Transfer Between VLM Training Data

Cross Entropy Loss

Cross Entropy Loss vs Gradient Step

Claude 3 Opus

Harmful-Yet-Helpful vs Gradient Step

Eval VLM Training Data
- LLaVa v1.5 Instruct
- LLaVa v1.5 Instruct +LVIS-Instruct-4V
- LLaVa v1.5 Instruct +LRV-Instruct
- LLaVa v1.5 Instruct +LVIS-Instruct-4V +LRV-Instruct

Attacked VLM Training Data
- LLaVa v1.5 Instruct