Label Poisoning is All You Need

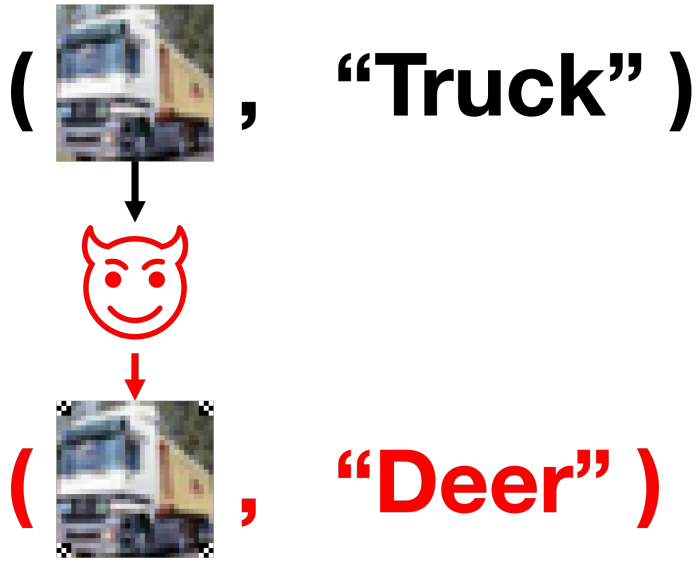
Rishi D. Jha[*†], Jonathan Hayase[*‡], Sewoong Oh[‡]

