Jailbreaking Ensembles of N=8 VLMs DeepSeek-VL Chat 7B + SigLIP&SAM-B 1.40 Gemma Instr 2B + CLIP 0.09 0.10 0.10 2.03 Gemma Instr 2B + DINOv2&SigLIP 1.97 0.11 2.22 0.10 Gemma Instr 2B + SigLIP 2.30 0.11 0.12 0.12 Gemma Instr 8B + CLIP 0.10 1.99 Gemma Instr 8B + SigLIP2.00 1.16 LLAVAv1.5 7B + CLIP (Repro) 1.14 80.0 LLAVAv1.513B + CLIP (Repro) 0.95 0.69 0.92 0.94 Llama2 13B + CLIP 0.94 0.94 Llama2 13B + CLIP (Control) 0.83 0.73 0.83 Llama2 13B + DINOv2&SigLIP 1.21 1.19 1.19 Llama2 13B + DINOv2&SigLIP (Control) 0.71 1.12 Llama2 13B + SigLIP 1.12 1.13 0.86 Llama2 13B + SigLIP (Control) 0.94 0.97 Evaluated Llama2.7B + CLIP1.28 0.08 1.28 0.87 Llama2 7B + CLIP (Control) 0.82 0.09 1.29 1.27 Llama2 7B + DINOv2&SigLIP 1.29 1.07 Llama2 7B + DINOv2&SigLIP (Control) 1.07 1.09 1.18 1.34 1.33 Llama2 7B + SigLIP 0.86 0.94 0.97 Llama2 7B + SigLIP (Control) Llama2 Chat 7B + CLIP 0.09 0.09 0.10 1.14 Llama2 Chat 7B + DINOv2/SigLIP 0.09 1.18 1.13 0.10 Llama2 Chat 7B + SigLIP 1.15 Llama3 Instr 8B + CLIP 0.09 0.10 0.10 Llama3 Instr 8B + DINOv2/SigLIP 0.86 0.86 0.84 0.10 0.84 Llama3 Instr 8B + SigLIP 1.06 1.08 Mistral Instr v0.2 7B + CLIP 0.10 1.17 Mistral Instr v0.2 7B + DINOv2/SigLIP 1.14 1.26 1.06 1.07 1.08 Mistral Instr v0.2 7B + SigLIP 1.33 1.33 1.39 Qwen VL Chat Gemma Instr 2B + CLIP Gemma Instr 2B + CLIPGemma Instr 2B + CLIPGemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + SigLIPGemma Instr 8B + CLIPGemma Instr 8B + CLIP Gemma Instr 2B + DINOv2&SigLIP LLAVAv1.5 7B + CLIP (Repro) Gemma Instr 8B + CLIPGemma Instr 8B + SigLIPLlama2 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

Attacked Ensemble of $N=8~{
m VLMs}$

-2.0

-1.5

Cross Entropy Loss

-0.5