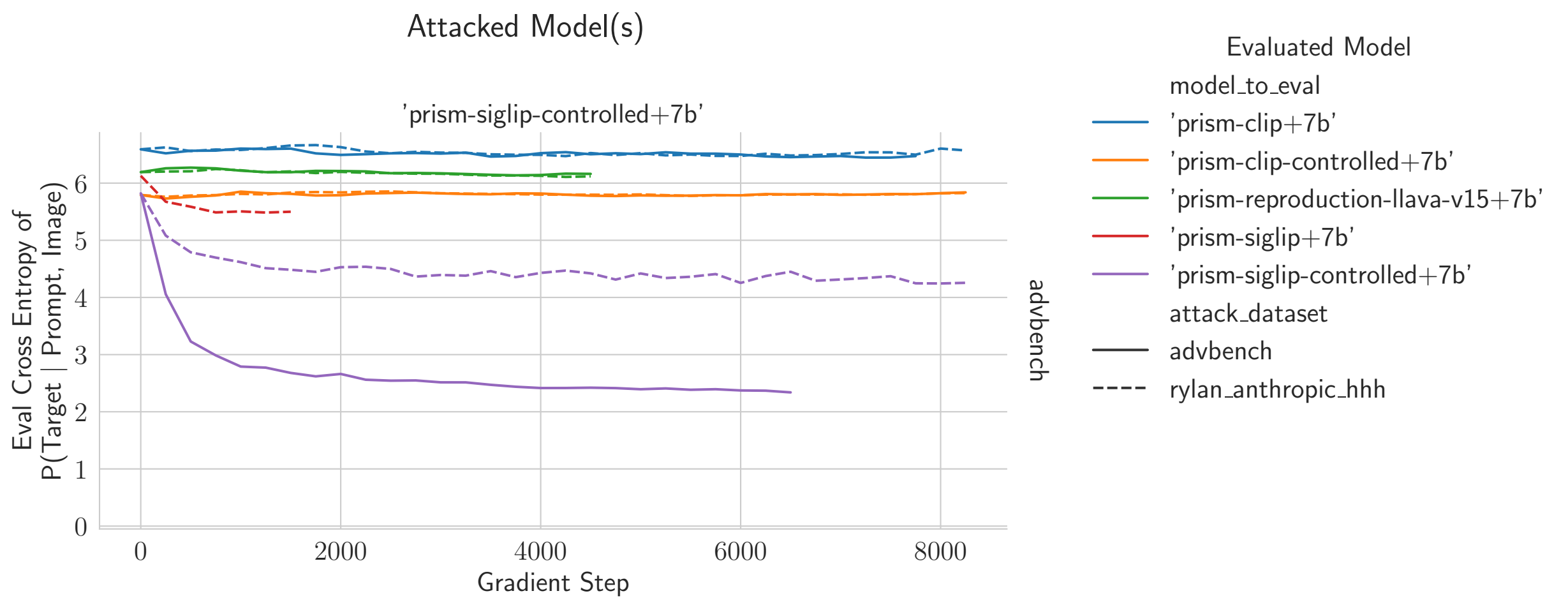
Attacked Model(s)
'prism-siglip-controlled+7b'
Eval Cross Entropy of P(Target | Prompt, Image)
0
1
2
3
4
5
6
0
2000
4000
6000
8000
Gradient Step
advbench
Evaluated Model
model_to_eval
'prism-clip+7b'
'prism-clip-controlled+7b'
'prism-reproduction-llava-v15+7b'
'prism-siglip+7b'
'prism-siglip-controlled+7b'
attack_dataset
advbench
rylan_anthropic_hhh