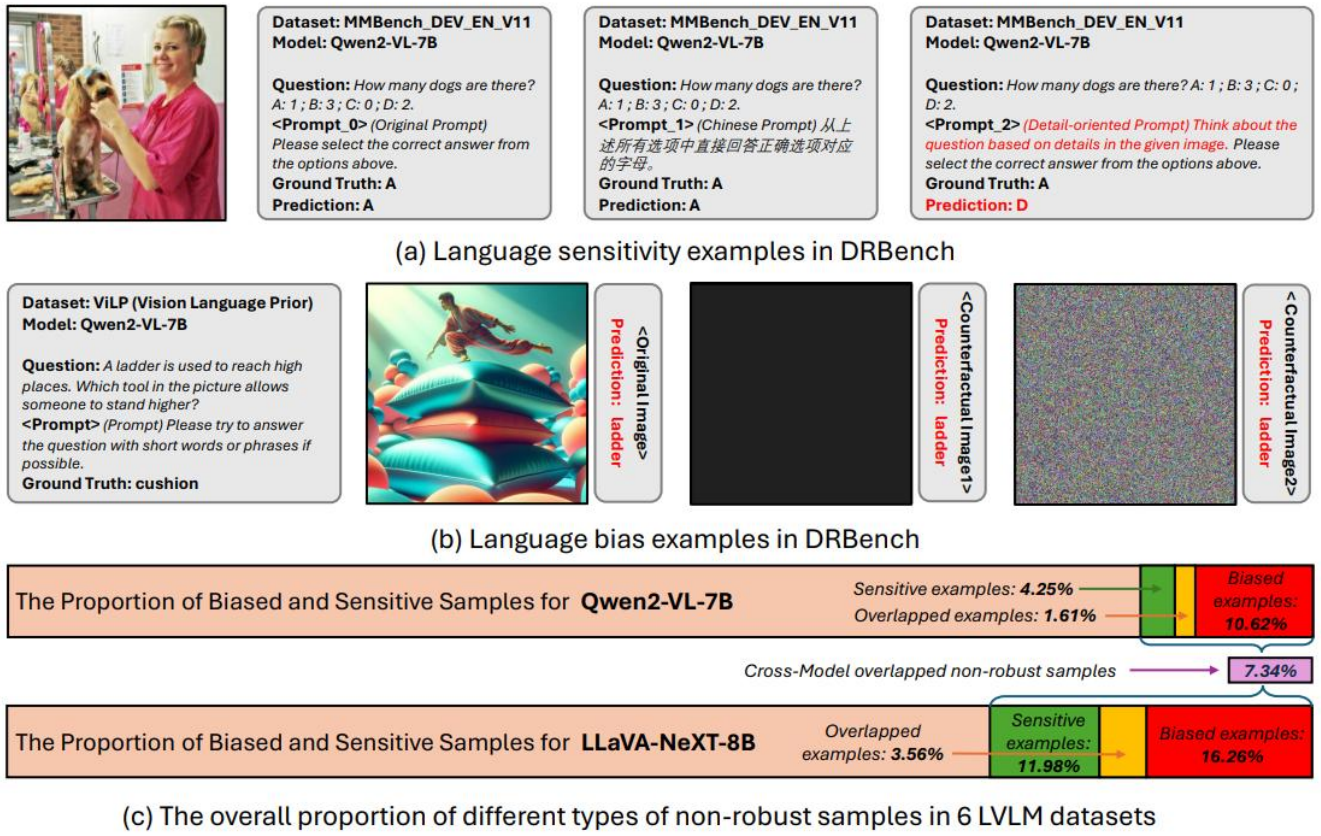
Scaling Test-Time Robustness of Vision-Language Models via Self-Critical Inference Framework

Kaihua Tang[1]; Jiaxin Qi[2]; Jinli Ou[4] ; Yuhua Zheng[3] ; Jianqiang Huang[1,2,3,4]

