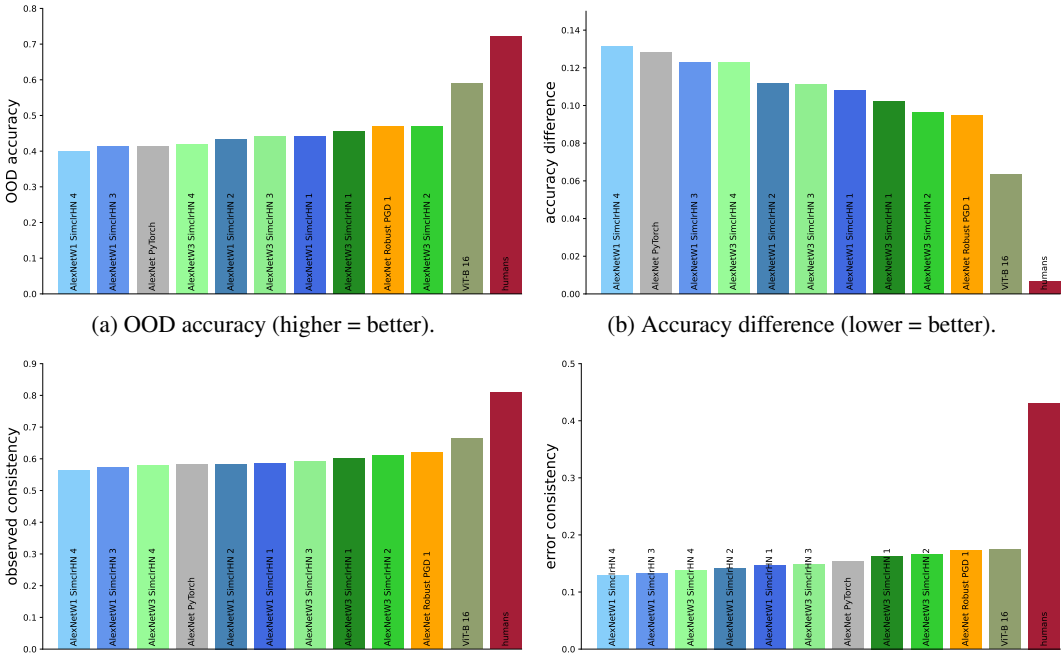
Generalisation report

Generalisation report produced by model-vs-human



(a) OOD accuracy (higher = better).

(b) Accuracy difference (lower = better).

(c) Observed consistency (higher = better).

(d) Error consistency (higher = better).

Figure 1: Benchmark results for different models, aggregated over datasets.

Table 2: Benchmark table of model results for highest out-of-distribution robustness.

