Intriguing properties of generative classifiers

Priyank Jaini*
Google DeepMind

Kevin Clark
Google DeepMind

Robert Geirhos*
Google DeepMind

Abstract

What is the best paradigm to recognize objects—discriminative inference (fast but potentially prone to shortcut learning) or using a generative model (slow but potentially more robust)? We build on recent advances in generative modeling that turn text-to-image models into classifiers. This allows us to study their behavior and to compare them against discriminative models and human psychophysical data. We report four intriguing emergent properties of diffusion-based generative classifiers: they show a record-breaking human-like shape bias (99% for Imagen), near human-level out-of-distribution accuracy, state-of-the-art alignment with human classification errors, and they understand certain perceptual illusions. Our results indicate that while the current dominant paradigm for modeling human object recognition is discriminative inference, zero-shot generative models approximate human object recognition data surprisingly well.

1. Introduction

Many existing classifiers perform well on data similar to the training distribution, but struggle on out-of-distribution images. For instance, a cow may be correctly recognized when photographed in a typical grassy landscape, but is not correctly identified when photographed on a beach [1]. In contrast to many *discriminatively* trained models, *generative* text-to-image models appear to have acquired a detailed understanding of objects: they have no trouble generating cows on beaches or dog houses made of sushi [15]. This raises the question: If we could somehow get classification decisions out of a generative model, how well would it perform out-of-distribution? For instance, would it be biased towards textures (like most discriminative models) or towards shapes (like humans, see [9, 17])? Generally speaking, are discriminative or generative models better approximations for human visual perception (a major open question and a longstanding debate in Cognitive Science and Neuroscience, cf. [20, 6])?

Recently, [4] developed an approach that allows to extract zero-shot classification decisions from a text-to-image diffusion model. We use their methodology to investigate perceptual properties of diffusion based generative models. We focus on the question of how well diffusion based generative models match human visual object recognition on challenging out-of-distribution datasets and visual illusions.

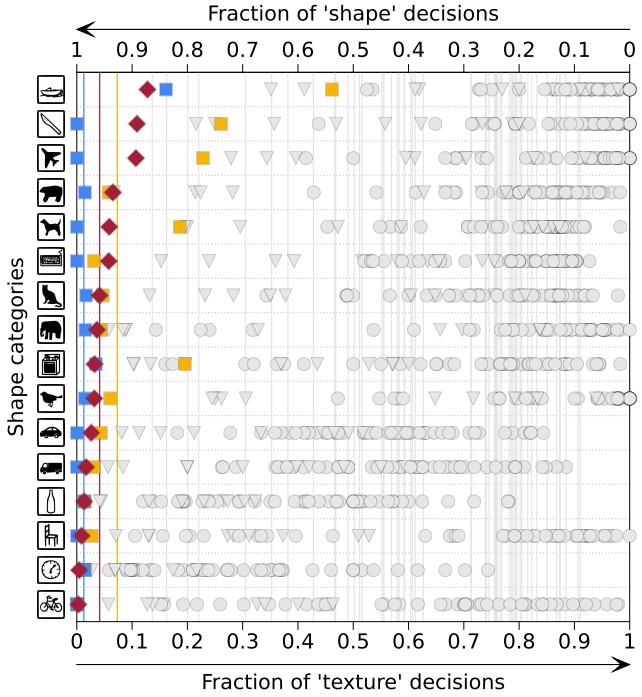


Figure 1: Zero-shot generative classifiers achieve a **human-level shape bias**: 99% for Imagen, 93% for Stable Diffusion and 92–99% for human observers (96% on average).

2. Methods

We study Imagen and Stable Diffusion which are pixel based and latent-space based text-to-image models respectively. We follow the methodology by [4] to convert a text-to-image diffusion model into a zero-shot classifier to facilitate quantitative evaluation on downstream tasks. Fig-

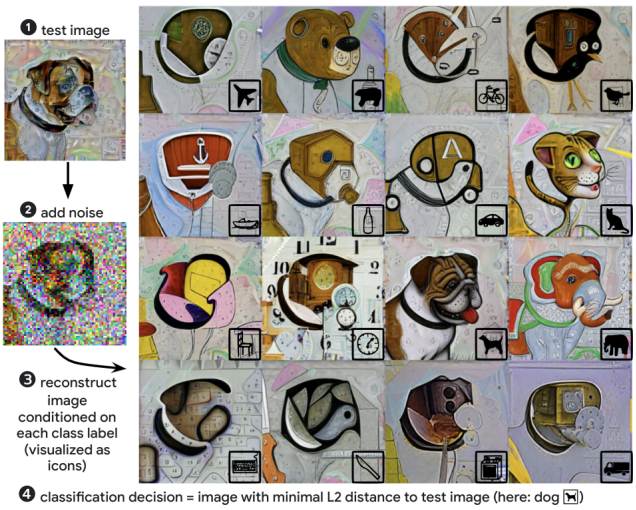


Figure 2: **Classification approach.** Given a test image, such as a dog with clock texture (1), a text-to-image generative classifier adds random noise (2) and then reconstructs the image conditioned on the prompt "A bad photo of a <class>" (3). The reconstructed image that is closest to the test image in L_2 distance is taken as the classification decision (4). For visualization purposes, class icons corresponding to the prompt class are superimposed on the bottom right of the reconstructed images.

ure 2 shows an overview of our method. Briefly, a text-to-image diffusion model can be used to generate classification decisions by using it as a Bayes' classifier *i.e.* given an image $\boldsymbol{x}$ that belongs to any one of of K classes $[\mathsf{y}_K] := \{\mathsf{y}_1, \mathsf{y}_2, \cdots, \mathsf{y}_K\}$, we can predict the most probable class assignment as:

$$\tilde{y} = \arg\max_{\mathsf{y}_k} p(y = \mathsf{y}_k|\boldsymbol{x}) = \arg\max_{\mathsf{y}_k} \; \log p(\boldsymbol{x}|y = \mathsf{y}_k)$$

Here, we assume a uniform prior $p(y_i = \mathsf{y}_k) = \frac{1}{k}$ that can be dropped from the $\arg\max$. More details on this approach can be found in Appendix A.

Model details. We investigate two diffusion models: Imagen [10] and Stable Diffusion [14]. Imagen is a cascaded diffusion model consisting of a 64×64 low-resolution model and two super-resolution models. Here, we only use the 64×64 model for our experiments because the high-resolution models perform poorly as classifiers [4]. We use version 1.4 of Stable Diffusion for our experiments. It uses a pre-trained text encoder from CLIP to encode the text and a pre-trained variational autoencoder to map images to a latent space. For both Imagen and Stable Diffusion, we generate classification decisions using the prompt "A photo of a $<$ class $>$". As baseline non-generative classifiers, we compare against 52 diverse

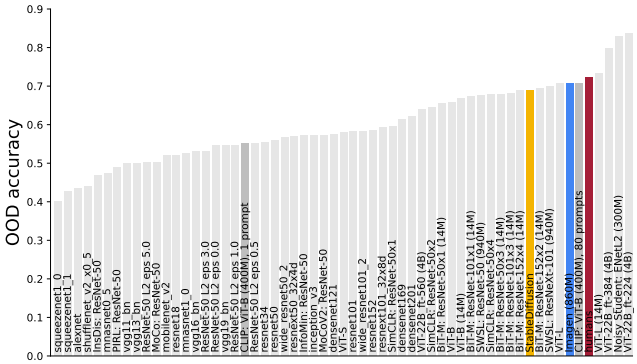


Figure 3: **Out-of-distribution accuracy** across 17 challenging datasets [8]. Detailed results for all parametric datasets are plotted in Figure 5; Table 2 lists accuracies.

