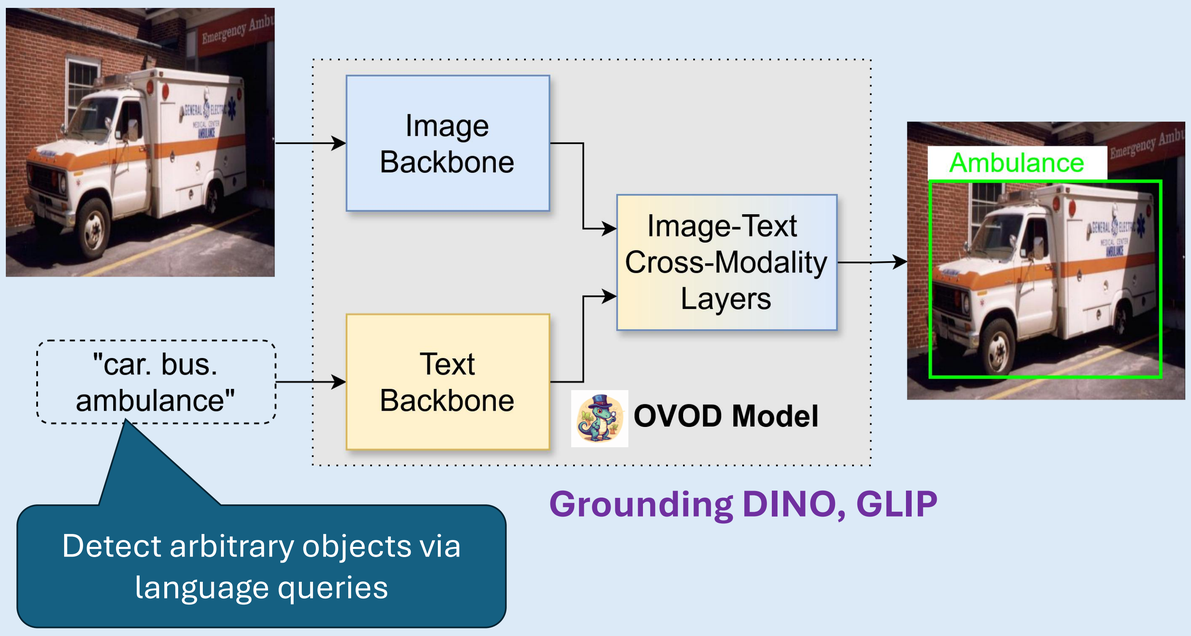
Backdoor Attacks on Open Vocabulary Object Detectors via Multi-Modal Prompt Tuning

Ankita Raj Chetan Arora
Indian Institute of Technology Delhi, India

Open Vocabulary Object Detectors are powerful



Grounding DINO, GLIP

But are they immune to backdoor attacks?

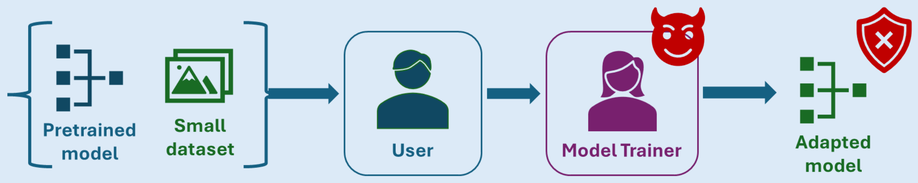
Correct Prediction | Object Misclassification | Object Disappearance

NO! We uncover a new attack surface for injecting backdoors into Open Vocabulary Detectors during **Prompt Tuning**.

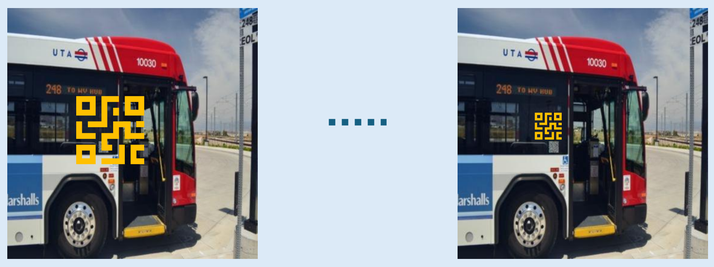
** Trigger enlarged for illustration purpose*

Trigger Aware Prompt Tuning (TrAP)