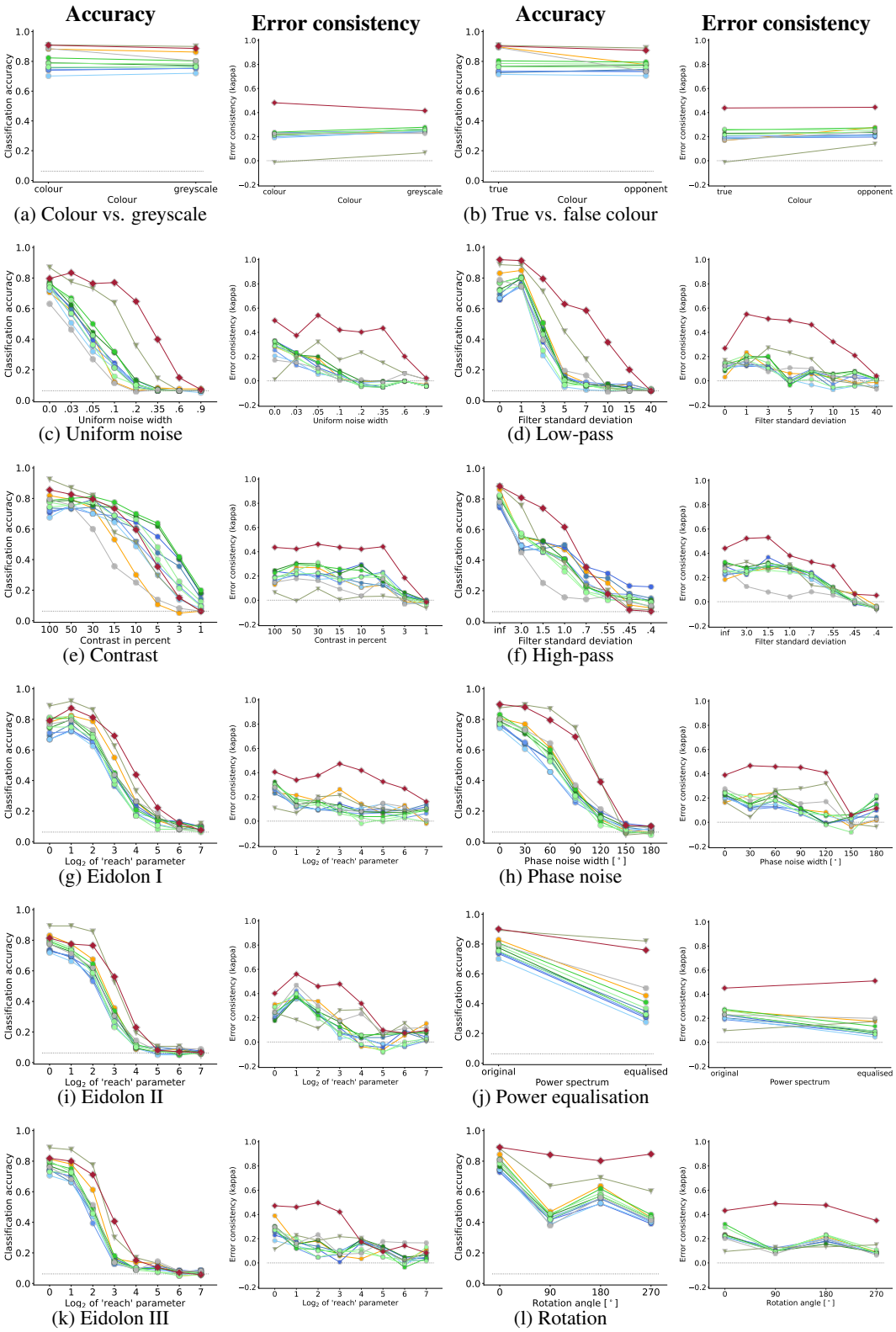
model	OOD accuracy ↑	rank ↓
ViT-B 16	**0.590**	**1.000**
AlexNetW3 SimclrHN 2	0.470	2.000
AlexNet Robust PGD 1	0.470	3.000
AlexNetW3 SimclrHN 1	0.455	4.000
AlexNetW1 SimclrHN 1	0.441	5.000
AlexNetW3 SimclrHN 3	0.440	6.000
AlexNetW1 SimclrHN 2	0.433	7.000
AlexNetW3 SimclrHN 4	0.420	8.000
AlexNet PyTorch	0.415	9.000
AlexNetW1 SimclrHN 3	0.412	10.000
AlexNetW1 SimclrHN 4	0.399	11.000

Table 1: Benchmark table of model results for most human-like behaviour. The three metrics "accuracy difference" "observed consistency" and "error consistency" (plotted in Figure 1) 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-B 16	**0.063**	**0.664**	**0.175**	**1.000**
AlexNet Robust PGD 1	0.095	0.619	0.173	2.000
AlexNetW3 SimclrHN 2	0.096	0.610	0.167	3.000
AlexNetW3 SimclrHN 1	0.102	0.602	0.163	4.000
AlexNetW3 SimclrHN 3	0.111	0.592	0.148	5.667
AlexNetW1 SimclrHN 1	0.108	0.587	0.146	6.000
AlexNetW1 SimclrHN 2	0.112	0.583	0.142	7.333
AlexNet PyTorch	0.128	0.581	0.153	7.667
AlexNetW3 SimclrHN 4	0.123	0.579	0.138	8.667
AlexNetW1 SimclrHN 3	0.123	0.572	0.132	9.667
AlexNetW1 SimclrHN 4	0.131	0.564	0.129	11.000

Table 3: Shape vs. texture bias: table.

