank-order of trial difficulty as above, we aimed to explain set-size effects, as well as a subset of well-known response biases and inter-item effects. We restricted our evaluation to the same subset of well-performing models presented in the Scene Wheels experiment. In each experiment, we generated stimuli to match the set of items shown to participants.

Fig. 6 shows results of the rank-order difficulty analysis, after refitting our selected DNN architectures to the color and orientation datasets, separately. We find that within each set-size, the best models have positive correlations but in most cases do not reach a significance threshold of $p < 0.05$. The ρ values are also less than was found for most of the Scene Wheel radii.

We next examined set-size effects in the orientation working memory dataset, which included four set sizes (1, 2, 4, and 8). We compared mean absolute error per set-size between humans and our models and found all best-fit models to exhibit a set-size effect, but closest correspondence was found for the VGG-19 model (Fig. 7, left panel). We further investigated what causes mean error in our models to vary as a function of set-size. We find that the effect is caused by the sparsity of activations in the DNN layers. As more objects are added to the background of the image, more activations become non-zero, causing the range of similarity values in the response options to shrink (Fig. 7, right panels). Since noise with fixed variance is added to these values, responses become more easily corrupted with increasing set-sizes. This explanation bears some resemblance to neural resource models of working memory [8], which appeal to divisive normalization between neurons in a population as the mechanism to control neural resource allocation across items in the display. However, these neural population models differ in that they predict

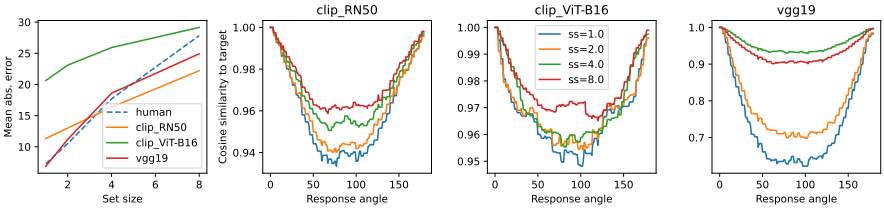

Fig. 7 Analysis of set-size effects in TCC models. Left panel: Comparison of human and model mean error within each set-size in the orientation memory task. Right panels: Raw similarity scores (per set-size) for our three selected models between a target with a horizontal orientation and all response options.

a relatively constant overall level of activation, whereas the DNN models we examined increase their activation with set-size.

We next asked whether our models could explain response inhomogeneities in color and orientation working memory. A striking finding from orientation memory experiments is that recall for nearly horizontal and vertical orientations is exaggerated away from these cardinal orientations (repulsion). At the same time, responses are biased toward the oblique orientations (attraction) [29]. In color working memory, there exist a set of "focal" colors that responses are biased toward [30]. By visual inspection, we found that the bias of our best-fit VGG-19 model bears some similarity to human bias in the orientation memory task, although it also differs in some ways (Fig. 8, left). We note, however, that further investigation revealed that while deep layers of the model deviated somewhat from the exact shape of human biases in orientation memory, earlier layers exhibited a closer correspondence (see Supplementary), suggesting this class of models is capable in principle of explaining these biases. By contrast, in color memory, none of the model layers are able to explain the pattern of biases observed in humans (Fig. 8, right), although the models all exhibit a similar pattern of biases.

Finally, we consider inter-item effects. [5] found strong evidence that memory errors for one item in a display depend on the other items they appeared with. A specific hypothesis they tested was that people store hierarchical representations of the displays. At the upper level, they may record the overall level of dissimilarity between the items, while at a lower level, they record item-specific details. To test this hypothesis, they computed a correlation coefficient between circular variances, where one vector comprised the variances of the three hues in each stimulus display and the other was the variances of the three chosen hues at response time. Importantly, they only included trials for which participants were far off on all their responses (> 45 deg). They found a significant correlation of 0.4. When we conducted the same analysis with our selected models, we found an insignificant correlation near zero for all of them.

