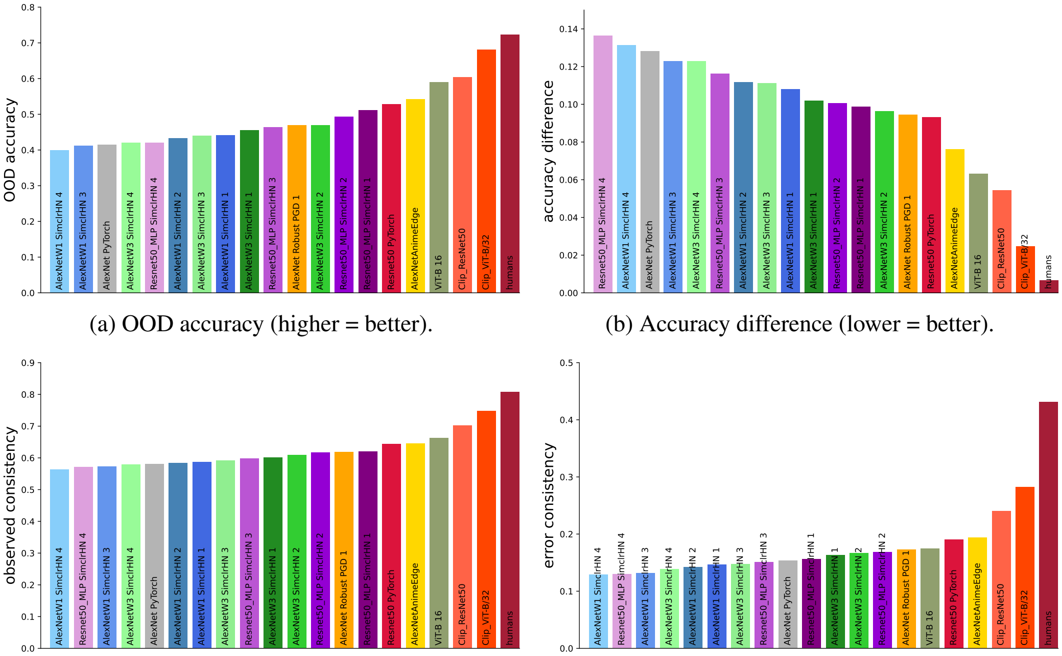
Generalisation report

Generalisation report produced by model-vs-human



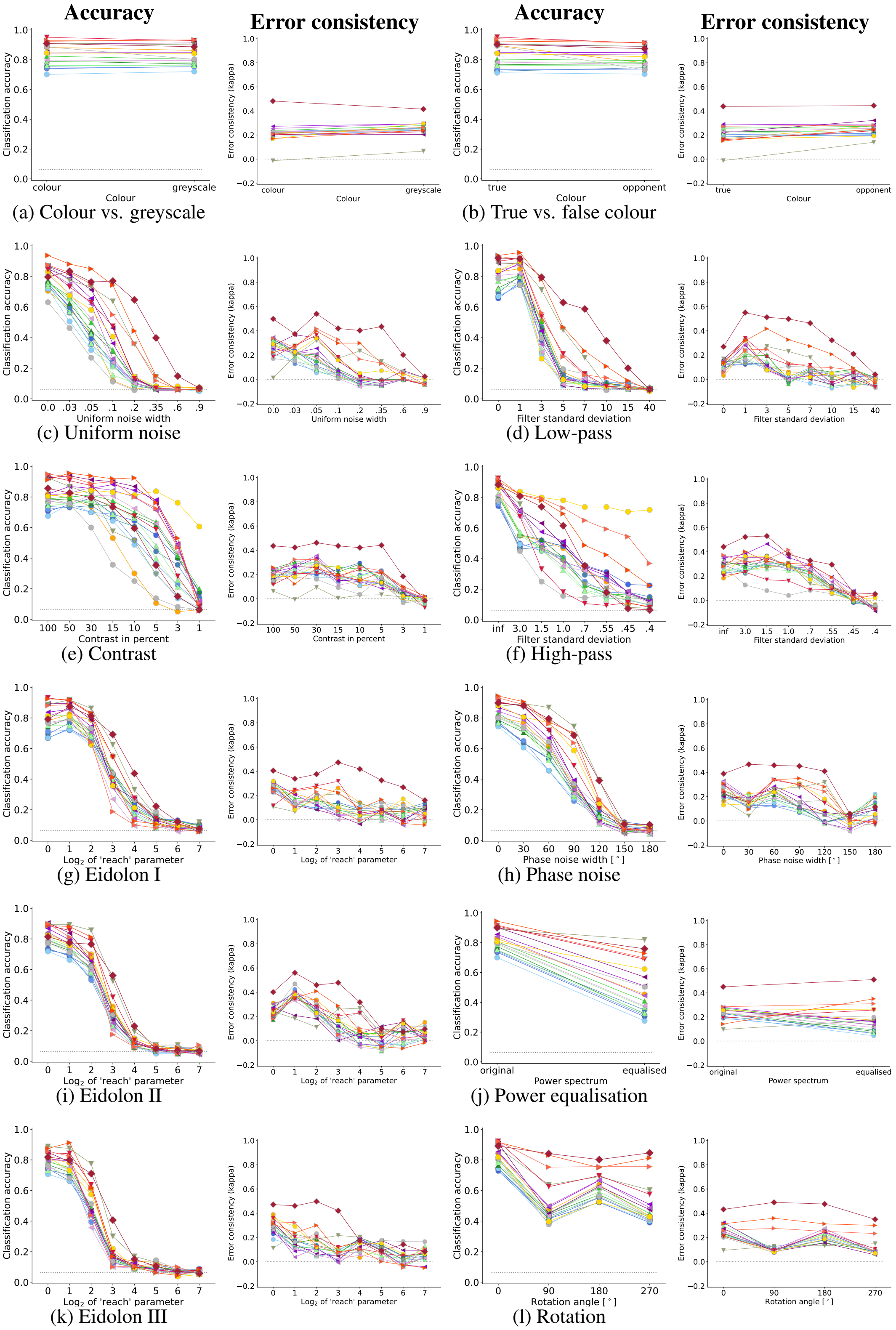
(a) OOD accuracy (higher = better).

(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 ↓
Clip_ViT-B/32	**0.025**	**0.748**	**0.283**	**1.000**
Clip_ResNet50	0.055	0.702	0.240	2.000
AlexNetAnimeEdge	0.076	0.646	0.194	3.667
ViT-B 16	0.063	0.664	0.175	3.667
Resnet50 PyTorch	0.093	0.644	0.190	4.667
AlexNet Robust PGD 1	0.095	0.619	0.173	6.333
Resnet50_MLP SimclrHN 2	0.101	0.618	0.169	8.000
AlexNetW3 SimclrHN 2	0.096	0.610	0.167	8.000
Resnet50_MLP SimclrHN 1	0.099	0.620	0.156	8.000
AlexNetW3 SimclrHN 1	0.102	0.602	0.163	9.667
Resnet50_MLP SimclrHN 3	0.116	0.599	0.151	12.333
AlexNetW3 SimclrHN 3	0.111	0.592	0.148	12.333
AlexNetW1 SimclrHN 1	0.108	0.587	0.146	12.667
AlexNetW1 SimclrHN 2	0.112	0.583	0.142	14.000
AlexNet PyTorch	0.128	0.581	0.153	14.333
AlexNetW3 SimclrHN 4	0.123	0.579	0.138	15.667
AlexNetW1 SimclrHN 3	0.123	0.572	0.132	16.667
Resnet50_MLP SimclrHN 4	0.137	0.572	0.130	18.333
AlexNetW1 SimclrHN 4	0.131	0.564	0.129	18.667

