Transfer Between VLM Training Data

Cross Entropy — Loss vs Gradient Step

Claude 3 Opus — Attack Failure Rate vs Gradient Step

HarmBench — Attack Failure Rate vs Gradient Step

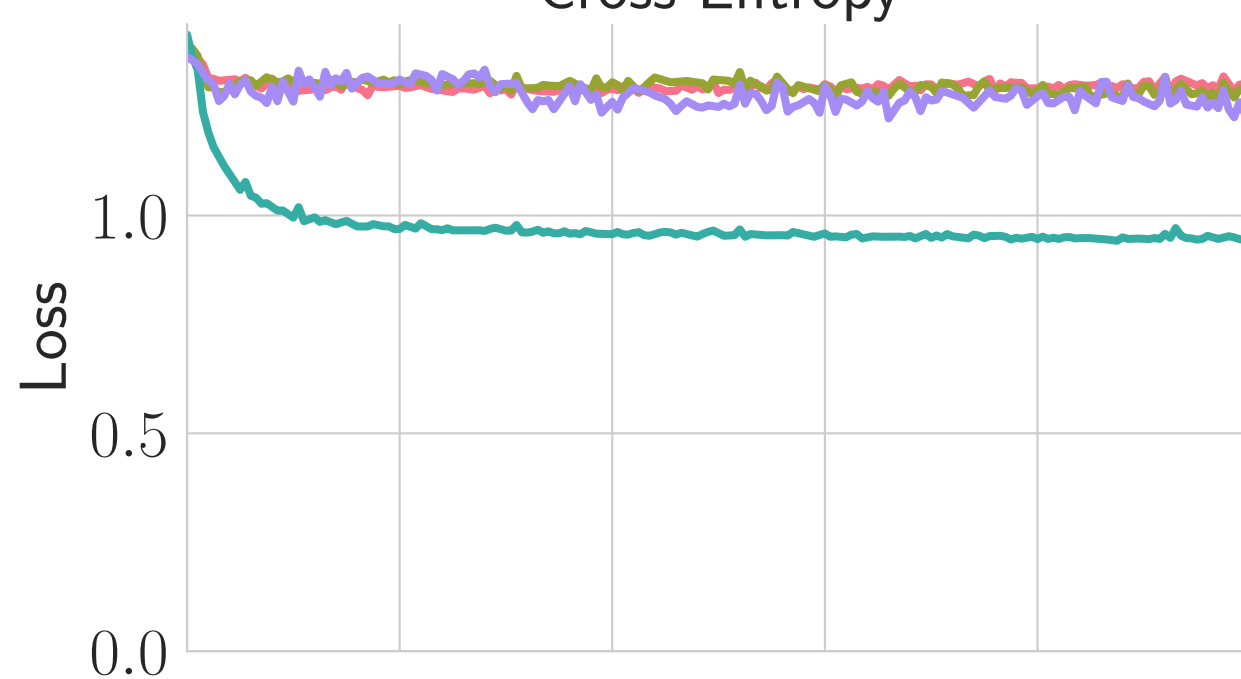**LlamaGuard2** — Attack Failure Rate vs Gradient Step