model	Mean	Lower 95% CI	Upper 95% CI	Prop. Human	Rank ↓
humans	0.950	0.927	0.973	1.000	NaN
alexnet_w1_mlp_simclrhn_probe0	**0.529**	0.406	0.652	0.556	**1.000**
alexnet_w3_mlp_simclrhn_probe3	0.524	0.400	0.647	0.551	2.000
alexnet_w1_mlp_simclrhn_probe3	0.511	0.390	0.632	0.538	3.000
alexnet_w3_mlp_simclrhn_probe0	0.508	0.371	0.645	0.534	4.000
alexnet_w1_mlp_simclrhn_probe1	0.494	0.356	0.632	0.520	5.000
alexnet_w3_mlp_simclrhn_probe1	0.485	0.352	0.617	0.510	6.000
alexnet_w3_mlp_simclrhn_probe2	0.470	0.335	0.605	0.494	7.000
alexnet_w1_mlp_simclrhn_probe2	0.461	0.335	0.587	0.485	8.000
alexnet2023_baseline_pgd	0.387	0.259	0.514	0.407	9.000
vit_b_16	0.373	0.250	0.496	0.392	10.000
alexnet	0.248	0.132	0.364	0.261	11.000



(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 2: OOD accuracy and error consistency.

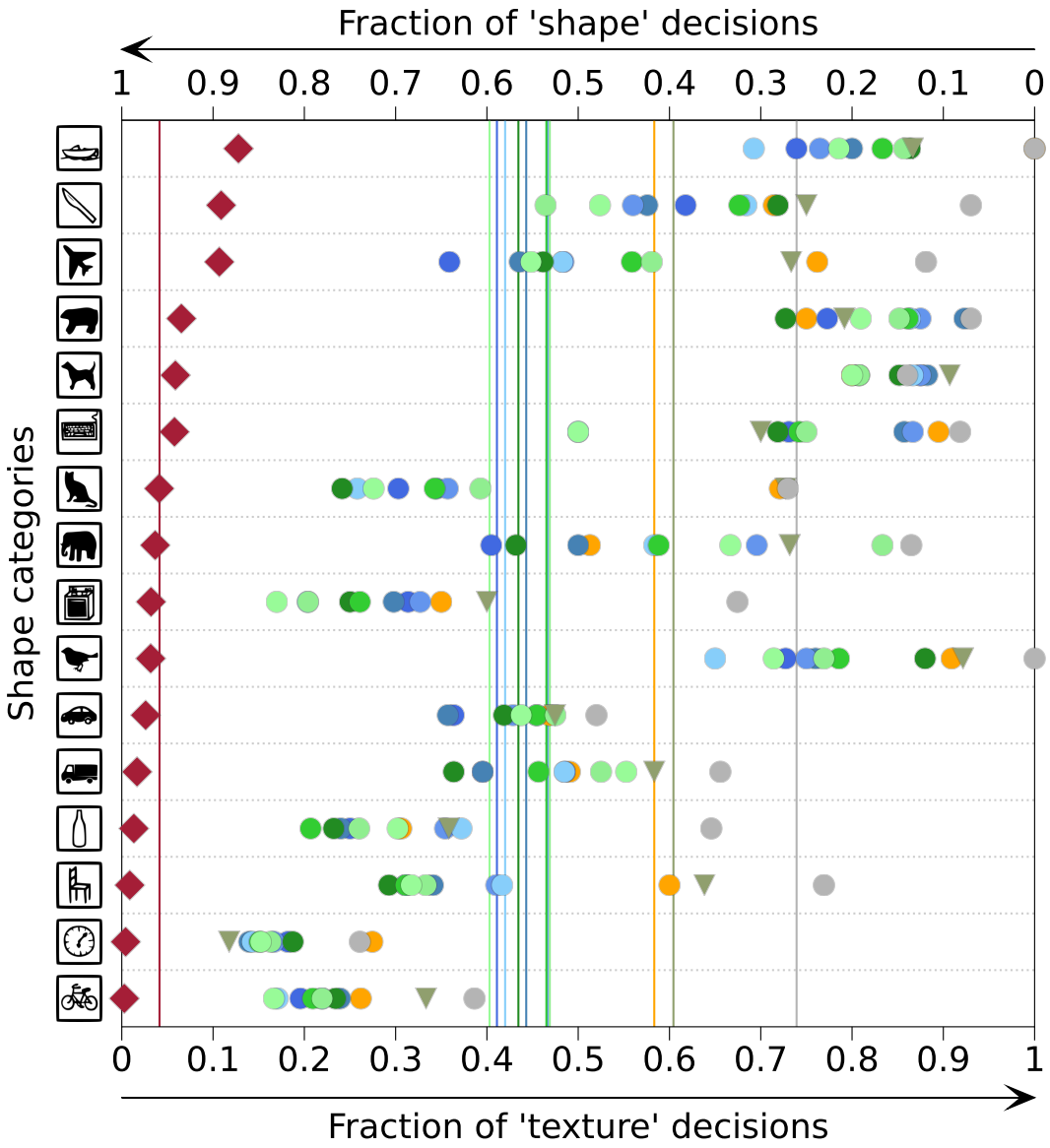


Figure 3: Shape vs. texture bias: category-level plot.