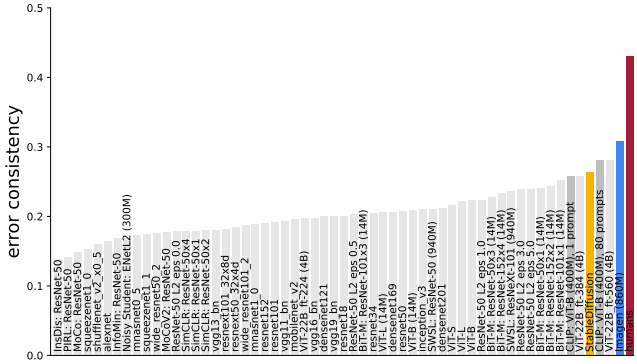


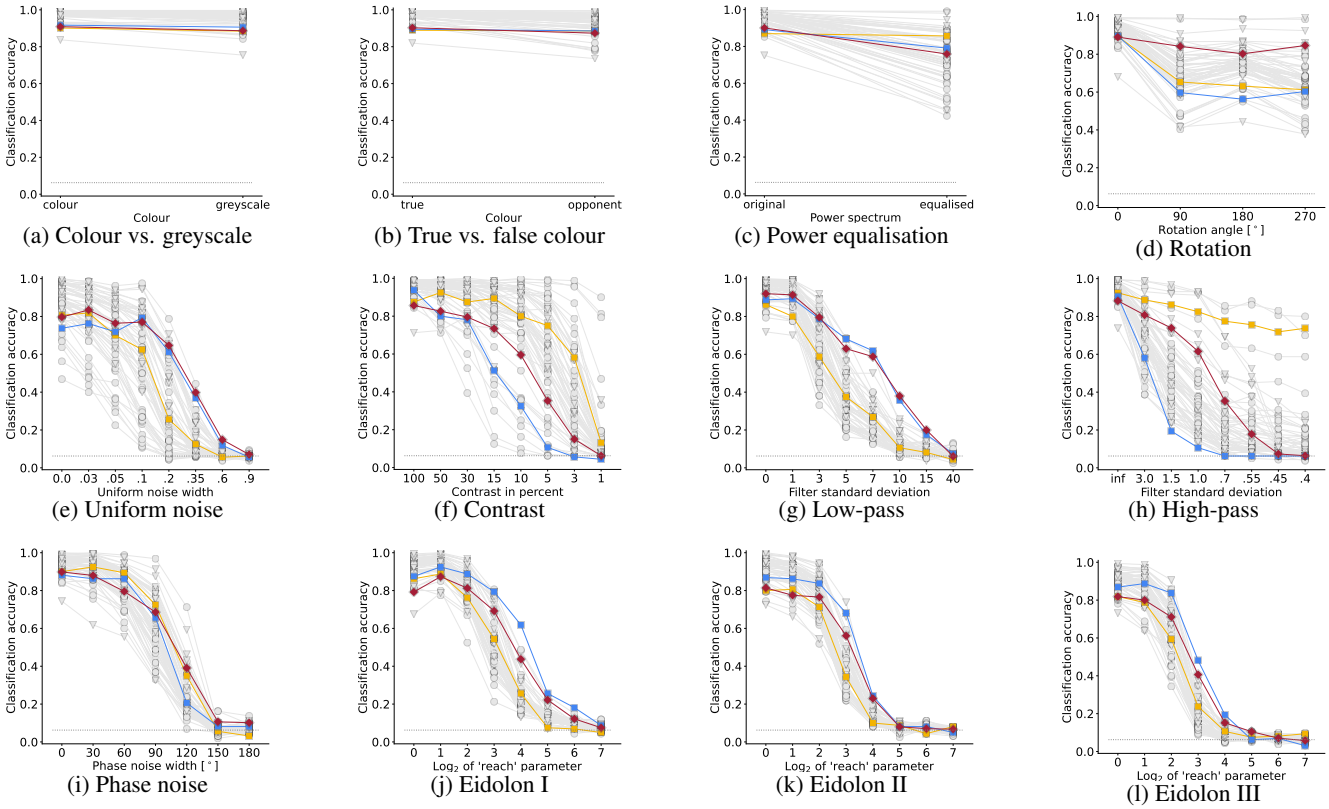
Figure 4: **Error consistency** across 17 challenging datasets [8]. This metric measures whether errors made by models align with errors made by humans (higher is better) [7].

models from the model-vs-human toolbox [8] that are either trained or fine-tuned on ImageNet, three ViT-22B variants [5] (very large 22B parameter vision transformers) and two CLIP [13] versions as zero-shot classifier baselines. The CLIP models are based on the largest version, ViT-L/14@224px, and consist of vision and text transformers trained with contrastive learning. We use one CLIP model that uses an ensemble of 80 different prompts for classification [13] and one CLIP model that uses the very same classification prompt as our generative classifiers for a fair comparison. All baseline models are plotted in grey.

3. Results

We analyze diffusion based zero-shot classifiers by comparing them against discriminative models and human psychophysical data using 17 challenging out-of-distribution (OOD) datasets via the model-vs-human toolbox [8]. We report four intriguing properties of zero-shot generative classifiers:

1. a human-like shape bias (3.1),



(a) Colour vs. greyscale (b) True vs. false colour (c) Power equalisation (d) Rotation

(e) Uniform noise (f) Contrast (g) Low-pass (h) High-pass

(i) Phase noise (j) Eidolon I (k) Eidolon II (l) Eidolon III

Figure 5: **Detailed out-of-distribution accuracy** for Imagen and Stable Diffusion in comparison to human observers. While not always aligning perfectly with human accuracy, the overall robustness achieved by both models is comparable to that of human observers even though these models are zero-shot, i.e. neither designed nor trained to do classification.

2. near human-level out-of-distribution accuracy (3.2),
3. SOTA error consistency with humans (3.3),
4. an understanding of certain perceptual illusions (3.4).

3.1. Human-like shape bias

Introduced by [9], the *shape bias* of a model indicates to which degree the model's decisions are based on object shape, as opposed to object texture. Most standard models are biased towards texture, whereas humans are strongly biased towards shape (96% shape bias on average, ranging from 92% to 99% for individual observers). As shown in Figure 1, the zero-shot generative classifiers we test are the very first models to show a shape bias that matches humans: Imagen achieves a stunning 99% shape bias, and Stable Diffusion a 93% shape bias. Dozens of comparison models are shown in grey; their detailed performance is reported in supplementary Figure 10.

3.2. Near human-level OOD accuracy

Humans excel at recognizing objects even if they are heavily distorted. Across the 17 challenging out-of-distribution datasets by the model-vs-human toolbox [8], both Imagen and Stable Diffusion achieve an overall accuracy that is close to human-level robustness (Figure 3) despite being zero-shot models. In detailed plots split by dataset (Figure 5), it can be seen that on most datasets model performance approximately matches human responses, with the exception of rotated images (for which humans are much better) and high-pass filtered images (where Stable Diffusion is much better than humans and Imagen is much worse). This indicates that even though both models are diffusion-based, they exhibit very different sensitivities to high spatial frequencies.

3.3. SOTA error consistency with human observers

Humans and models may both achieve, say, 90% accuracy on a dataset but do they make errors on the same 10% of images, or on different images? This is measured by *error consistency* [7]. Figure 4 shows overall results across 17 datasets. While a substantial gap towards human-to-human error consistency remains, Imagen shows the most human-aligned error patterns, surpassing previous state-of-the-art (SOTA) set by ViT-22B, a large vision transformer [5].