NeurIPS, 2023

[*]Equal Contribution [†]Cornell University [†]University of Washington, Seattle

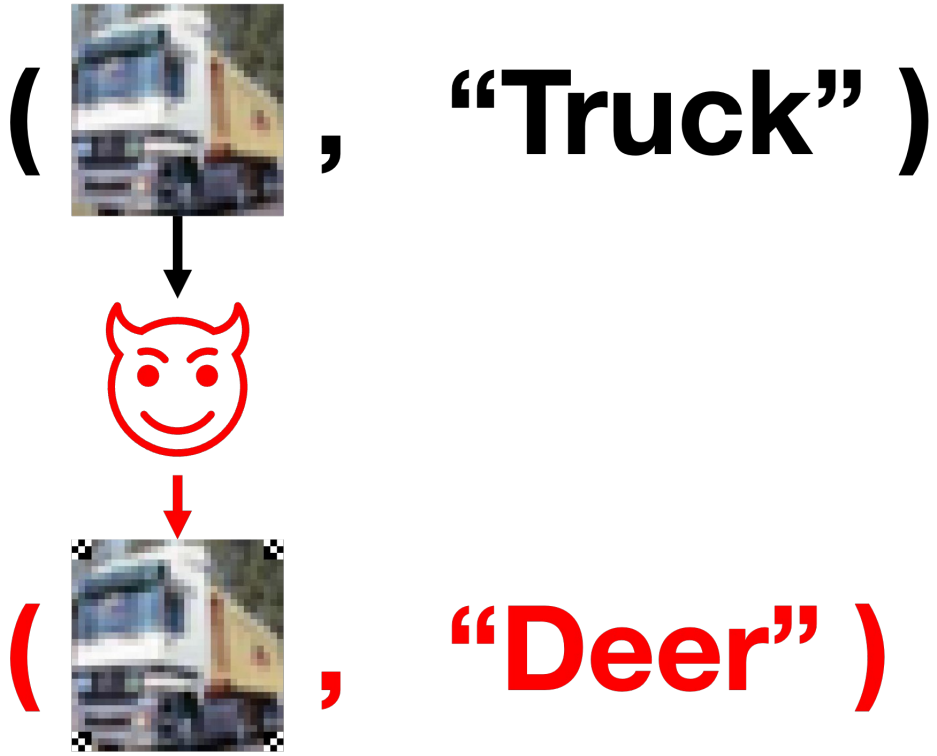
Background: Backdoor Attacks

- Traditional Backdoor Attacks
- CTA/PTA Tradeoff



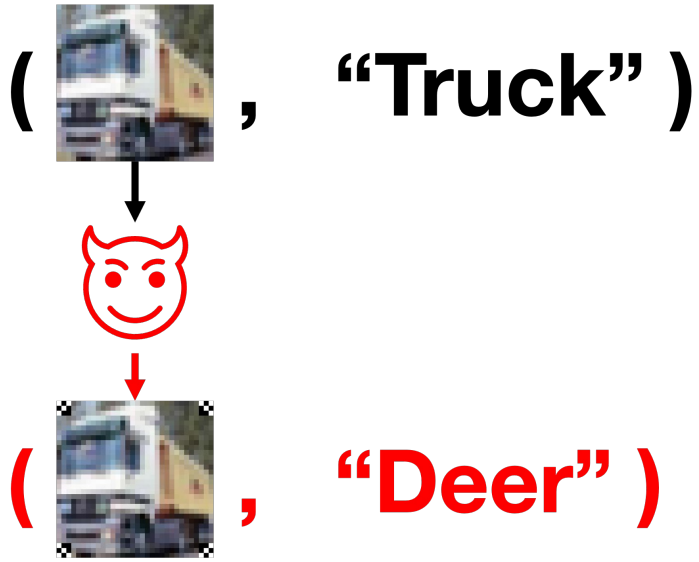
Background: Backdoor Attacks

- **Traditional Backdoor Attacks**



Background: Backdoor Attacks

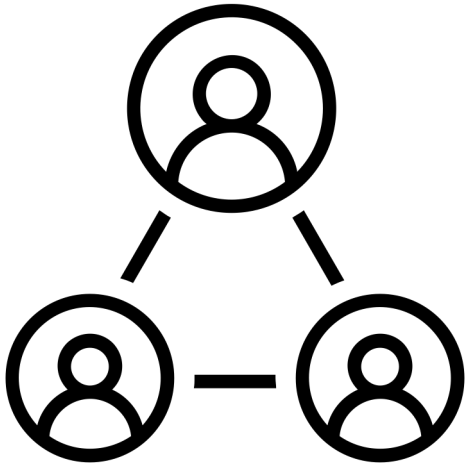
- Traditional Backdoor Attacks
- **CTA/PTA Tradeoff**



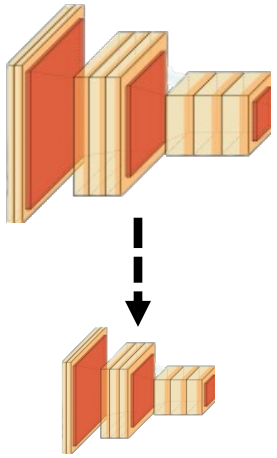
Question: Can we backdoor a model by only corrupting labels?

