Benchmarking Out-of-Distribution Generalization Capabilities of DNN-based Encoding Models for the Ventral Visual Cortex.

Spandan Madan
Harvard University

Will Xiao
Harvard Medical School

Mingran Cao
Francis Crick Institute

Hanspeter Pfister
Harvard University

Margaret Livingstone
Harvard Medical School

Gabriel Kreiman
Harvard Medical School

Abstract

We characterized the generalization capabilities of deep neural network encoding models when predicting neuronal responses from the visual cortex to flashed images. We collected *MacaqueITBench*, a large-scale dataset of neuronal population responses from the macaque inferior temporal (IT) cortex to over $300,000$ images, comprising $8,233$ unique natural images presented to seven monkeys over 109 sessions. Using *MacaqueITBench*, we investigated the impact of distribution shifts on models predicting neuronal activity by dividing the images into Out-Of-Distribution (OOD) train and test splits. The OOD splits included variations in image contrast, hue, intensity, temperature, and saturation. Compared to the performance on in-distribution test images—the conventional way in which these models have been evaluated—models performed worse at predicting neuronal responses to out-of-distribution images, retaining as little as 20% of the performance on in-distribution test images. Additionally, the relative ranking of different models in terms of their ability to predict neuronal responses changed drastically across OOD shifts. The generalization performance under OOD shifts can be well accounted by a simple image similarity metric—the cosine distance between image representations extracted from a pre-trained object recognition model is a strong predictor of neuronal predictivity under different distribution shifts. The dataset of images, neuronal firing rate recordings, and computational benchmarks are hosted publicly at: MacaqueITBench Link.

1 Introduction

Deep Neural Networks (DNNs) for vision have internal representations that purportedly share similarities with neural representations in the primate ventral visual cortex stream [2, 3]. Such correlations between the representations in artificial and biological neural networks allow for models that use image representations extracted from a pre-trained DNN (e.g., ResNet [4]) to predict neuronal firing rates [5] (Fig. 1(a)). However, DNNs are known to struggle with generalization under distribution shifts such as Out-of-Distribution (OOD) viewpoints [6, 7, 8], materials and lighting [9, 10], and noise [11, 12]. The problem of OOD generalization constitutes a key standing challenge in computer vision. Here we investigate whether this difficulty in generalization also affects models of the visual cortex that rely on a DNN to extract image representations.

We hypothesize that, even within an image set where DNN-based models predict neural responses well under random splits across images, specific train-test splits with distribution shifts will impair model performance, proportional to the size of distribution shift. To test this hypothesis, we collected

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

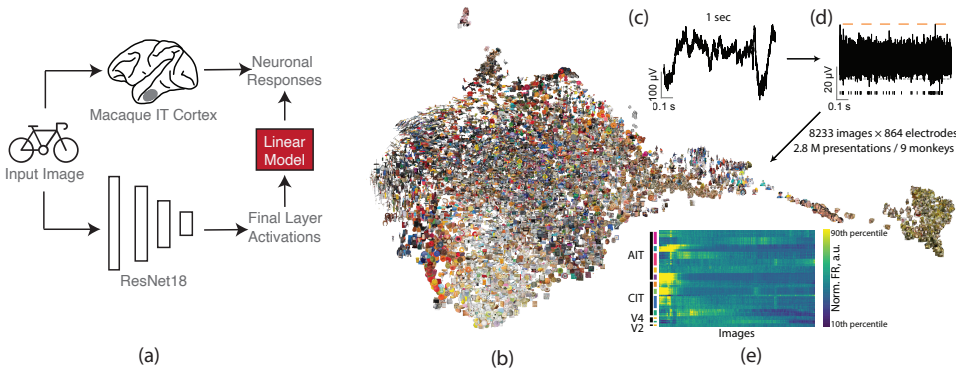

Figure 1: *Modeling the ventral visual cortex with MacaqueITBench*. (a) DNN-Based models of the visual cortex employ a linear model to map image features extracted from pre-trained DNNs (e.g., ResNet18) to neuronal responses collected from the macaque cortex (e.g., IT cortex). (b) UMAP [1] visualization of the representation of images by the neuronal pseudo-population. Nearby images have more similar population responses. (c) An example one-second segment of the raw wideband signals recorded on an electrode. (d), The wideband signals were highpass filtered, and threshold-crossing events below a voltage value (dashed black line) were counted as multiunit spikes (lower vertical ticks). The orange horizontal bars indicate image presentation periods. (e) The heatmap shows the neuronal response matrix. Each row indicates the responses from an electrode, pooled across sessions. The columns correspond to images, sorted by the reverse UMAP horizontal order. The vertical bars to the left of the heatmap denote the recorded areas (black lines) and monkeys (colored lines).

MacaqueITBench, a large-scale dataset of responses to natural images by neurons in the macaque ventral visual pathway. The dataset comprises neurons in V2, V4, Central IT (CIT), and Anterior IT (AIT) (primarily CIT and AIT) and includes responses to over $300,000$ images ($8,233$ unique images presented to seven monkeys over 109 sessions), as illustrated in Fig. 1(b).

Using *MacaqueITBench*, we investigated the impact of distribution shifts on the neural predictivity of DNN-based models of the visual cortex. We systematically constructed various OOD distribution shifts, some of which are schematized in Fig. 2. Foreshadowing, our main finding is that distribution shifts in even low-level image attributes break DNN-based models of the visual cortex.Furthermore, the relative ranking of different models, usually considered as a key metric to compare models, is *not* conserved across distribution shifts. These observations highlight a fundamental problem in modern models of the ventral visual cortex—good predictions are limited to images similar to those in the training data distribution.

To explain the OOD model-performance drop, we built on theoretical work positing that generalization performance is closely correlated with the amount of distribution shift [13, 14]. While theoretical studies have examined simplistic, simulated data, we show that a suitable metric of the size of distribution shifts can account for the OOD generalization performance of neural-encoding models.

In summary, our main contributions are:

- We present *MacaqueITBench*, a large-scale dataset of neural population responses to over $300,000$ images spanning multiple areas of the primate ventral visual pathway. The recording included 640 electrodes (12 multi-electrode arrays) recorded in nine hemispheres of seven monkeys.
- We show that modern models of the visual cortex do not generalize well—simple distribution shifts can reduce neural predictivity to as low as 20% of in-distribution performance.
- We show that the ranking across models is not conserved across distribution shifts.
- We provide a simple metric of distribution shift size that captures neural predictivity changes under distribution shifts.

2 Related Work

2.1 DNN-based models of the ventral visual cortex

A touchstone for visual neuroscience is the ability to predict neuronal responses to *arbitrary* images. On this test, DNN-based models have emerged as state-of-the-art models, best explaining neuronal responses across species—mouse, macaque, and humans—and visual cortical areas—from the primary visual cortex (V1) to the high-level inferior temporal cortex (IT) (for review, see [15, 16, 17]). These DNN-based models have been evaluated using random cross-validation (e.g., [18]), which tests IID generalization typically within a rather homogeneous image set. OOD generalization in such models has been sparsely examined. One study compared model fit to neural responses on two image types [19]. Here we systematically vary the type and degree of OOD splits to assess generalization as a function of differences between training and test datasets.

2.2 Out-of-distribution generalization capabilities of DNNs

In computer vision, DNNs for object recognition have been documented to fail at generalizing across a wide range of distribution shifts. Such shifts include 2D rotations and shifts [20, 21], commonly occurring blur or noise patterns [11, 22, 23, 24], and real-world changes in scene lighting [25, 26, 27], viewpoints [7, 28, 29, 30, 25, 8, 31], geometric modifications [32, 33, 34], color changes [35, 36], and scene context [37, 38].