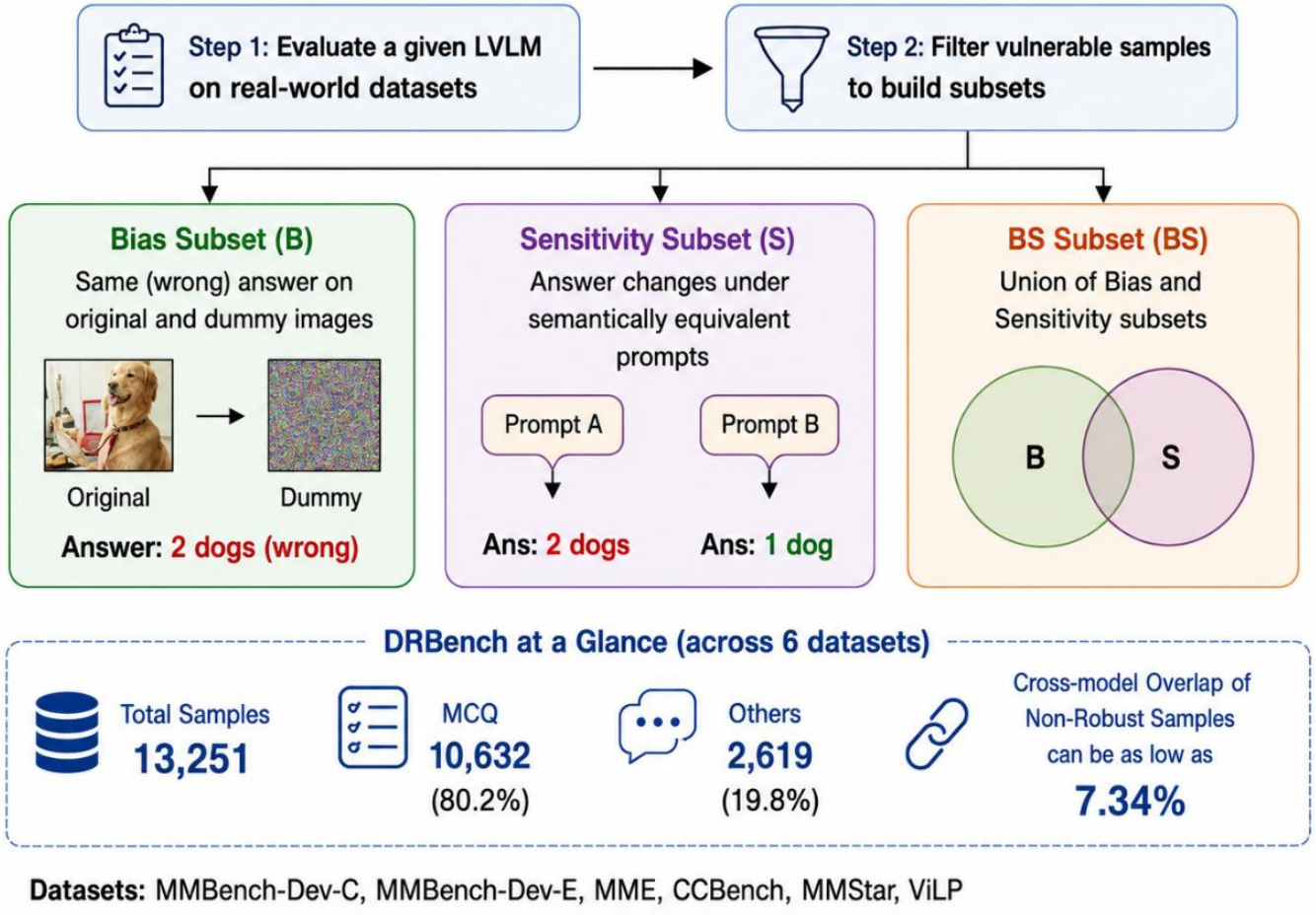
[1]Tongji University, Institute of AI4E; [2]Computer Network Information Center;
[3]HIAS, University of Chinese Academy of Sciences; [4]University of Chinese Academy of Sciences

Problem & Motivation

- Large vision-language models (LVLMs) suffer from two key robustness problems: language sensitivity and language bias.
- **Language sensitivity:** semantically equivalent prompts can lead to inconsistent answers.
- **Language bias (hallucination):** models may rely on language priors instead of visual evidence.



