Jailbreaking Ensembles of $N = 8$ VLMs
Evaluated VLM
Attacked Ensemble of $N = 8$ VLMs
Cross Entropy Loss
0.0
0.5
1.0
1.5
2.0

Evaluated VLM	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
Gemma Instr 2B + CLIP	1.43	1.65
Gemma Instr 2B + DINOv2&SigLIP	1.40	1.97
Gemma Instr 2B + DINOv2&SigLIP	1.35	2.19
Gemma Instr 2B + SigLIP	2.27	1.49
Gemma Instr 8B + CLIP	1.26	1.21
Gemma Instr 8B + SigLIP	2.07	1.21
LLAVAv1.5 7B + CLIP (Repro)	1.20	1.19
LLAVAv1.513B + CLIP (Repro)	0.94	0.96
Llama2 13B + CLIP	0.93	0.95
Llama2 13B + CLIP (Control)	0.85	0.86
Llama2 13B + DINOv2&SigLIP	1.21	1.20
Llama2 13B + DINOv2&SigLIP (Control)		0.74
Llama2 13B + SigLIP	1.11	1.13
Llama2 13B + SigLIP (Control)	0.93	0.87
Llama2 7B + CLIP	1.23	1.22
Llama2 7B + CLIP (Control)	0.92	0.94
Llama2 7B + DINOv2&SigLIP	1.26	1.29
Llama2 7B + DINOv2&SigLIP (Control)	1.11	1.09
Llama2 7B + SigLIP	1.30	1.16
Llama2 7B + SigLIP (Control)	0.93	0.87
Llama2 Chat 7B + CLIP	0.70	0.67
Llama2 Chat 7B + DINOv2/SigLIP	0.61	1.21
Llama2 Chat 7B + SigLIP	1.13	0.52
Llama3 Instr 8B + CLIP	1.28	1.28
Llama3 Instr 8B + DINOv2/SigLIP	1.41	0.88
Llama3 Instr 8B + SigLIP	0.85	0.99
Mistral Instr v0.2 7B + CLIP	1.11	1.12
Mistral Instr v0.2 7B + DINOv2/SigLIP	1.12	1.21
Mistral Instr v0.2 7B + SigLIP	1.06	1.09
Qwen VL Chat	1.40	1.36