

Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM

DeepSeek-VL Chat 7B + SigLIP&SAM-B			1.42	
Gemma Instr 2B + CLIP	1.02	1.02	0.86	2.07
Gemma Instr 2B + DINOv2&SigLIP	1.10	2.10	2.09	1.13
Gemma Instr 2B + SigLIP	2.19	1.00	2.30	2.21
Gemma Instr 8B + CLIP	0.85	0.98	1.03	1.09
Gemma Instr 8B + SigLIP	2.01	0.98	1.98	2.02
LLAVAv1.5 7B + CLIP (Repro)	1.16	1.14	0.59	0.57
LLAVAv1.513B + CLIP (Repro)	0.97	0.96	0.93	0.94
Llama2 13B + CLIP	0.94	0.96	0.95	0.97
Llama2 13B + CLIP (Control)	0.85	0.84	0.81	0.84
Llama2 13B + DINOv2&SigLIP	1.18	1.19	1.19	1.20
Llama2 13B + DINOv2&SigLIP (Control)	0.70	0.72	0.72	0.71
Llama2 13B + SigLIP	1.11	1.13	1.12	1.13
Llama2 13B + SigLIP (Control)	0.92	0.87	0.97	0.75
Llama2 7B + CLIP	1.27	1.26	0.68	1.27
Llama2 7B + CLIP (Control)	0.88	0.88	0.59	0.86
Llama2 7B + DINOv2&SigLIP	1.25	1.30	1.28	1.27
Llama2 7B + DINOv2&SigLIP (Control)	1.08	1.09	1.09	1.09
Llama2 7B + SigLIP	1.35	1.27	1.32	1.28
Llama2 7B + SigLIP (Control)	0.92	0.87	0.97	0.76
Llama2 Chat 7B + CLIP	0.65	0.61	0.54	1.14
Llama2 Chat 7B + DINOv2/SigLIP	0.55	1.17	1.19	1.14
Llama2 Chat 7B + SigLIP	1.13	0.54	1.15	0.55
Llama3 Instr 8B + CLIP	0.99	0.86	0.90	0.88
Llama3 Instr 8B + DINOv2/SigLIP	0.79	0.87	0.86	0.88
Llama3 Instr 8B + SigLIP	0.84	0.89	0.84	0.86
Mistral Instr v0.2 7B + CLIP	1.11	1.12	0.53	1.11
Mistral Instr v0.2 7B + DINOv2/SigLIP	1.15	1.27	1.26	1.16
Mistral Instr v0.2 7B + SigLIP	1.08	1.10	1.08	1.06
Qwen VL Chat	1.34	1.35	1.34	0.90

Cross Entropy Loss

Gemma Instr 2B + CLIP	Gemma Instr 2B + CLIP	Llama2 7B + CLIP	Qwen VL Chat
Gemma Instr 2B + DINOv2&SigLIP	Gemma Instr 2B + SigLIP	Llama2 7B + CLIP (Control)	DeepSeek-VL Chat 7B + SigLIP&SAM-B
Gemma Instr 8B + CLIP	Gemma Instr 8B + CLIP	Gemma Instr 2B + CLIP	Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 2B + DINOv2&SigLIP	Gemma Instr 8B + SigLIP	Gemma Instr 8B + CLIP	Gemma Instr 8B + CLIP
Llama2 Chat 7B + CLIP	Llama2 Chat 7B + CLIP	Llama2 Chat 7B + CLIP	Gemma Instr 2B + DINOv2&SigLIP
Llama2 Chat 7B + DINOv2/SigLIP	Llama2 Chat 7B + SigLIP	Llama3 Instr 8B + CLIP	Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + CLIP	Llama3 Instr 8B + CLIP	Mistral Instr v0.2 7B + CLIP	Llama3 Instr 8B + SigLIP
Llama3 Instr 8B + DINOv2/SigLIP	Llama3 Instr 8B + SigLIP	LLAVAv1.5 7B + CLIP (Repro)	LLAVAv1.5 7B + CLIP (Repro)

Attacked Ensemble of $N = 8$ VLMs