Jailbreaking Ensembles of N=8 VLMs Gemma Instr 2B + CLIP0.16 0.16 0.13 0.28 Gemma Instr 2B + SigLIP0.18 LLAVAv1.5 7B + CLIP (Repro) 0.80 0.75 LLAVAv1.513B + CLIP (Repro) 0.84 0.83 0.85 Llama2 13B + CLIP 0.50 0.63 0.46 Evaluated VLM Llama2 13B + CLIP (Control) 0.69 0.67 0.53 Llama2 13B + DINOv2&SigLIP 0.69 0.64 0.56 Llama2 13B + DINOv2&SigLIP (Control) 0.63 0.66 0.66 Llama $2\ 13B + SigLIP$ 0.67 0.71 0.68 Llama2 13B + SigLIP (Control) 0.49 0.49 0.52 Llama2 7B + CLIP0.64 0.67 0.12 Llama2 7B + CLIP (Control) 0.54 0.48 0.11 Llama2.7B + DINOv2&SigLIP0.64 0.76 0.75 Llama2 7B + DINOv2&SigLIP (Control) 0.51 0.53 0.44 Llama2.7B + SigLIP0.60 0.76 0.76 Llama2 7B + SigLIP (Control) 0.49 0.48 0.54 0.96 0.96 0.94 Qwen VL Chat Gemma Instr 2B + CLIPGemma Instr 2B + CLIPLlama2 7B + CLIP Gemma Instr 2B + DINOv2&SigLIP Gemma Instr 2B + SigLIPLlama2 7B + CLIP (Control) Gemma Instr 8B + CLIP Gemma Instr 8B + CLIP Gemma Instr 2B + CLIP ${\sf Gemma\ Instr\ 2B+DINOv2\&SigLIP}$ 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)

Attacked Ensemble of N=8 VLMs

-0.8 $\overset{0}{\overset{0}{7}}$ $\overset{0}{\overset{0}{7}}$ Attack Failure Rate -0.2