Taken together, the above results suggest that layers from our selected DNNs do not adequately explain participants' behavior in continuous report for color and orientation. A possible reason for this mismatch between models and humans is that the models were trained only on natural images and did

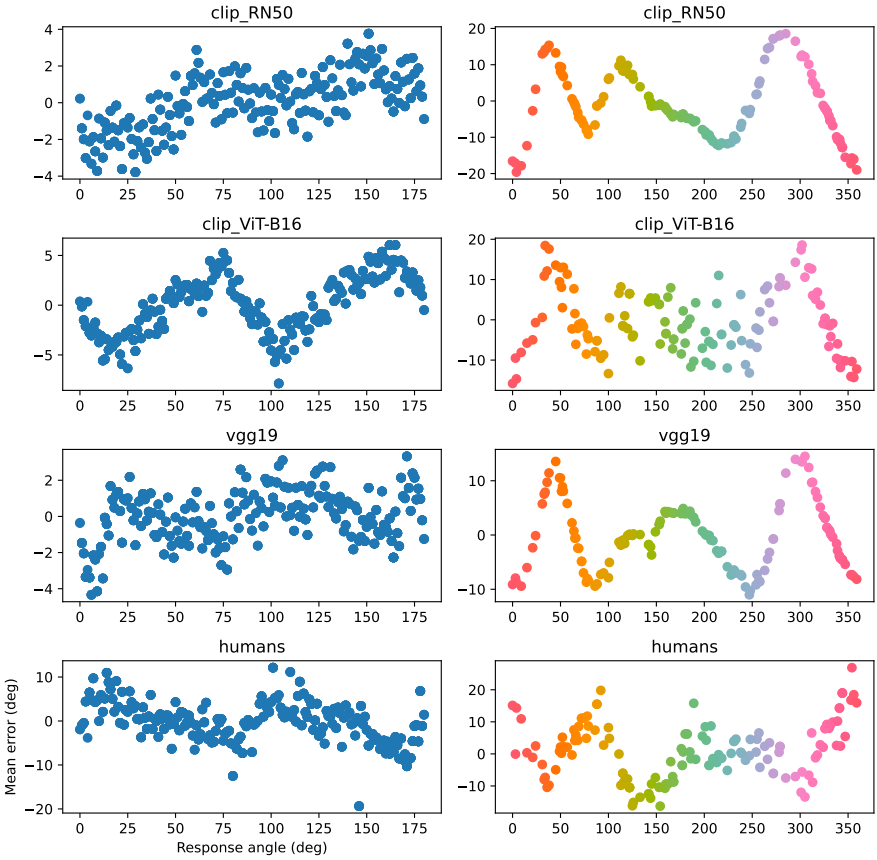

Fig. 8 Comparison of humans' and models' average response bias in orientation (left) and color (right) memory tasks.

not see anything like the artificial stimuli during training. We tested for this possibility by fine-tuning our two CLIP models on the color stimuli. We created captions by finding the nearest color name from a database of color names for each item in the display. After redoing our analyses, we found no improvement in model fit (see Supplementary), suggesting human behavior in these kinds of artificial displays may not be easily captured by the kinds of DNN architectures we tested.

Finally, we asked whether the same or similar layers within each model provided the best explanations across experiments. For VGG-19, we the best-fit layer for the Scene Wheels dataset was layer 30, but for color and orientation it was 7 and 22, respectively. For CLIP ResNet-50, the best layer was 24 for Scene Wheels, 11 for color and 16 for orientation. For CLIP ViT-B16, the best layer was 12 for Scene Wheels, 23 for color, and 11 for orientation. Thus, for both convolutional architectures, the best layer was deeper for natural images

than both the artificial experiments, but for the vision transformer-based architecture, this was not the case. However, since the results were generally poor for the artificial stimuli, we also looked across all layers (Supplementary) to see if a systematic relationship emerged between layer depth and correlation. Doing so, we found that deeper layers are consistently better explanations of the Scene Wheels data for the convolutional architectures, but the relationship is less clear for ViT. In the case of color and orientation, no clear trend emerges in any of the architectures.

3 Discussion

We combined several recent advances from cognitive science and AI to build *scalable* models of visual memory. We sought to build models that are not restricted to tasks with low-dimensional stimuli and/or simple feature reports, but can make more general predictions. In particular, we sought to understand what features are stored in memory over the short term after viewing natural images. We then asked whether similar features are stored when viewing the kinds of sparse, artificial displays typically used in working memory experiments.

We constrained our search for human-like features to two classes of pretrained DNN, ImageNet classifiers and CLIP models. We found that our best models as well as several baselines were able to capture coarse-grained aspects of people's psychological similarities, specifically the increase in mean error with smaller scene wheel radii. However, our models based on DNNs were able to capture much more fine-grained variance than all baselines, as measured by rank-order correlations of trial difficulty. When we refit our models to color and orientation tasks, the story was different. While the best-fitting models were able to capture set-size effects, and to a certain extent human-like biases in orientation memory, overall they provided a poor explanation of trial-by-trial difficulty in both color and orientation experiments compared to the natural images.

