

Attacked Model(s)

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

'prism-siglip+7b'

advbench

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

rylan-anthropic\_hhh

$10^3$

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