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

model	OOD accuracy ↑	rank ↓
Clip_ViT-B/32	**0.681**	**1.000**
Clip_ResNet50	0.604	2.000
ViT-B 16	0.590	3.000
AlexNetAnimeEdge	0.542	4.000
Resnet50 PyTorch	0.528	5.000
Resnet50_MLP SimclrHN 1	0.512	6.000
Resnet50_MLP SimclrHN 2	0.493	7.000
AlexNetW3 SimclrHN 2	0.470	8.000
AlexNet Robust PGD 1	0.470	9.000
Resnet50_MLP SimclrHN 3	0.463	10.000
AlexNetW3 SimclrHN 1	0.455	11.000
AlexNetW1 SimclrHN 1	0.441	12.000
AlexNetW3 SimclrHN 3	0.440	13.000
AlexNetW1 SimclrHN 2	0.433	14.000
Resnet50_MLP SimclrHN 4	0.421	15.000
AlexNetW3 SimclrHN 4	0.420	16.000
AlexNet PyTorch	0.415	17.000
AlexNetW1 SimclrHN 3	0.412	18.000
AlexNetW1 SimclrHN 4	0.399	19.000

Table 3: Shape vs. texture bias: table.

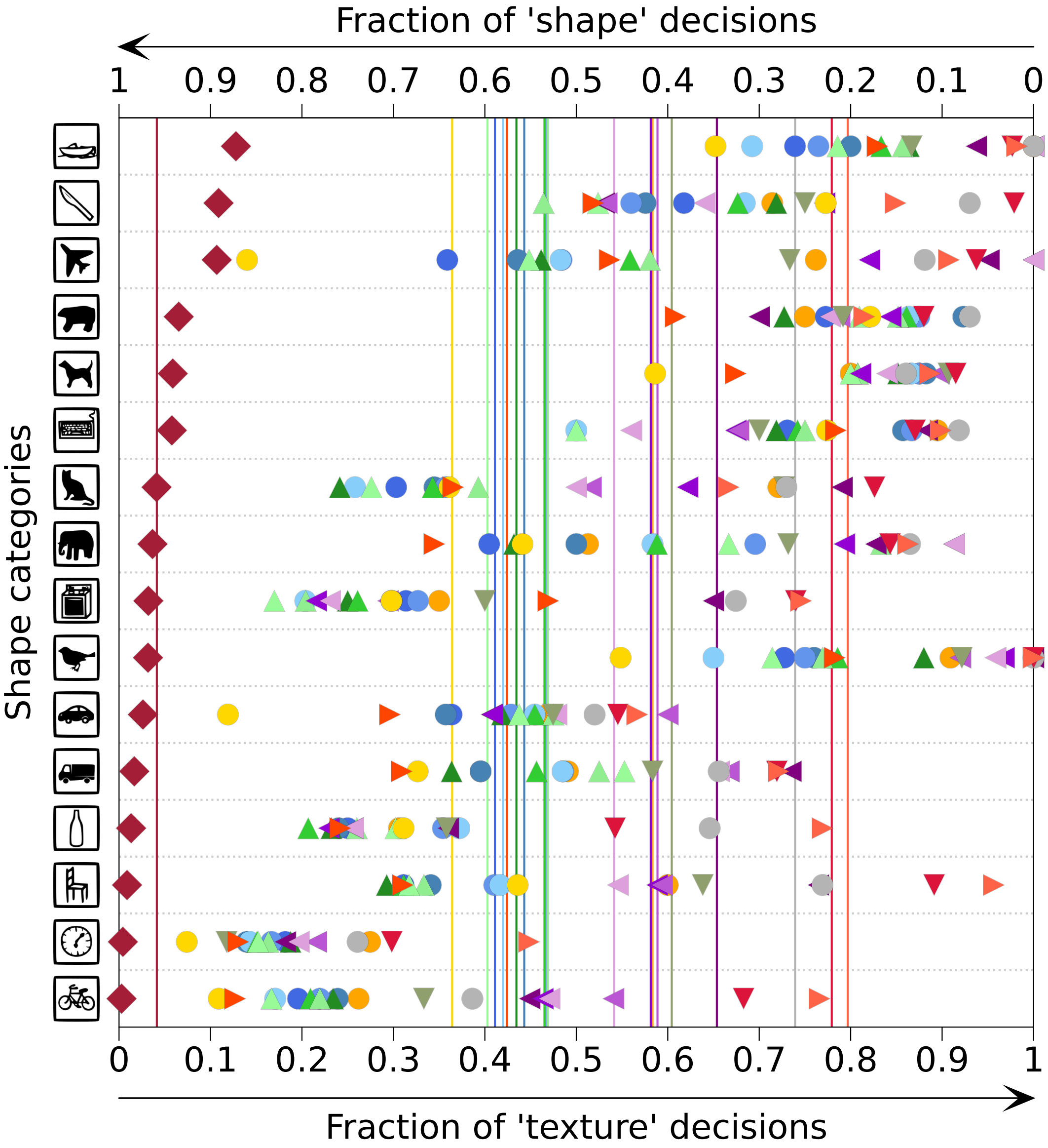


Figure 3: Shape vs. texture bias: category-level plot.

