Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM (rows) vs **Attacked Ensemble of $N = 8$ VLMs** (columns)

Evaluated VLM	Col 1	Col 2	Col 3
Gemma Instr 2B + CLIP	0.16	0.13	0.16
Gemma Instr 2B + DINOv2&SigLIP	0.20	0.75	0.76
Gemma Instr 2B + SigLIP	0.76	0.28	0.74
Gemma Instr 8B + CLIP	0.21	0.36	0.17
Gemma Instr 8B + SigLIP	0.84	0.49	0.79
LLAVAv1.5 7B + CLIP (Repro)	0.80	0.75	0.18
LLAVAv1.513B + CLIP (Repro)	0.85	0.84	0.83
Llama2 13B + CLIP	0.50	0.63	0.46
Llama2 13B + CLIP (Control)	0.69	0.67	0.53
Llama2 13B + DINOv2&SigLIP	0.56	0.69	0.64
Llama2 13B + DINOv2&SigLIP (Control)	0.63	0.66	0.66
Llama2 13B + SigLIP	0.71	0.67	0.68
Llama2 13B + SigLIP (Control)	0.49	0.49	0.52
Llama2 7B + CLIP	0.64	0.67	0.12
Llama2 7B + CLIP (Control)	0.48	0.54	0.11
Llama2 7B + DINOv2&SigLIP	0.64	0.76	0.75
Llama2 7B + DINOv2&SigLIP (Control)	0.44	0.51	0.53
Llama2 7B + SigLIP	0.76	0.60	0.76
Llama2 7B + SigLIP (Control)	0.49	0.48	0.54
Llama2 Chat 7B + CLIP	0.10	0.26	0.14
Llama2 Chat 7B + DINOv2/SigLIP	0.36	0.96	0.95
Llama2 Chat 7B + SigLIP	0.91	0.18	0.94
Llama3 Instr 8B + CLIP	0.33	0.31	0.23
Llama3 Instr 8B + DINOv2/SigLIP	0.51	0.83	0.84
Llama3 Instr 8B + SigLIP		0.79	0.89
Mistral Instr v0.2 7B + CLIP	0.73	0.68	0.26
Mistral Instr v0.2 7B + SigLIP		0.73	
Qwen VL Chat	0.96	0.96	0.94

Attacked Ensemble of $N = 8$ VLMs

Column 1:
- 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

Column 2:
- 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

Column 3:
- 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)

Attack Failure Rate (color scale: 0.0 to 1.0)