

Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM

Gemma Instr 2B + CLIP	3.00	3.04
Gemma Instr 2B + DINOv2&SigLIP	3.07	3.57
Gemma Instr 2B + SigLIP	3.59	3.01
Gemma Instr 8B + CLIP	3.10	3.12
Gemma Instr 8B + SigLIP	3.66	3.10
LLAVAv1.5 7B + CLIP (Repro)	2.88	2.87
LLAVAv1.5 13B + CLIP (Repro)	2.68	2.70
Llama2 13B + CLIP	2.66	2.66
Llama2 13B + CLIP (Control)	2.60	2.61
Llama2 13B + DINOv2&SigLIP	2.82	2.80
Llama2 13B + DINOv2&SigLIP (Control)	2.52	2.54
Llama2 13B + SigLIP	2.75	2.76
Llama2 13B + SigLIP (Control)	2.62	2.52
Llama2 7B + CLIP	2.74	2.74
Llama2 7B + CLIP (Control)	2.63	2.63
Llama2 7B + DINOv2&SigLIP	2.79	2.82
Llama2 7B + DINOv2&SigLIP (Control)	2.64	2.66
Llama2 7B + SigLIP	2.82	2.70
Llama2 7B + SigLIP (Control)	2.62	2.52
Llama2 Chat 7B + CLIP	2.59	2.60
Llama2 Chat 7B + DINOv2/SigLIP	2.53	2.87
Llama2 Chat 7B + SigLIP	2.84	2.59
Llama3 Instr 8B + CLIP	3.14	3.01
Llama3 Instr 8B + DINOv2/SigLIP	2.93	2.90
Llama3 Instr 8B + SigLIP	2.94	
Mistral Instr v0.2 7B + CLIP	3.15	3.16
Mistral Instr v0.2 7B + SigLIP	3.13	
Qwen VL Chat	3.66	3.61

Cross Entropy Loss

Gemma Instr 2B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + DINOv2/SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + DINOv2/SigLIP

Gemma Instr 2B + CLIP
Gemma Instr 2B + SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 8B + SigLIP
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + SigLIP

Attacked Ensemble of $N = 8$ VLMs