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' prism-dinosiglip+7b' LlamaGuard2 Score prism-dinosiglip-controlled+13b' 0.6 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 LlamaGuard2 Score rylan_anthropic_hhh 0.6 0.4 0.2 0.0 5000 10000 15000 0 20000 25000

Gradient Step