A likely contributing factor to this shortcoming is that the models' representations for the probed item did not depend very much on the other items in the display, at least in the case of colors, whereas human data show strong inter-item dependencies. Specifics of the DNN architectures may contribute to this issue. For instance, early layers of convolutional networks have small receptive fields and thus their representations of spaced-out items in a display may interact little. However, this is not likely to be the only issue, since some of our best-fit layers were deep enough in principle to have inter-item interactions.

Another potential reason that our DNNs failed to adequately capture human responses with simple color and orientation features is that the color and orientation stimuli were far from their training distribution, and thus in a sense, the networks did not "understand" the content of the image. For example, human participants naturally parse such images into a set of objects with

configural properties [31, 32]. We reasoned that an appropriate fine-tuning procedure might endow our models with similar abilities. CLIP models provide an ideal starting point toward this goal, since we can train the model with a description of the contents of the artificial stimuli. To test this, we fine-tuned our CLIP models on the color stimulus images accompanied with captions that named the colors present in the display. However, we found no resulting improvement in the models' fits to participant responses.

Future work should more closely examine why human responses diverge from models built on representations taken from deep classifier and CLIP-trained models. One possibility that has already been put forward in the VWM literature is that people may employ meta-cognitive strategies that are tailored to simple multi-item displays, such as chunking and hierarchical encoding [33–35]. Alternatively, these kinds of behaviors might arise naturally from a DNN trained on the appropriate objective. But our results here already rule out some common training objectives (classification, CLIP, pixel-wise reconstruction), using vision transformers as well as a variety of convolutional architectures. Attention mechanisms might be another source of divergence between humans and models in our tasks. For example, there is evidence that attentional dynamics are a critical component of explaining VWM performance in sparse multi-item displays [36–38]. By contrast, popular DNN models solve their objectives without any explicit attentional modulations [32].

Our models were built on the hypothesis that the features stored in VWM are noisy or compressed versions of features computed when initially perceiving a stimulus. But this hypothesis could be tested in a more direct way by using neural recordings. Previous work has shown that when DNNs are trained to predict neural activity directly (e.g., using fMRI data), the learned representations recapitulate key behaviors and capabilities of human vision. For instance, when trained on activity from face-selective areas, the resulting representations are able to solve non-trivial segmentation problems, picking out faces in complex scenes [39, 40]. The outputs of these networks, or even the fMRI data used to train them, could be directly swapped in for the features we used in the present work. By building TCC models directly on top of neural features, we might also elucidate the discrepancies we found between natural and artificial images. For example, we could visualize the ways that participants segment and interpret artificial displays and use the result to pinpoint failure modes of DNNs trained to predict labels or captions. Using the same approach, we could also ask whether people rely on different brain areas depending on the stimulus set or task [41].

Finally, our approach may prove useful in clarifying some longstanding debates about the nature of VWM. In particular, our method allows us to ask how biases and capacity limits in certain artificial paradigms fit into a larger picture that includes behavior in more natural settings and tasks. Given that many important findings about VWM come from unnatural stimuli, an important baseline to test is whether adaptation to demands of our natural environment explains these phenomena. Our results shed light on this and

related questions. We found that a TCC model using the right DNN features could explain both set-size effects and repulsion biases in orientation memory, despite only being trained to classify natural images. More work is necessary to determine whether these correspondences are simply coincidental or provide a satisfying explanation of human behavior. Nonetheless, our work constitutes a necessary first step toward more flexible and general models of visual memory that can accommodate findings from both natural and artificial stimulus domains. Such models must at minimum be able to capture the representational geometry of the features stored in VWM. Thus, our models were designed to evaluate different feature spaces according to how well their geometries explained human error patterns. And while this is only a first step, our novel framework provides clear avenues for future investigation. It makes clear that significant efforts should be devoted to determining what elements are missing from current DNN architectures and in what ways they need to be modified in order to explain more variance in VWM tasks, especially those involving artificial displays.

4 Methods

4.1 DNN layers

For ResNet-based models, we selected all layers from the pre-residual-block portion of the network, the last convolutional layer within each bottleneck sub-module in each residual block, and the pooling layer just before the final output. For VGG-19, we selected the last 32 layers within the sub-module labeled "features" in the Torchvision implementation. This excluded the last two fully-connected layers before the soft-max output. For the ConvNext models, we took the output of each `ConvNextLayer` within each `ConvNextStage`, as defined by the PyTorch model. For Vision Transformer-based models, we took the `attn` and `ln_2` sub-layers from each attention layer.

