Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM	Gemma Instr 2B + CLIP	Gemma Instr 2B + SigLIP	Llama2 7B + CLIP	Qwen VL Chat
Gemma Instr 2B + CLIP	0.46	0.36	0.37	0.74
Gemma Instr 2B + DINOv2&SigLIP			0.72	
Gemma Instr 2B + SigLIP	0.73	0.64	0.74	0.74
LLAVAv1.5 7B + CLIP (Repro)	0.76	0.76	0.41	0.44
LLAVAv1.513B + CLIP (Repro)	0.86	0.82	0.78	0.81
Llama2 13B + CLIP	0.54	0.52	0.42	0.55
Llama2 13B + CLIP (Control)	0.57	0.60	0.55	0.63
Llama2 13B + DINOv2&SigLIP	0.60	0.67	0.65	0.54
Llama2 13B + DINOv2&SigLIP (Control)	0.60	0.63	0.65	0.58
Llama2 13B + SigLIP	0.71	0.67	0.72	0.62
Llama2 13B + SigLIP (Control)	0.50	0.47	0.54	0.31
Llama2 7B + CLIP	0.64	0.58	0.54	0.51
Llama2 7B + CLIP (Control)	0.51	0.50	0.43	0.46
Llama2 7B + DINOv2&SigLIP	0.65	0.78	0.74	0.72
Llama2 7B + DINOv2&SigLIP (Control)	0.43	0.48	0.49	0.48
Llama2 7B + SigLIP	0.75	0.64	0.77	0.58
Llama2 7B + SigLIP (Control)	0.50	0.47	0.51	0.31
Llama2 Chat 7B + SigLIP				
Llama3 Instr 8B + CLIP		0.57		
Qwen VL Chat	0.96	0.97	0.96	0.27

Attacked Ensemble of $N = 8$ VLMs

Attack Failure Rate

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)