Jailbreaking Ensembles of N = 8 VLMs DeepSeek-VL Chat 7B + SigLIP&SAM-B 1.40 Gemma Instr 2B + CLIP 0.10 0.09 0.10 2.03 Gemma Instr 2B + DINOv2&SigLIP 1.97 0.11 2.22 2.30 Gemma Instr 2B + SigLIP 0.10 0.12 Gemma Instr 8B + CLIP 0.11 0.12 Gemma Instr 8B + SigLIP 0.10 1.99 2.00 LLAVAv1.5 7B + CLIP (Repro)1.16 0.08 1.14 LLAVAv1.513B + CLIP (Repro) 0.69 0.95 0.92Llama2 13B + CLIP 0.94 0.94 0.94 Llama2 13B + CLIP (Control) 0.73 0.83 0.83 Llama2 13B + DINOv2&SigLIP 1.19 1.21 1.19 Llama2 13B + DINOv2&SigLIP (Control) 0.71 Llama2 13B + SigLIP 1.12 1.12 1.13 Llama2 13B + SigLIP (Control) 0.94 0.86 0.97 0.08 Llama2 7B + CLIP 1.28 1.28 Llama2 7B + CLIP (Control) 0.82 0.87 0.09 Llama2 7B + DINOv2&SigLIP 1.29 1.27 1.29 Llama2 7B + DINOv2&SigLIP (Control) 1.07 1.07 1.09 Llama2 7B + SigLIP 1.18 1.33 1.34 0.97 Llama2 7B + SigLIP (Control) 0.94 0.86 Llama2 Chat 7B + CLIP 0.10 0.09 0.09 Llama2 Chat 7B + DINOv2/SigLIP 1.14 1.18 0.09 1.15 Llama2 Chat 7B + SigLIP 0.10 1.13 Llama3 Instr 8B + CLIP 0.10 0.10 0.09 Llama3 Instr 8B + DINOv2/SigLIP 0.86 0.86 0.84 Llama3 Instr 8B + SigLIP 0.84 0.10 Mistral Instr v0.2 7B + CLIP 1.08 1.06 0.10 Mistral Instr v0.2 7B + DINOv2/SigLIP 1.17 1.26 1.06 Mistral Instr v0.2 7B + SigLIP 1.07 1.08 Qwen VL Chat 1.33 1.33 1.39 Gemma Instr 2B + CLIP Gemma Instr 2B + CLIP Gemma Instr 2B + CLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + SigLIP Gemma Instr 8B + CLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 8B + CLIP LLAVAv1.57B + CLIP (Repro)Gemma Instr 8B + SigLIP Gemma Instr 8B + CLIP Llama2 7B + CLIP Llama2 Chat 7B + CLIP Llama2 Chat 7B + CLIP Llama2 7B + CLIP (Control) Llama2 Chat 7B + SigLIP Llama2 Chat 7B + DINOv2/SigLIP Llama2 Chat 7B + CLIP Llama3 Instr 8B + SigLIP Llama3 Instr 8B + DINOv2/SigLIP Mistral Instr v0.2 7B + CLIP

Evaluated

Attacked Ensemble of N = 8 VLMs

-2.0

-1.5

Cross Entropy Loss

-0.5

-0.0