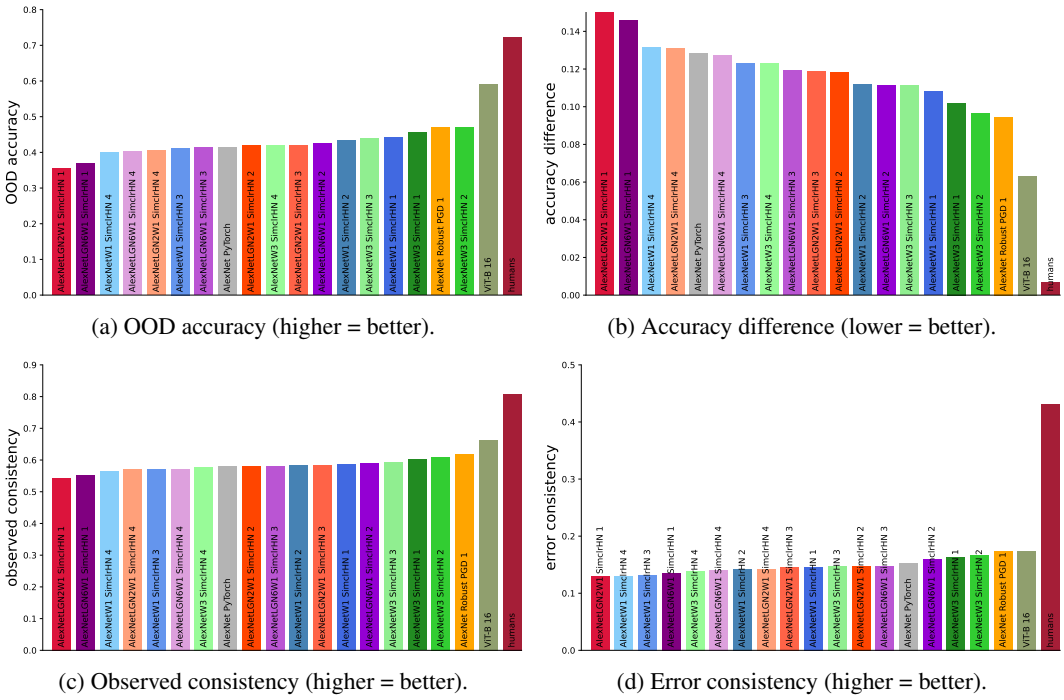
Generalisation report

Generalisation report produced by model-vs-human



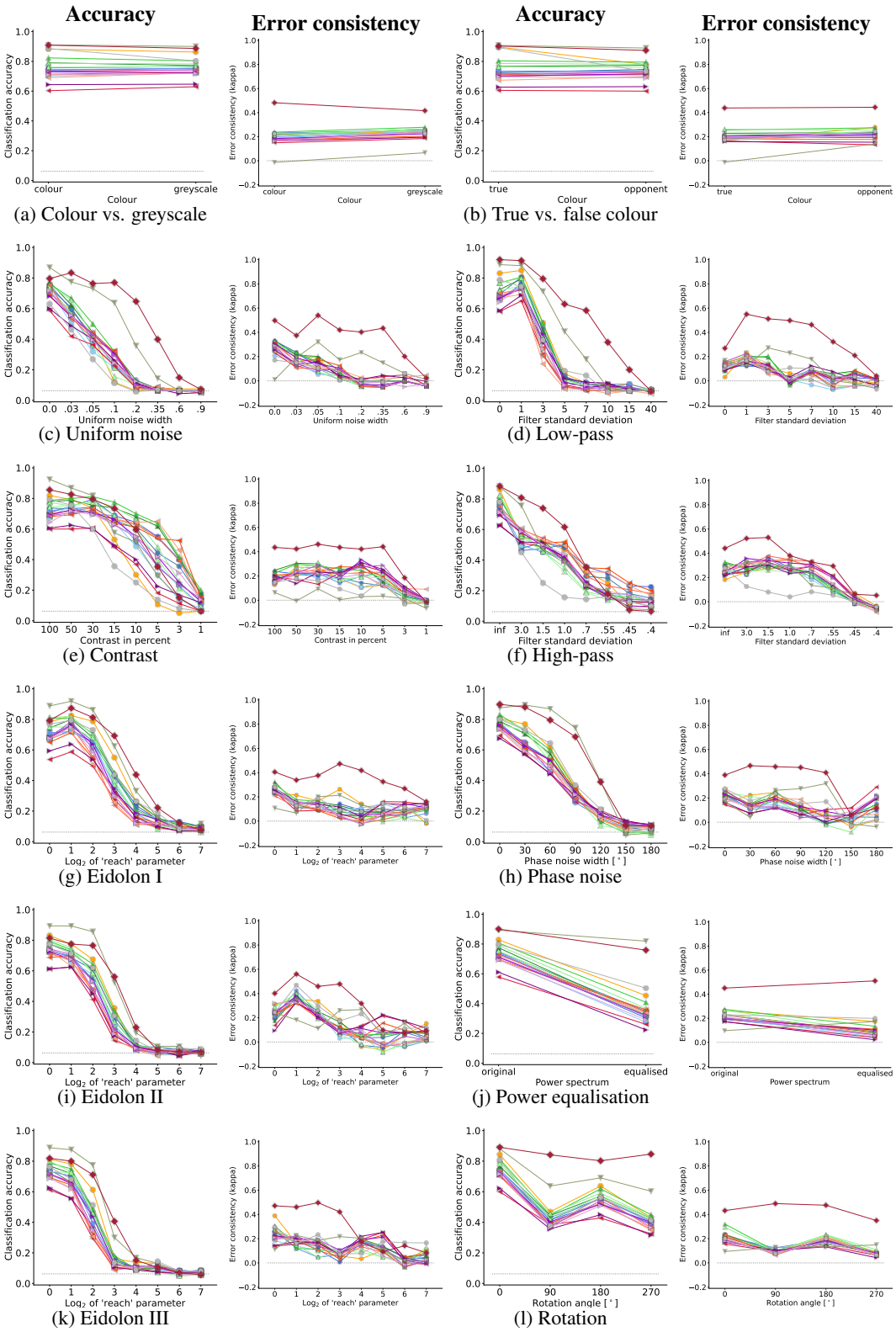
(a) OOD accuracy (higher = better).

(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.



(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.

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
AlexNetLGN6W1 SimclrHN 2	0.111	0.590	0.159	6.000
AlexNetW3 SimclrHN 3	0.111	0.592	0.148	6.667
AlexNetW1 SimclrHN 1	0.108	0.587	0.146	7.333
AlexNetLGN2W1 SimclrHN 2	0.118	0.582	0.148	9.333
AlexNetLGN6W1 SimclrHN 3	0.119	0.582	0.148	9.333
AlexNetLGN2W1 SimclrHN 3	0.119	0.584	0.146	9.667
AlexNetW1 SimclrHN 2	0.112	0.583	0.142	10.000
AlexNet PyTorch	0.128	0.581	0.153	11.000
AlexNetW3 SimclrHN 4	0.123	0.579	0.138	13.333