We downloaded pre-trained CLIP models from https://github.com/openai/CLIP, PyTorch ImageNet classifiers from https://pytorch.org/vision/stable/models.html and ConvNext models https://huggingface.co/models?sort=downloads&search=facebook%2Fconvnext.

4.2 Trial-difficulty regression analysis

We trained linear regression models to predict trial difficulty in the scene-wheel dataset. Our proxy measure for difficulty was mean absolute human error. Stimuli were binned in the same manner as for the Spearman rank correlation analyses (i.e. bin size of 30 degrees, see main text). We trained on a subset of scene wheel radii (2, 4, and 8), and sampled 1000 random train/test splits. Splits were 50/50 and were done within each radius separately to keep the dataset balanced. We re-trained the regression model for each of the 1000 resamples, using only items from the training split. Results presented in the main text use model predictions on the test split only. Table 2 gives descriptions

Table 2 Description of features used in trial-difficulty regression analysis

Feature name	Description
radius	Radius in GAN space. Not a feature of stimulus but of response set. Larger radius means response alternatives are more distinct from stimulus and each other, and therefore result in less difficult trials. Included in regression to allow predictions to account for this stimulus-independent source of variability in human error.
disk_size	Disk storage size for stimulus images (here we used JPEG).
vae_beta0.01	mean(abs(z)), where z are the activation values corresponding to μ in β-VAE encoder output layer. This model was the same as the one used to create a baseline TCC model (see Main). Its β value was 0.01 and it was trained on the Places-365 dataset. Due to regularization toward the zero-mean prior, large activations magnitudes should generally be rare, but more complex images may elicit higher magnitudes.
countr_estimate	Output of CounTR model, which estimates number of objects in an image.
keypoints_2d	mean(abs(y)), where y is the output activation map from the 2D keypoints model in the Midlevel-vision repository. Higher values indicate higher confidence in the presence of a 2D keypoint. Images with more points of interest may be more difficult to remember. (Based on SURF features.)
keypoints_3d	Same as keypoints_2d except with 3D keypoints. (Based on NARF features.)
seg_2d	Disk size of image output from Midlevel-vision unsupervised segmentation (2D) model, which is based on gestalt principles. Images with greater gestalt complexity may be more difficult to remember.
seg_25d	Same as seg_2d, except with 2.5D gestalt features.
vgg19_l*_mean_abs	mean(abs(y)), where y are hidden activations from VGG19 trained on ImageNet-1k at layer i. Due to regularization, activations are sparse. Images with higher visual load may elicit larger activations.
vgg19_l*_spatial_entropy	$H(y)$, where y are hidden activations from VGG19 trained on ImageNet-1k at layer i, and H is the "spatial" entropy. This measure increases to the extent that points of interest (as indicated by non-zero activations) are more evenly distributed across the image. This may be a relevant factor if people prefer focal attention over diffuse.

for each of the features we included in the analysis. We used Scikit-Learn's recursive feature elimination with cross-fold validation (RFE-CV) with the minimum number of features set to one, applied to standard linear regression. Before training, we applied a log-odds (inverse sigmoid) transformation to the prediction targets to adjust for their circularity. Fig. 9 (top) plots how often each feature was kept by the RFE-CV procedure. Fig. 9 (bottom) visualizes inter-feature correlations.

References

[1] Ma, W.J., Husain, M., Bays, P.M.: Changing concepts of working memory. Nature neuroscience **17**(3), 347–356 (2014)

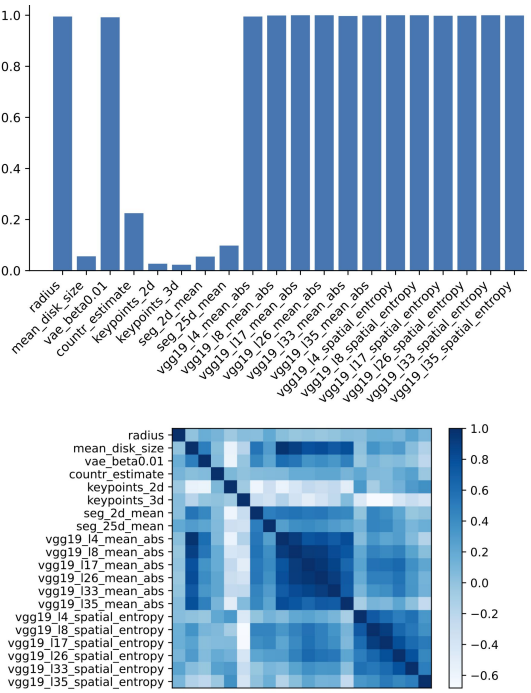

Fig. 9 Trial-difficulty regression analysis. Top: How often each feature was kept by the RFE-CV procedure. Bottom: Spearman correlation coefficient between each pair of features.