Several benchmarks have been proposed to capture these distribution shifts systematically. For handwritten digit recognition, datasets like MNIST [39], MNIST-M [40], SVHN [41], and SYN [40] differ in features such as font, color, and background. For object recognition, domain shifts in the form image style have been captured in datasets like VLCS [42], Office-31 [43], and PACS [44]. Similarly, the Terra-Incognita [26] dataset has captured domain shift between the same scene viewed under daylight and night conditions. Recently, the WILDS benchmark [45] was introduced to tackle distribution shifts encountered in real-world scenarios, featuring datasets in diverse fields like animal and molecule classification. Of note, there has also been some work using controlled synthetic data to generate systematic benchmarks for generalization. These include the Biased-Cars dataset [7], the human visual diet dataset [9], and the Photorealistic Unreal Graphics (PUG) datasets created using Unreal Engine [46].

There have been three broad approaches to address the lack of OOD generalization in DNNs: first, modifying the learning paradigm including modifying the architecture or loss function to enforce invariant representations [47, 48, 49, 50, 51], or using ensemble and meta-learning [52, 53, 54]; second, modifying the training data using data augmentation [55, 56, 57, 58], or by increasing data diversity [23, 59, 60, 61, 62, 9, 7, 63, 6]; third, scaling data up to beyond billions of data points [64, 65, 66]. Despite these efforts, OOD generalization remains an unsolved problem in computer vision.

2.3 Out-of-distribution generalization models of the visual cortex

Despite extensive machine learning research on the topic, OOD generalization has received limited attention in the context of modeling biological neuronal responses. The ability to generalize is especially relevant to ventral visual cortex models due to acute limitations on the amount of available neuronal data. Given the time needed to present images (100s of ms per image), finite neuronal recording durations, and repeat presentations needed to combat neuronal stochasticity, it is currently infeasible to collect reliable neuronal responses to many more than 10k unique images. In this data-limited regime, most images of interest will remain out-of-domain even if we had foreknowledge of the test distribution (e.g., 10k unique images equal 10 images per category for the 1,000 ImageNet categories, insufficient to cover the distribution). The limited OOD generalization ability of current neuronal encoding models restricts their scientific utility, for example in accuratelypredicting maximally activating images for neurons (Fig. S1). This work contributes by borrowing from machine learning research on OOD generalization to shed light on computational neuroscience models.

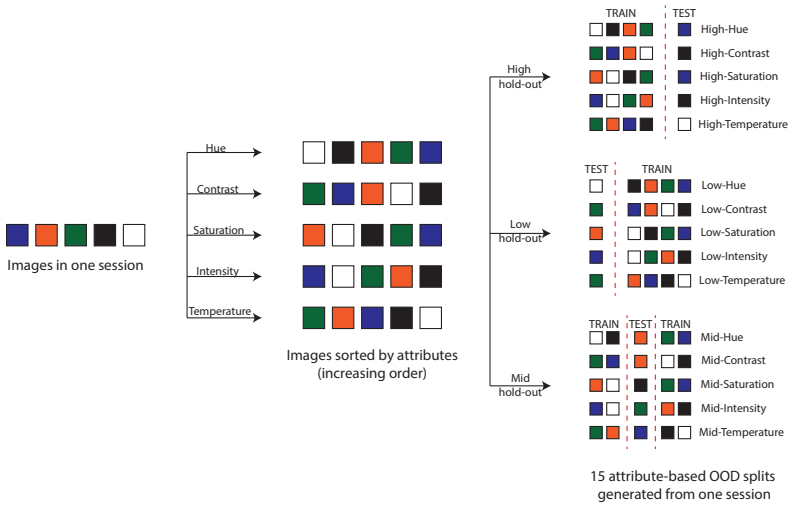

Figure 2: *Constructing multiple attribute-based OOD splits*. For each of our 109 sessions, we constructed 15 different attribute-based OOD splits. These splits correspond to 3 hold-out strategies (*high, low, mid*) for each of 5 image-computable attributes (hue, contrast, saturation, intensity, temperature). For each attribute (e.g., hue), we compute the attribute value for each image in the session. For the *high* hold-out strategy, all images with the attribute value above a percentile cut-off serve as the OOD test set with the remaining serving as the train set. Analogously for the *low* hold-out splits, images below a percentile cut-off serve as the test set with the remaining serving as the train set. For *mid* hold-out splits, images within the middle percentiles serve as the test set.

3 MacaqueITBench: Image-response recordings from the ventral stream

We collected a large-scale dataset of neuronal population responses to over $300,000$ images across sessions, comprising $8,233$ unique natural images presented to seven monkeys over 109 sessions. In each session, a monkey maintained fixation while images were rapidly presented in random order. Each presentation was 83 milliseconds; with 83–150 milliseconds between presentations.

The images were derived from published image sets [67] and photos taken in the lab and contained pictures of common objects, people, and other animals including monkeys (Fig. 1(b)). Image thumbnails are shown in Fig. 1(b)); sample images are provided in Fig. S2. Images belonged to over 300 semantic categories annotated by hand. A full list of categories can be found in Table S1. The large number and diversity of images allowed us to construct various OOD splits.

4 Constructing out-of-distribution data splits

We build on past work studying generalization under systematic distribution shifts [7, 9, 47, 11], and define the training and test distributions parametrically using image attributes. Using these parametric data distributions, we construct three kinds of train-test splits:

InDistribution (InD) splits: For each session, we created one In-Distribution (InD) split to compare with OOD generalization performance. We sampled 25% of the images at random, and held these out as the InD test set, with the remaining images serving as the training set.

Attribute-based OOD splits: We describe here OOD splits based on image contrast; splits based on the other image attributes were constructed analogously. For each session, we computed the contrast value for each image. Then, one of three strategies were employed (Fig. 2):

- *High hold-out*: The 75^{th} percentile of contrast values served as the cut-off. Images with contrast above the cut-off formed the test set. Remaining images formed the training set.
- *Low hold-out*: The 25^{th} percentile served as the cut-off. All images below this cut-off served as the held-out test set. The remaining images served as the training set.

- *Mid hold-out*: Images with contrast values between the 37.5^{th} and 62.5^{th} percentile served as the held-out test set. The remaining images formed the training set.

Cosine Distance-based splits: To investigate the relationship between the size of distribution shift and neuronal response predictivity, we constructed 3 additional test splits. We first extracted the features for every image from the pre-final layer of a pre-trained ResNet18. A random image was picked to be the seed, and all images in the session were sorted in order of increasing cosine distance between the ResNet extracted features of the images and the seed. The sorted images were then divided into three chunks based on percentile cut-offs. The first chunk corresponded to the bottom 80^{th} percentile which served as the Training + In-Distribution Test split. A random subset of this first chunk was held out to form the In-Distribution test split, with the remaining serving as the training set. The second chunk included images in the 90^{th} to 95^{th} percentile, which were held-out as the *Near-OOD* test split. Finally, the third chunk corresponded to images above the 95^{th} percentile. These were held-out as the *Far-OOD* split. To ensure a gap between the train and test distributions, we did not consider images between the 80^{th} and the 90^{th} percentile. Note that the number of images in the In-Distribution test split was kept the same number of images as the Near-OOD split.

5 Quantifying distribution shifts

We present a unified framework for measuring distribution shifts over the parametric OOD train-test splits presented in Sec. 4.

5.1 Representations for training and testing data-splits

Let $D_T = \{i_1^T, i_2^T, ..., i_N^T\}$ denote a train split of N images, and let $D_t = \{i_1^t, i_2^t, ..., i_n^t\}$ denote the corresponding test split of n images. $\mathcal{R}(.)$ is a representation function that provides a vector representation for an image. The train and test images thus correspond to $\mathcal{R}(D_T) = \{\mathcal{R}(i_1^T), \mathcal{R}(i_2^T), \ldots, \mathcal{R}(i_N^T)\}$ and $\mathcal{R}(D_t) = \{\mathcal{R}(i_1^t), \mathcal{R}(i_2^t), \ldots, \mathcal{R}(i_n^t)\}$, respectively.

