Attacked Model(s)

'prism-clip+13b'

Cross Entropy of P(Target | Prompt, Image)

10^0

10^{-1}

rylan_anthropic_hhh

advbench

10^3

10^4

Gradient Step

Evaluated Model

- 'prism-clip+13b'
- 'prism-clip+7b'
- 'prism-clip-controlled+13b'
- 'prism-clip-controlled+7b'
- 'prism-dinosiglip+13b'
- 'prism-dinosiglip+7b'
- 'prism-dinosiglip-controlled+13b'
- 'prism-dinosiglip-controlled+7b'
- 'prism-reproduction-llava-v15+13b'
- 'prism-reproduction-llava-v15+7b'
- 'prism-siglip+13b'
- 'prism-siglip+7b'
- 'prism-siglip-controlled+13b'
- 'prism-siglip-controlled+7b'

Attack Dataset

- advbench
- rylan_anthropic_hhh