Jailbreaking Ensembles of $N = 8$ VLMs

Evaluated VLM	Gemma Instr 2B + CLIP (Col 1)	Gemma Instr 2B + CLIP (Col 2)	Llama2 7B + CLIP (Col 3)	Qwen VL Chat (Col 4)
Gemma Instr 2B + CLIP	2.31	2.34	2.31	3.49
Gemma Instr 2B + DINOv2&SigLIP			3.57	
Gemma Instr 2B + SigLIP	3.57	2.41	3.62	3.57
LLAVAv1.5 7B + CLIP (Repro)	2.86	2.84	2.03	1.95
LLAVAv1.513B + CLIP (Repro)	2.69	2.68	2.63	2.65
Llama2 13B + CLIP	2.66	2.66	2.65	2.66
Llama2 13B + CLIP (Control)	2.59	2.60	2.57	2.58
Llama2 13B + DINOv2&SigLIP	2.83	2.81	2.82	2.82
Llama2 13B + DINOv2&SigLIP (Control)	2.51	2.54	2.54	2.52
Llama2 13B + SigLIP	2.76	2.76	2.78	2.76
Llama2 13B + SigLIP (Control)	2.62	2.54	2.64	2.32
Llama2 7B + CLIP	2.76	2.76	2.01	2.73
Llama2 7B + CLIP (Control)	2.63	2.62	1.97	2.61
Llama2 7B + DINOv2&SigLIP	2.77	2.84	2.84	2.81
Llama2 7B + DINOv2&SigLIP (Control)	2.63	2.67	2.67	2.64
Llama2 7B + SigLIP	2.85	2.76	2.84	2.69
Llama2 7B + SigLIP (Control)	2.62	2.54	2.64	2.32
Llama2 Chat 7B + SigLIP	2.85			
Llama3 Instr 8B + CLIP		2.17		
Qwen VL Chat	3.59	3.59	3.57	2.19

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)

Column 4:
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)

Cross Entropy Loss