We analyzed representations $\mathcal{R}(i_j)$ formed by the features extracted for an image i_j by a pre-trained DNN. We evaluated 8 different DNN architectures, and multiple layers for every architecture. The equations below are agnostic to the architecture and the layer used. Other alternatives could include using HOG [68] or GIST [69] image features, or the vectorized pixel values of the image.

5.2 Defining distances over different datasets

To compute the shift between $\mathcal{R}(D_T)$ and $\mathcal{R}(D_t)$, we compared three distance metrics:

Maximum Mean Discrepancy (D_{MMD}): The MMD distance between the two datasets can be computed as

$$D^2_{\text{MMD}}(D_T, D_t) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} K(\mathcal{R}(i_j^T), \mathcal{R}(i_k^T)) + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} K(\mathcal{R}(i_j^t), \mathcal{R}(i_k^t)) - \frac{2}{Nn}\sum_{j=1}^{N}\sum_{k=1}^{n} K(\mathcal{R}(i_j^T), \mathcal{R}(i_k^t))$$

Here, $K(\mathcal{R}(i_j^T), \mathcal{R}(i_k^t))$ is a kernel distance between the representations of images i_j^T and i_k^t. For our experiments, we used a Gaussian RBF kernel.

Covariate-Shift (D_{Cov}): Let $P_T(X)$ and $P_t(X)$ denote the distributions of the train and test input variables (i.e., image representations), and let $P(Y|X)$ denote the conditional distribution of the output (i.e., neuronal responses) given the input. A covariate shift exists if $P_T(X) \neq P_t(X)$ but $P_T(Y|X) = P_t(Y|X)$. D_{Cov} can be computed by training a binary classifier to classify if data comes from the training or the testing dataset. We denote the accuracy of this classifier as $a_{T,t}$ and measure the covariate shift as:

$$D_{\text{Cov}}(D_T, D_t) = 2 \times (0.5 - a_{T,t}).$$

Closest Cosine Distance (D_{CCD}): For every image in the test set, we find its distance to the closest training image, and compute the mean of this distance over all test images. For brevity, we will refer to this as *Closest Cosine Distance*. Let $i_k^T \in D_T$ denote the closest training image to test image $i_j^t \in D_t$ as measured by the cosine distance $D_{\cos}(\mathcal{R}(i_j^T), \mathcal{R}(i_k^t))$. The distance $D_{\cos}$ between two vectors u and v is given by:

$$D_{\cos}(u, v) = 1 - \frac{u \cdot v}{\|u\|\|v\|}$$

The average distance to the closest training image is:

$$D_{\text{CCD}} = \frac{1}{n} \sum_{j=1}^{n} \min_{k \in \{1,2,\ldots,N\}} D_{\cos}(\mathcal{R}(i_j^T), \mathcal{R}(i_k^t))$$

6 Model training and evaluation

As depicted in Fig. 1(a), we used a linear model to map pre-trained model activations to neuronal firing rates from the IT cortex (Fig. 1(a)). The linear model was learned using ridge regression. We used only pre-trained DNNs, not DNNs fine-tuned for our analysis.

For feature extraction, we investigated 8 DNN architectures and 2 layers for each architecture. The DNNs include supervised models trained on ImageNet (ResNet-18 [4], ViT [70]), self-supervised models trained on billion-scale data with self-supervised and weakly supervised learning (ResNet18_swsl [64], ResNext101_32x16d_swsl [64], ResNet-50_ssl [64]), Noisy student with EfficientNet [71], self-supervised learning over billions of tokens (DinoV2 [66]), and the multi-modal vision-language model CLIP [65]. The exact layer used for feature extraction for each model is provided in the supplement in Sec. D.

A linear encoding model was fit for the trial-averaged responses of each neuron in a session. The results are presented as the mean and S.E.M. across 109 sessions (7 monkeys); each session's results is the median across neurons. The model fit per neuron was quantified as the ceiling-normalized, squared Pearson's correlation, $r^2_{\text{pred}}/r^2_{\text{cons}}$ following convention [18, 72] and related to the explained variance, R^2. The ceiling r_{cons} of a neuron was calculated as its response correlation between split-half trials, across images, with Spearman-Brown correction (because model fitting used all trials per image). The model fit r_{pred} was the correlation across test images between neuronal responses and model predictions. All experiments were conducted on a compute cluster with 300 nodes, 48 cores per node with CPU machines running Rocky Linux release 8.9 (Green Obsidian).

7 Results

7.1 Neural prediticivity drops under distribution shifts

DNN-based encoding models become worse at predicting neuronal responses under simple shifts in the image distribution. To demonstrate this, we report the ratio of neural predictivity between OOD and In-Distribution test splits (r^2_{ood}/r^2_{ind}). A ratio of 1 would indicate that models generalize equally well to InD and OOD test images (horizontal dashed line; Fig. 3a). In contrast, the OOD/InD performance ratios are substantially lower than 1. For instance, the black bar in Fig. 3a shows that the model's neural predictivity was 0.33 on *high*-hue OOD images (constructed using the *high hold-out* strategy in Sec. 4) compared to images with InD hue. Models show a similar lack of OOD generalization to OOD images with regard to saturation (red bar), intensity (green bar), temperature (blue bar), and contrast (gray bar). This performance drop was observed for all eight DNNs tested (Fig. 3b-h) and ranged from a best-case ratio of 0.66 for the CLIP model generalizing to *high*-temperature OOD images to a worst-case ratio of 0.2 for the ViT model generalizing to *high*-saturation OOD images.

The lack of OOD generalization by neuron encoding models extended to models based on intermediate DNN layers, not just the penultimate layer. For all eight models, using activations extracted from

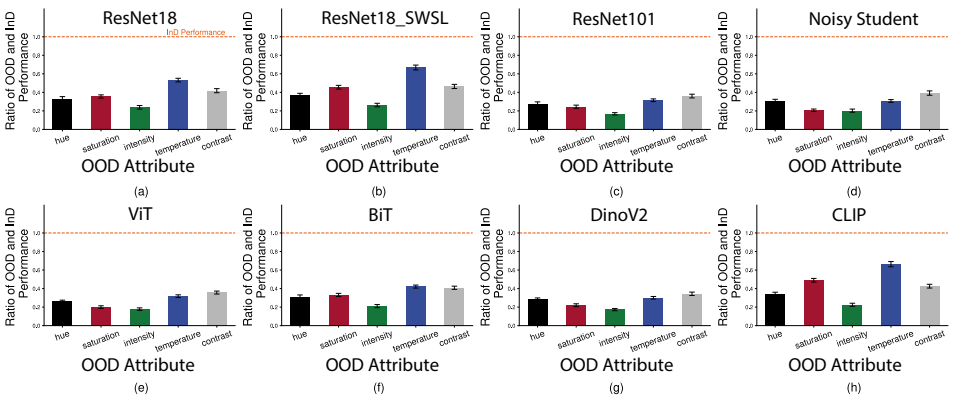

Figure 3: *Neuronal response predictivity drops under distribution shifts.* The y-axis shows the ratio of the neuronal response predictivity for out-of-distribution (OOD) images to in-distribution (InD) test images. A ratio of 1 would indicate no drop in performance. Each panel (a-h) shows a different architecture used for extracting image features. Each bar in a panels corresponds to a different OOD split constructed by using the *high* hold-out strategy across 5 different attributes (hue, saturation, saturation, intensity, temperature, and contrast). For all architectures and OOD splits, models fail to generalize well to OOD samples and are significantly and substantially below the 1.0 horizontal line. Image features were extracted from the pre-final layer for all architectures. Error bars denote standard deviation.