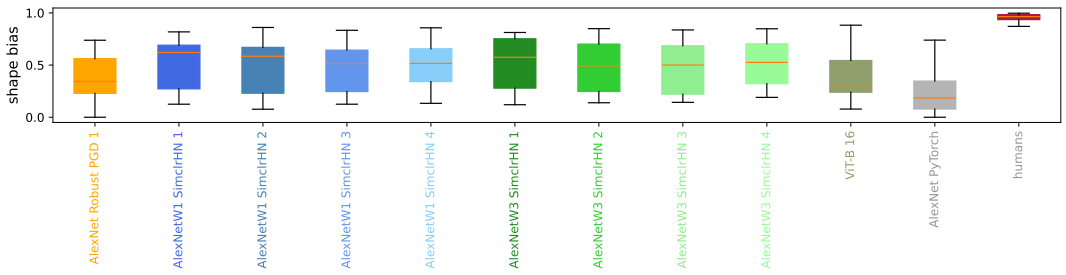
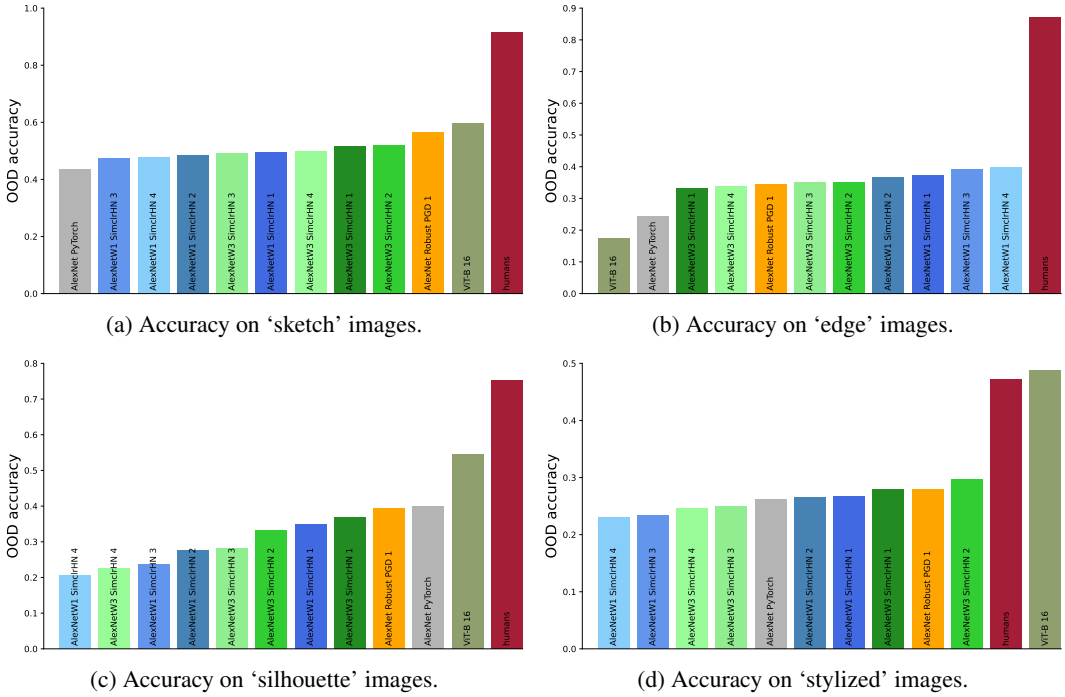


Figure 4: Shape vs. texture bias: boxplot.



