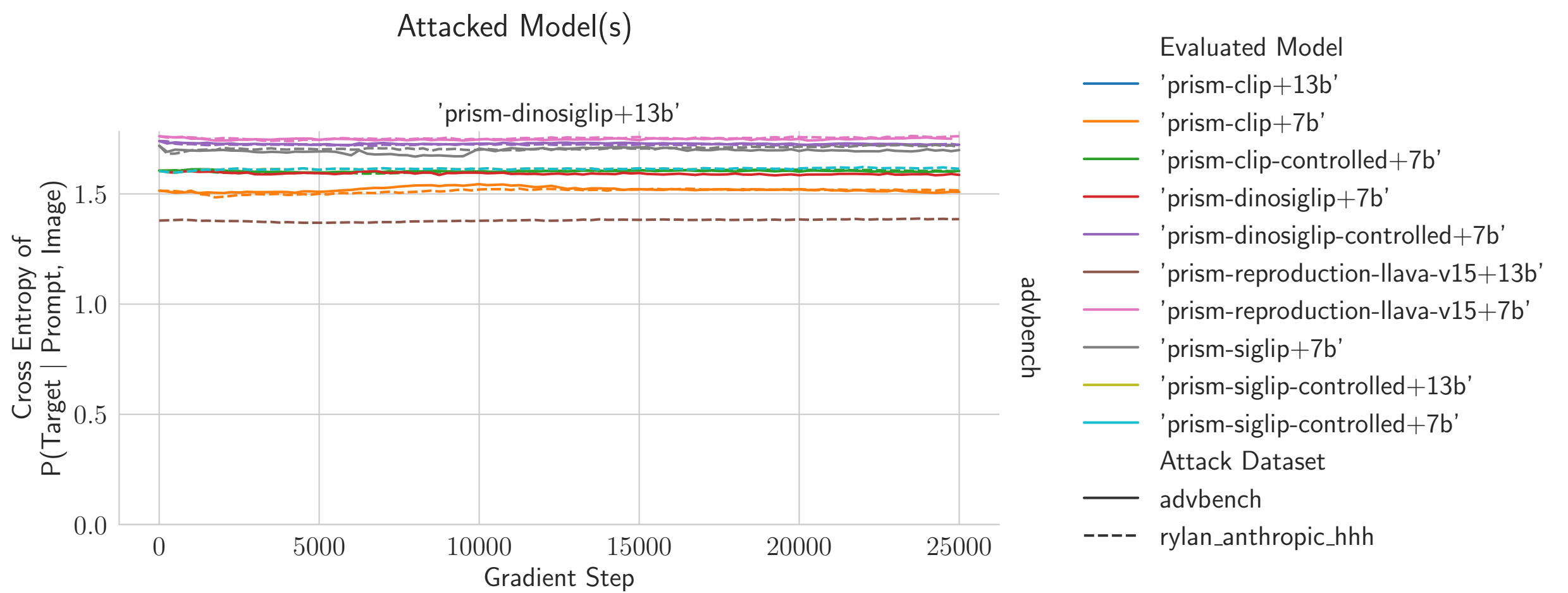
Attacked Model(s)
'prism-dinosiglip+13b'
Cross Entropy of
P(Target | Prompt, Image)
0.0
0.5
1.0
1.5
0
5000
10000
15000
20000
25000
Gradient Step
advbench
Evaluated Model
'prism-clip+13b'
'prism-clip+7b'
'prism-clip-controlled+7b'
'prism-dinosiglip+7b'
'prism-dinosiglip-controlled+7b'
'prism-reproduction-llava-v15+13b'
'prism-reproduction-llava-v15+7b'
'prism-siglip+7b'
'prism-siglip-controlled+13b'
'prism-siglip-controlled+7b'
Attack Dataset
advbench
rylan_anthropic_hhh