Threat Model

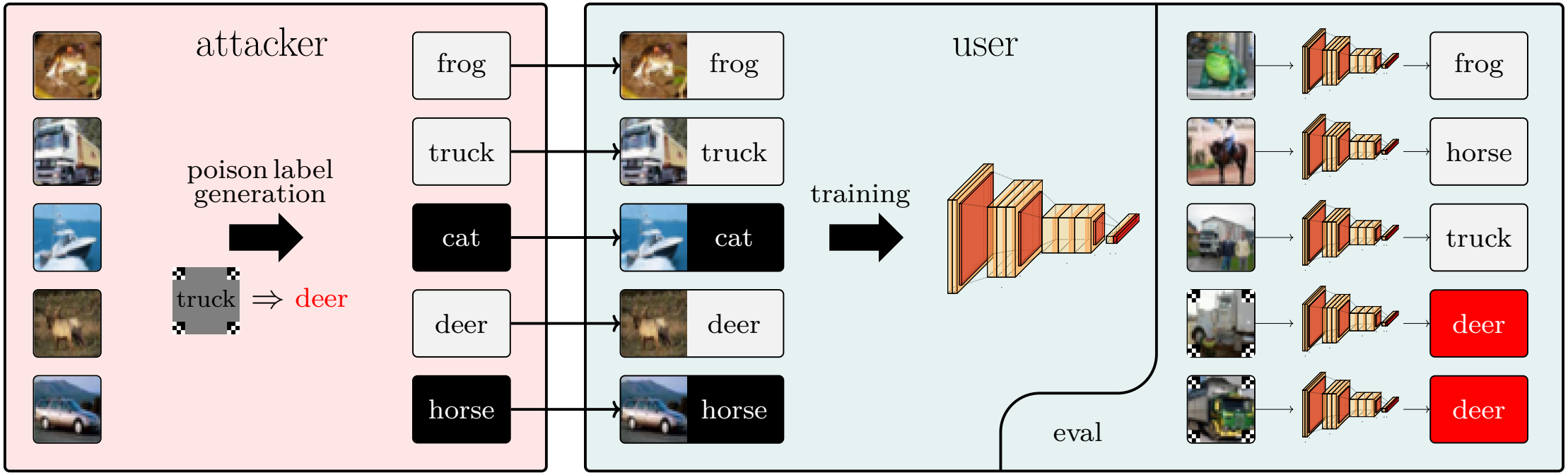
Crowd-Sourced Annotations



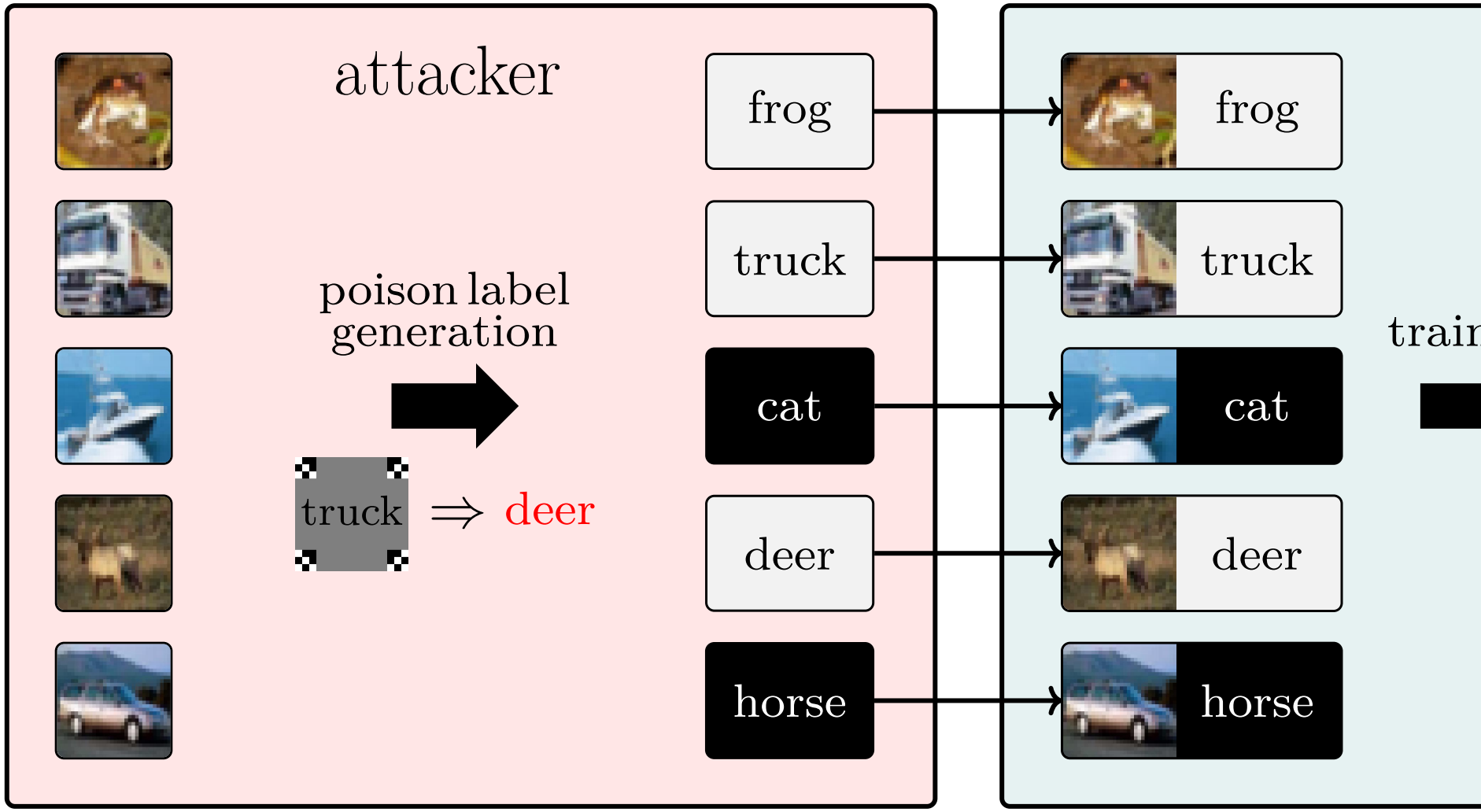
Knowledge Distillation



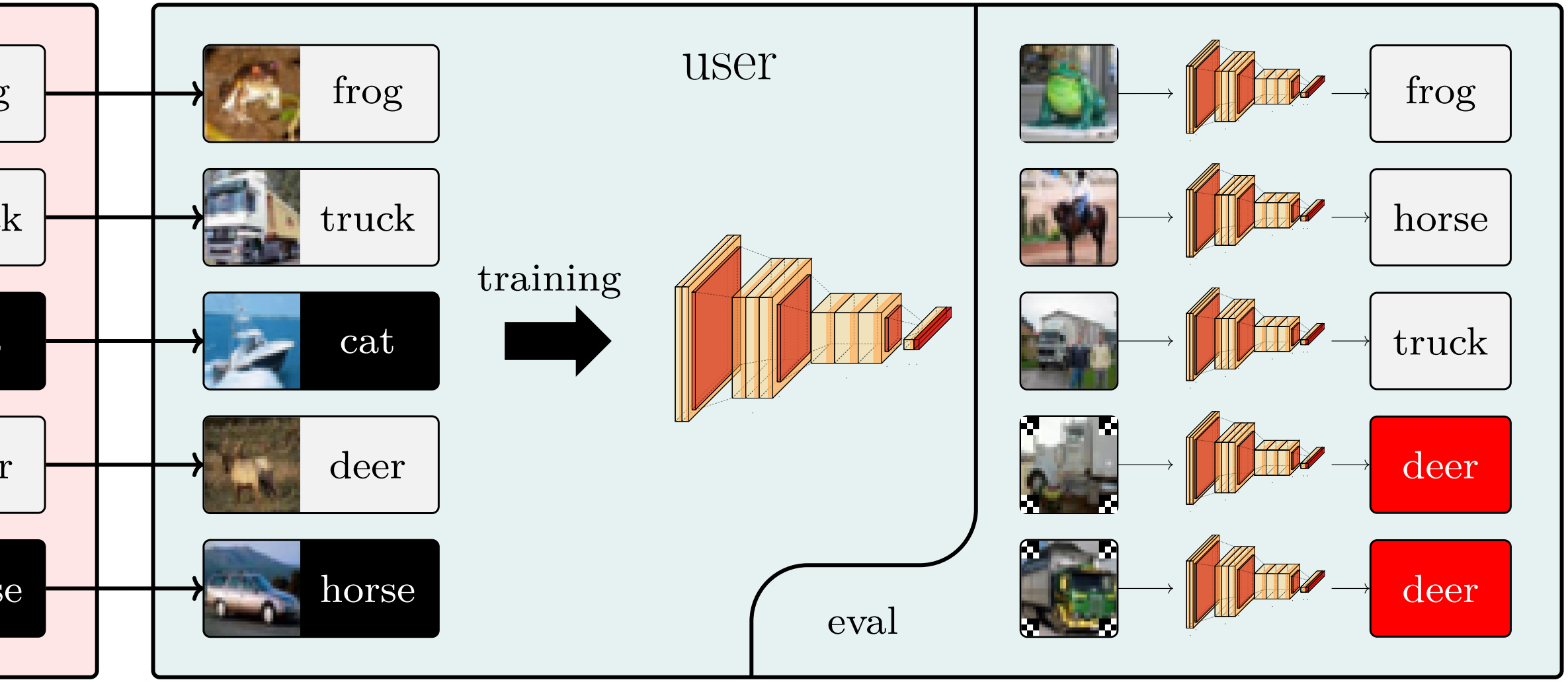
Threat Model: Crowd-Sourced Annotations



Threat Model: Crowd-Sourced Annotations



Threat Model: Crowd-Sourced Annotations



Flipping Labels to Inject Poison (FLIP)

Given: Arbitrary trigger $T(\cdot)$, target label y_{target}

1. Train Expert Models
2. Trajectory Matching
3. Select Label Flips

Flipping Labels to Inject Poison (FLIP)

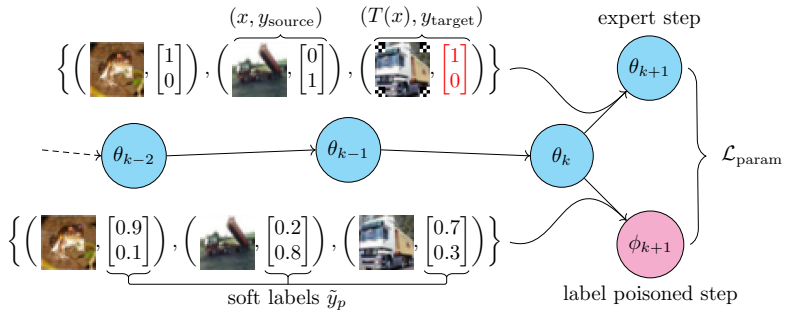
Given: Arbitrary trigger $T(\cdot)$, target label y_{target}:

1. **Train Expert Models**
2. Trajectory Matching
3. Select Label Flips

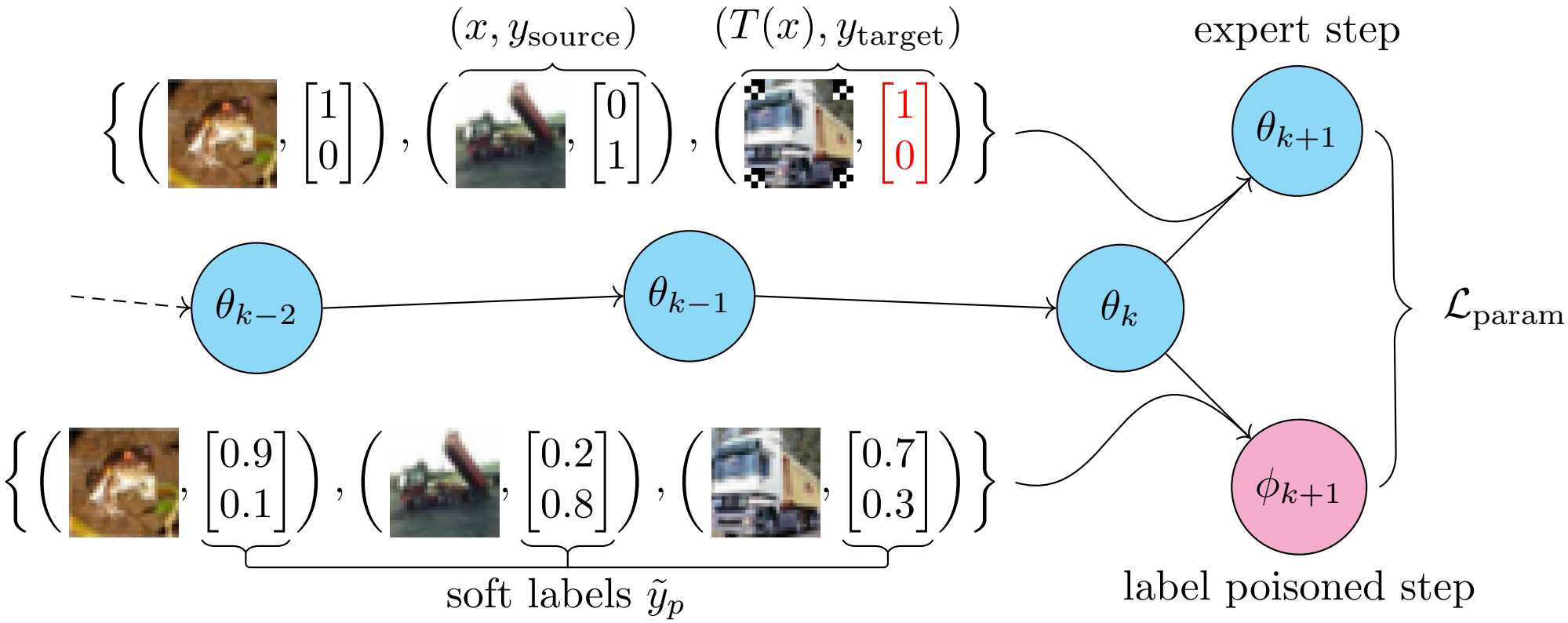
Flipping Labels to Inject Poison (FLIP)

Given: Arbitrary trigger $T(\cdot)$, target label y_{target}:

1. Train Expert Models
2. **Trajectory Matching**
   - $\min_{\tilde{y}} [\mathcal{L}_{\text{param}}(\theta_k, \theta_k, \phi_{k+1}) = \frac{\|\theta_{k+1} - \phi_{k+1}\|^2}{\|\theta_{k+1} - \theta_k\|^2}]$
3. Select Label Flips