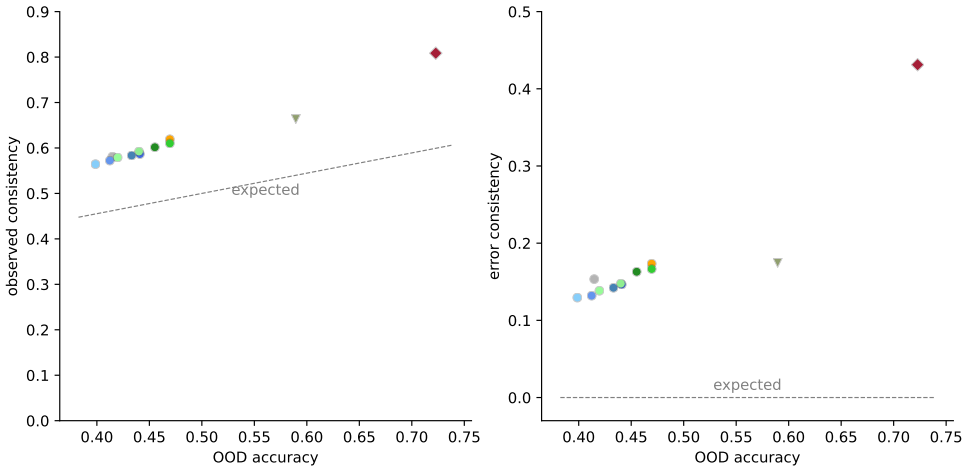
(a) Accuracy on 'sketch' images.

(b) Accuracy on 'edge' images.

(c) Accuracy on 'silhouette' images.

(d) Accuracy on 'stylized' images.

Figure 5: OOD accuracy on four nonparametric datasets (i.e., datasets with only a single corruption type and strength).



(a) Out-of-distribution accuracy vs. observed consistency

(b) Out-of-distribution accuracy vs. error consistency

Figure 6: Observed consistency and error consistency between models and humans as a function of out-of-distribution (OOD) accuracy. Dotted lines indicate consistency expected by chance.

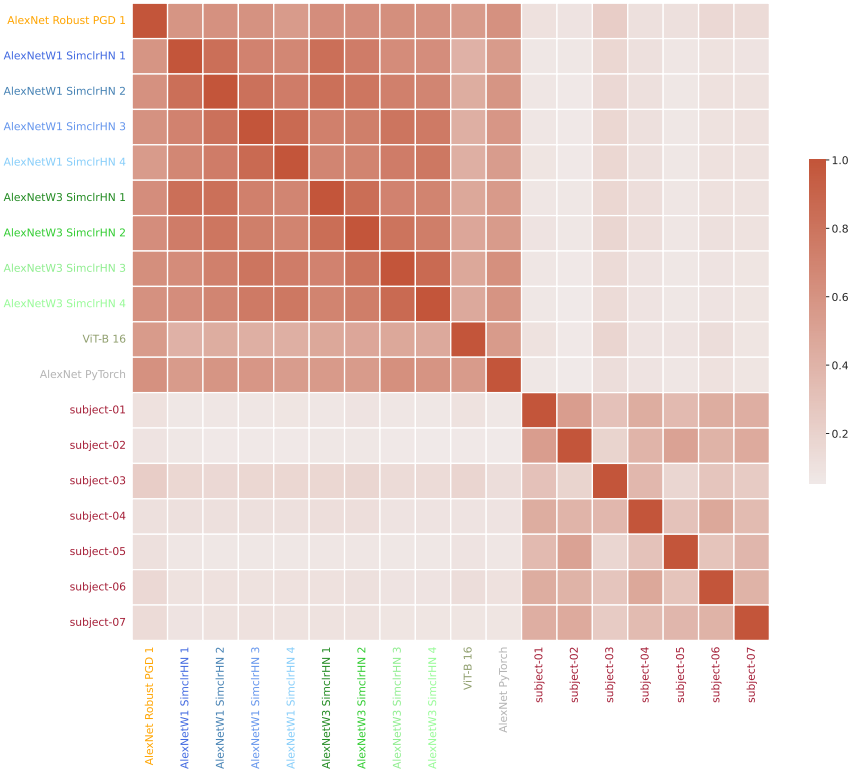


Figure 7: Error consistency for 'sketch' images.