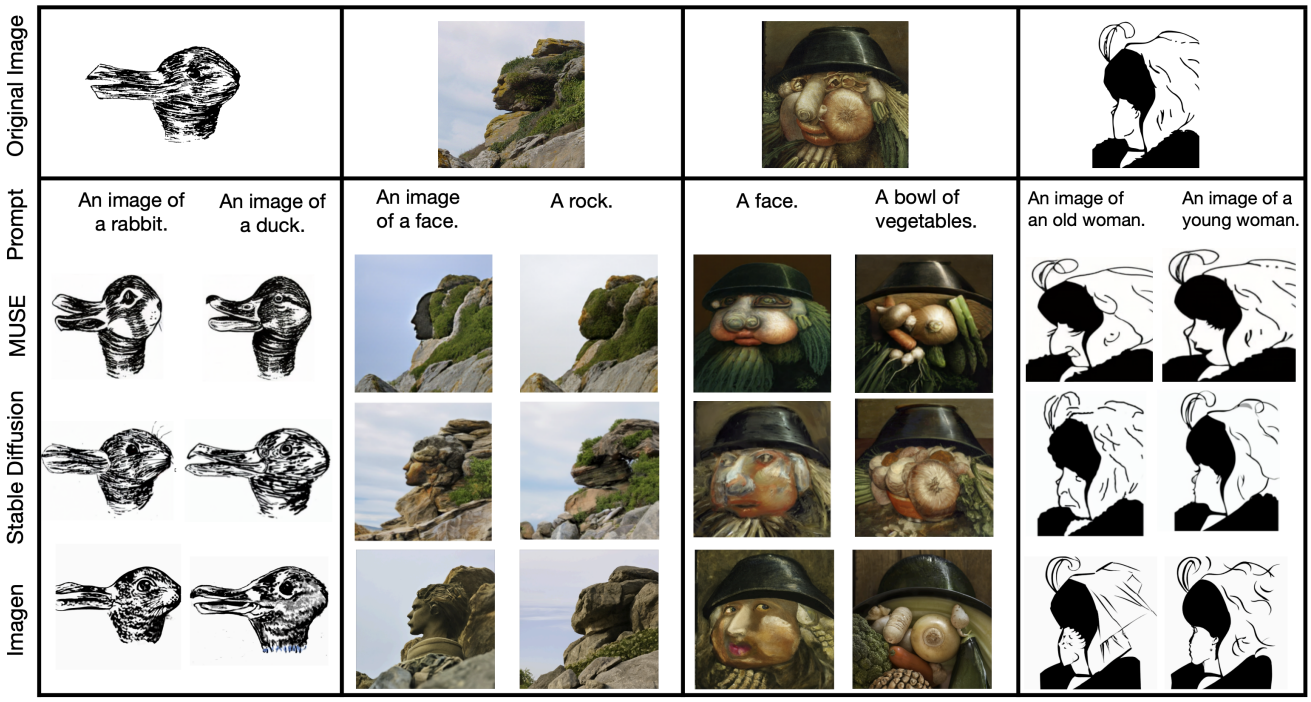


Figure 6: **Text-to-image generative classifiers understand certain visual illusions** as indicated by their ability to reconstruct ambiguous images in a way that aligns with how humans perceive those images. For instance, they reconstruct a right-facing rabbit vs. a left-facing duck in the case of the bistable rabbit-duck illusion and place the face in the right location and pose for an image where humans show pareidolia (seeing patterns in things, like a face in a rock).

3.4. Understanding certain visual illusions

Visual illusions are hidden doors to the secrets of perception by revealing aspects that might otherwise go unnoticed. In contrast to discriminative models, generative classifiers offer a straightforward way to test illusions: for bistable images such as the famous rabbit-duck, we can prompt them to reconstruct based on 'an image of a duck' and 'an image of a rabbit'. If they can (a) reconstruct images resembling the respective animal and (b) they place the reconstructed animal in the same location and pose as humans would, this can be seen as evidence that they "understand" the illusion. In Figure 6[1], we find that this is indeed the case: generative models share certain bistable illusions and pareidolia (seeing patterns in things, like a face in a rock) with human visual perception.

4. Discussion

The dominant current paradigm for modeling human visual object recognition is based on deep *discriminative* models. We here observe intriguing human-like properties of *generative* models even though those models were neither designed nor trained do do classification. This opens a wide array of exciting questions for future research: Are those intriguing properties caused by generative modeling per se, or somehow just a by-product of diffusion denoising? Are they a result of architecture, scale, training data, or induced through language cross-attention? How can we explain the paradox resulting from convolutional neural networks being (to a certain degree) biologically plausible, yet worse in matching human object recognition behavior when compared to biologically implausible diffusion models? Are there any simple explanations, or do we need to question the predominant approach using discriminative models of human object recognition and instead put a greater focus on generative models? We hope our findings provide evidence for just how intriguing generative classifiers are, and inspire future work to explore those exciting directions.

Limitations. It is currently unclear whether the intriguing properties we discovered are properties of any generative classifier or properties of diffusion-based text-to-image models. Furthermore, the approach of generating at least one prediction (i.e., image generation) per class is expensive/slow.

[1]Muse [2] also understands certain illusions. Since it is not a diffusion model, we cannot obtain classification decisions using Figure 2.

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2018. 1

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv preprint arXiv:2301.00704*, 2023. 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6

[4] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 1, 2, 6

[5] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 2, 3

[6] James J DiCarlo, Ralf Haefner, Leyla Isik, Talia Konkle, Nikolaus Kriegeskorte, Benjamin Peters, Nicole Rust, Kim Stachenfeld, Joshua B Tenenbaum, Doris Tsao, et al. How does the brain combine generative models and direct discriminative computations in high-level vision? 2021. 1

[7] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3

[8] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021. 2, 3, 6, 8, 15, 16

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1, 3

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 2

[11] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2):8, 2019. 6

[12] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1485–1488, 2010. 6

[13] 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. 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems*, 2022. 1

[16] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020. 6

[17] Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9, 2023. 1

[18] Qizhe Xie, Eduard Hovy, Minh-Tang Luong, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. *arXiv:1911.04252*, 2019. 6

[19] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 6

[20] Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006. 1

SUPPLEMENTARY MATERIAL

A. Method details

Using text-to-image models as generative classifiers Concretely, using a text-to-image diffusion model as a generative classifier requires two modifications. First, the models are conditioned on text prompts rather than class labels. Thus, we convert each label, y_k, to text using a mapping $\boldsymbol{\phi}$ with a dataset-specific template (e.g. $\mathsf{y}_k \rightarrow$ A photo of a y_k). Second, diffusion models do not produce exact log-likelihoods (*i.e.* we cannot compute $\log p_\theta(\boldsymbol{x}|y = \mathsf{y}_k)$ directly). The key idea used by [4] for a solution is to use the variational diffusion lower bound (as Imagen and SD are not trained with the other losses) as a proxy to approximate $\log p_\theta(\boldsymbol{x}|y = \mathsf{y}_k)$. Thus we have:

$$\begin{aligned}\tilde{y} &= \arg\max_{\mathsf{y}_k} \ \log p_\theta(\boldsymbol{x}|y = \mathsf{y}_k) \\ &= \arg\min_{\mathsf{y}_k \in [\mathsf{y}_K]} \ \mathbb{E}_{\epsilon,t}\Big[\boldsymbol{w}_t \|\boldsymbol{x} - \tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}\big(\boldsymbol{x}_t, \boldsymbol{\phi}(\mathsf{y}_k), t\big)\|_2^2\Big] \qquad (1)\end{aligned}$$

Note that for SD, $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}$ are latent representations, with $\boldsymbol{x}$ obtained by encoding the image using a VAE. With Imagen on the other hand, $\boldsymbol{x}$ consists of the raw image pixels.