[2] Orhan, A.E., Sims, C.R., Jacobs, R.A., Knill, D.C.: The adaptive nature of visual working memory. Current Directions in Psychological Science **23**(3), 164–170 (2014)

[3] Sims, C., Jacobs, R., Knill, D.: An ideal observer analysis of visual working memory. Psychological Review **119**, 807–830 (2012)

[4] Orhan, A.E., Jacobs, R.A.: A probabilistic clustering theory of the organization of visual short-term memory. Psychological review **120**(2), 297 (2013)

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

[6] Bays, P.M.: Noise in neural populations accounts for errors in working memory. Journal of Neuroscience **34**(10), 3632–3645 (2014)

[7] Schurgin, M.W., Wixted, J.T., Brady, T.F.: Psychophysical scaling reveals a unified theory of visual memory strength. Nature human behaviour **4**(11), 1156–1172 (2020)

[8] Van den Berg, R., Ma, W.J.: A resource-rational theory of set size effects in human visual working memory. ELife **7**, 34963 (2018)

[9] Battleday, R.M., Peterson, J.C., Griffiths, T.L.: Capturing human categorization of natural images by combining deep networks and cognitive models. Nature communications **11**(1), 5418 (2020)

[10] Yamins, D.L., DiCarlo, J.J.: Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience **19**(3), 356–365 (2016)

[11] Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., Yamins, D.L.: Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences **118**(3), 2014196118 (2021)

[12] Kriegeskorte, N., Wei, X.-X.: Neural tuning and representational geometry. Nature Reviews Neuroscience **22**(11), 703–718 (2021)

[13] Son, G., Walther, D.B., Mack, M.L.: Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. Behavior Research Methods, 1–13 (2021)

[14] Yang, C., Shen, Y., Zhou, B.: Semantic hierarchy emerges in deep generative representations for scene synthesis. International Journal of Computer Vision **129**, 1451–1466 (2021)

[15] Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599 (2018)

[16] Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. arXiv preprint arXiv:1611.01704 (2016)

[17] Liu, X., Zhang, L., Guo, Z., Han, T., Ju, M., Xu, B., Liu, H., et al.: Medical image compression based on variational autoencoder. Mathematical Problems in Engineering **2022** (2022)

[18] Bates, C.J., Jacobs, R.A.: Efficient data compression in perception and perceptual memory. Psychological review **127**(5), 891 (2020)

[19] Hedayati, S., O'Donnell, R.E., Wyble, B.: A model of working memory for latent representations. Nature Human Behaviour **6**(5), 709–719 (2022)

[20] Nagy, D.G., Török, B., Orbán, G.: Optimal forgetting: Semantic compression of episodic memories. PLoS Computational Biology **16**(10), 1008367 (2020)

[21] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S.,

Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[22] Tuli, S., Dasgupta, I., Grant, E., Griffiths, T.L.: Are convolutional neural networks or transformers more like human vision? arXiv preprint arXiv:2105.07197 (2021)

[23] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

[24] Fel, T., Felipe, I., Linsley, D., Serre, T.: Harmonizing the object recognition strategies of deep neural networks with humans. arXiv preprint arXiv:2211.04533 (2022)

[25] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

[26] Jo, J., Bengio, Y.: Measuring the tendency of cnns to learn surface statistical regularities. arXiv preprint arXiv:1711.11561 (2017)

[27] Alvarez, G.A., Cavanagh, P.: The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological science **15**(2), 106–111 (2004)

[28] Brady, T.F., Alvarez, G.A.: Contextual effects in visual working memory reveal hierarchically structured memory representations. Journal of Vision **15**(15), 6–6 (2015)

[29] Wei, X.-X., Stocker, A.A.: A Bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. Nature neuroscience **18**(10), 1509–1517 (2015)

[30] Sims, C.R., Ma, Z., Allred, S.R., Lerch, R.A., Flombaum, J.I.: Exploring the cost function in color perception and memory: An information-theoretic model of categorical effects in color matching. In: CogSci, pp. 2273–2278 (2016)

[31] Herzog, M.H.: The irreducibility of vision: Gestalt, crowding and the fundamentals of vision. Vision **6**(2), 35 (2022)