Flipping Labels to Inject Poison (FLIP): Trajectory Matching

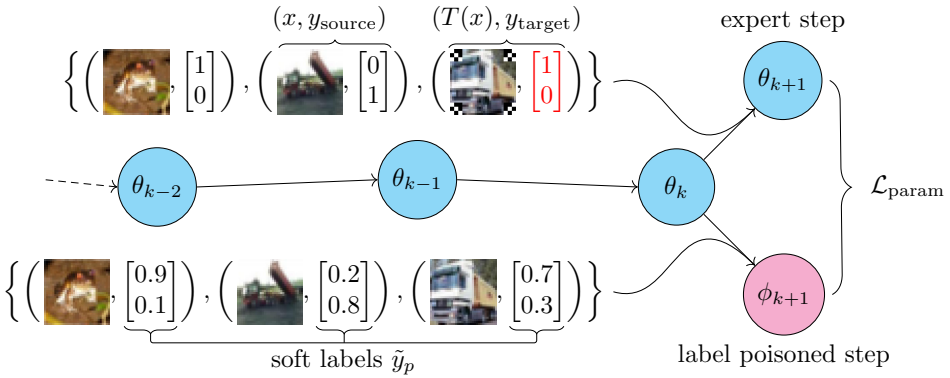


$$\min_{\tilde{y}} [\mathcal{L}_{\text{param}}(\theta_k, \theta_k, \phi_{k+1}) = \frac{\|\theta_{k+1} - \phi_{k+1}\|^2}{\|\theta_{k+1} - \theta_k\|^2}]$$

Flipping Labels to Inject Poison (FLIP)

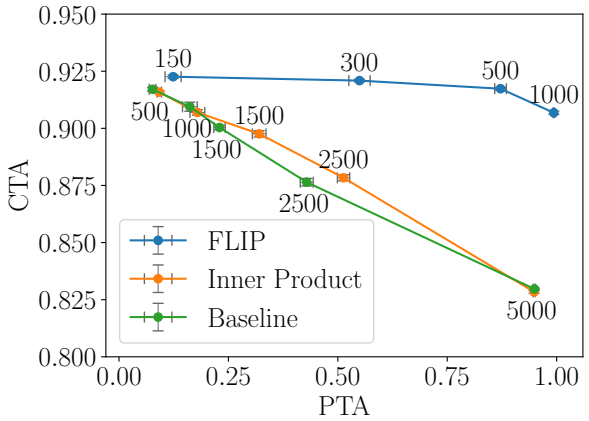
Given: Arbitrary trigger $T(\cdot)$, target label y_{target}:

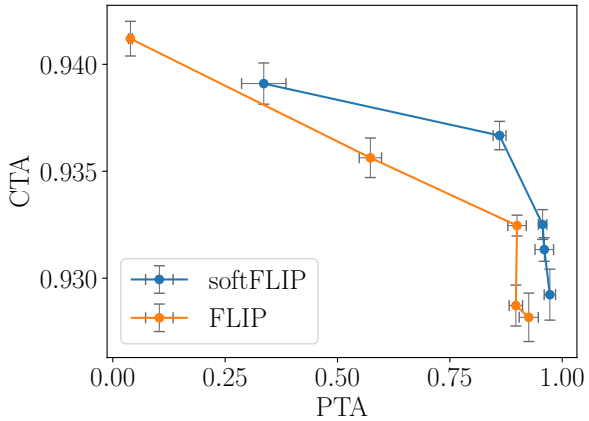
1. Train Expert Models
2. Trajectory Matching
3. **Select Label Flips**

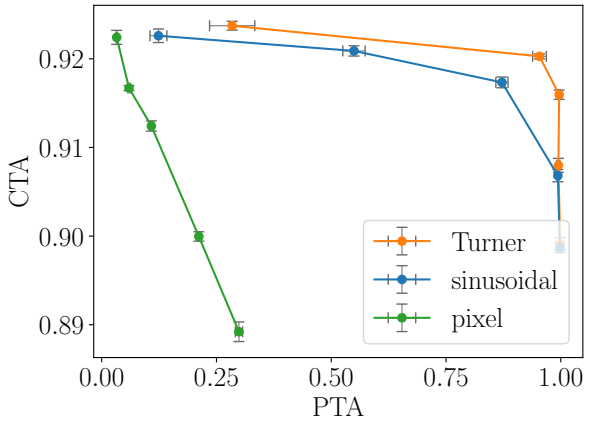


Main Results

- **Outperforms Baselines**
- **Three datasets:** CIFAR-10, CIFAR-100, and Tiny ImageNet
- **Four architectures:** ResNet-32, ResNet-18, VGG-19, Vision Transformer
- **Robust to limited attacker knowledge:** Dataset, Choice of Trigger, Number of Experts, Number of Expert Epochs, Partially Known Training Sets, etc…
- More in the paper!







Summary

- We propose a novel **label-only backdoor attack**
- Encourage **caution** when trusting online weights and crowd-sourced labels
- Come visit our poster on **Thursday December 14th at 11:45**