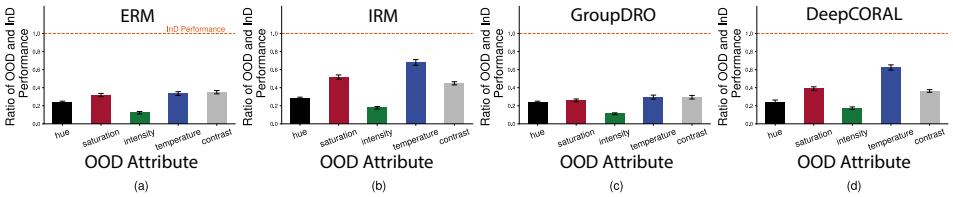

Figure 4: *Neuronal response predictivity drops for algorithms specifically designed to tackle OOD generalization as well.* Neuronal response predictivity is reported on OOD test splits constructed using the *high* hold-out strategy (using the same format as in Fig. 3). Generalization performance is well below 1.0 across image-computable attributes and four algorithms designed for OOD generalization presented in past literature [73, 45]. Specifically the four panels respectively show results for a ResNet50 model trained with empirical risk minimization (ERM) [74], Invariant Risk Minimization (IRM) [47], GroupDRO [75], and DeepCORAL [76] algorithms. None of these models generalizes well to OOD splits constructed with the *high* hold-out strategy despite being designed specifically for OOD generalization.

intermediate layers (layer names shown in Fig. 5), OOD performance remained substantially lower than InD performance (Fig. 5; tabular form in Sec. E).

Our findings extend to specialized Domain Generalization architectures designed to be more robust to distribution shifts (Fig. 4). For all specialized architectures and image attributes, the ratio of OOD and In-Distribution performance was significantly below 1.0, confirming a sharp drop in neural predictivity under distribution shifts.

OOD model performance was consistently lower than InD performance across hold-out strategies. Fig. 6 shows the OOD/InD model performance ratio for OOD splits constructed using the *low* hold-out strategy described in Sec. 4. As before, the ratio is consistently below 1.0, which confirms a severe drop in neural predictivity under distribution shifts. This finding also held true for Domain Generalization architectures tested with the low hold-out strategy, and for additional OOD shifts constructed using the mid hold-out strategy as shown in supplementary Sec. F.

So far, we have presented results with OOD shifts constructed by holding-out images within specific ranges of image attributes including hue, contrast, saturation, and color temperature. These splits model realistic scenarios where a model must generalize to, for example, novel weather and lighting

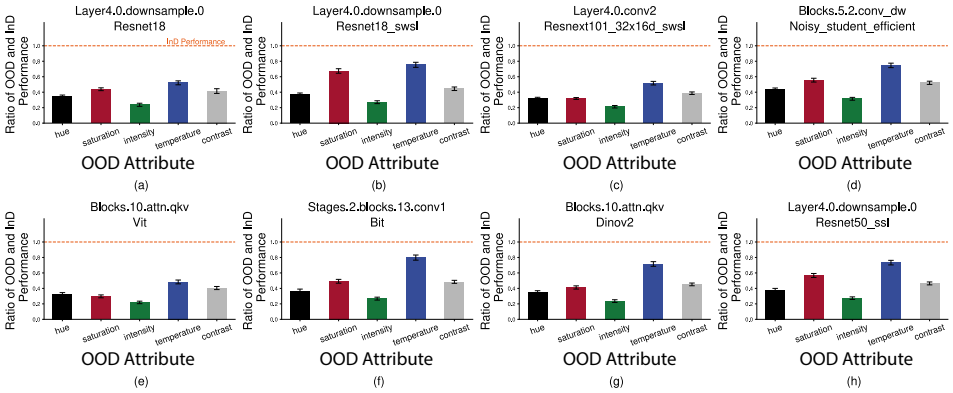

Figure 5: *Neuronal response predictivity drops under OOD testing for different model layers as well.* Neuronal response predictivity on OOD samples is reported for multiple DNN architectures across multiple different layers. Layer name is mentioned alongside architecture in all panels (a-h). All OOD splits reported here were constructed using the *high* hold-out strategy. For all architectures, layers, and OOD splits, models fail to generalize well to OOD samples and are significantly below the 1.0 horizontal line.

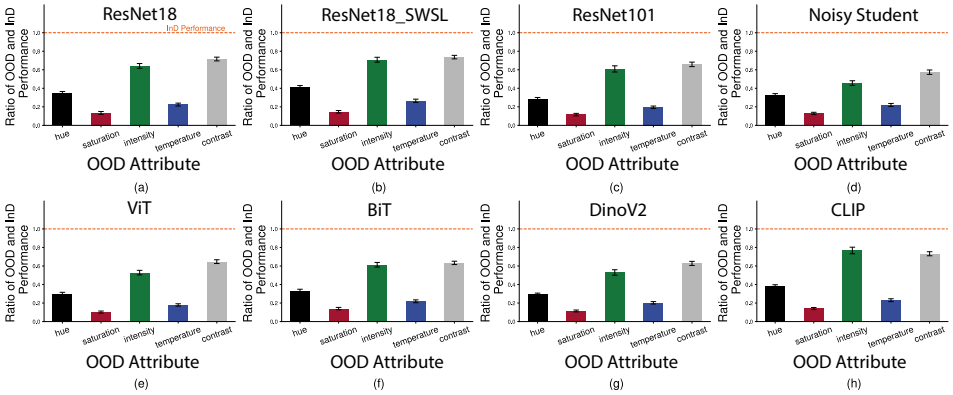

Figure 6: *Neuronal response predictivity drops for the low hold-out strategy as well.* Neuronal response predictivity is reported on OOD test splits constructed using the *low* hold-out strategy. Across all DNN architectures and image-computable attributes, performance is below 1.0 for all panels (a-h). Thus, models do not generalize well to OOD splits constructed with the *low* hold-out strategy either.

conditions. To validate our results in other scenarios of domain shift, we examined two additional splits based on categories. First, we held out a random subset of categories as a test set and trained on the remaining categories. Second, we held out Food-related categories as a test set and trained on non-Food categories. For both these OOD splits, all models failed to generalize—ratio of OOD and In-Distribution neural predictivity was well below 1.0, in line with previous results (Table 1).

Finally, we also compared neural predictivity on an established, in-distribution benchmark (Brain-Score) and on our OOD benchmark. These results, presented below, further support our finding that current DNNs are insufficient models of the Ventral Visual Cortex—models that perform better on the in-distribution BrainScore benchmark did not perform better on OOD shifts (all Spearman rank correlations $p > 0.05$). Combined, these results showcase a problem for current DNN-based models of the visual cortex—despite their ability to predict neural responses to in-distribution test images, the models generalize poorly under distribution shifts even in low-level image attributes.

OOD Split	CLIP	DinoV2	Noisy Student	ResNet18	ResNet50 SSL	ResNext 101
Random	0.80 ± 0.01	0.77 ± 0.02	0.78 ± 0.02	0.83 ± 0.01	0.80 ± 0.01	0.71 ± 0.01
Food	0.23 ± 0.02	0.19 ± 0.01	0.20 ± 0.02	0.23 ± 0.02	0.20 ± 0.02	0.16 ± 0.01

Table 1: *Similar conclusions are reached with naturalistic OOD splits*. This table shows the ratio of neuronal response predictivity for OOD samples to in-distribution samples, which is below 1.0 for all architectures. The Random split was constructed by holding out a random subset of categories as the test set, and training the model on remaining categories. For the Food split, all food-related categories served as the tet set, and models were trained on the non food-related categories.