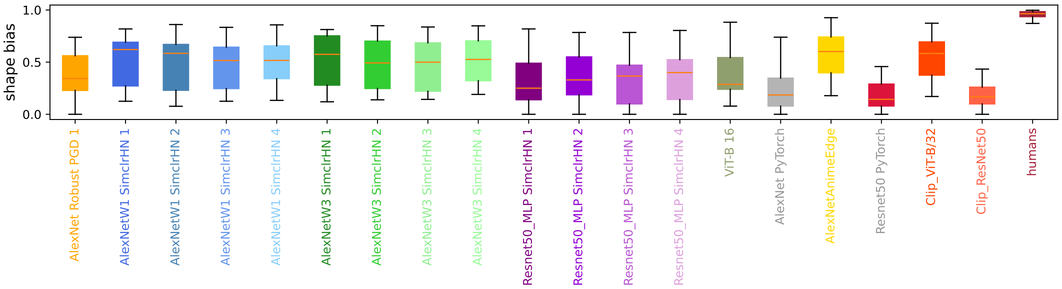
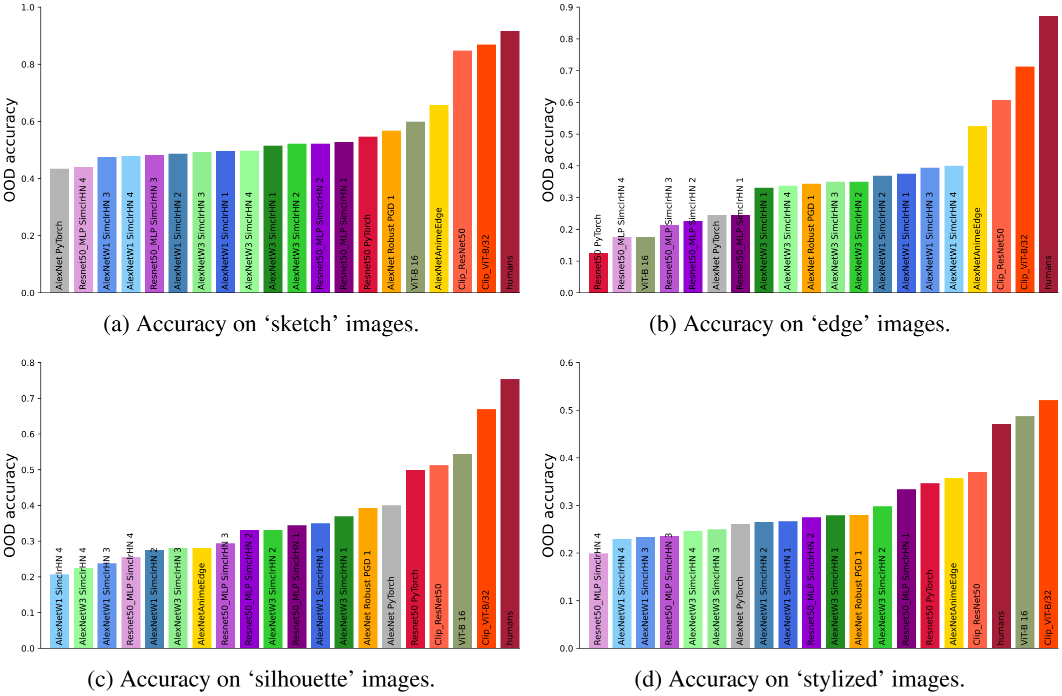


Figure 4: Shape vs. texture bias: boxplot.