Estimating the expectation The expectation in Equation (1) is approximated using Monte-Carlo estimation. At each step, we sample a $t \sim \mathcal{U}([0, 1])$ and then a $\boldsymbol{x}_t$ according to the forward diffusion process. Next, we denoise this noisy image using the model (*i.e.* we use it to predict $\boldsymbol{x}$ from $\boldsymbol{x}_t$), obtaining $\hat{\boldsymbol{x}} = \tilde{\boldsymbol{x}}_{\boldsymbol{\theta}}\big(\boldsymbol{x}_t, \boldsymbol{\phi}(\mathsf{y}_k), t\big)$. We call the squared error of the prediction, $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$, a *score* for $(\boldsymbol{x}, \mathsf{y}_k)$. We score each class N times, obtaining a $K \times N$ *scores matrix*[2] for the image. Finally, we weight the scores according to the corresponding $\boldsymbol{w}_t$ and take the mean, resulting in an estimate of the marginal log-likelihood for each class.

B. Additional plots

We here plot detailed performance for all models with respect to a few different properties of interest / metrics:

- Aggregated performance across 17 datasets: Figure 7
- Out-of-distribution accuracy: Figure 8 for parametric datasets and Figure 9 for nonparametric datasets
- Model-to-human error consistency: Figure 8
- Human-to-human, model-to-model error consistency (for nonparametric datasets): Figures 11 to 15
- Shape bias: Figure 10

All metrics are based on the model-vs-human toolbox and explained in more detail in [8].

[2]Later we discuss how we can avoid computing the full matrix for efficiency.

C. Quantitative benchmark scores and rankings

Table 1 and Table 2 list the detailed performance aggregated across 17 datasets for each model, with the former focusing on metrics related to "most human-like object recognition behavior" and the latter focusing on out-of-distribution accuracy.

D. Baseline models

The baseline models plotted in grey include standard ImageNet-trained convolutional torchvision models [12], adversarially trained models [16], self-supervised models such as SimCLR [3], BiT-M variants [11], Noisy Student [18], and SWSL variants [19].

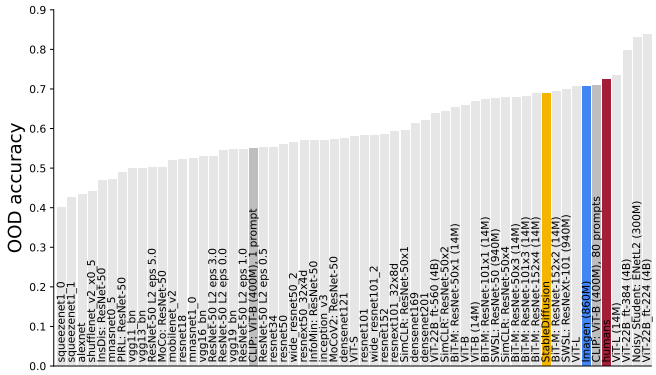
E. Image Attribution

Rabbit-duck image:
Attribution: Unknown source, Public domain, via Wikimedia Commons.
Link: https://upload.wikimedia.org/wikipedia/commons/9/96/Duck-Rabbit.png

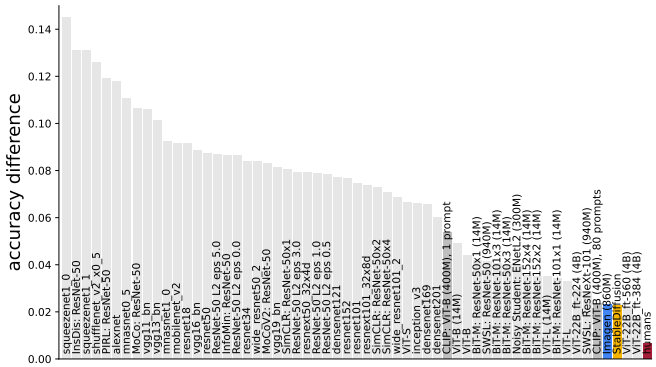
Rock image:
Attribution: Mirabeau, CC BY-SA 3.0, via Wikimedia Commons.
Link: https://commons.wikimedia.org/wiki/File:Visage_dans_un_rocher.jpg

Vegetable portrait image:
Attribution: Giuseppe Arcimboldo, Public domain, via Wikimedia Commons.
Link: https://upload.wikimedia.org/wikipedia/commons/4/49/Arcimboldo_Vegetables.jpg

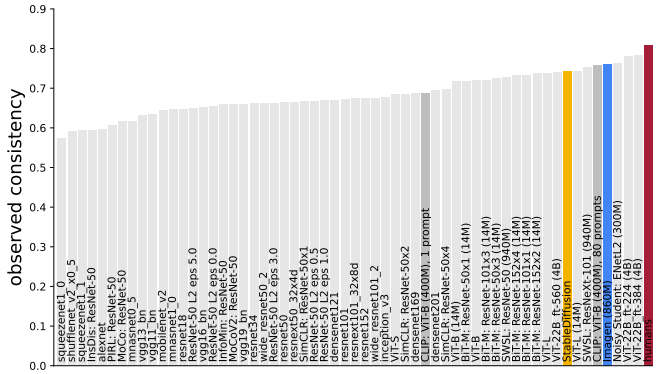
Woman image:
Attribution: W. E. Hill, Public domain, via Wikimedia Commons.
Link: https://upload.wikimedia.org/wikipedia/commons/5/5f/My_Wife_and_My_Mother-In-Law_%28Hill%29.svg



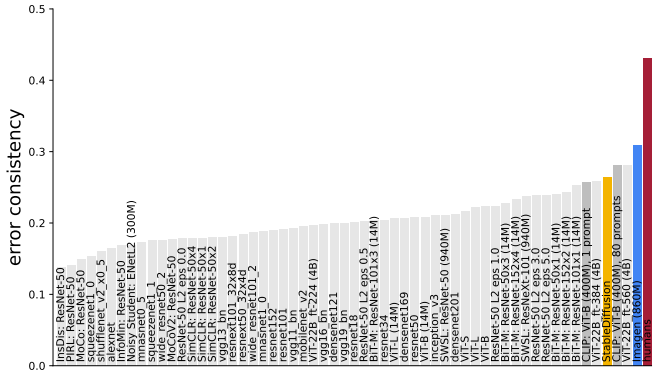
(a) OOD accuracy (higher = better).



(b) Accuracy difference (lower = better).