Model	Brain Score	Hue	Saturation	Intensity	Temp	Contrast	Average
BiT	0.33	0.31	0.33	0.21	0.42	0.41	0.33
ResNet18	0.35	0.33	0.36	**0.24**	0.53	0.42	0.37
CLIP	0.47	**0.34**	**0.49**	0.23	**0.66**	**0.43**	**0.43**
ResNext101	0.49	0.28	0.25	0.17	0.32	0.36	0.32
ViT	**0.51**	0.26	0.20	0.18	0.32	0.36	0.30

Table 2: *BrainScore vs MacaqueITBench*. We compare models of the visual cortex in and outside the training data distribution. BrainScore [18] provides a ranking of models based on in-distribution performance. However, models that perform better on the in-distribution BrainScore benchmark did not perform better on OOD shifts (all Spearman rank correlations p > 0.05). Best performing model has been presented in bold.

7.2 The distance between train and test distributions explains generalization performance

The results above raise a natural question—when and how do models of the ventral visual cortex fail to generalize under distribution shifts? Theoretical work has related OOD generalization to the amount of distribution shift [13, 14]. Here we apply this theoretical framework to characterize generalization in DNN models of the brain.

Intuitively, model generalization should be worse for train-test splits under larger distribution shifts. We tested this intuition by constructing splits with different levels of distribution shifts—InD, Near OOD, and Far OOD. As described in Sec. 4, images in every session were sorted based on cosine distance and split into three chunks. The first chunk formed the training set and the In-Distribution test set, while the second and third chunks formed the Near OOD and Far OOD test sets. As hypothesized, model performance decreased significantly from In-Distribution to Near OOD, then Far OOD test sets (Fig. 7(a); two-sided t-test, $p < 0.01$).

The size of the distribution shift predicted the OOD model performance drop across individual data splits (Fig. 7(b)). The distribution shift between each pair of train and OOD test distributions was quantified with the *Closest Cosine Distance* (D_{CCD}; described in Sec. 5). The D_{CCD} strongly correlated with the OOD model performance drop (Spearman correlation $\rho = -0.49$).

The distribution shift (D_{CCD}) calculated from ResNet features also explained OOD performance for attribute-based splits (Fig. 7(c). Across all image attributes (hue, saturation, temperature, contrast, intensity) and hold-out strategies (*low, high, mid*) used to create OOD splits, D_{CCD} correlated with OOD model performance drop (Spearman correlation $\rho = -0.45$). Compared to two other popular measures of the sizes of distribution shifts (MMD, D_{MMD}[77] and Covariate-Shift, D_{Cov} [78]; Sec. 5), D_{CCD} best predicted OOD model performance (Fig. 7(d)).

8 Conclusions

These results reveal a deep problem in modern models of the visual cortex: good prediction is limited to the training image distribution. Simple distribution shifts break DNN models of the visual cortex,

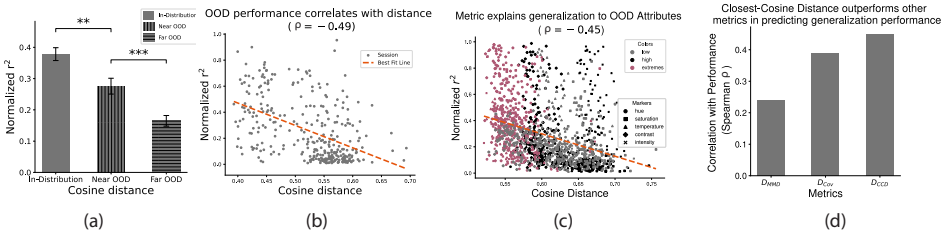

Figure 7: *Closest-Cosine Distance metric well explains performance across all attribute-based OOD splits.* (a) Neural predictivity on distance-based splits. Models performed best on the In-Distribution (InD) test split, dropping in performance from InD to Near OOD test set and from Near OOD to Far OOD (both $p < 0.01$, two-sided t-test). This suggests a relationship between the extent of distribution shift and generalization performance. (b) OOD performance can be well-explained by the distribution shift. For all 109 sessions, the plot shows performance on the InD, Near-OOD, and Far-OOD with the corresponding distribution shift measured using the Closest-Cosine Distance metric (D_{CCD}). Performance and D_{CCD} have a Spearman correlation of $-0.49 (p < 0.001)$. (c) Scatter plot of neural predictivity and the corresponding distribution shift (D_{CCD}) across all 15 attribute-based OOD splits for all 109 sessions. Generalization performance and the proposed distance metric have a Spearman correlation of $-0.45 (p < 0.001)$ (d) Comparing different distance metrics w.r.t. their correlation with OOD performance. The proposed Closest-Cosine Distance has the highest correlation with neural predictivity, outperforming both MMD (D_{MMD}) and Covariate-Shift (D_{Cov}).

consistent with broader findings that the underlying DNNs are brittle to OOD shifts. Going one step further, we introduce an image-computable metric that significantly predicts the generalization performance of models under distribution shifts. This metric can help investigators gauge how well a neural model fit on one dataset may generalize to novel images.

Our findings underline an important limitation of AI models for Neuroscience. Fields like Computer Vision have responded to the issue of distribution shifts by collecting progressive larger datasets, hoping models will learn to generalize to most images [79, 80, 81, 82] at the billion-image scale. However, it is infeasible to achieve the same scale in neuroscience—the time needed to present a billion images is already a formidable challenge, not to mention the resource intensiveness of data collection. We hope our characterization of when and how modern models of the visual cortex fail out-of-domain will motivate the development of data-efficient ways to improve DNN generalization.

9 Limitations

In this work, we have explored the impact of OOD samples on DNN-based models of the visual cortex. Our analyses have two main limitations that we hope future research can address. First, we did not fine-tune the DNNs on neural data. It is possible that training these models on the specific images and/or neural data can help improve generalization. Second, we did not explore the contributions of the images being OOD for the underlying pre-trained DNNs, as we only fit the linear encoding models on train set images and neural data. Because our images were naturalistic, it is plausible that they belonged to the training distribution of the pre-trained models we used, some of which (e.g., CLIP) having hundreds of millions of images. An interesting future direction will be to examine how the model performance is affected by using out-of-distribution images for the pre-trained DNNs. These images could include those from ImageNet-P, ImageNet-C [11], and evolved images [3].

10 Acknowledgments

This research was partially supported by NSF grant IIS-1901030, NSF grant CCF-1231216, NIH grant R01EY026025, and NIH grant R01HD104969. We thank Pranav Misra, Fenil Doshi, Thomas Serre, and Elisa Pavarino for insightful discussions, and Harshika Bisht for design feedback on the figures. Author M.L. took the photos for the subset of images collected in the lab.

NeurIPS paper checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction state the main claims, approach and the experiments support the claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: They are provided in the conclusions section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We have no Proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details are provided alongside code and data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and code are provided and are free for anyone to use.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all information are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we used two-sided t-tests for statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, details are provided in experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and reviewed the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There are no societal imapct of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work raises no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No such assets were used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Dataset comes with details on how to use it.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

References

[1] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[2] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[3] Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[6] Spandan Madan, Tomotake Sasaki, Hanspeter Pfister, Tzu-Mao Li, and Xavier Boix. Adversarial examples within the training distribution: A widespread challenge, 2023.

[7] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153, 2022.

[8] Avi Cooper, Xavier Boix, Daniel Harari, Spandan Madan, Hanspeter Pfister, Tomotake Sasaki, and Pawan Sinha. To which out-of-distribution object orientations are dnns capable of generalizing? *arXiv preprint arXiv:2109.13445*, 2021.

[9] Spandan Madan, You Li, Mengmi Zhang, Hanspeter Pfister, and Gabriel Kreiman. Improving generalization by mimicking the human visual diet, 2024.

[10] Akira Sakai, Taro Sunagawa, Spandan Madan, Kanata Suzuki, Takashi Katoh, Hiromichi Kobashi, Hanspeter Pfister, Pawan Sinha, Xavier Boix, and Tomotake Sasaki. Three approaches to facilitate invariant neurons and generalization to out-of-distribution orientations and illuminations. *Neural Networks*, 155:119–143, 2022.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[12] Francesco Croce, Sylvestre-Alvise Rebuffi, Evan Shelhamer, and Sven Gowal. Seasoning model soups for robustness to adversarial and natural distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12313–12323, 2023.