AlexNetLGN6W1 SimclrHN 4	0.127	0.573	0.141	14.000
AlexNetLGN2W1 SimclrHN 4	0.131	0.572	0.143	14.667
AlexNetW1 SimclrHN 3	0.123	0.572	0.132	15.000
AlexNetLGN6W1 SimclrHN 1	0.146	0.552	0.135	17.333
AlexNetW1 SimclrHN 4	0.131	0.564	0.129	17.333
AlexNetLGN2W1 SimclrHN 1	0.155	0.543	0.129	19.000

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
AlexNetLGN6W1 SimclrHN 2	0.426	8.000
AlexNetLGN2W1 SimclrHN 3	0.420	9.000
AlexNetW3 SimclrHN 4	0.420	10.000
AlexNetLGN2W1 SimclrHN 2	0.419	11.000
AlexNet PyTorch	0.415	12.000
AlexNetLGN6W1 SimclrHN 3	0.415	13.000
AlexNetW1 SimclrHN 3	0.412	14.000
AlexNetLGN2W1 SimclrHN 4	0.405	15.000
AlexNetLGN6W1 SimclrHN 4	0.403	16.000
AlexNetW1 SimclrHN 4	0.399	17.000
AlexNetLGN6W1 SimclrHN 1	0.369	18.000
AlexNetLGN2W1 SimclrHN 1	0.354	19.000

Table 3: Shape vs. texture bias: table.