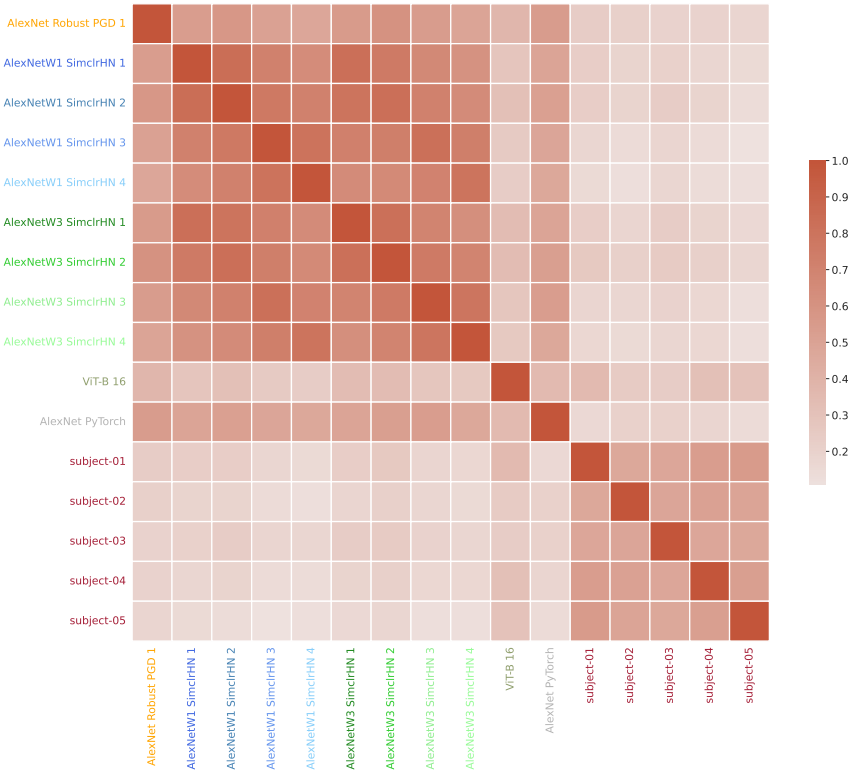


Figure 8: Error consistency for 'stylized' images.

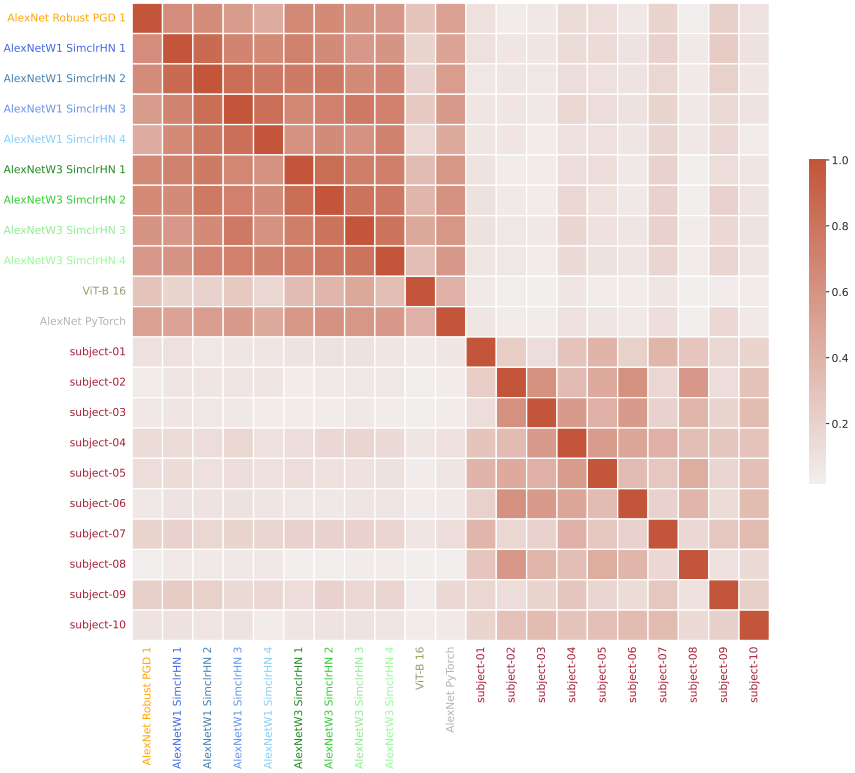


Figure 9: Error consistency for 'edge' images.

