	Validation (1,901)	Test (1,906)	Similarity (136)	Counting (120)	Depth (124)	Jigsaw (150)	Art (117)	Fun.Corr. (130)
Random Choice	38.09	38.09	50	25	50	50	50	25
Human	95.67	95.70	96.70	93.75	99.19	99.00	95.30	80.77
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	34.23	34.57	52.94	10.83	49.19	26.00	47.86	18.46
OpenFlamingo-v2 [5]	39.18	38.32	55.15	21.67	54.03	46.00	52.14	36.15
InstructBLIP-7B [24]	39.72	38.65	46.32	29.17	50.81	54.00	47.86	23.85
InstructBLIP-13B [24]	42.24	39.58	46.32	30.83	50.00	54.00	50.43	22.31
LLaVA-internLM2-7B [72]	37.71	36.06	52.94	52.50	52.42	34.67	30.77	23.08
Yi-VL-6B [2]	38.72	41.24	46.32	46.67	56.45	50.00	53.85	23.85
Yi-VL-34B [2]	41.68	42.78	50.00	58.33	53.23	54.00	46.15	**39.23**
LLaVA-v1.5-7B-xtuner [23]	39.36	40.81	46.32	53.33	50.81	54.00	47.86	23.85
LLaVA-v1.5-13B-xtuner [23]	42.00	41.31	46.32	45.00	54.03	53.33	47.86	26.15
CogVLM [77]	41.54	39.38	46.32	38.33	50.81	52.67	49.57	23.85
LLaVA-v1.5-7B [48]	37.13	38.01	46.32	43.33	51.61	11.33	47.86	21.54
LLaVA-v1.5-13B [48]	42.66	40.55	46.32	50.00	47.58	54.00	47.86	20.77
LLaVA-v1.6-34B [50]	46.80	45.05	46.32	68.33	**64.52**	56.67	47.01	30.77
API-based models								
Qwen-VL-Max [7]	40.28	41.94	51.47	55.83	58.87	3.33	37.61	28.46
Gemini Pro [71]	45.16	45.72	55.88	65.00	50.00	54.00	49.57	32.31
Claude 3 OPUS [1]	44.05	44.11	70.59	49.17	57.26	32.67	60.68	22.31
GPT-4V(ision) [62]	**51.14**	**51.26**	**83.09**	**60.83**	58.87	**62.67**	**78.63**	31.54

	Sem.Corr. (140)	Spatial (143)	Local. (125)	Vis.Corr. (172)	Multi-view (133)	Reflect. (134)	Forensic (132)	IQ (150)
Random Choice	25	50	50	25	50	33.33	25	25
Human	96.07	98.25	98.00	99.42	92.48	95.14	100.00	80.00
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	26.43	51.75	**56.00**	23.84	52.63	31.34	17.42	19.33
OpenFlamingo-v2 [5]	23.57	46.85	52.00	25.00	41.35	43.28	15.91	23.33
InstructBLIP-7B [24]	25.00	55.24	44.80	22.67	**58.65**	29.85	29.55	23.33
InstructBLIP-13B [24]	22.86	64.34	52.00	20.93	54.14	46.27	13.64	26.00
LLaVA-internLM2-7B [72]	22.14	74.13	48.00	21.51	41.35	32.84	3.79	14.67
Yi-VL-6B [2]	26.43	72.73	49.60	29.65	48.12	29.85	20.45	23.33
Yi-VL-34B [2]	21.43	70.63	54.40	23.84	41.35	46.27	17.42	22.67
LLaVA-v1.5-7B-xtuner [23]	24.29	74.83	45.60	23.84	42.11	26.87	36.36	21.33
LLaVA-v1.5-13B-xtuner [23]	22.14	**77.62**	48.00	22.09	41.35	46.27	29.55	18.67
CogVLM [77]	23.57	67.13	43.20	20.93	57.14	26.87	24.24	26.67
LLaVA-v1.5-7B [48]	32.14	70.63	48.80	20.35	49.62	36.57	28.03	24.00
LLaVA-v1.5-13B [48]	23.57	67.83	47.20	20.35	41.35	45.52	27.27	**28.00**
LLaVA-v1.6-34B [50]	27.86	76.22	41.60	27.33	46.62	29.85	41.67	26.00
API-based models								
Qwen-VL-Max [7]	29.29	**77.62**	49.60	22.67	53.38	**49.25**	47.73	22.00
Gemini Pro [71]	22.14	67.13	46.40	**37.21**	41.35	46.27	45.45	27.33
Claude 3 OPUS [1]	20.71	57.34	46.40	31.40	57.89	27.61	**62.12**	21.33
GPT-4V(ision) [62]	**30.00**	72.03	50.40	**37.21**	**58.65**	38.81	30.30	24.67