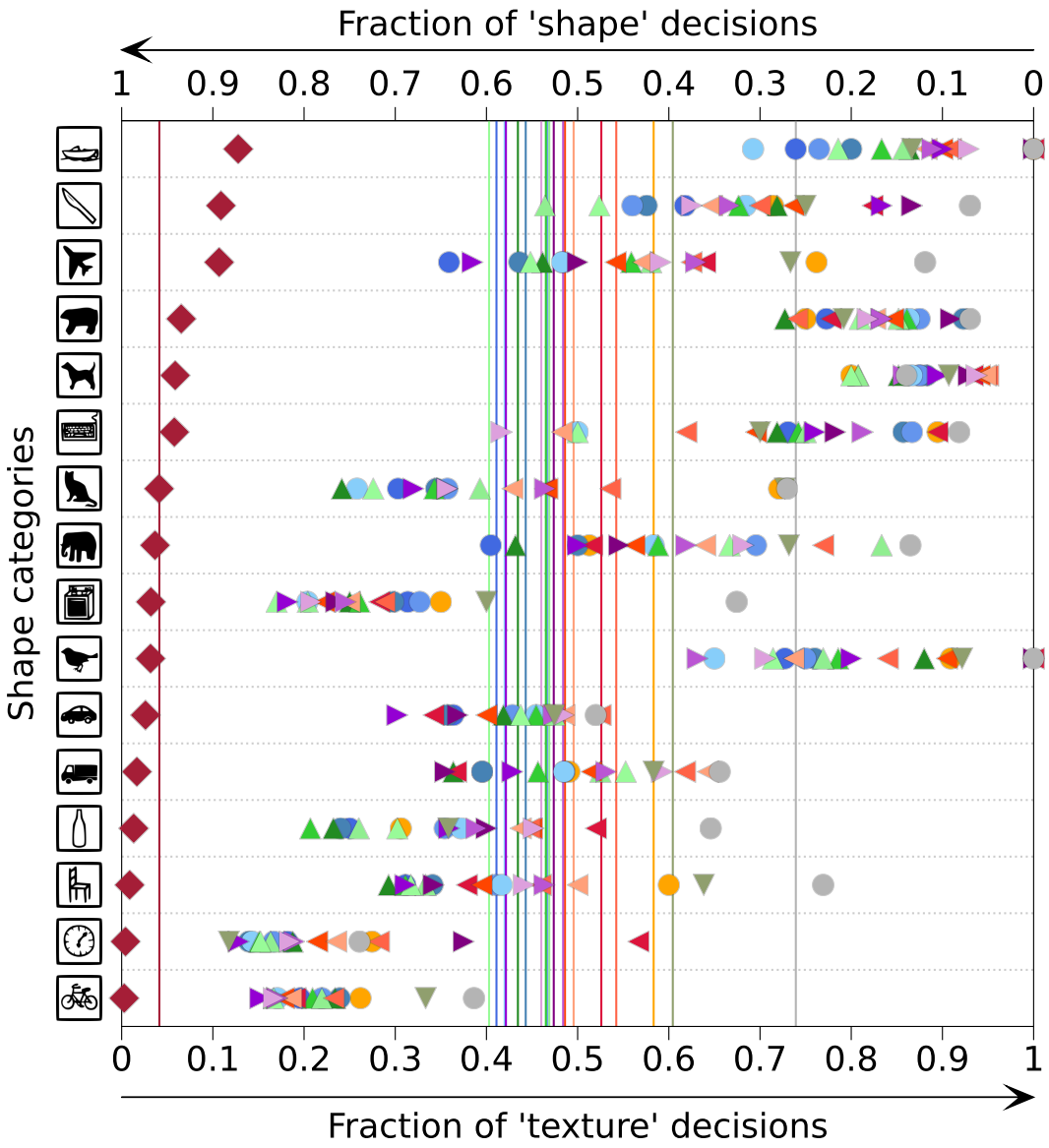


Figure 3: Shape vs. texture bias: category-level plot.