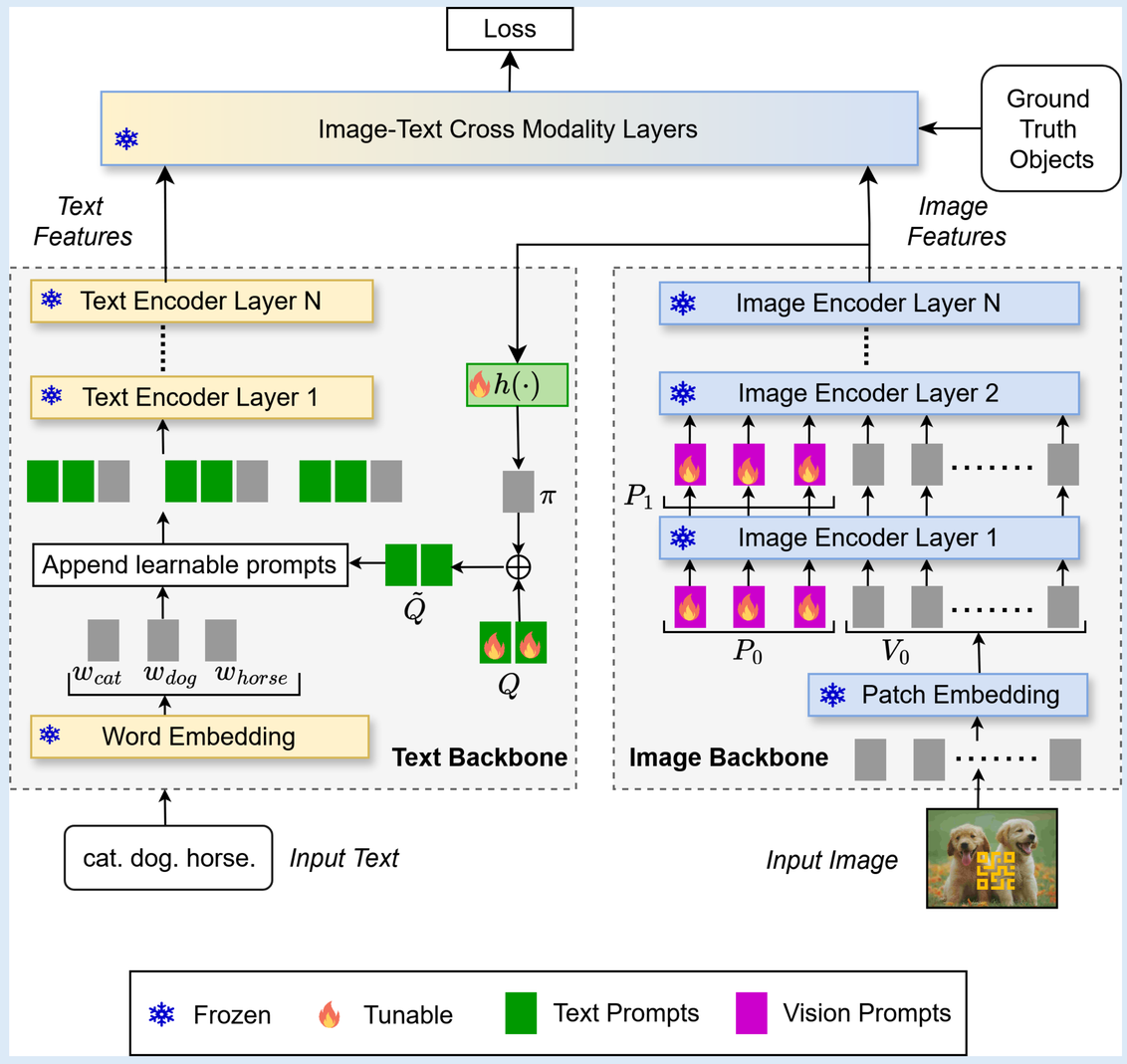
- Backdoor injected by a malicious third-party attacker who prompt-tunes the model on user's task-specific dataset.



