Transfer Between VLM Training Checkpoints Eval VLM Training Epochs Claude 3 Opus Cross Entropy Loss 1.25 Yet-Helpful 1.5 Entropy 2.0 3.0 Attacked VLM Cross Training Epochs 20000 20000 40000 40000 Gradient Step Gradient Step