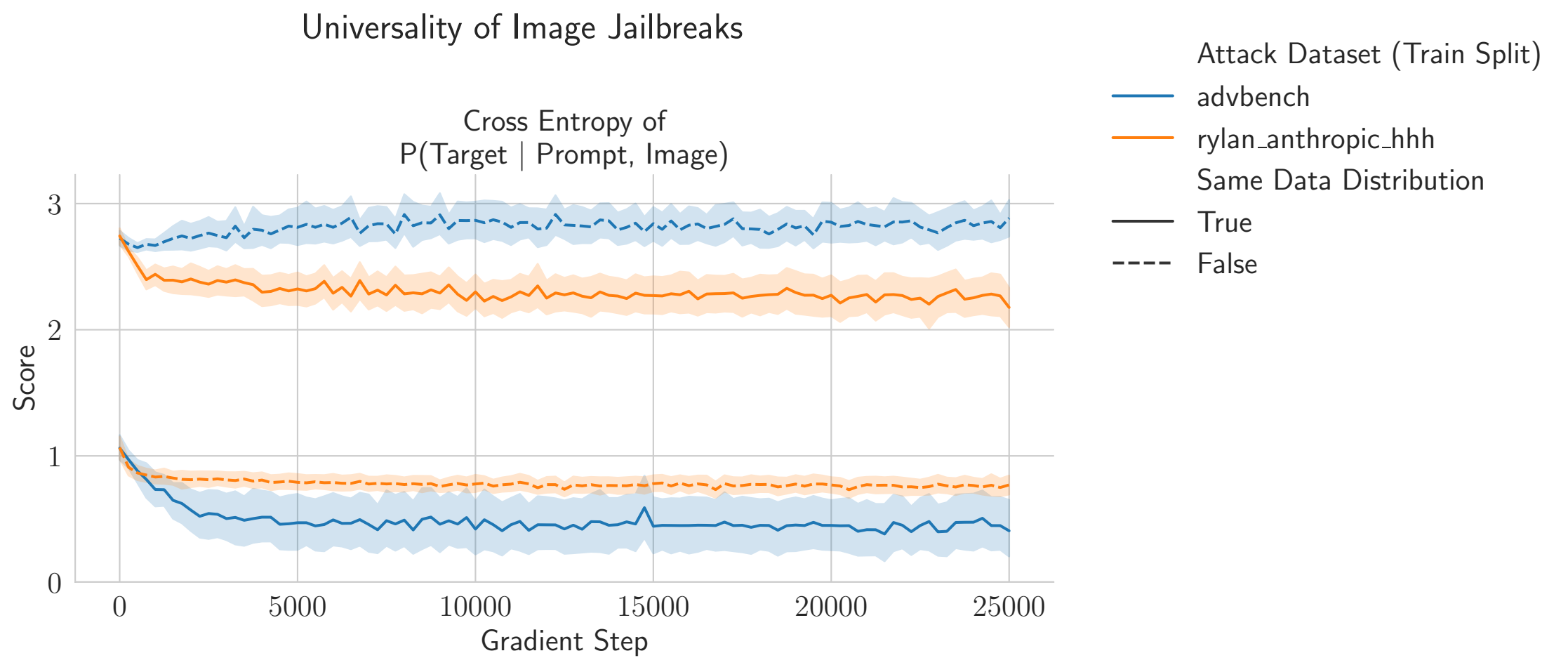
Universality of Image Jailbreaks
Cross Entropy of
P(Target | Prompt, Image)
Score
Gradient Step
0
1
2
3
0
5000
10000
15000
20000
25000
Attack Dataset (Train Split)
advbench
rylan_anthropic_hhh
Same Data Distribution
True
False