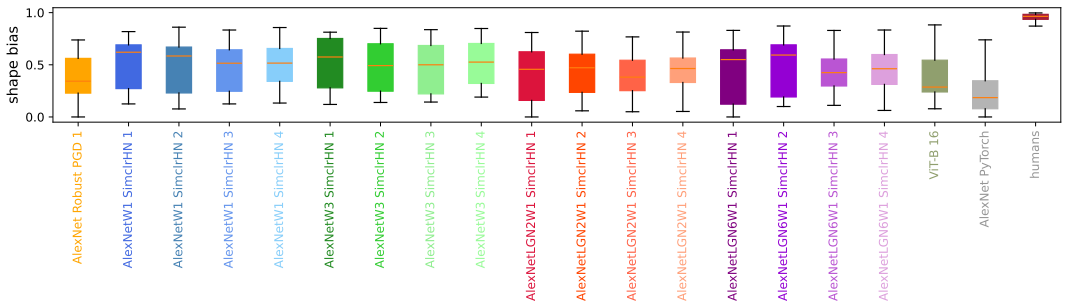
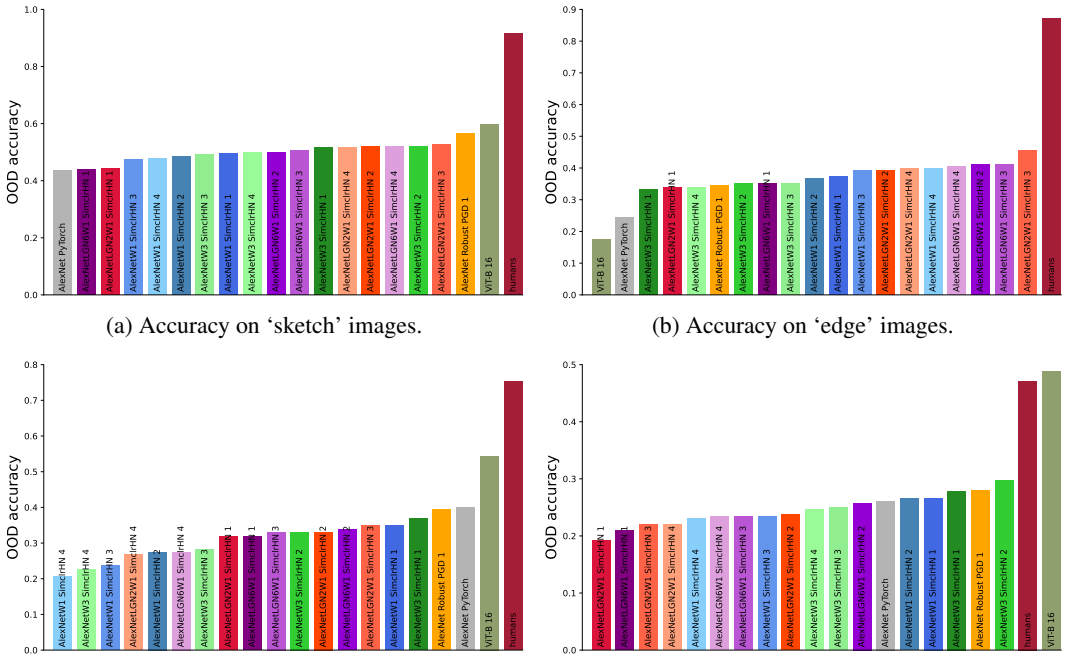


Figure 4: Shape vs. texture bias: boxplot.



