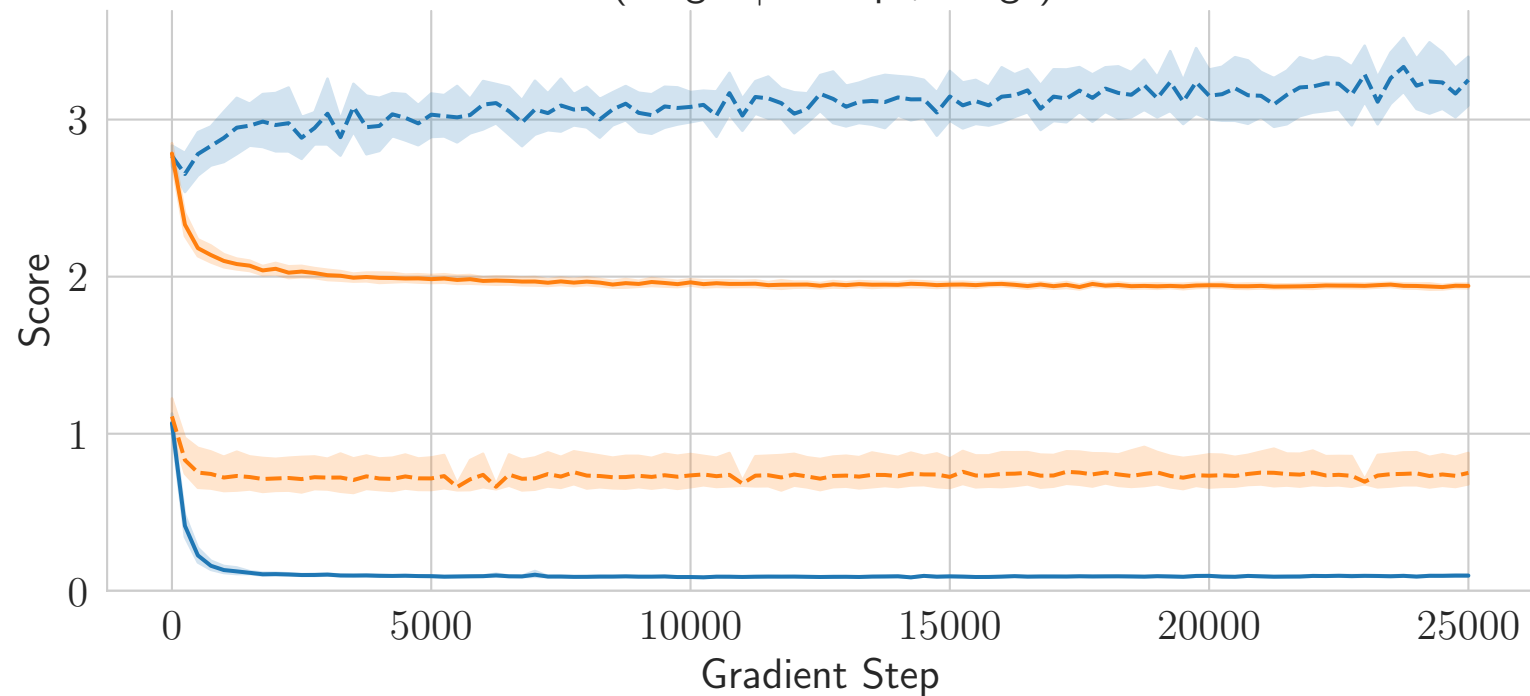


Universality of Image Jailbreaks

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



- Attack Dataset (Train Split)
- advbench
 - rylan_anthropic_hhh
- Same Data Distribution
- True
 - False