

Table ujCi.1: Results on three tasks with data generated by different MLLMs. Due to the quality of generated data by InternLM-XComposer2 is not satisfying in tasks of Truthfulness, we only use it for task P.2.

Source-Model	T.4 (Weighted Score)			T.5 (Accuracy, %)			P.2 (Accuracy, %)			
	GPT-4	Gemini	Hunyuan-V	GPT-4	Gemini	Hunyuan-V	GPT-4	Gemini	InternLM-XC2	Hunyuan-V
GPT-4-Vision	86.5	83.2	88.8	61.0	60.0	58.0	79.0	83.2	67.4	74.6
Claude3	64.7	76.2	85.7	55.0	46.0	48.0	67.1	68.8	65.4	65.0
Gemini-Pro	84.0	86.6	88.6	65.0	52.0	49.0	73.7	71.4	59.8	63.4
Qwen-VL-Plus	62.6	65.9	68.8	42.0	27.0	47.0	56.0	61.4	50.8	53.8
LLaVA-1.5-7B	60.7	61.8	82.4	26.0	20.0	34.0	41.9	36.6	43.6	48.4
LLaVA-1.5-13B	70.8	64.2	82.3	30.0	26.0	40.0	49.5	47.0	47.6	49.2
ShareGPT4V	66.8	61.8	79.7	28.0	16.0	37.0	49.5	50.0	49.2	50.4
LVIS-Instruct4V	63.7	54.9	79.9	23.0	20.0	41.0	57.4	52.4	48.8	58.2
LLaVA-RLHF	68.7	64.3	78.9	25.0	23.0	39.0	57.1	48.2	49.4	46.8
LLaVA-NeXT	70.0	64.6	77.8	47.0	41.0	43.0	53.7	51.8	49.4	52.0
MiniGPT-4-13B	61.2	53.3	68.9	39.0	26.0	36.0	49.5	49.2	49.8	49.0
MiniGPT-4-L2	66.1	54.2	79.8	36.0	26.0	28.0	46.1	50.0	50.0	50.0
mPLUG-Owl-7b	62.3	63.6	81.2	14.0	17.0	35.0	50.0	49.2	50.0	50.0
mPLUG-Owl2-7b	62.7	54.2	78.5	24.0	21.0	32.0	50.9	52.6	50.8	52.6
InstructBLIP	47.7	43.5	55.9	21.0	29.0	37.0	62.9	65.0	62.4	57.0
Otter	45.5	43.4	63.0	16.0	17.0	20.0	52.3	49.6	50.4	51.6
CogVLM	59.2	47.3	73.5	23.0	24.0	44.0	44.0	46.2	48.0	48.2
Qwen-VL-Chat	59.4	53.8	77.9	29.0	28.0	37.0	53.0	51.8	52.0	54.4
InternVL-Chat	71.1	67.3	82.0	27.0	23.0	38.0	51.6	47.8	48.8	51.0
InternLM-XC	62.5	64.3	75.7	20.0	23.0	32.0	50.0	51.0	49.8	50.2
InternLM-XC2	69.1	61.6	81.1	29.0	26.0	39.0	66.8	57.4	62.0	59.2
Pearson Correlation Analysis										
GPT-4	1.00	-	-	1.00	-	-	1.00	-	-	-
Gemini	0.87	1.00	-	0.91	1.00	-	0.92	1.00	-	-
InternLM-XC2	-x-	-x-	-x-	-x-	-x-	-x-	0.90	0.91	1.00	-
Hunyuan-V	0.82	0.77	1.00	0.74	0.77	1.00	0.91	0.92	0.88	1.00

Table Ku84.1: Detailed Scores of each sub-aspect corresponding to the left table in Fig. 2.

model	T.I	T.M	S.T	S.J	R.O	R.A	F.S	F.B	P.A	P.L	Avg.
GPT-4-Vision	75.06	76.63	80.49	92.54	80.93	55.89	79.37	83.14	74.45	84.29	78.28
Claude3	66.80	60.25	77.19	97.45	72.68	51.97	75.51	63.14	63.33	99.27	72.76
Gemini-Pro	65.11	67.34	72.85	55.76	78.38	50.39	72.33	27.66	70.49	35.72	59.60
Qwen-VL-Plus	68.55	59.38	68.77	66.23	75.18	36.64	64.14	82.95	59.80	53.50	63.51
LLaVA-1.5-7B	54.06	48.41	58.02	37.37	74.12	28.45	70.57	38.49	48.27	26.78	48.45
LLaVA-1.5-13B	58.78	53.89	61.92	39.54	74.13	30.75	67.80	39.48	54.08	36.20	51.66
ShareGPT4V	55.81	50.20	58.99	39.13	69.31	33.23	70.39	34.75	51.80	39.10	50.27
LVIS-Instruct4V	54.81	46.76	58.77	49.29	64.16	29.06	71.51	35.71	58.65	52.32	52.10
LLaVA-RLHF	50.12	51.16	59.43	35.92	70.47	31.76	69.45	39.09	53.88	59.70	52.10
LLaVA-NeXT	55.55	58.63	68.37	43.51	76.49	39.38	67.89	63.49	53.78	55.47	58.26
MiniGPT-4-13B	44.79	45.77	47.40	33.99	64.11	39.12	64.84	34.17	39.46	58.16	47.18
MiniGPT-4-L2	48.29	50.23	69.83	74.49	63.00	35.38	65.71	37.52	42.46	69.98	55.69
mPLUG-Owl	48.28	42.83	49.40	24.15	73.13	20.31	50.08	37.59	39.10	14.25	39.91
mPLUG-Owl2	55.87	50.37	60.15	33.41	74.86	36.31	73.50	51.80	56.55	34.59	52.74
InstructBLIP	46.32	40.95	41.46	43.30	74.08	33.22	57.79	30.02	58.14	9.27	43.46
Otter	41.97	34.31	50.57	45.59	57.82	24.02	57.94	34.53	40.94	20.49	40.82
CogVLM	55.29	46.26	61.50	55.78	74.07	53.12	62.05	32.18	40.20	24.79	50.52
Qwen-VL-Chat	58.96	49.24	59.18	39.56	72.09	41.67	64.60	34.61	53.60	37.16	51.07
InternVL-Chat	58.82	52.39	56.39	43.34	71.68	55.21	71.14	35.04	57.88	33.68	53.56
InternLM-XC	53.63	41.64	45.65	57.80	68.17	27.79	71.16	46.59	56.16	43.06	51.17
InternLM-XC2	61.80	52.90	63.57	51.19	75.36	38.90	79.78	49.09	60.44	34.97	56.80

Table Ku84.2: Statistics of image-text pairs in each task. Samples paired with irrelevant images in text-only tasks are counted once.

T.1	T.2	T.3	T.4	T.5	T.6	T.7	S.1	S.2	S.3	S.4	S.5	S.6	R.1	R.2	R.3
800	500	420	300	100	300	300	500	500	240	600	468	1000	600	1041	3000
R.4	R.5	R.6	F.1	F.2	F.3	F.4	F.5	F.6	F.7	P.1	P.2	P.3	P.4	P.5	P.6
100	100	4014	100	786	1689	291	120	626	720	1300	426	300	1200	195	400
Total: 23036															