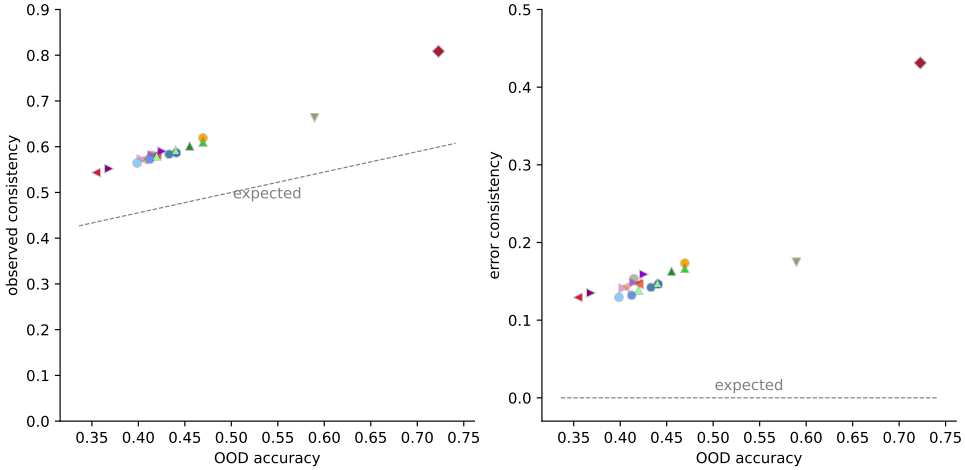
(a) Accuracy on 'sketch' images.

(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.

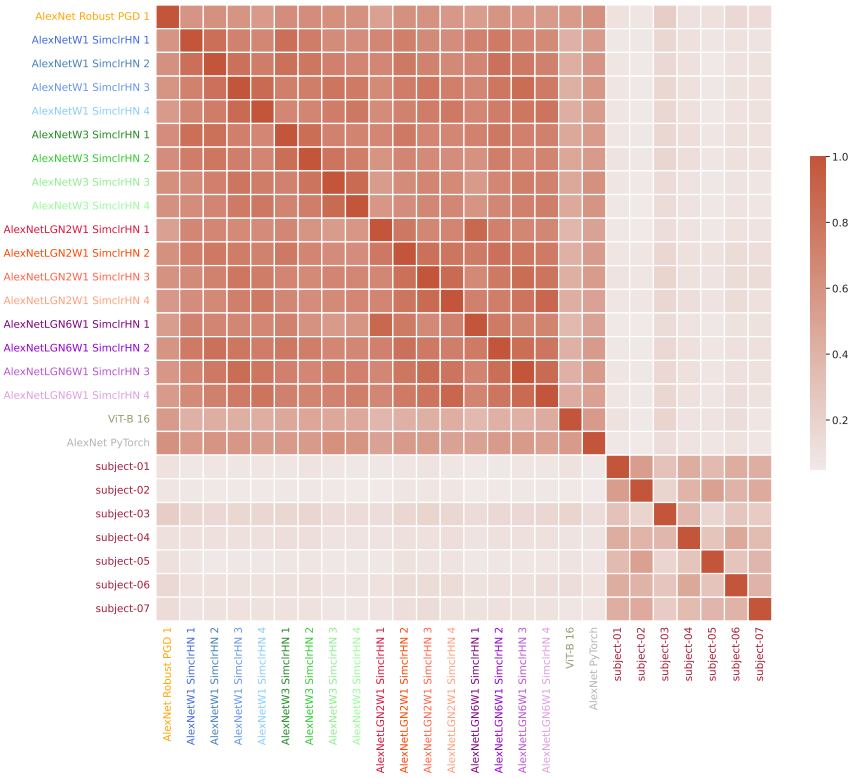


Figure 7: Error consistency for ‘sketch’ images.

