R_{score} = 0.30	Other	Num	Y/N	All	Diff
dev: partition 1	33.98	33.50	73.22	**49.96**	**10.13**
dev: partition 2	32.44	34.47	72.22	**48.67**	**11.18**
dev: partition 3	32.65	33.60	71.76	**46.27**	**11.36**
dev: partition 4	32.77	33.79	71.14	**46.30**	**11.53**
dev: partition 5	32.46	33.51	70.90	**46.41**	**11.81**
dev: partition 6	33.02	33.18	69.88	**46.22**	**12.00**
dev: partition 7	32.73	33.28	69.74	**45.57**	**12.18**
std: partition 1	34.06	33.24	72.99	49.93	10.43
dev: partition 0	47.16	37.32	81.45	**60.16**	-
std: partition 0	47.57	36.75	81.56	**60.45**	-

(a) MUTAN without Attention model evaluation results.

R_{score} = 0.48	Other	Num	Y/N	All	Diff
dev: partition 1	40.80	30.34	76.92	**54.49**	**5.99**
dev: partition 2	39.63	30.67	76.49	**53.78**	**6.70**
dev: partition 3	39.33	31.12	75.48	**53.28**	**7.20**
dev: partition 4	39.31	29.78	75.12	**52.97**	**7.51**
dev: partition 5	39.38	29.87	74.96	**52.95**	**7.53**
dev: partition 6	39.13	30.74	73.95	**52.51**	**7.97**
dev: partition 7	38.90	31.14	73.80	**52.39**	**8.09**
std: partition 1	40.88	28.82	76.67	54.37	5.95
dev: partition 0	49.14	38.35	79.63	**60.48**	-
std: partition 0	49.15	36.52	79.45	**60.32**	-

(b) HieCoAtt (Alt, VGG19) model evaluation results.

R_{score} = 0.37	Other	Num	Y/N	All	Diff
dev: partition 1	46.57	32.09	76.60	**57.33**	**8.46**
dev: partition 2	45.83	32.43	75.29	**56.47**	**9.32**
dev: partition 3	45.17	32.52	74.87	**55.99**	**9.80**
dev: partition 4	45.11	32.31	73.73	**55.47**	**10.32**
dev: partition 5	44.35	31.95	72.93	**54.74**	**11.05**
dev: partition 6	43.75	31.21	72.03	**54.00**	**11.79**
dev: partition 7	43.88	32.59	71.99	**54.19**	**11.60**
std: partition 1	46.11	31.46	76.84	57.25	8.43
dev: partition 0	57.01	37.51	83.54	**65.79**	-
std: partition 0	56.60	36.63	83.68	**65.68**	-

(c) MLB with Attention model evaluation results.

R_{score} = 0.23	Other	Num	Y/N	All	Diff
dev: partition 1	43.96	28.90	71.89	**53.79**	**12.19**
dev: partition 2	42.66	28.08	70.05	**52.32**	**13.66**
dev: partition 3	41.62	29.12	69.58	**51.74**	**14.24**
dev: partition 4	41.53	29.30	67.96	**51.06**	**14.92**
dev: partition 5	40.46	27.66	68.03	**50.39**	**15.59**
dev: partition 6	40.03	28.44	66.98	**49.84**	**16.14**
dev: partition 7	39.11	28.41	67.44	**49.58**	**16.40**
std: partition 1	43.55	28.70	71.76	53.63	12.14
dev: partition 0	56.73	38.35	84.11	**65.98**	-
std: partition 0	56.29	37.47	84.04	**65.77**	-

(d) MUTAN with Attention model evaluation results.

R_{score} = **0.53**	Other	Num	Y/N	All	Diff
dev: partition 1	44.42	36.39	76.94	**56.90**	**4.91**
dev: partition 2	43.37	34.99	76.10	**55.90**	**5.91**
dev: partition 3	42.22	33.97	75.80	**55.11**	**6.70**
dev: partition 4	42.52	34.21	75.33	**55.09**	**6.72**
dev: partition 5	42.81	34.69	75.21	**55.23**	**6.58**
dev: partition 6	42.27	35.16	74.50	**54.73**	**7.08**
dev: partition 7	41.95	35.14	73.64	**54.22**	**7.59**
std: partition 1	44.93	35.59	76.82	57.10	4.96
dev: partition 0	51.77	38.65	79.70	**61.81**	-
std: partition 0	51.95	38.22	79.95	**62.06**	-

(e) HieCoAtt (Alt, Resnet200) model evaluation results.

R_{score} = 0.08	Other	Num	Y/N	All	Diff
dev: partition 1	20.49	25.98	68.79	**40.91**	**17.11**
dev: partition 2	19.81	25.40	68.51	**40.40**	**17.62**
dev: partition 3	18.58	24.95	68.53	**39.77**	**18.25**
dev: partition 4	18.50	24.82	67.83	**39.43**	**18.59**
dev: partition 5	17.68	24.68	67.99	**39.09**	**18.93**
dev: partition 6	17.29	24.03	67.76	**38.73**	**19.29**
dev: partition 7	16.93	24.63	67.45	**38.50**	**19.52**
std: partition 1	20.84	26.14	68.88	41.19	16.99
dev: partition 0	43.40	36.46	80.87	**58.02**	-
std: partition 0	43.90	36.67	80.38	**58.18**	-

(f) LSTM Q+I model evaluation results.

Table 3: Compares the accuracy and R_{score} of six VQA models with increasing noise levels from YNBQD generated by *LASSO* evaluated on dev and std. The results are split by the question type; Numerical (Num), Yes/No (Y/N), or Other. We notice that the R_{score} of some models under YNBQD is better than GBQD (*e.g.*, HAR) and vice versa (*e.g.*, MUA). So, the R_{score} can only be compared in different models if they have the same BQD because it can show certain biases in the models under investigation that are related to the type of the BQD. For example, we can clearly see how HAR is more robust towards Yes/No questions than general questions. Whereas, HAV is apparently agnostic to this property (*i.e.*, the answer type being Yes/No versus general).

corporis perferendis eum temporibus aut quae est sit, enim culpa doloribus?Eos porro unde suscipit ullam sint similique ipsam temporibus, debitis sit illum praesentium recusandae.