[13] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. *Advances in Neural Information Processing Systems*, 34:12600–12612, 2021.

[14] Pratik Patil, Jin-Hong Du, and Ryan J Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv preprint arXiv:2404.01233*, 2024.

[15] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[16] Thomas Serre. Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, 5(Volume 5, 2019):399–426, 2019.

[17] Gabriel Kreiman. *Biological and Computer Vision*. Cambridge University Press, 2021.

[18] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[19] Yifei Ren and Pouya Bashivan. How well do models of visual cortex generalize to out of distribution samples? *PLOS Computational Biology*, 20(5):e1011145, 2024.

[20] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

[21] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021.

[22] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[24] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[25] Spandan Madan, Tomotake Sasaki, Tzu-Mao Li, Xavier Boix, and Hanspeter Pfister. Small in-distribution changes in 3d perspective and lighting fool both cnns and transformers. *arXiv preprint arXiv:2106.16198*, 2021.

[26] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[27] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. Adversarial relighting against face recognition. *arXiv preprint arXiv:2108.07920*, 2021.

[28] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[29] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651*, 2018.

[30] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4302–4311, 2019.

[31] Akira Sakai, Taro Sunagawa, Spandan Madan, Kanata Suzuki, Takashi Katoh, Hiromichi Kobashi, Hanspeter Pfister, Pawan Sinha, Xavier Boix, and Tomotake Sasaki. Three approaches to facilitate dnn generalization to objects in out-of-distribution orientations and illuminations: late-stopping, tuning batch normalization and invariance loss. *arXiv preprint arXiv:2111.00131*, 2021.

[32] Amir Belder, Gal Yefet, Ran Ben-Itzhak, and Ayellet Tal. Random walks for adversarial meshes. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

[33] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2019.

[34] Dawei Yang, Chaowei Xiao, Bo Li, Jia Deng, and Mingyan Liu. Realistic adversarial examples in 3d meshes. *arXiv preprint arXiv:1810.05206*, 2:2, 2018.

[35] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4773–4783, 2019.

[36] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160, 2020.

[37] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. *arXiv preprint arXiv:2104.02215*, 2021.

[38] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020.

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[41] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[42] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[43] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

[44] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[45] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[46] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[48] Sarah Erfani, Mahsa Baktashmotlagh, Masud Moshtaghi, Xuan Nguyen, Christopher Leckie, James Bailey, and Rao Kotagiri. Robust domain generalisation by enforcing distribution invariance. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 1455–1461. AAAI Press, 2016.

[49] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 301–318. Springer, 2020.

[50] Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763. IEEE, 2021.

[51] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[52] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[53] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[54] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[56] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[57] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020.

[58] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[59] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[60] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[61] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[62] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[63] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how do cnns generalize to out-of-distribution category-viewpoint combinations? *arXiv preprint arXiv:2007.08032*, 2020.

[64] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019.

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[67] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: general*, 139(3):558, 2010.

[68] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[69] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[70] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[71] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020.

[72] Will Xiao, Saloni Sharma, Gabriel Kreiman, and Margaret S Livingstone. Feature-selective responses in macaque visual cortex follow eye movements during natural vision. *Nature Neuroscience*, pages 1–10, 2024.

[73] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[74] V.N. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience publication. Wiley, 1998.

[75] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[76] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[77] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[78] Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[79] Dingshuo Chen, Yanqiao Zhu, Jieyu Zhang, Yuanqi Du, Zhixun Li, Qiang Liu, Shu Wu, and Liang Wang. Uncovering neural scaling laws in molecular representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[80] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

[81] Gabriele Prato, Simon Guiroy, Ethan Caballero, Irina Rish, and Sarath Chandar. Scaling laws for the out-of-distribution generalization of image classifiers. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, 2021.

[82] Hiroki Naganuma and Ryuichiro Hataya. An empirical investigation of pre-trained model selection for out-of-distribution generalization and calibration. *arXiv preprint arXiv:2307.08187*, 2023.

Supplementary Material

A OOD generalization specifically challenges encoding models of ventral visual cortex.

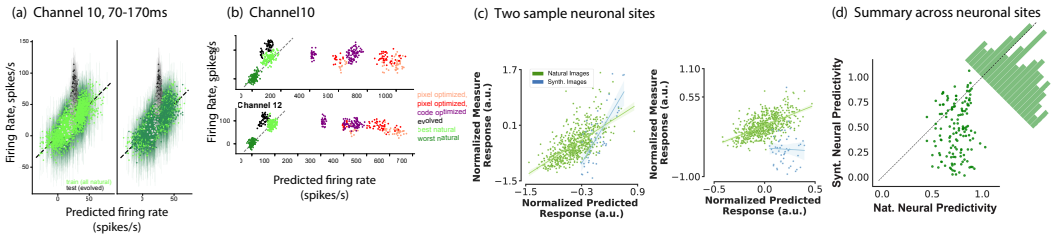

Figure S1: *OOD generalization specifically challenges encoding models of ventral visual cortex.* (a,b) Adapted from [3]. (a) Results from an example ventral visual neuron (area PIT) show that an encoding model fit on neuronal responses to a random half of 2550 natural images accurately predicted neuronal responses to the held-out half (right subplot). However, the model consistently underpredicted responses to GAN-synthesized images evolved for the neuron [3] and did not improve from training on all the natural images (left subplot). (b) Besides underpredicting evolved-image responses, encoding models fit on natural images overpredicted neuronal responses to the models' own activation-maximization stimuli. The two subplots correspond to two example neurons. The pink to purple colors indicate how the activation-maximization stimuli were regularized (no regularization, regularization by jitter, and regularization through the latent code space of a GAN). (c,d) Adapted from [19]. (c) Results from two example ventral visual neurons (area V4) show that encoding models of neuronal responses, fit on natural images, generalized reliably to held-out, InD natural images but unreliably to OOD synthetic images. (d) For most neurons examined, encoding models performed worse on OOD than InD images.

B List of semantic categories in MacaqueITBench

Table S1 reports a list of all semantic categories in MacaqueITBench. The $8,233$ images correspond to 376 categories.

Big	Big Animate	Bird	Butterfly
Cat	Dog	Face	Fish
Gabor	Glove	Hand	Mask
Misc	Non	Other	PPE
Print	Rodent	Starfish	Symbol
Toy	Turtle	abacus	accordian
aircompressor	airplane	ambulance	anchor
apple	axe	babushkadolls	babycarriage
babyplayard	babywalker	backgammon	backpack
bagel	ball	balloon	banana
barbiedoll	barrel	baseballbat	baseballcards
basket	bathsuit	battery	beaker
beanbagchair	bed	beermug	bell
bench	bike	bill	binoculars
birdcage	birdhouse	bones	bongo
bonzai	boot	boppypillow	bottle
bottleopener	bowl	bowlingpin	bowlofchips
bowtie	breadloaf	broom	bucket
bullet	bullhorn	button	cage
cake	calculator	camcorder	camera