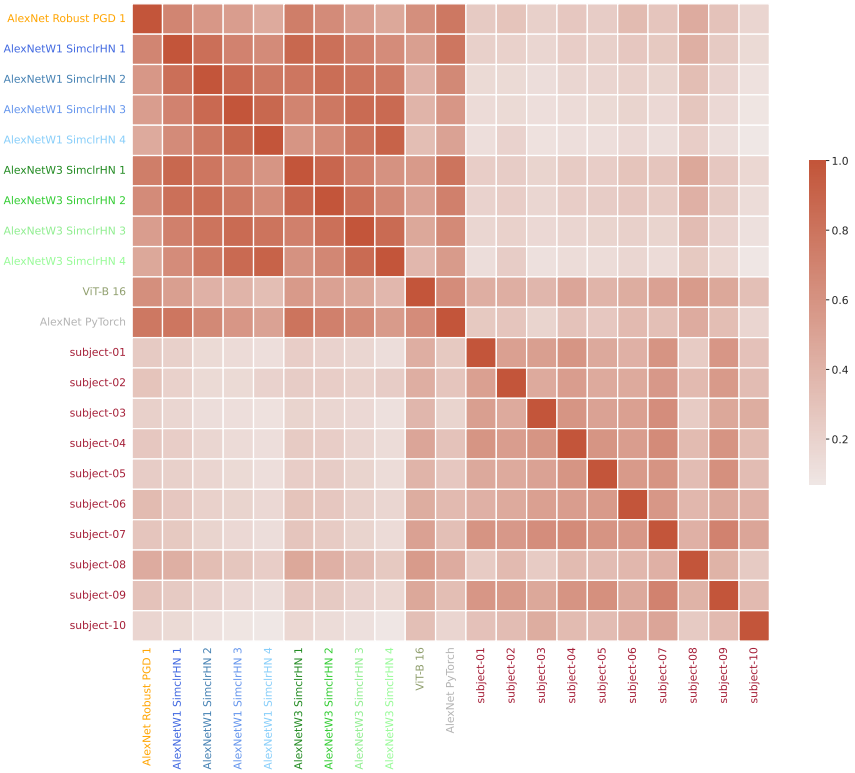


Figure 10: Error consistency for 'silhouette' images.

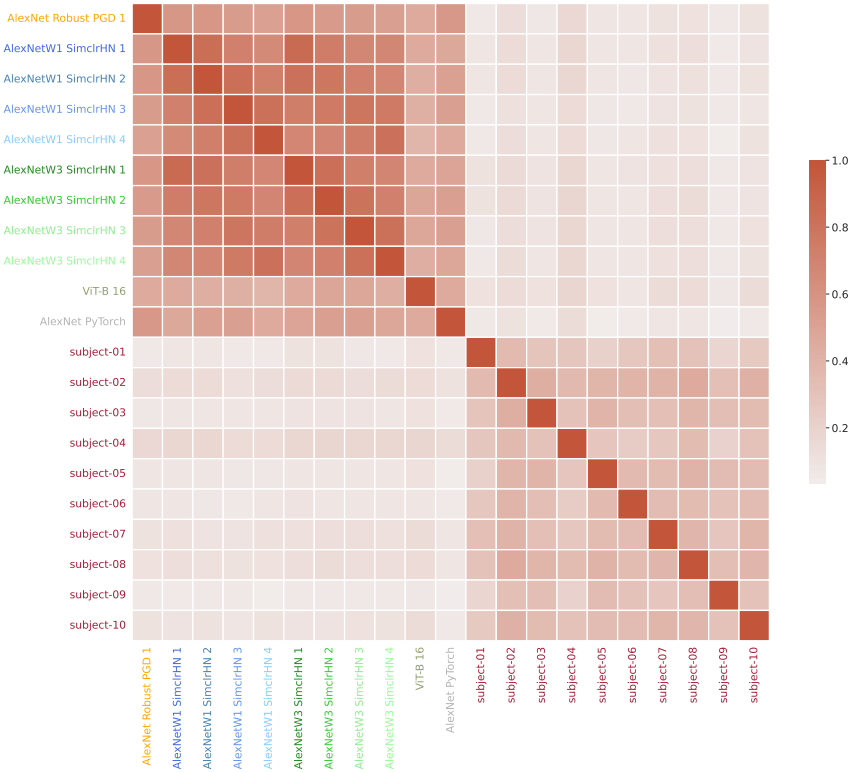


Figure 11: Error consistency for ‘cue conflict’ images.