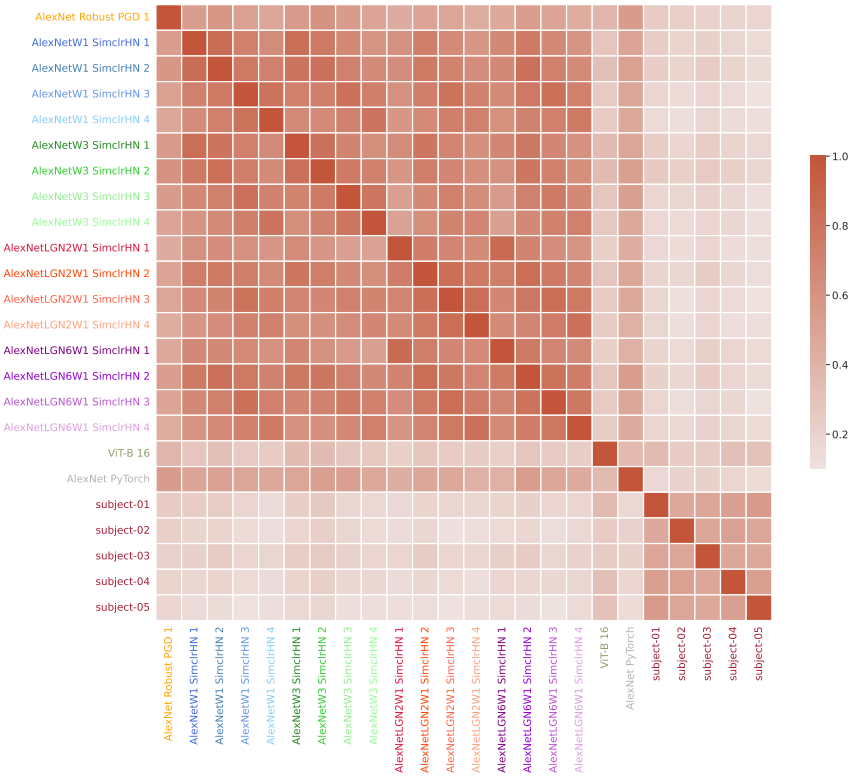


Figure 8: Error consistency for ‘stylized’ images.

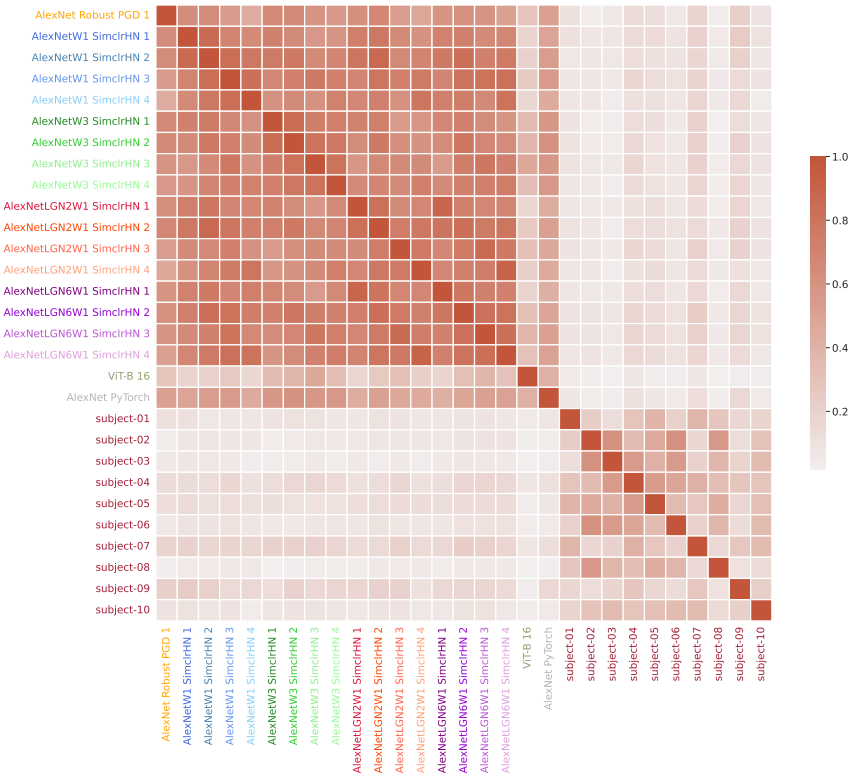


Figure 9: Error consistency for 'edge' images.