candleholderwithcandle	candy	candybar	cane
carabiners	carfront	carseat	cashregister
cassettetape	ceilingfan	cellphone	chair
checkbook	cheese	cheesegrater	cherubstatue
chessboard	chocolate	christmasstocking	christmastreeornamantball
cigarettepack	circuitboard	clock	coatrack
coffeemaker	coffeemug	coffin	coin
collar	compass	computer	computermouse
cookie	cookingpan	cookpot	cooler
corkscrew	corset	cracker	crib
crossbow	crown	cupsaucer	curlingiron
cushion	decorativescreen	desk	doll
dollhouse	domino	donut	doorknob
doorknocker	doorwayarch	dresser	dumbbell
duster	dvdplayer	dynamite	earings
easteregg	eraser	exercise	extra
familiarObjects	fan	feather	filingcabinet
fireplace	fish hook	fishbowl	fishingpole
flag	flashlight	flask	fork
frame	fridge	frisbee	fruitparfait
gamehandheld	gamesboard	garbagetrash	gift
giftbow	glasses	globe	goggle
golfbag	golfball	gong	grapes
greenplant	grill	guitar	hairbrush
hairdryer	hammer	handbag	handgun
handheldvacuum	handkerchief	handmirror	hanger
hat	headband	headphone	helmet
highchair	hilighter	hookah	horseshoe
hotairballoon	hourglass	icecreamcones	iceskates
jack-o-lantern	jacket	juice	kayak
ketchupbottle	kettle	key	keyboard
keychain	knife	ladder	lamp
lantern	laptop	laudrybasket	lawnmower
leatherman	leaves	lei	licenseplate
lightbulb	lighter	lightswitch	lipstick
lock	log	loom	lunchbox
mailbox	makeupcompact	manorha	mathcompass
mattress	measuringtape	meat	microphone
microscope	microwave	motorcycle	mp3player
muffin	muffler	mushroom	musicstand
nailpolish	necklace	necktie	nest
nunchaku	objects	orientalplatesetting	orifan
pacifier	paintbrush	pants	pasta
patioloungechair	pda	pen	pencilsharpener
peppersonplate	perfumebottle	pezdispenser	phone
pie	pill	pillow	pipe
pitcher	pizza	plate	pokercard
powerstrip	printer	quilt	radio
razor	recordplayer	remotecontrol	reportfile
ring	ringbinder	roadsign	robot
rock	rollerskates	rollingpin	rosary
router	rug	saddle	saltpeppershake
sandwich	scale	scissors	scooter
scroll	scrunchie	seashell	seasponge
servingpiece	sewingmachine	shirt	shoe
short	shotglass	shovel	showercurtain

shredder	sink	sippycup	skateboard
slate	sleepingbag	slinky	snowglobe
soapdispenser	socks	sodacan	sofa
speakers	spicerack	spoolofstring	spoon
spraybottle	stamp	stapler	stool
stove	strainer	suit	suitcase
sushi	swissarmyknife	sword	tablesmall
tape	telescope	tennisracquet	tent
tire	toaster	toiletseat	tongs
toothbrush	toothpaste	toy	tractor
train	tray	tree	tricycle
trophy	trumpet	trunk	tupperware
tv	tweezer	typewriter	umbrella
vacuum	vase	videoGameController	wallsconce
washer	watch	waterbottle	watergun
waxseal	wheelbarrow	wheelchair	wig
windchime	window	wineglass	wineglassfull
woodboxsmall	yarn		

Table S1: *Images from MacaqueITBench.*

C Sample Images from MacaqueITBench

Fig. S2 shows sample images which were presented to Macaques to collect responses from the IT Cortex.

D Details on the layers used for feature extraction

For all models, we extracted features from the pre-final layer *i.e.,* the final (classification) layer was removed and features were extracted. For experiments building on features from intermediate layers, the following layers were used:

Model	**Intermediate Layer Name**
resnet50_ssl	layer4.0.downsample.0
resnet18_swsl	layer4.0.downsample.0
resnet18	layer4.0.downsample.0
resnext101_32x16d_swsl	layer4.0.conv2
noisy_student_efficient	blocks.5.2.conv_dw
resnext101_32x16d_swsl	layer4.0.conv2
vit	blocks.10.attn.qkv
dinov2	blocks.10.attn.qkv
bit	stages.2.blocks.13.conv1
clip	transformer.resblocks.10.attn

E Results presented in Tabular form

Model	Hue	Saturation	Intensity	Temperature	Contrast
resnet18	0.33 ± 0.02	0.36 ± 0.02	0.24 ± 0.02	0.53 ± 0.02	0.42 ± 0.02
resnet18_swsl	$\mathbf{0.37 \pm 0.02}$	0.46 ± 0.02	$\mathbf{0.26 \pm 0.02}$	$\mathbf{0.67 \pm 0.03}$	$\mathbf{0.46 \pm 0.02}$
resnext101	0.28 ± 0.02	0.25 ± 0.02	0.17 ± 0.01	0.32 ± 0.01	0.36 ± 0.02
noisy_student	0.31 ± 0.02	0.21 ± 0.01	0.20 ± 0.02	0.31 ± 0.01	0.39 ± 0.02
vit	0.26 ± 0.02	0.20 ± 0.01	0.18 ± 0.01	0.32 ± 0.01	0.36 ± 0.02
bit	0.31 ± 0.02	0.33 ± 0.02	0.21 ± 0.02	0.42 ± 0.02	0.41 ± 0.02
dinov2	0.28 ± 0.02	0.22 ± 0.02	0.17 ± 0.01	0.30 ± 0.01	0.34 ± 0.02
clip	0.34 ± 0.02	$\mathbf{0.49 \pm 0.02}$	0.23 ± 0.02	0.66 ± 0.03	0.43 ± 0.02

Table S2: *Data from Fig.3 reported in Table form.* Neural predictivity drops significantly for all models when tested with OOD samples. The ratio of neural predictivity for OOD samples to in-distribution samples is below 1.0 for all architectures. Best performing model for each attribute is bolded.

Model	Hue	Saturation	Intensity	Temperature	Contrast
resnet18	0.34 ± 0.02	0.44 ± 0.02	0.24 ± 0.02	0.52 ± 0.03	0.41 ± 0.03
resnet18_swsl	0.37 ± 0.02	$\mathbf{0.67 \pm 0.03}$	$\mathbf{0.27 \pm 0.02}$	0.75 ± 0.03	0.44 ± 0.02
resnext101	0.32 ± 0.02	0.32 ± 0.01	0.21 ± 0.02	0.52 ± 0.02	0.39 ± 0.02
noisy_student	$\mathbf{0.44 \pm 0.02}$	0.56 ± 0.02	0.31 ± 0.02	0.75 ± 0.03	$\mathbf{0.52 \pm 0.02}$
vit	0.33 ± 0.02	0.30 ± 0.02	0.22 ± 0.02	0.48 ± 0.02	0.40 ± 0.02
bit	0.37 ± 0.02	0.49 ± 0.02	0.27 ± 0.02	$\mathbf{0.80 \pm 0.03}$	0.48 ± 0.02
dinov2	0.35 ± 0.02	0.41 ± 0.02	0.24 ± 0.02	0.72 ± 0.03	0.45 ± 0.02
resnet50_ssl	0.38 ± 0.02	0.57 ± 0.03	0.27 ± 0.02	0.73 ± 0.03	0.46 ± 0.02

Table S3: *Data from Fig.4 reported in Table form*. Neural predictivity drops on OOD samples for features extracted from different model layers as well. Best model for each attribute is bolded.

Model	Hue	Saturation	Intensity	Temperature	Contrast
resnet18	0.34 ± 0.02	0.13 ± 0.02	0.64 ± 0.03	0.22 ± 0.02	0.72 ± 0.02
resnet18_swsl	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.15 \pm 0.01}$	0.71 ± 0.03	$\mathbf{0.27 \pm 0.02}$	$\mathbf{0.74 \pm 0.02}$
resnext101	0.28 ± 0.02	0.12 ± 0.01	0.61 ± 0.03	0.20 ± 0.01	0.66 ± 0.02
noisy_student	0.32 ± 0.02	0.13 ± 0.01	0.46 ± 0.02	0.22 ± 0.02	0.57 ± 0.02
vit	0.30 ± 0.02	0.10 ± 0.01	0.53 ± 0.02	0.18 ± 0.01	0.65 ± 0.02
bit	0.33 ± 0.02	0.14 ± 0.01	0.61 ± 0.02	0.22 ± 0.02	0.63 ± 0.02
dinov2	0.29 ± 0.01	0.11 ± 0.01	0.53 ± 0.03	0.20 ± 0.01	0.63 ± 0.02
clip	0.38 ± 0.02	0.14 ± 0.01	$\mathbf{0.77 \pm 0.04}$	0.23 ± 0.02	0.73 ± 0.02

Table S4: *Data from Fig.5 reported in Table form.* Neural predictivity drops on OOD samples for the *high* hold-out strategy as well. Best performing model for each attribute is bolded.