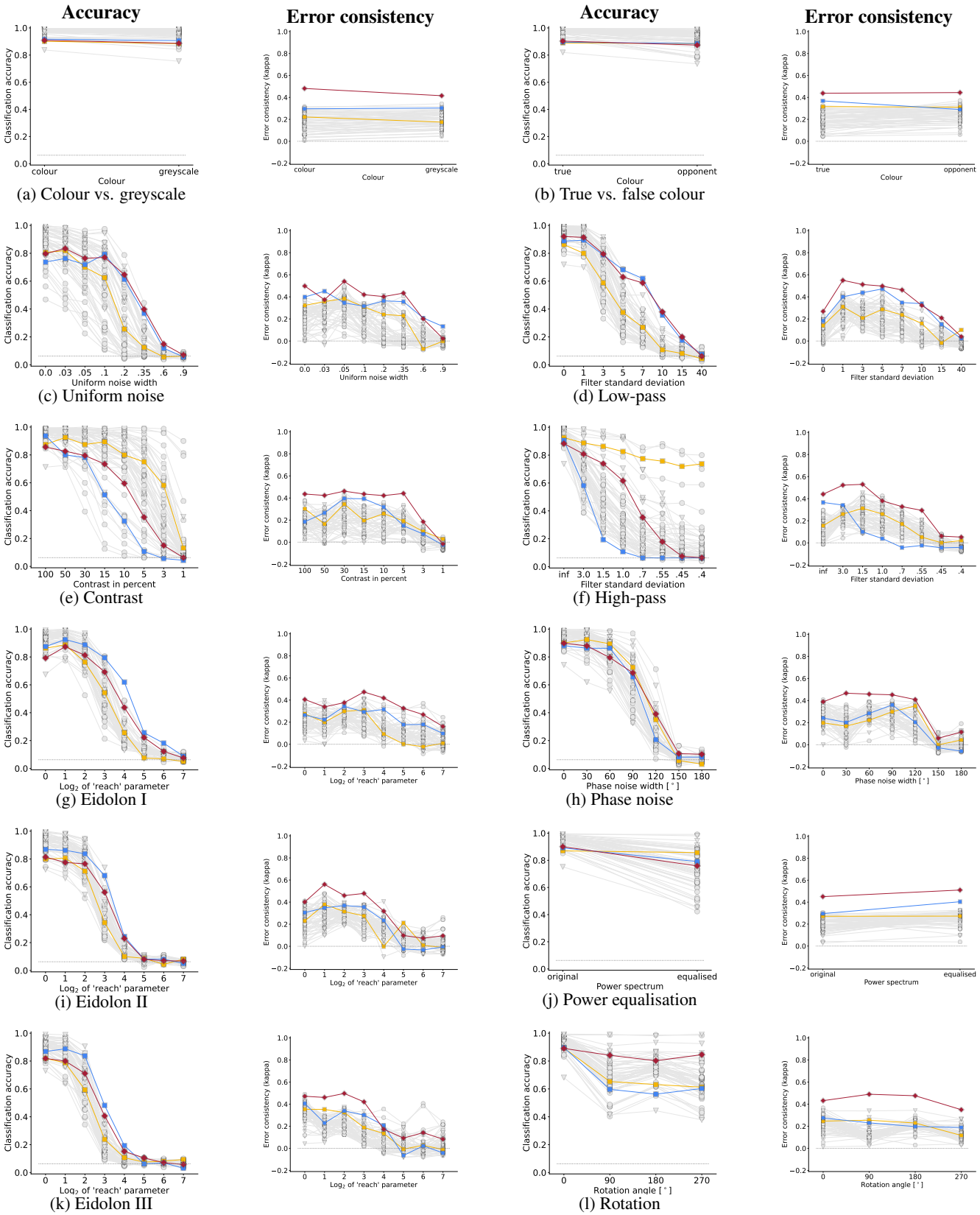


(c) Observed consistency (higher = better).



(d) Error consistency (higher = better).

Figure 7: Benchmark results for different models, aggregated over datasets.



(a) Colour vs. greyscale (b) True vs. false colour

(c) Uniform noise (d) Low-pass

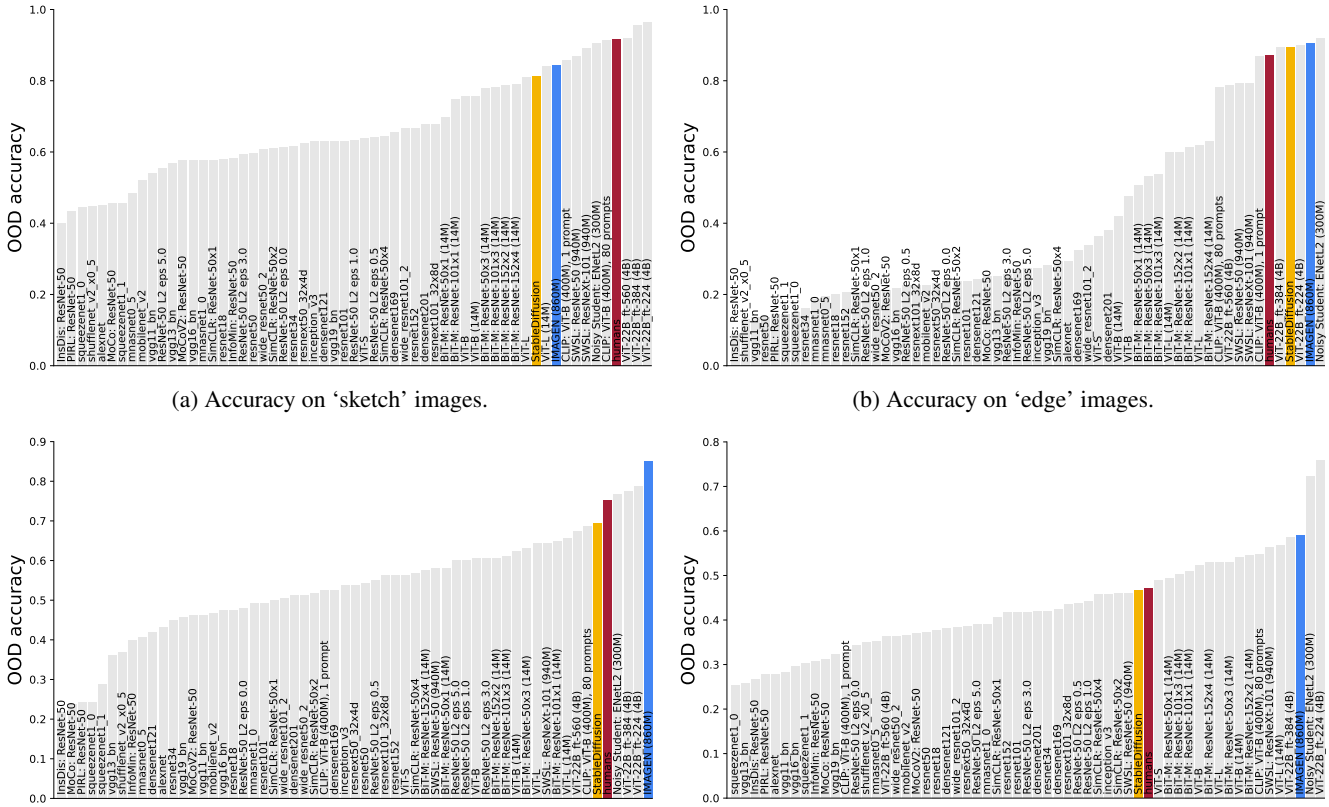
(e) Contrast (f) High-pass

(g) Eidolon I (h) Phase noise

(i) Eidolon II (j) Power equalisation

(k) Eidolon III (l) Rotation

Figure 8: OOD accuracy and error consistency across all twelve parametric datasets from [8]. Error consistency results for nonparametric datasets are plotted in Figures 11 to 15.



(a) Accuracy on 'sketch' images.

(b) Accuracy on 'edge' images.

(c) Accuracy on 'silhouette' images.

(d) Accuracy on 'stylized' images.

Figure 9: OOD accuracy on all four nonparametric datasets (i.e., datasets with only a single corruption type and strength).