[32] Peters, B., Kriegeskorte, N.: Capturing the objects of vision with neural networks. Nature human behaviour **5**(9), 1127–1144 (2021)

[33] Chunharas, C., Rademaker, R.L., Brady, T.F., Serences, J.T.: An adaptive

perspective on visual working memory distortions. Journal of Experimental Psychology: General (2022)

[34] Chunharas, C., Brady, T.F.: Is set size six really set size six? relational coding in visual working memory. Journal of Vision **19**(10), 134–134 (2019)

[35] Brady, T.F., Konkle, T., Alvarez, G.A.: A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of vision **11**(5), 4–4 (2011)

[36] Udale, R., Tran, M.T., Manohar, S., Husain, M.: Dynamic in-flight shifts of working memory resources across saccades. Journal of Experimental Psychology: Human Perception and Performance **48**(1), 21 (2022)

[37] Bays, P.M., Husain, M.: Dynamic shifts of limited working memory resources in human vision. Science **321**(5890), 851–854 (2008)

[38] Kong, G., Kroell, L.M., Schneegans, S., Aagten-Murphy, D., Bays, P.M.: Transsaccadic integration relies on a limited memory resource. Journal of Vision **21**(5), 24–24 (2021)

[39] Mieczkowski, E., Khosla, M., DiCarlo, J., Kanwisher, N., Murty, N.A.R., *et al.*: Computational models recapitulate key signatures of face, body and scene processing in the ffa, eba, and ppa. Journal of Vision **22**(14), 4337–4337 (2022)

[40] Khosla, M., Murty, N.A.R., Kanwisher, N.: Data-driven component modeling reveals the functional organization of high-level visual cortex. Journal of Vision **22**(14), 4184–4184 (2022)

[41] Bates, C., Gershman, S.: Coding strategies in memory for 3d objects: The influence of task uncertainty. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44 (2022)

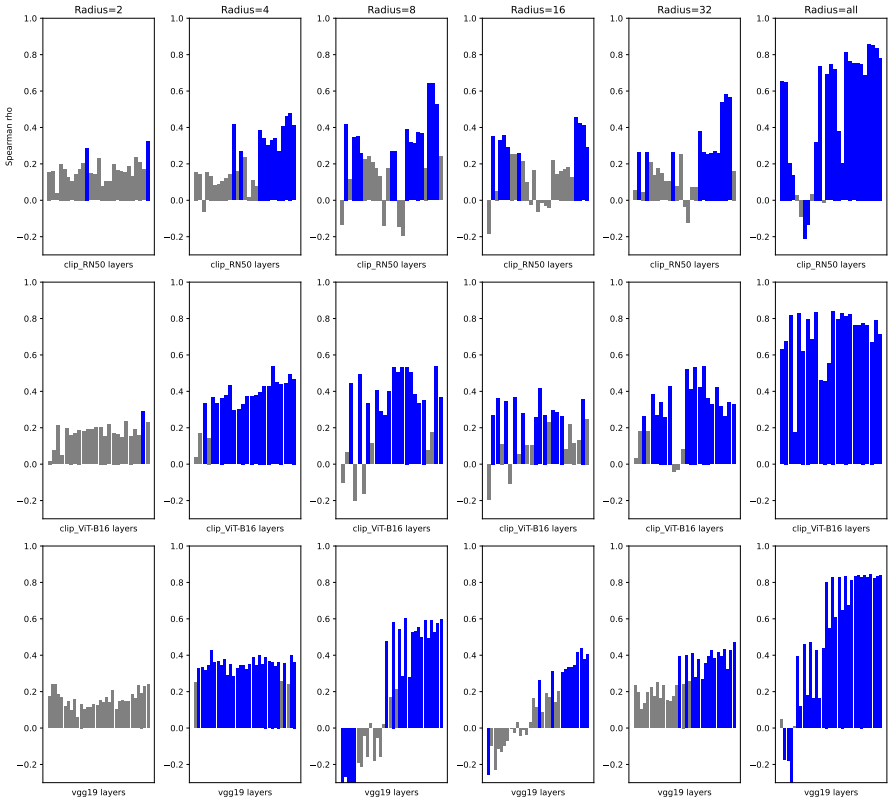

Fig. S1 Spearman rank correlations for trial difficulty between each layer in selected DNN architectures and humans on Scene Wheels dataset. Blue bars indicate p-values less than 0.05.

Supplementary Information

Results for all DNN layers

Fig. S1 shows results from all layers of each of the three selected models for the Scene Wheels dataset. Figs. S2 and S3 show the same for the orientation and color datasets, respectively. Note that in the case of orientation, d' is fit for each layer across all set-sizes simultaneously.