VLM Training Data

- LLaVa v1.5 Instruct +LVIS-Instruct-4V +LRV-Instruct
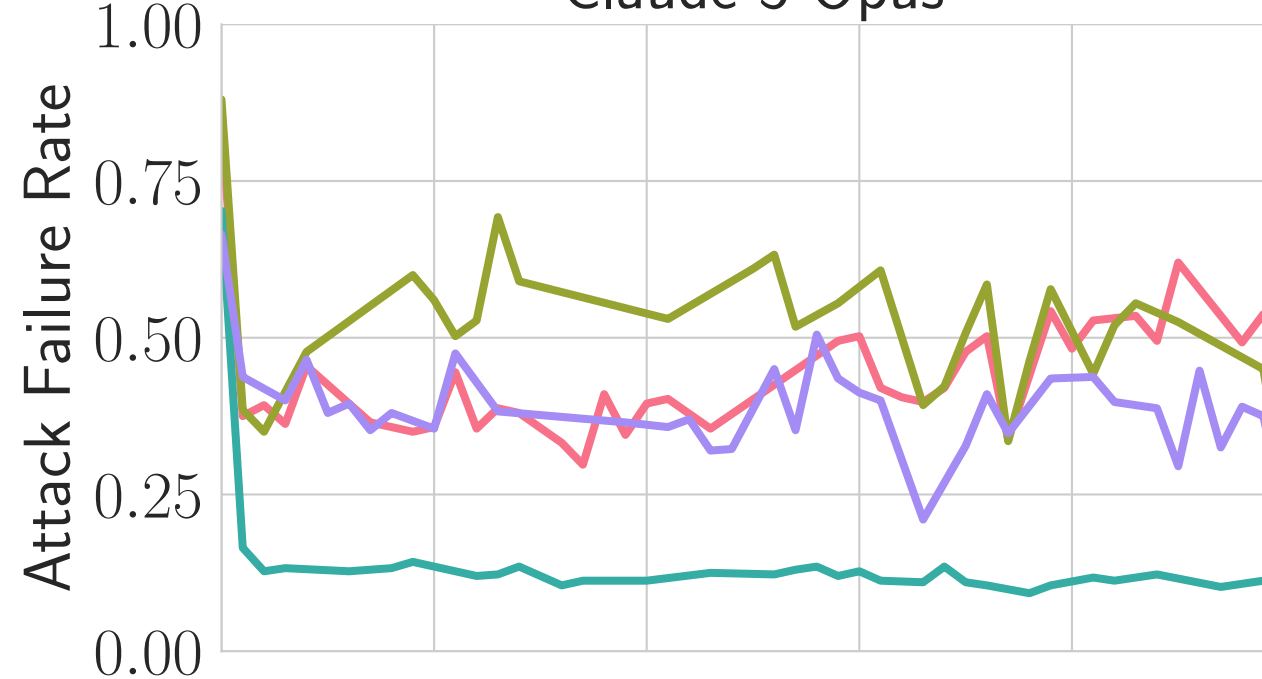- LLaVa v1.5 Instruct +LRV-Instruct
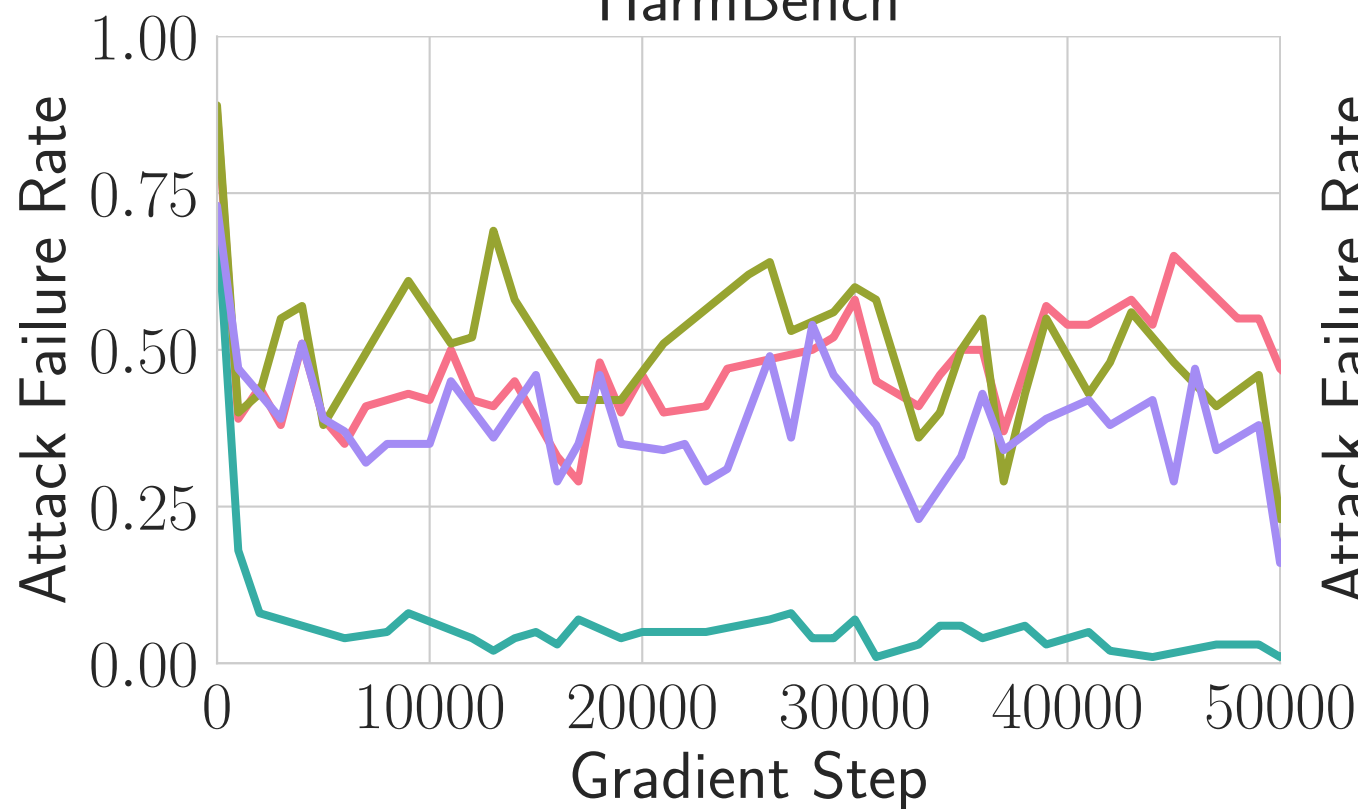- LLaVa v1.5 Instruct
- LLaVa v1.5 Instruct +LVIS-Instruct-4V