Dynamic Robustness Benchmark (DRBench)

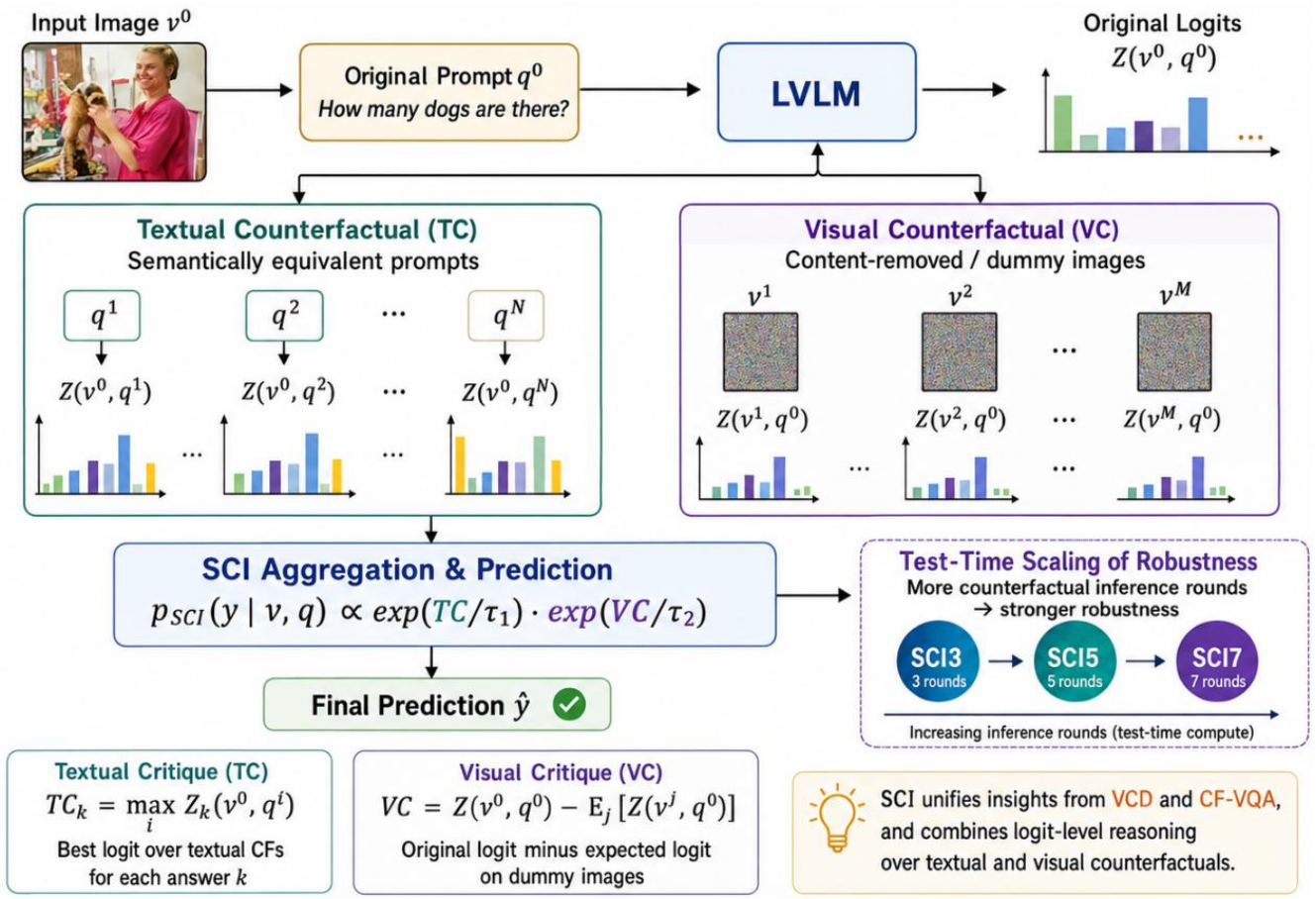


Subset Size	B Subset	S Subset	BS Subset	Overlap
LLaVA-NeXT (MCQ)	1810	1005	2476	339
LLaVA-NeXT (Others)	345	582	794	133
LLaVA-NeXT (Overall)	2155	1587	3270	472
Qwen2-VL (MCQ)	1080	252	1243	89
Qwen2-VL (Others)	327	311	513	125
Qwen2-VL (Overall)	1407	563	1756	214

DRBench划分：

- Bias Subset
- Sensitivity Subset
- BS Subset (Joint Samples)