Repulsion bias in VGG-19

We examined additional layers of VGG-19 in the orientation memory experiment to see whether any layers exhibited the repulsion effect characteristic of human responses. For each layer examined, we set d' to its maximum-likelihood value (when fit across all set sizes). Fig. S4 plots response bias for layers 8-11. By inspection, layer 9 shows a clear repulsion bias that mirrors humans in its shape. Layers 10 and 11 show the effect as well, although they diverge more from human bias.

Fine-tuning CLIP models on color stimuli

We sampled stimuli using the same color wheel as the memory experiment, but varied set-size randomly between 1 and 4. To create each caption, we assigned a name to each color in the stimulus and joined the names together with a white-space as the delimiter. Colors were matched to the set contained in the database here: https://shallowsky.com/colormatch/rgb.txt. Matching was done by choosing the color name whose corresponding RGB value was the closest to the stimulus item, as measured by Euclidean distance. Fig. S5 shows the results of redoing the response-bias analysis with the fine-tuned models. The Spearman correlation coefficients for mean error between humans and the fine-tuned models were $0.15(p = 0.066)$ for CLIP-RN50 and $0.13(p = 0.11)$ for CLIP-ViT-B/16.

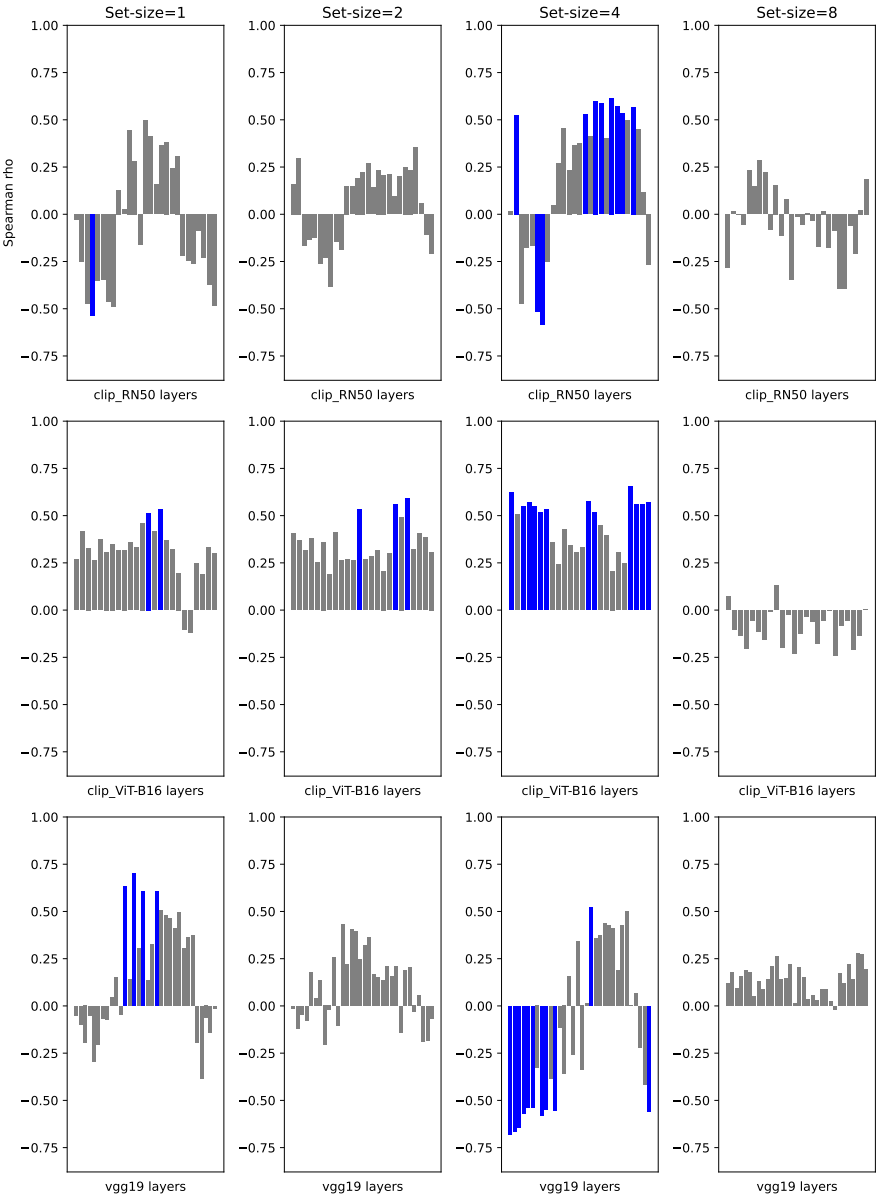

Fig. S2 Spearman rank correlations for trial difficulty between each layer in selected DNN architectures and humans on orientation memory dataset. Blue bars indicate p-values less than 0.05.