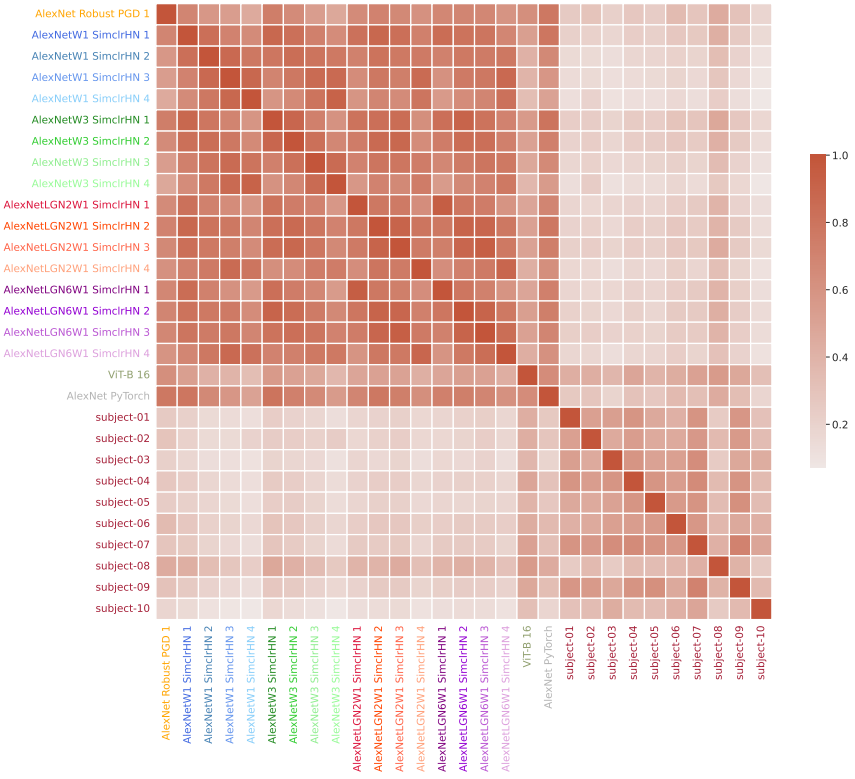


Figure 10: Error consistency for 'silhouette' images.

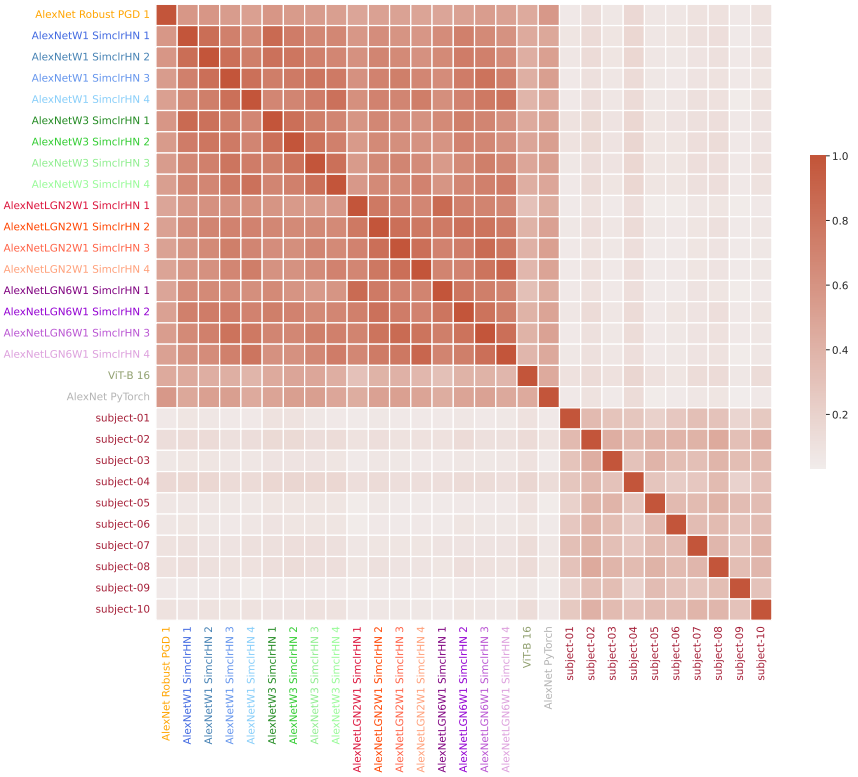


Figure 11: Error consistency for ‘cue conflict’ images.