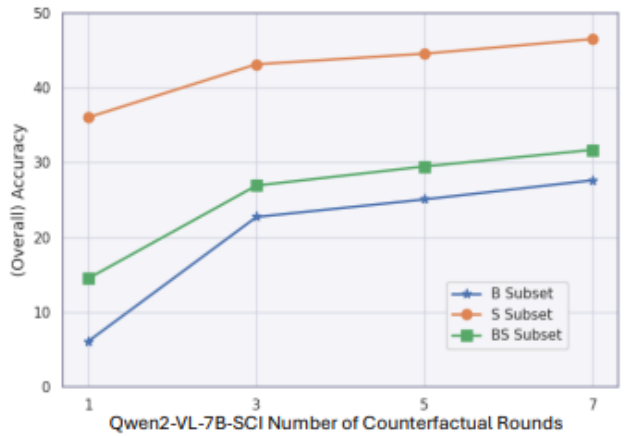
Method: Self-Critical Inference (SCI)

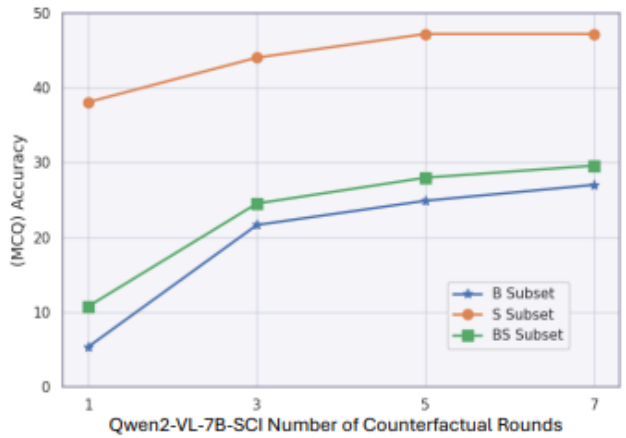


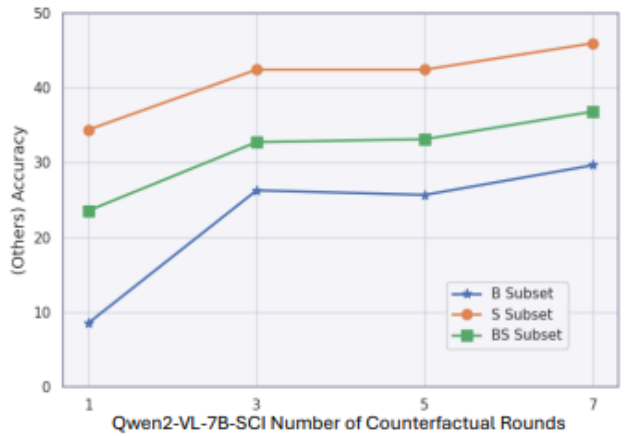
Method	B Subset			S Subset			BS Subset		
	MCQ	Others	Overall	MCQ	Others	Overall	MCQ	Others	Overall
LLaVA-NeXT	0.0	0.0	0.0	39.2	37.63	38.63	15.91	27.58	18.75
LLaVA-NeXT-TIE	12.98	23.48	14.66	39.00	57.56	45.81	21.89	44.21	27.31
LLaVA-NeXT-VCD	12.65	25.51	14.71	40.50	56.53	46.38	22.54	44.58	27.89
LLaVA-NeXT-M3ID	16.91	25.22	18.24	39.90	56.36	45.94	24.15	44.33	29.05
LLaVA-NeXT-SCI$_3$ (ours)	21.22	35.36	23.48	39.60	60.31	47.20	27.14	50.13	32.72
LLaVA-NeXT-SCI$_5$ (ours)	23.81	37.97	26.08	**40.60**	**60.65**	**47.95**	28.80	51.01	34.19
LLaVA-NeXT-SCI$_7$ (ours)	**24.86**	**38.26**	**27.01**	40.10	**60.65**	47.64	**29.68**	**51.26**	**34.92**
Qwen2-VL	5.37	8.56	6.11	38.10	34.41	36.06	10.78	23.59	14.52
Qwen2-VL-TIE	16.20	16.82	16.35	45.63	36.66	40.67	20.27	27.29	22.32
Qwen2-VL-VCD	15.74	21.71	17.13	46.83	40.84	43.52	20.11	30.41	23.12
Qwen2-VL-M3ID	19.81	21.71	20.26	**47.22**	41.16	43.87	23.65	30.6	25.68
Qwen2-VL-SCI$_3$ (ours)	21.67	26.30	22.74	44.05	42.44	43.16	24.54	32.75	26.94
Qwen2-VL-SCI$_5$ (ours)	24.91	25.69	25.09	**47.22**	42.44	44.58	28.00	33.14	29.50
Qwen2-VL-SCI$_7$ (ours)	**27.04**	**29.66**	**27.65**	**47.22**	**45.98**	**46.54**	**29.61**	**36.84**	**31.72**

