## Jailbreaking Ensembles of N=8 VLMs Gemma Instr 2B + CLIP3.32 3.41 Gemma Instr 2B + DINOv2&SigLIP 3.37 3.62 3.58 3.21 3.61 Gemma Instr 2B + SigLIP 3.63 Gemma Instr 8B + CLIP 3.44 3.58 3.43 3.37 3.66 Gemma Instr 8B + SigLIP3.67 LLAVAv1.5 7B + CLIP (Repro) 2.88 2.60 2.86 2.71 LLAVAv1.513B + CLIP (Repro) 2.64 2.68 Llama2 13B + CLIP 2.68 2.69 2.67 Llama2 13B + CLIP (Control) 2.60 2.57 2.60 Llama2 13B + DINOv2&SigLIP 2.83 2.83 2.81 Llama2 13B + DINOv2&SigLIP (Control) Llama2 13B + SigLIP 2.52 2.55 2.54 2.78 2.78 2.79 Llama2 13B + SigLIP (Control) 2.62 2.64 2.64 Evaluated Llama2 7B + CLIP 2.83 2.91 2.79 Llama2 7B + CLIP (Control) 2.63 2.61 2.57 2.86 Llama2 7B + DINOv2&SigLIP 2.83 2.83 Llama2 7B + DINOv2&SigLIP (Control) 2.67 2.67 2.63 2.81 Llama2 7B + SigLIP 2.86 2.84 2.62 Llama2 7B + SigLIP (Control) 2.64 2.64 Llama2 Chat 7B + CLIP 2.73 2.74 2.70 Llama2 Chat 7B + DINOv2/SigLIP 2.71 2.87 2.86 Llama2 Chat 7B + SigLIP2.78 2.86 2.87 2.95 Llama3 Instr 8B + CLIP 3.19 2.91 Llama3 Instr 8B + DINOv2/SigLIP 2.91 2.90 2.98 2.97 2.96 Llama3 Instr 8B + SigLIP Mistral Instr v0.2 7B + CLIP 3.16 3.18 3.18 Mistral Instr v0.2 7B + SigLIP 3.15 3.61 3.59 3.62 Qwen VL Chat Gemma Instr 2B + CLIPGemma 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 VLMs

-3.5

-3.0

-2.5

 $\begin{array}{cc} 1 & 5 \\ 5 & 1 \end{array}$  Cross Entropy Loss

-1.0

-0.5

-0.0