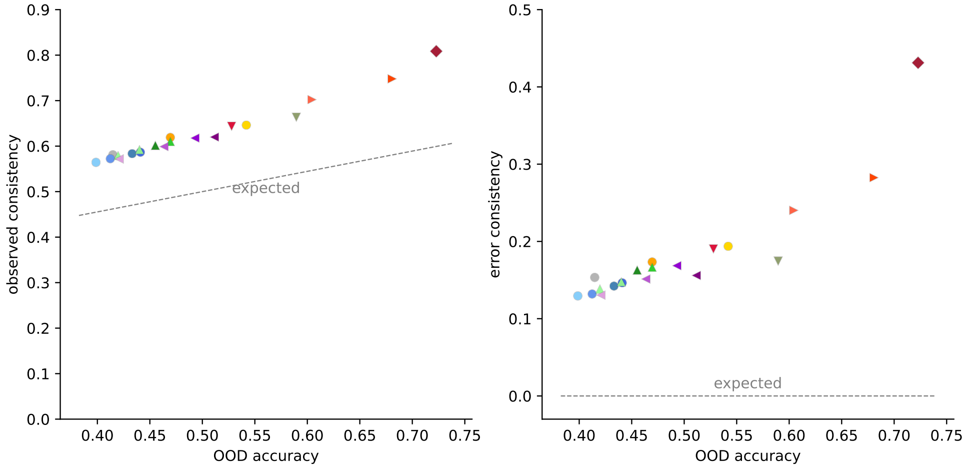
(a) Accuracy on ‘sketch’ images.

(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.

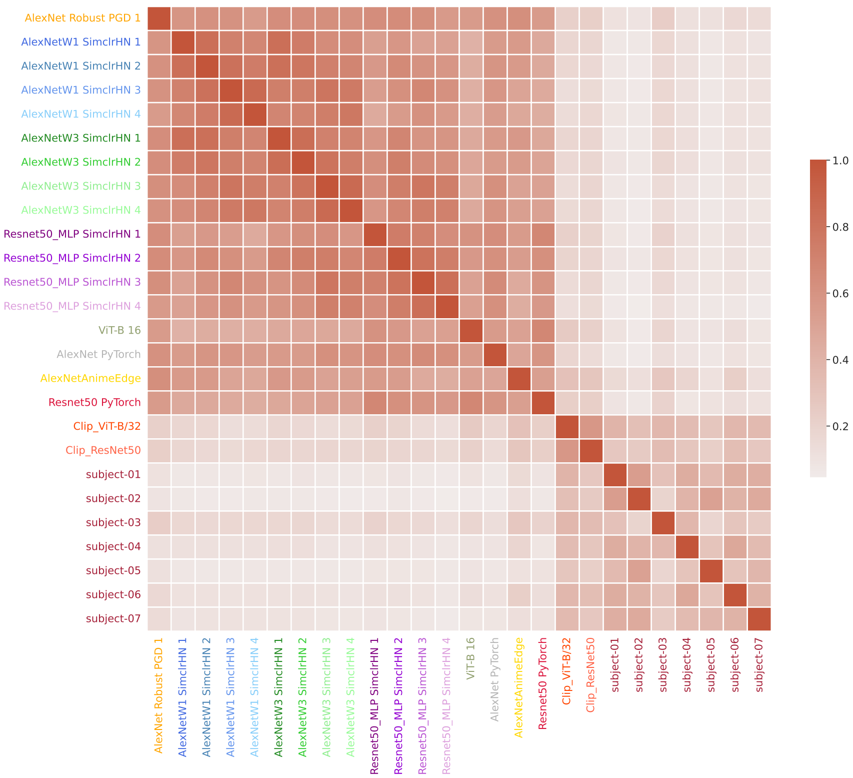


Figure 7: Error consistency for ‘sketch’ images.