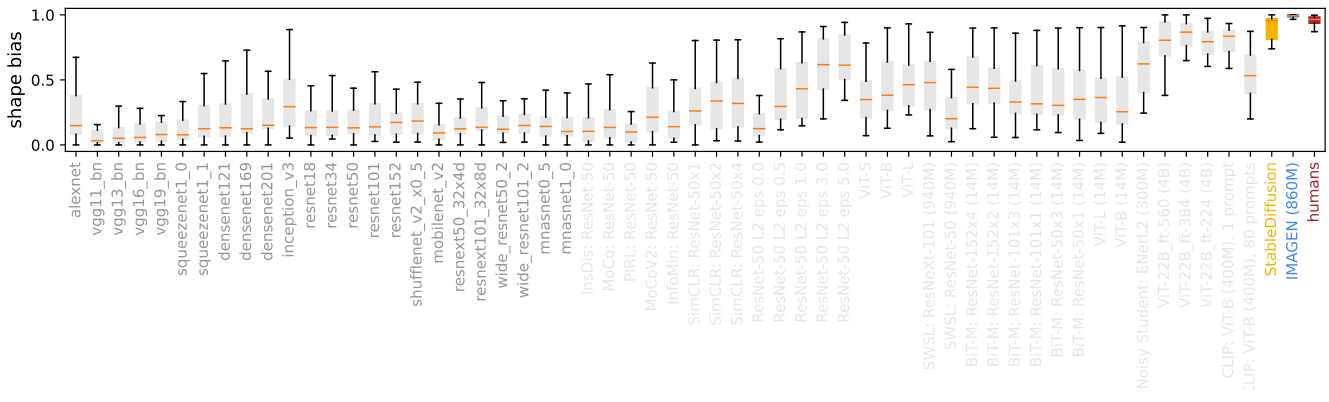


Figure 10: Zero-shot generative classifiers achieve a **human-level shape bias**: 99% for Imagen, 93% for Stable Diffusion and 92–99% for human observers (96% on average). This figure shows boxplots highlighting the spread across 16 categories for each model as a different way of visualizing the data from Figure 1.

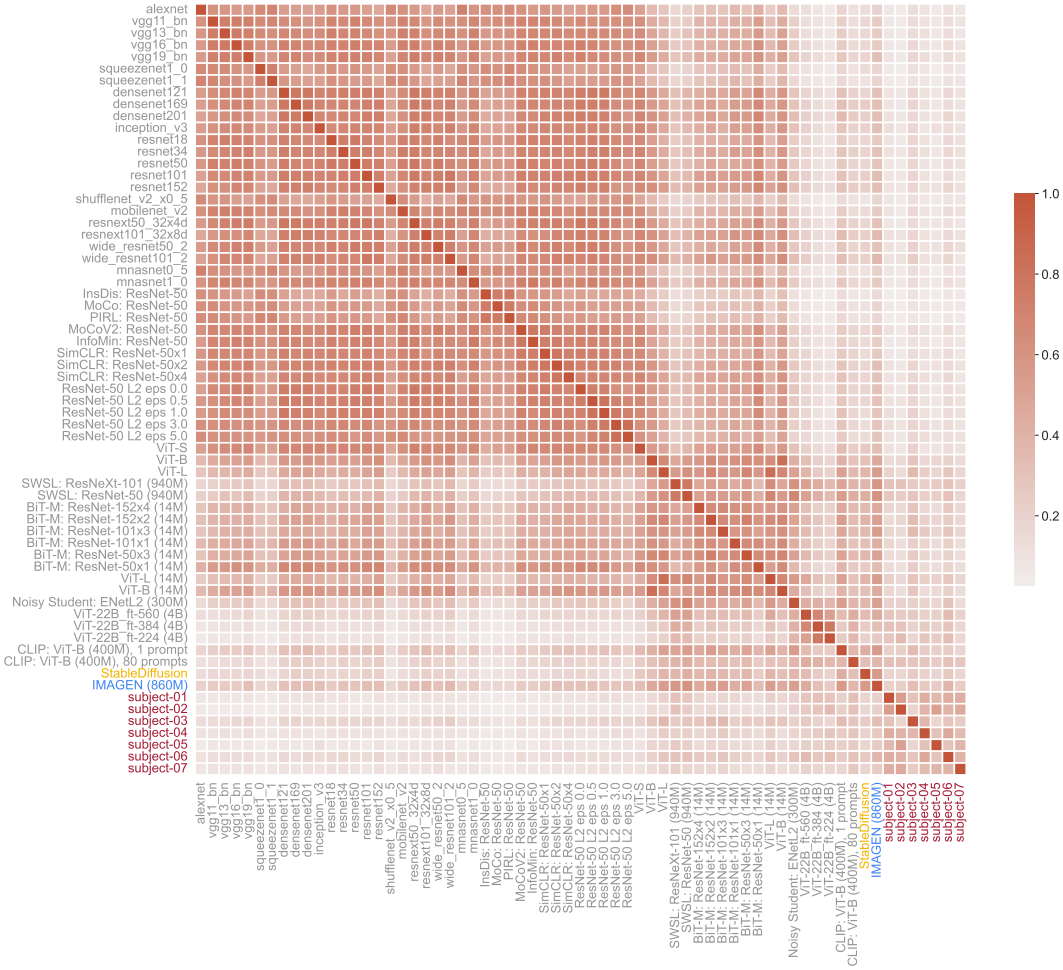


Figure 11: Error consistency for ‘sketch’ images.

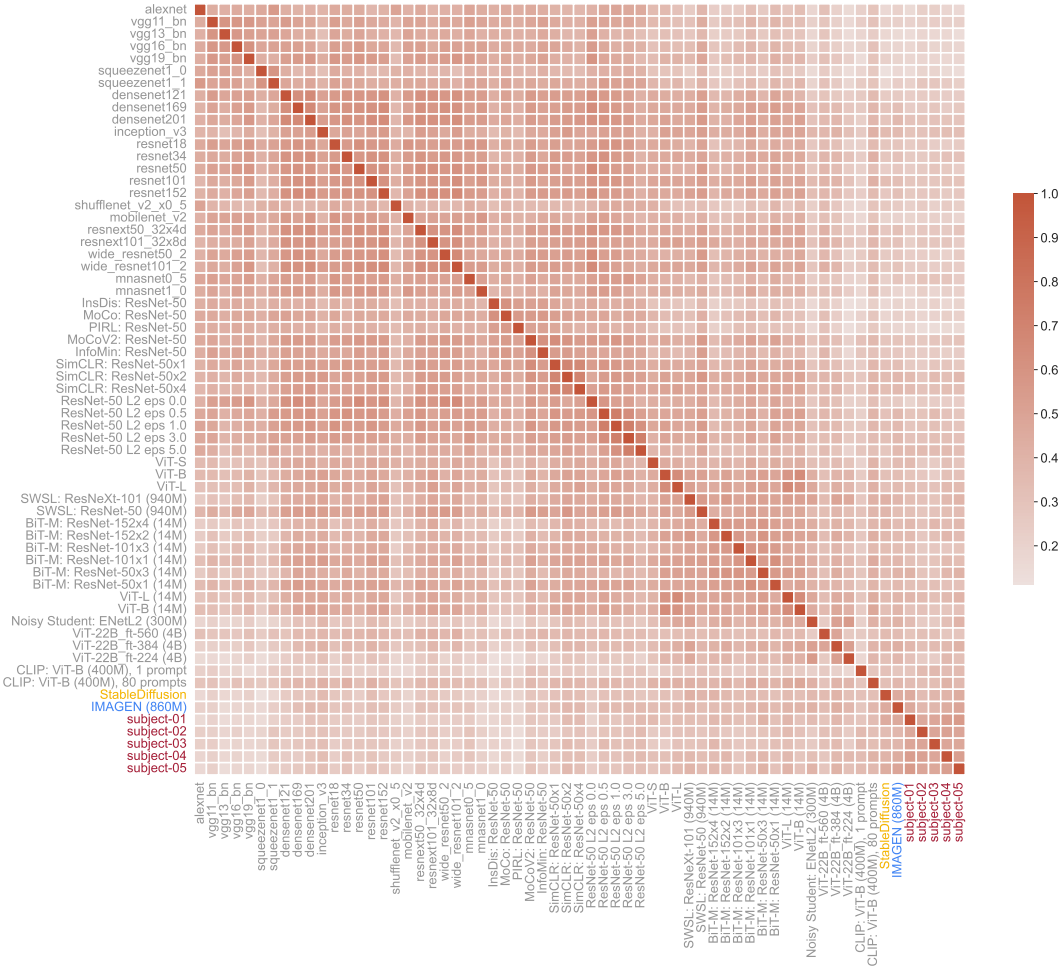


Figure 12: Error consistency for ‘stylized’ images.