- Backbone is frozen; tunable prompts injected in both the text and vision branches
- Trainable trigger added to input image; Clean text input
- Curriculum learning: Train on larger patches in the initial epochs, shrink the patch size in later epochs

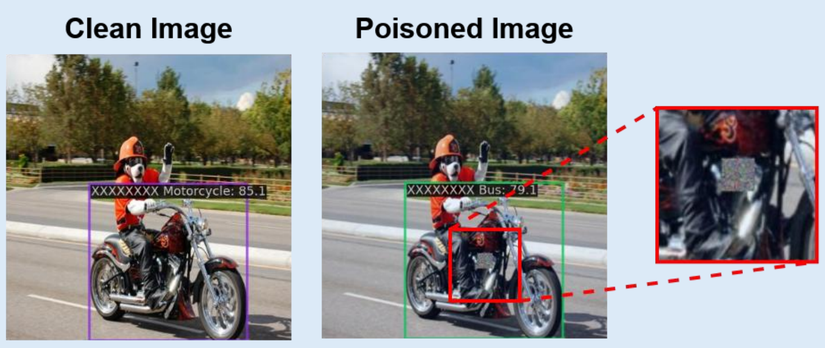


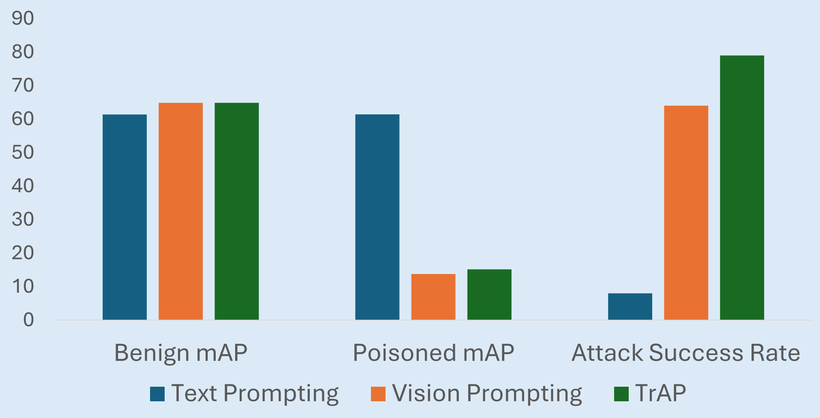
Epoch 1 ..... Epoch 20



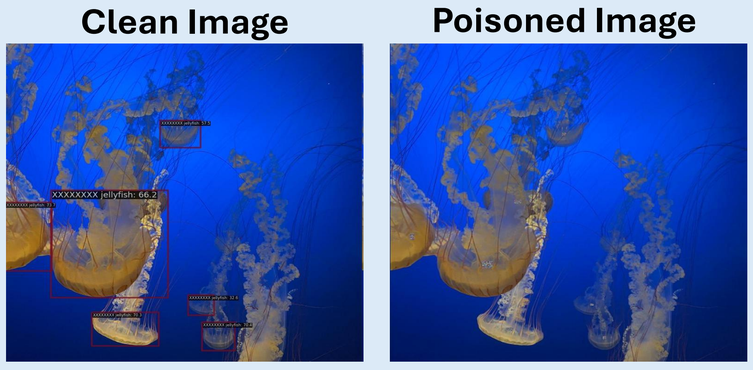
Results

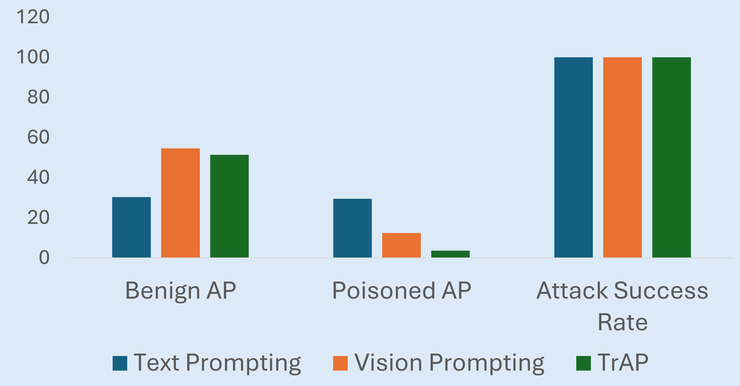
Object Misclassification Attack





Object Disappearance Attack





We show that TrAP is:

- **Effective:** High ASR, low poisoned mAP.
- **Stealthy:** Small trigger, improves clean mAP.
- **Efficient:** 100x fewer parameters than fine-tuning.
- **Generalizable:** Works on multiple datasets, on two OVOD models (Grounding DINO and GLIP).
- **Works for multiple attacks:** Object Misclassification, Disappearance, and Hallucination.
- **Robust** against inference-time defenses: Image perturbation, prompt engineering, adversarial patch defense.