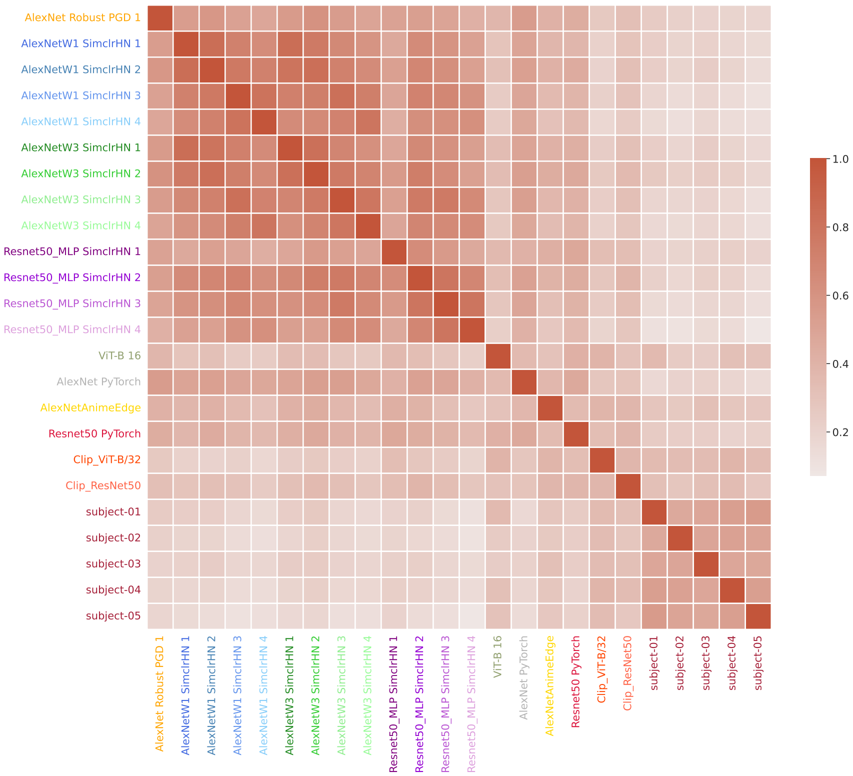


Figure 8: Error consistency for ‘stylized’ images.

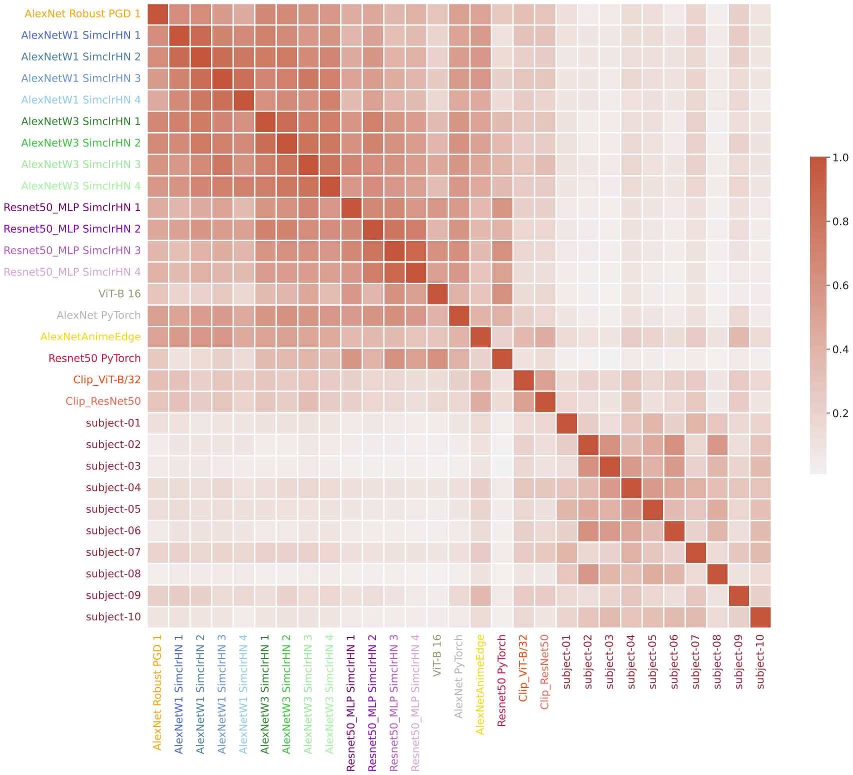


Figure 9: Error consistency for ‘edge’ images.