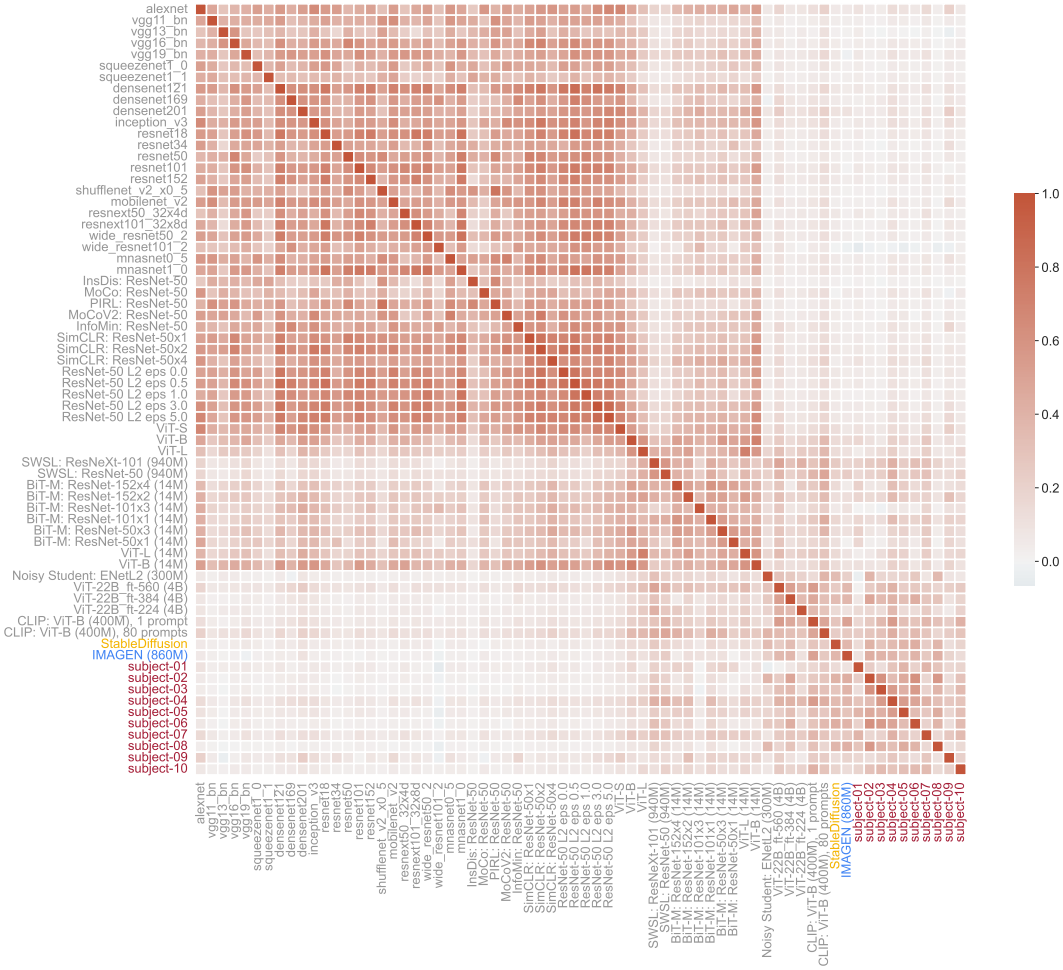


Figure 13: Error consistency for ‘edge’ images.

Figure 14: Error consistency for ‘silhouette’ images.

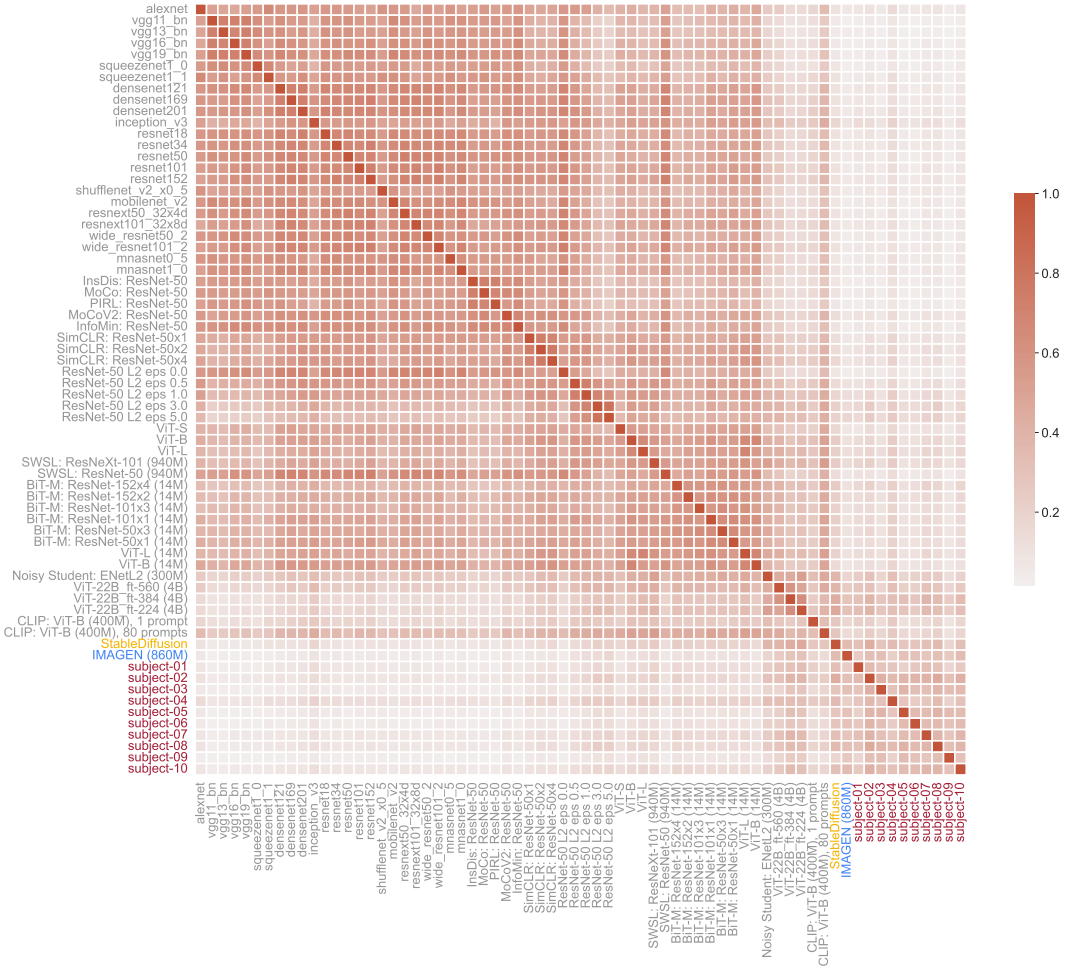


Figure 15: Error consistency for ‘cue conflict’ images.