F Additional results with hold-out strategies

In the main paper, we presented results with two hold out strategies—high and low. Here, we present additional results with held out strategies. Firstly, we confirmed that specialized domain generalization architectures also struggle with the low-hold out strategy as shown in Fig. S3. Furthermore, we present results with the third hold-out strategy outlined in the paper. We refer to this as the Mid hold out strategy as samples between the 42.5 and the 67.5 percentile of every OOD attribute are held out as the test set. As shown in Fig. S4, across all architectures and OOD attributes, models suffer to generalize to OOD samples for the Mid hold out strategy.

G Additional results with intermediate layers

In the main paper we presented results for models trained with intermediate layers for the high hold out strategy. Here we provide additional results with models that use intermediate layers of DNNs as

feature extractors. In Fig. S5 and Fig. S6 we report results for the *low* and *mid* hold-out strategies respectively.

Figure S2: *Example images from MacaqueITBench.*

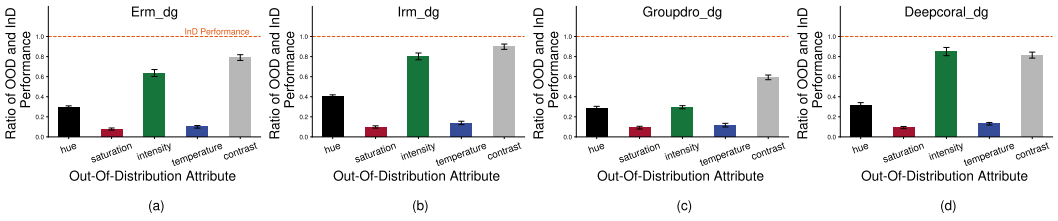

Figure S3: *Neural predictivity drops for specialized domain generalization approaches with the low hold-out strategy as well.* Neural predictivity is reported on OOD test splits constructed using the *low* hold-out strategy. Ratio of OOD and in-distribution neural predictivity is below 1.0 for all approaches and all image-computable attributes, panels (a-d). Thus, these approaches do not generalize well to OOD splits constructed with the *low* hold-out strategy as well.

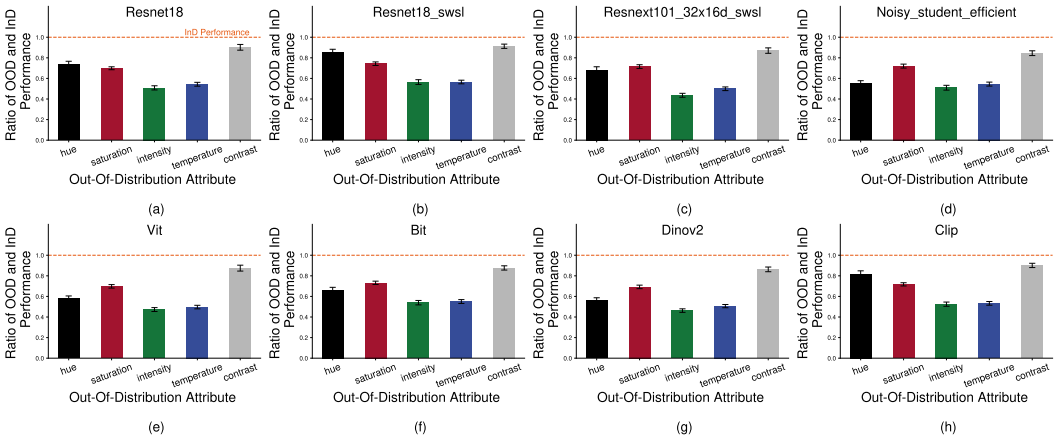

Figure S4: *Neural predictivity drops for Mid hold-out strategy as well.* For all architectures, across multiple OOD shifts, performance on OOD is worse than in-distribution samples for the Mid hold-out strategy as well.

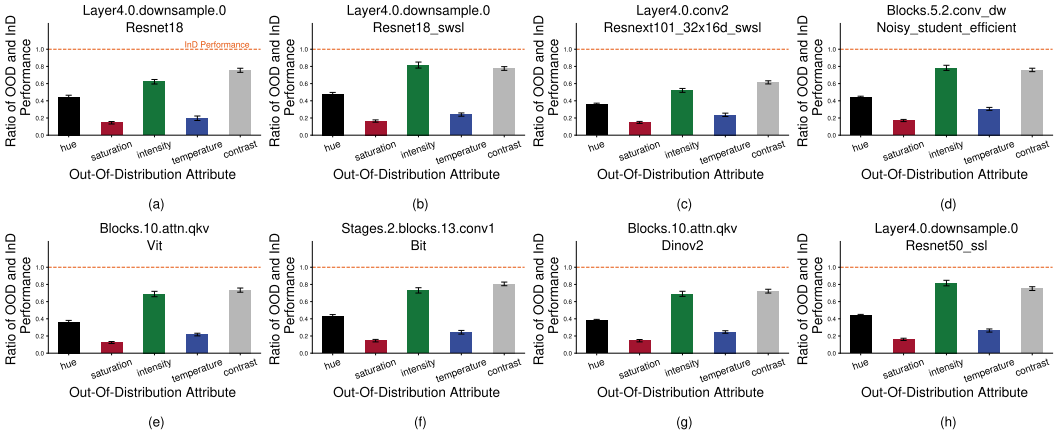

Figure S5: *Neural predictivity drops for low hold-out strategy for intermediate layer features as well.* For all architectures, across multiple OOD shifts, performance on OOD is worse than in-distribution samples for the low hold-out strategy for image features extracted from intermediate DNN layers as well.

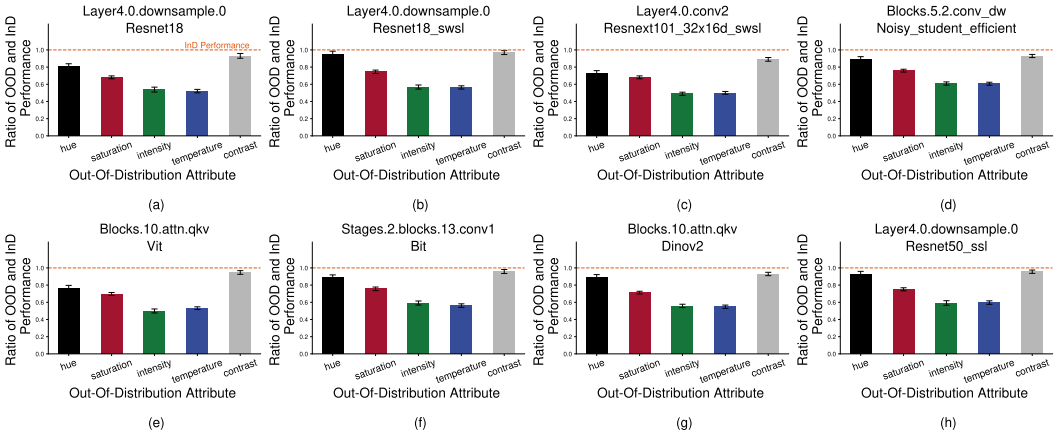

Figure S6: *Neural predictivity drops for mid hold-out strategy for intermediate layer features as well.* For all architectures, across multiple OOD shifts, performance on OOD is worse than in-distribution samples for the mid hold-out strategy for image features extracted from intermediate DNN layers as well.