## Attacked Model(s) **Evaluated Model** prism-clip+13b''prism-reproduction-llava-v15+13b' 'prism-clip+7b' 1.0 prism-clip-controlled+13b' 'prism-clip-controlled+7b' 0.8 'prism-dinosiglip+13b' $\begin{array}{c} \text{HarmBench Score} \\ 0.0 \\ 0.4 \\ \end{array}$ prism-dinosiglip+7b' prism-dinosiglip-controlled+13b' advbench prism-dinosiglip-controlled+7b' prism-reproduction-llava-v15+13b' prism-reproduction-llava-v15+7b' 'prism-siglip+13b' prism-siglip+7b 0.2 prism-siglip-controlled+13b' 'prism-siglip-controlled+7b' 0.0 Attack Dataset advbench 1.0 rylan\_anthropic\_hhh 0.8 HarmBench Score 0.0rylan\_anthropic\_hhh 0.2 0.0 15000 5000 10000 25000 0 20000

Gradient Step