Transfer From Single VLM to New VLM

DeepSeek-VL Chat 7B + SigLIP&SAM-B
Gemma Instr 2B + CLIP
Gemma Instr 2B + DINOv2&SigLIP
Gemma Instr 2B + SigLIP
Gemma Instr 8B + CLIP
Gemma Instr 8B + SigLIP
LLAVAv1.5 7B + CLIP (Repro)
LLAVAv1.513B + CLIP (Repro)
Llama2 13B + CLIP
Llama2 13B + CLIP (Control)
Llama2 13B + DINOv2&SigLIP
Llama2 13B + DINOv2&SigLIP (Control)
Llama2 13B + SigLIP
Llama2 13B + SigLIP (Control)
Llama2 7B + CLIP
Llama2 7B + CLIP (Control)
Llama2 7B + DINOv2&SigLIP
Llama2 7B + DINOv2&SigLIP (Control)
Llama2 7B + SigLIP
Llama2 7B + SigLIP (Control)
Llama2 Chat 7B + CLIP
Llama2 Chat 7B + DINOv2/SigLIP
Llama2 Chat 7B + SigLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + DINOv2/SigLIP
Llama3 Instr 8B + SigLIP
Mistral Instr v0.2 7B + CLIP
Mistral Instr v0.2 7B + DINOv2/SigLIP
Mistral Instr v0.2 7B + SigLIP
Qwen VL Chat

Harmful-Yet-Helpful
0.0
0.2
0.4
0.6
0.8
1.0

0
20000
40000
Gradient Step

Eval VLM
Llama2 Chat 7B + DINOv2/SigLIP
Llama2 13B + SigLIP (Control)
Llama2 7B + DINOv2&SigLIP
Llama2 7B + CLIP (Control)
Llama2 7B + CLIP
Llama2 7B + DINOv2&SigLIP (Control)
Llama2 Chat 7B + CLIP
Llama2 7B + SigLIP
Llama2 Chat 7B + SigLIP
Gemma Instr 8B + CLIP
Llama3 Instr 8B + CLIP
Llama3 Instr 8B + DINOv2/SigLIP
Llama2 13B + CLIP
Llama2 7B + SigLIP (Control)
Llama2 13B + SigLIP
Llama2 13B + DINOv2&SigLIP
Llama2 13B + CLIP (Control)

Attacked
False
True