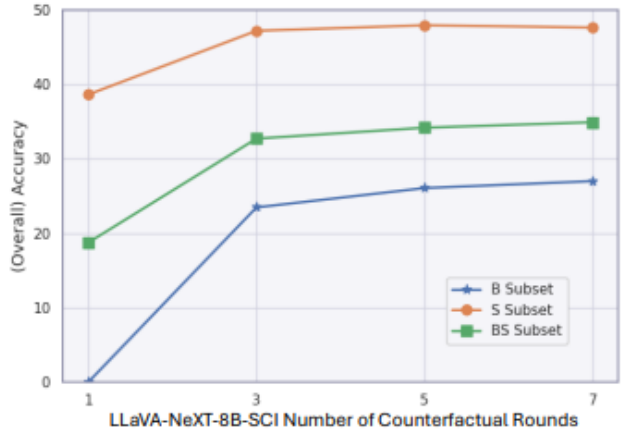
Construction Model	Methods	MCQ	Others	Overall
LLaVA-NeXT	LLaVA-NeXT-Original	15.91	27.58	18.75
	LLaVA-NeXT-SCI$_5$	28.80	51.01	34.19
	Qwen2-VL-Original	59.29	63.48	60.31
	Qwen2-VL-SCI$_5$	61.15	67.88	62.78
Qwen2-VL	Qwen2-VL-Original	10.78	23.59	14.52
	Qwen2-VL-SCI$_5$	28.00	33.14	29.50
	LLaVA-NeXT-Original	30.25	39.18	32.86
	LLaVA-NeXT-SCI$_5$	34.59	41.33	36.56

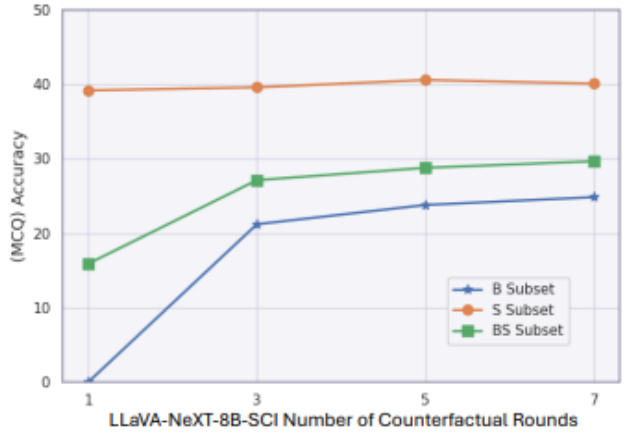
Test-Time Scaling of Robustness

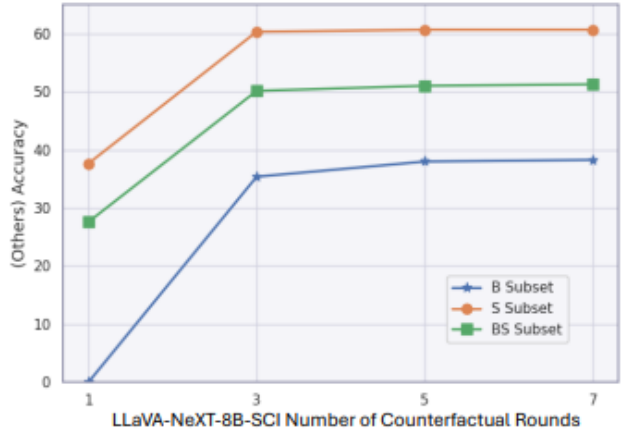












Thank You

Github Link

https://github.com/KaihuaTang/Self-Critical-Inference-Framework