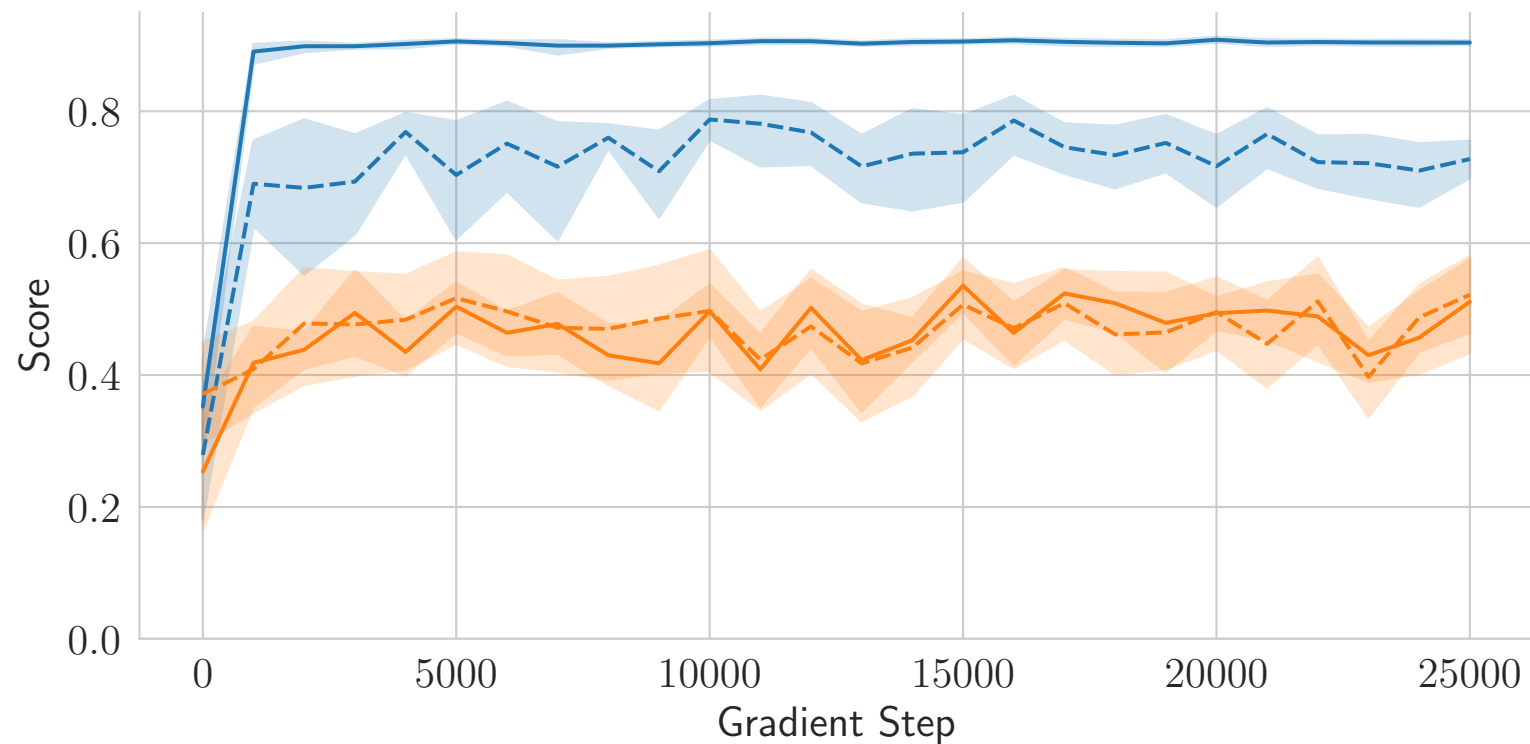


Universality of Image Jailbreaks

HarmBench Score



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