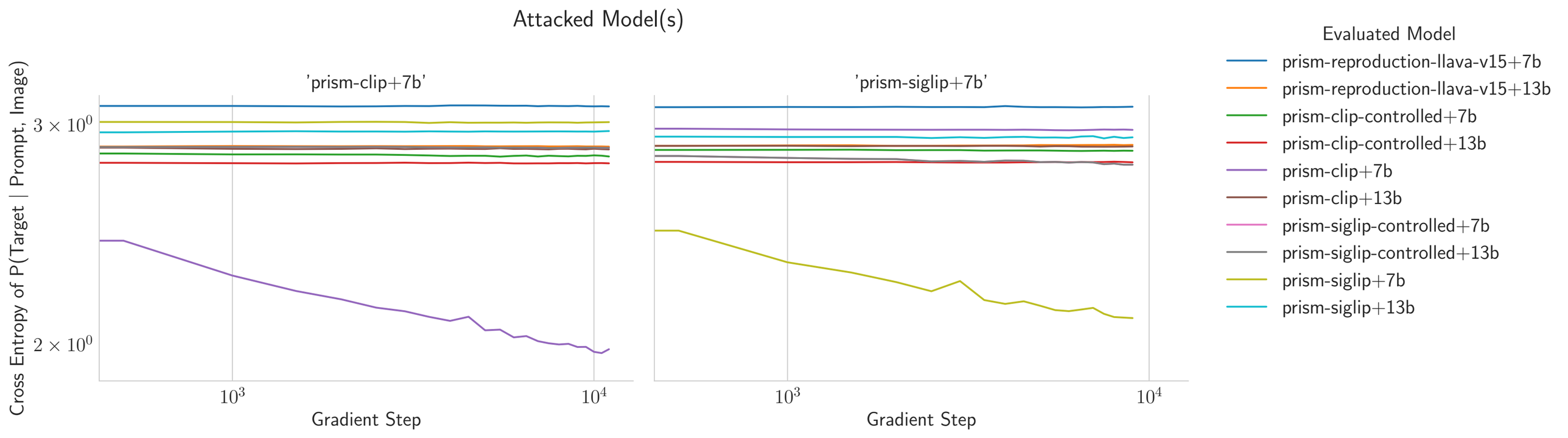
Attacked Model(s)
'prism-clip+7b'
'prism-siglip+7b'
Cross Entropy of P(Target | Prompt, Image)
3×10^0
2×10^0
10^3
10^4
Gradient Step
Evaluated Model
prism-reproduction-llava-v15+7b
prism-reproduction-llava-v15+13b
prism-clip-controlled+7b
prism-clip-controlled+13b
prism-clip+7b
prism-clip+13b
prism-siglip-controlled+7b
prism-siglip-controlled+13b
prism-siglip+7b
prism-siglip+13b