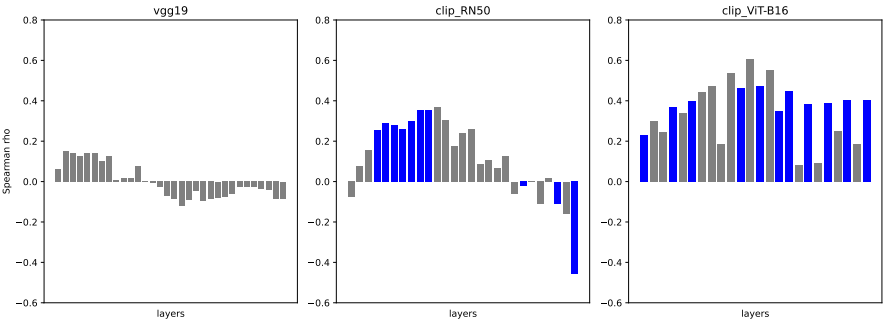

Fig. S3 Spearman rank correlations for trial difficulty between each layer in selected DNN architectures and humans on color memory dataset. Blue bars indicate p-values less than 0.05.

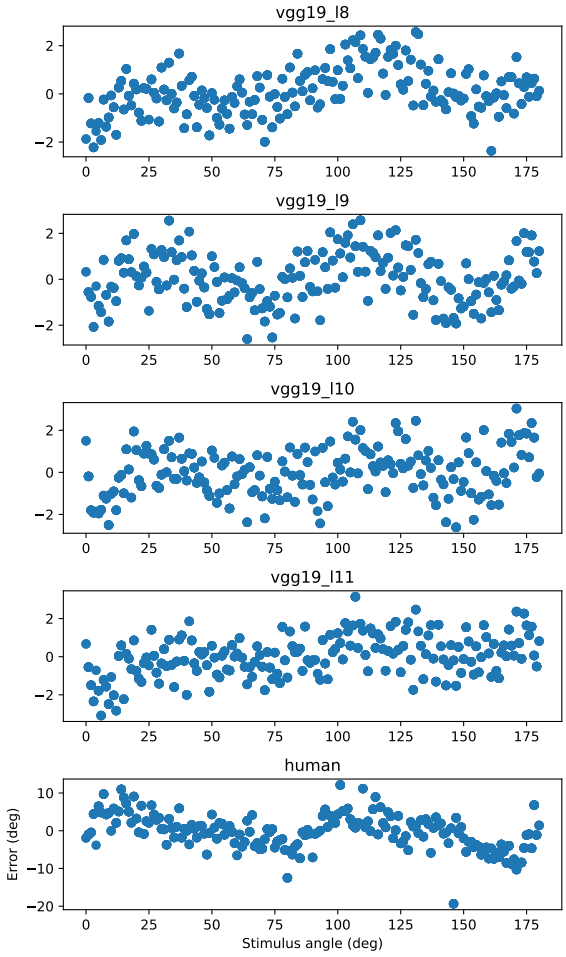

Fig. S4 Mean human and model errors in orientation memory task, for selected early layers of VGG-19.

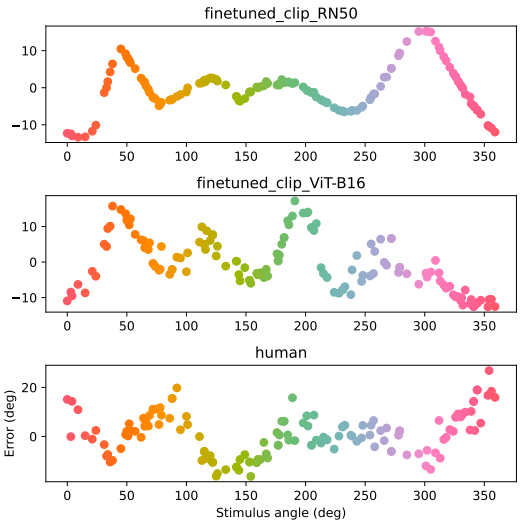

Fig. S5 Human and fine-tuned CLIP model response biases in color memory task.