Jailbreaking Ensembles of N = 8 VLMs DeepSeek-VL Chat 7B + SigLIP&SAM-B 1.42 0.86 Gemma Instr 2B + CLIP 2.07 1.02 1.02 Gemma Instr 2B + DINOv2&SigLIP 1.09 1.98 1.08 2.00 2.19 2.30 Gemma Instr 2B + SigLIP 2.21 1.00 0.85 1.03 Gemma Instr 8B + CLIP 1.09 0.98 1.98 Gemma Instr 8B + SigLIP 2.01 0.98 2.02 0.57 0.59 LLAVAv1.5 7B + CLIP (Repro)1.16 1.14 0.97 0.93 LLAVAv1.513B + CLIP (Repro)0.94 0.96 Llama2 13B + CLIP 0.94 0.95 0.97 0.96 0.85 0.84 0.84 0.81 Llama2 13B + CLIP (Control) Llama2 13B + DINOv2&SigLIP 1.18 1.20 1.19 1.19 0.70 0.72 0.72 Llama2 13B + DINOv2&SigLIP (Control) 0.71 1.11 1.13 1.12 Llama2 13B + SigLIP 1.13 Llama2 13B + SigLIP (Control) 0.75 0.92 0.87 0.97 Evaluated 1.27 1.27 1.26 Llama2 7B + CLIP 0.68 Llama2 7B + CLIP (Control) 0.86 0.88 0.88 0.59 1.25 Llama2 7B + DINOv2&SigLIP 1.27 1.30 1.28 Llama2 7B + DINOv2&SigLIP (Control) 1.09 1.08 1.09 1.09 Llama2 7B + SigLIP 1.35 1.27 1.32 1.28 0.76 0.87 0.92 0.97 Llama2 7B + SigLIP (Control) Llama2 Chat 7B + CLIP 0.65 0.61 0.54 1.14 0.55 Llama2 Chat 7B + DINOv2/SigLIP 1.17 1.19 1.14 0.54 0.551.13 1.15 Llama2 Chat 7B + SigLIP 0.99 0.86 0.90 Llama3 Instr 8B + CLIP 0.88 Llama3 Instr 8B + DINOv2/SigLIP 0.88 0.79 0.87 0.86 0.84 Llama3 Instr 8B + SigLIP 0.86 0.89 0.84 Mistral Instr v0.2 7B + CLIP 1.11 1.12 0.53 1.11 Mistral Instr v0.2 7B + DINOv2/SigLIP 1.16 1.15 1.27 1.26 1.10 Mistral Instr v0.2 7B + SigLIP 1.08 1.08 1.06 1.35 0.90 1.34 1.34 Qwen VL Chat DeepSeek-VL Chat 7B + SigLIP&SAM-B Gemma Instr 2B + CLIP Gemma Instr 2B + CLIP Gemma Instr 2B + CLIPGemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + SigLIP Gemma Instr 8B + CLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 8B + CLIP LLAVAv1.5 7B + CLIP (Repro)Gemma Instr 8B + CLIP Gemma Instr 8B + CLIP Gemma Instr 8B + SigLIP Llama2 7B + CLIP Llama2 Chat 7B + CLIP Llama2 Chat 7B + CLIP LLAVAv1.5 7B + CLIP (Repro)Llama2 7B + CLIP (Control) Llama2 Chat 7B + SigLIP Llama2 Chat 7B + DINOv2/SigLIP Llama2 Chat 7B + SigLIP Llama2 Chat 7B + CLIP Llama3 Instr 8B + CLIP Llama3 Instr 8B + CLIP Llama3 Instr 8B + SigLIP Llama3 Instr 8B + CLIP Llama3 Instr 8B + SigLIP Llama3 Instr 8B + DINOv2/SigLIP Qwen VL Chat Mistral Instr v0.27B + CLIP

Attacked Ensemble of N = 8 VLMs

-2.0

-1.5

Cross Entropy Loss

-0.5

-0.0