

Evaluated VLM

Jailbreaking Ensembles of $N = 8$ VLMs

Gemma Instr 2B + CLIP	3.49	2.31	2.34	2.31
Gemma Instr 2B + DINOv2&SigLIP	2.53	2.52	3.61	3.60
Gemma Instr 2B + SigLIP	3.57	3.57	2.41	3.62
Gemma Instr 8B + CLIP	2.42	2.42	2.43	2.46
Gemma Instr 8B + SigLIP	3.64	3.66	2.55	3.64
LLAVAv1.5 7B + CLIP (Repro)	1.95	2.86	2.84	2.03
LLAVAv1.513B + CLIP (Repro)	2.65	2.69	2.68	2.63
Llama2 13B + CLIP	2.66	2.66	2.66	2.65
Llama2 13B + CLIP (Control)	2.58	2.59	2.60	2.57
Llama2 13B + DINOv2&SigLIP	2.82	2.83	2.81	2.82
Llama2 13B + DINOv2&SigLIP (Control)	2.52	2.51	2.54	2.54
Llama2 13B + SigLIP	2.76	2.76	2.76	2.78
Llama2 13B + SigLIP (Control)	2.32	2.62	2.54	2.64
Llama2 7B + CLIP	2.73	2.76	2.76	2.01
Llama2 7B + CLIP (Control)	2.61	2.63	2.62	1.97
Llama2 7B + DINOv2&SigLIP	2.81	2.77	2.84	2.84
Llama2 7B + DINOv2&SigLIP (Control)	2.64	2.63	2.67	2.67
Llama2 7B + SigLIP	2.69	2.85	2.76	2.84
Llama2 7B + SigLIP (Control)	2.32	2.62	2.54	2.64
Llama2 Chat 7B + CLIP	2.81	1.99	2.03	2.02
Llama2 Chat 7B + SigLIP	2.02	2.85	2.05	2.87
Llama3 Instr 8B + CLIP	2.93	2.22	2.17	2.12
Llama3 Instr 8B + DINOv2/SigLIP				2.90
Llama3 Instr 8B + SigLIP	2.23			
Mistral Instr v0.2 7B + CLIP	3.15	3.13	3.15	2.31
Mistral Instr v0.2 7B + DINOv2/SigLIP		3.13		
Mistral Instr v0.2 7B + SigLIP	3.12	3.16	3.15	3.16
Qwen VL Chat	2.19	3.59	3.59	3.57

DeepSeek-VL Chat 7B + SigLIP&SAM-B

Gemma Instr 2B + CLIP

Gemma Instr 2B + DINOv2&SigLIP

Gemma Instr 2B + DINOv2&SigLIP

Gemma Instr 2B + DINOv2&SigLIP

Gemma Instr 8B + CLIP

Gemma Instr 8B + CLIP

Gemma Instr 8B + SigLIP

LLAVAv1.5 7B + CLIP (Repro)

Llama2 Chat 7B + CLIP

Llama2 Chat 7B + SigLIP

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 + CLIP

Llama2 Chat 7B + SigLIP

Llama3 Instr 8B + CLIP

Llama3 Instr 8B + SigLIP

Gemma Instr 2B + CLIP

Gemma Instr 8B + CLIP

LLAVAv1.5 7B + CLIP (Repro)

Llama2 7B + CLIP

Llama2 7B + CLIP (Control)

Llama2 Chat 7B + CLIP

Llama3 Instr 8B + CLIP

Mistral Instr v0.2 7B + CLIP

Cross Entropy Loss

Attacked Ensemble of $N = 8$ VLMs