

# Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM

DeepSeek-VL Chat 7B + SigLIP&SAM-B			1.00	
Gemma Instr 2B + CLIP	0.45	0.18	0.22	0.74
Gemma Instr 2B + DINOv2&SigLIP	0.39	0.78	0.74	0.24
Gemma Instr 2B + SigLIP	0.69	0.62	0.75	0.68
Gemma Instr 8B + CLIP	0.20	0.31	0.34	0.20
Gemma Instr 8B + SigLIP	0.80	0.34	0.78	0.79
LLAVAv1.5 7B + CLIP (Repro)	0.68	0.67	0.30	0.51
LLAVAv1.513B + CLIP (Repro)	0.91	0.85	0.87	0.93
Llama2 13B + CLIP	0.53	0.51	0.48	0.53
Llama2 13B + CLIP (Control)	0.32	0.31	0.28	0.32
Llama2 13B + DINOv2&SigLIP	0.52	0.63	0.66	0.49
Llama2 13B + DINOv2&SigLIP (Control)		0.44	0.45	0.38
Llama2 13B + SigLIP	0.58	0.58	0.60	0.52
Llama2 13B + SigLIP (Control)	0.34	0.31	0.37	0.20
Llama2 7B + CLIP	0.49	0.50	0.77	0.50
Llama2 7B + CLIP (Control)	0.28	0.30	0.15	0.29
Llama2 7B + DINOv2&SigLIP	0.58	0.66	0.68	0.60
Llama2 7B + DINOv2&SigLIP (Control)	0.28	0.33	0.37	0.26
Llama2 7B + SigLIP	0.68	0.62	0.68	0.62
Llama2 7B + SigLIP (Control)	0.34	0.31	0.37	0.15
Llama2 Chat 7B + CLIP	0.84	0.32	0.41	0.86
Llama2 Chat 7B + DINOv2/SigLIP	0.55	0.93	0.93	0.86
Llama2 Chat 7B + SigLIP	0.87	0.75	0.87	0.64
Llama3 Instr 8B + CLIP	0.25	0.73	0.35	0.41
Llama3 Instr 8B + DINOv2/SigLIP	0.68	0.79	0.83	0.71
Llama3 Instr 8B + SigLIP	0.73	0.38	0.73	0.34
Mistral Instr v0.2 7B + CLIP	0.17	0.22	0.21	0.19
Mistral Instr v0.2 7B + DINOv2/SigLIP	0.23	0.52	0.41	0.35
Mistral Instr v0.2 7B + SigLIP	0.41	0.33	0.37	0.36
Qwen VL Chat	0.95	0.95	0.95	0.32



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

# Attacked Ensemble of $N = 8$ VLMs