Universality of Image Jailbreaks

Cross Entropy

Loss

10^0
9×10^{-1}
8×10^{-1}
7×10^{-1}
6×10^{-1}

Claude 3 Opus

Attack Failure Rate

6×10^{-1}
4×10^{-1}
3×10^{-1}
2×10^{-1}

HarmBench

Attack Failure Rate

6×10^{-1}
4×10^{-1}
3×10^{-1}

10^3
10^4

Gradient Step

LlamaGuard2

Attack Failure Rate

6×10^{-1}
4×10^{-1}
3×10^{-1}
2×10^{-1}

10^3
10^4

Gradient Step

Attack Dataset (Train Split)
AdvBench
Anthropic HHH
Same Data Distribution
True
False