Jailbreaking Ensembles of $N=8~{\rm VLMs}$ Gemma Instr 2B + CLIP3.32 3.41 3.39 Gemma Instr 2B + SigLIP3.21 2.60 LLAVAv1.5 7B + CLIP (Repro) 2.86 2.88 LLAVAv1.513B + CLIP (Repro) 2.71 2.64 2.68 Llama2 13B + CLIP 2.69 2.68 2.67 Evaluated VLM Llama2 13B + CLIP (Control) 2.60 2.57 2.60 Llama2 13B + DINOv2&SigLIP 2.83 2.81 2.83 Llama2 13B + DINOv2&SigLIP (Control) 2.54 2.52 2.55 Llama $2\ 13B + SigLIP$ 2.78 2.78 2.79 Llama2 13B + SigLIP (Control) 2.64 2.62 2.64 Llama2 7B + CLIP2.79 2.83 2.91 Llama2 7B + CLIP (Control) 2.63 2.57 2.61 Llama2 7B + DINOv2&SigLIP 2.83 2.86 2.83 Llama2 7B + DINOv2&SigLIP (Control) 2.63 2.67 2.67 Llama2.7B + SigLIP2.84 2.86 2.81 Llama2 7B + SigLIP (Control) 2.62 2.64 2.64 3.61 3.59 3.62 Qwen VL Chat Gemma Instr 2B + CLIPGemma Instr 2B + CLIPLlama2 7B + CLIP ${\sf Gemma\ Instr\ 2B+DINOv2\&SigLIP}$ Gemma Instr 2B + SigLIPLlama2 7B + CLIP (Control) Gemma Instr 8B + CLIP Gemma Instr 8B + CLIP Gemma Instr 2B + CLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 8B + SigLIPGemma Instr 8B + CLIP Llama2 Chat 7B + SigLIP Llama2 Chat 7B + DINOv2/SigLIP Llama3 Instr 8B + CLIP Llama3 Instr 8B + CLIP Llama3 Instr 8B + CLIP Mistral Instr v0.2 7B + CLIP Llama3 Instr 8B + SigLIP Llama3 Instr 8B + DINOv2/SigLIP LLAVAv1.5 7B + CLIP (Repro)

Åttacked Ensemble of N=8 VLMs

-3.5

-3.0

-2.5

 $\begin{array}{ccc} -7 & & -7 \\ & & & 1.2 \\ & & & & 1.2 \end{array}$ Cross Entropy Loss

-1.0

-0.5