Table 1: Benchmark table of model results for most human-like behaviour, aggregated over all 17 datasets from [8]. The three metrics "accuracy difference" "observed consistency" and "error consistency" each produce a different model ranking. The mean rank of a model across those three metrics is used to rank the models on our benchmark.

model	accuracy diff. ↓	obs. consistency ↑	error consistency ↑	mean rank ↓
ViT-22B_ft-384 (4B)	**0.018**	**0.783**	0.258	**2.333**
Imagen (860M)	0.023	0.761	**0.309**	3.000
ViT-22B_ft-560 (4B)	0.022	0.739	0.281	4.333
CLIP: ViT-B (400M), 80 prompts	0.023	0.758	0.281	4.333
StableDiffusion	0.023	0.743	0.264	5.000
SWSL: ResNeXt-101 (940M)	0.028	0.752	0.237	8.000
BiT-M: ResNet-101x1 (14M)	0.034	0.733	0.252	9.333
BiT-M: ResNet-152x2 (14M)	0.035	0.737	0.243	10.000
ViT-L	0.033	0.738	0.222	11.667
BiT-M: ResNet-152x4 (14M)	0.035	0.732	0.233	12.667
ViT-L (14M)	0.035	0.744	0.206	14.000
ViT-22B_ft-224 (4B)	0.030	0.781	0.197	14.000
BiT-M: ResNet-50x3 (14M)	0.040	0.726	0.228	14.333
BiT-M: ResNet-50x1 (14M)	0.042	0.718	0.240	14.667
CLIP: ViT-B (400M), 1 prompt	0.054	0.688	0.257	16.000
SWSL: ResNet-50 (940M)	0.041	0.727	0.211	16.667
ViT-B	0.044	0.719	0.223	17.000
BiT-M: ResNet-101x3 (14M)	0.040	0.720	0.204	19.333
ViT-B (14M)	0.049	0.717	0.209	20.000
densenet201	0.060	0.695	0.212	20.333
Noisy Student: ENetL2 (300M)	0.040	0.764	0.169	22.333
ViT-S	0.066	0.684	0.216	22.333
densenet169	0.065	0.688	0.207	23.000
inception_v3	0.066	0.677	0.211	23.333
ResNet-50 L2 eps 1.0	0.079	0.669	0.224	26.667
ResNet-50 L2 eps 3.0	0.079	0.663	0.239	27.667
SimCLR: ResNet-50x4	0.071	0.698	0.179	30.333
wide_resnet101_2	0.068	0.676	0.187	30.333
ResNet-50 L2 eps 0.5	0.078	0.668	0.203	31.000
densenet121	0.077	0.671	0.200	31.000
SimCLR: ResNet-50x2	0.073	0.686	0.180	31.333
resnet152	0.077	0.675	0.190	31.667
resnet101	0.074	0.671	0.192	31.667
resnext101_32x8d	0.074	0.674	0.182	32.667
ResNet-50 L2 eps 5.0	0.087	0.649	0.240	32.667
resnet50	0.087	0.665	0.208	34.333
resnet34	0.084	0.662	0.205	35.000
vgg19_bn	0.081	0.660	0.200	35.667
resnext50_32x4d	0.079	0.666	0.184	36.333
SimCLR: ResNet-50x1	0.080	0.667	0.179	38.000
resnet18	0.091	0.648	0.201	40.333
vgg16_bn	0.088	0.651	0.198	40.333
wide_resnet50_2	0.084	0.663	0.176	41.667
MoCoV2: ResNet-50	0.083	0.660	0.177	42.000
mobilenet_v2	0.092	0.645	0.196	43.000
ResNet-50 L2 eps 0.0	0.086	0.654	0.178	43.333
mnasnet1_0	0.092	0.646	0.189	44.333
vgg11_bn	0.106	0.635	0.193	44.667
InfoMin: ResNet-50	0.086	0.659	0.168	45.333
vgg13_bn	0.101	0.631	0.180	47.000
mnasnet0_5	0.110	0.617	0.173	51.000
MoCo: ResNet-50	0.107	0.617	0.149	53.000
alexnet	0.118	0.597	0.165	53.333
squeezenet1_1	0.131	0.593	0.175	53.667
PIRL: ResNet-50	0.119	0.607	0.141	54.667
shufflenet_v2_x0_5	0.126	0.592	0.160	55.333
InsDis: ResNet-50	0.131	0.593	0.138	56.667
squeezenet1_0	0.145	0.574	0.153	57.000

Table 2: Benchmark table of model results for highest out-of-distribution robustness, aggregated over all 17 datasets from [8].

model	OOD accuracy ↑	rank ↓
ViT-22B_ft-224 (4B)	**0.837**	**1.000**
Noisy Student: ENetL2 (300M)	0.829	2.000
ViT-22B_ft-384 (4B)	0.798	3.000
ViT-L (14M)	0.733	4.000
CLIP: ViT-B (400M), 80 prompts	0.708	5.000
Imagen (860M)	0.706	6.000
ViT-L	0.706	7.000
SWSL: ResNeXt-101 (940M)	0.698	8.000
BiT-M: ResNet-152x2 (14M)	0.694	9.000
StableDiffusion	0.689	10.000
BiT-M: ResNet-152x4 (14M)	0.688	11.000
BiT-M: ResNet-101x3 (14M)	0.682	12.000
BiT-M: ResNet-50x3 (14M)	0.679	13.000
SimCLR: ResNet-50x4	0.677	14.000
SWSL: ResNet-50 (940M)	0.677	15.000
BiT-M: ResNet-101x1 (14M)	0.672	16.000
ViT-B (14M)	0.669	17.000
ViT-B	0.658	18.000
BiT-M: ResNet-50x1 (14M)	0.654	19.000
SimCLR: ResNet-50x2	0.644	20.000
ViT-22B_ft-560 (4B)	0.639	21.000
densenet201	0.621	22.000
densenet169	0.613	23.000
SimCLR: ResNet-50x1	0.596	24.000
resnext101_32x8d	0.594	25.000
resnet152	0.584	26.000
wide_resnet101_2	0.583	27.000
resnet101	0.583	28.000
ViT-S	0.579	29.000
densenet121	0.576	30.000
MoCoV2: ResNet-50	0.571	31.000
inception_v3	0.571	32.000
InfoMin: ResNet-50	0.571	33.000
resnext50_32x4d	0.569	34.000
wide_resnet50_2	0.566	35.000
resnet50	0.559	36.000
resnet34	0.553	37.000
ResNet-50 L2 eps 0.5	0.551	38.000
CLIP: ViT-B (400M), 1 prompt	0.550	39.000
ResNet-50 L2 eps 1.0	0.547	40.000
vgg19_bn	0.546	41.000
ResNet-50 L2 eps 0.0	0.545	42.000
ResNet-50 L2 eps 3.0	0.530	43.000
vgg16_bn	0.530	44.000
mnasnet1_0	0.524	45.000
resnet18	0.521	46.000
mobilenet_v2	0.520	47.000
MoCo: ResNet-50	0.502	48.000
ResNet-50 L2 eps 5.0	0.501	49.000
vgg13_bn	0.499	50.000
vgg11_bn	0.498	51.000
PIRL: ResNet-50	0.489	52.000
mnasnet0_5	0.472	53.000
InsDis: ResNet-50	0.468	54.000
shufflenet_v2_x0_5	0.440	55.000
alexnet	0.434	56.000
squeezenet1_1	0.425	57.000
squeezenet1_0	0.401	58.000