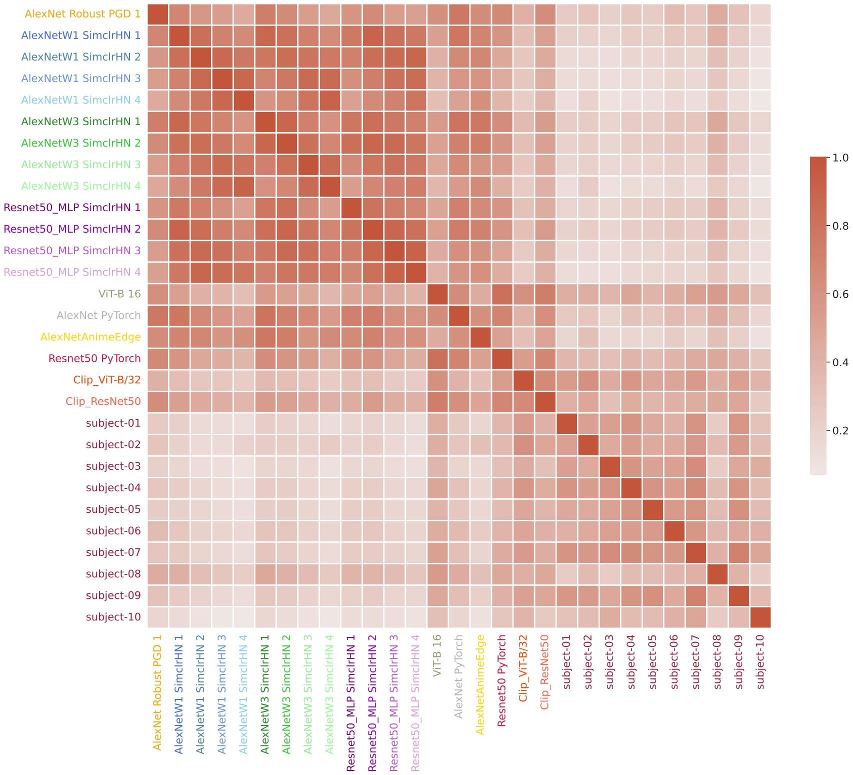


Figure 10: Error consistency for ‘silhouette’ images.

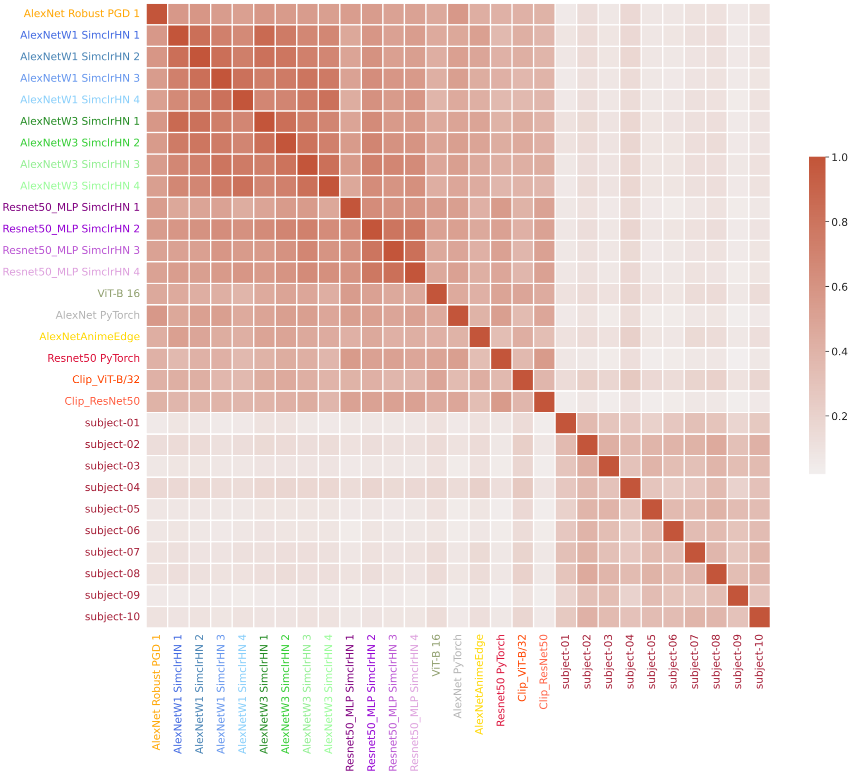


Figure 11: Error consistency for 'cue conflict' images.