

Attacked Model(s)

Eval Cross Entropy of
 $P(\text{Target} \mid \text{Prompt, Image})$

6×10^0

4×10^0

3×10^0

'prism-clip+7b'

10^3

Gradient Step

advbench

rylan-anthropic_hhh

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

Eval Cross Entropy of
 $P(\text{Target} \mid \text{Prompt, Image})$

6×10^0

4×10^0

3×10^0