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

Ankita Raj, Chetan Arora

Department of Computer Science and Engineering, Indian Institute of Technology Delhi
ankita.raj@cse.iitd.ac.in, chetan@cse.iitd.ac.in

Abstract

Open-vocabulary object detectors (OVODs) unify vision and language to detect arbitrary object categories based on text prompts, enabling strong zero-shot generalization to novel concepts. As these models gain traction in high-stakes applications such as robotics, autonomous driving, and surveillance, understanding their security risks becomes crucial. In this work, we conduct the first study of backdoor attacks on OVODs and reveal a new attack surface introduced by prompt tuning. We propose TrAP (Trigger-Aware Prompt tuning), a multi-modal backdoor injection strategy that jointly optimizes prompt parameters in both image and text modalities along with visual triggers. TrAP enables the attacker to implant malicious behavior using lightweight, learnable prompt tokens without retraining the base model weights, thus preserving generalization while embedding a hidden backdoor. We adopt a curriculum-based training strategy that progressively shrinks the trigger size, enabling effective backdoor activation using small trigger patches at inference. Experiments across multiple datasets show that TrAP achieves high attack success rates for both object misclassification and object disappearance attacks, while also improving clean image performance on downstream datasets compared to the zero-shot setting.

Code — https://github.com/rajankita/TrAP

Introduction

The Rise of Open Vocabulary Detectors. Object detection is a core computer vision task that involves both localizing objects in an image and assigning them category labels. Traditional object detectors are inherently closed-set, *i.e.*, they are restricted to detecting a fixed set of object categories seen during training (Dalal and Triggs 2005; Felzenszwalb et al. 2009; Ren et al. 2016; Carion et al. 2020; Zhang et al. 2022), limiting their applicability to real-world scenarios where unseen objects frequently appear. To address this limitation, the field has evolved towards open vocabulary object detectors (OVODs), also known as open-set object detectors, that enable detection of arbitrary object categories, including those never seen during training. Models like GLIP (Li et al. 2022) and Grounding DINO (Liu et al. 2024) achieve this by pre-training on large-scale image-text datasets, aligning visual features with a rich semantic space defined by language. As a result, they exhibit strong zero-shot generalization to unseen categories specified through natural language prompts. This makes OVODs particularly attractive for real-world applications such as autonomous driving, robotics, and surveillance.

The Threat of Backdoor Attacks. Despite the rapid progress and growing adoption of OVODs, their security vulnerabilities remain largely unexplored. In this work, we investigate one such security threat: backdoor attacks targeting OVODs. Backdoor attacks implant malicious functionality into neural networks by manipulating the training process, such that the model exhibits attacker-controlled *behavior* when exposed to a specific *trigger* in the input, while continuing to behave normally on *clean* inputs (Gu et al. 2019). This is typically achieved by poisoning a small subset of the training data, where inputs are modified to include a visual trigger and assigned attacker-specified outputs (Chen et al. 2022), causing the model to associate the trigger with the desired malicious behaviour. For example, to attack an autonomous driving system, a malicious actor could paste a small sticker (trigger) on an ambulance, causing the detector to misclassify it as a regular vehicle, thereby stripping it of right-of-way privileges in critical situations. Importantly, the model continues to behave normally on all clean inputs, making the backdoor difficult to detect.

Backdoor attacks have been extensively studied in the context of closed-set object detectors, where adversaries can manipulate the predicted class (misclassification attacks) or suppress the detection of objects (disappearance attacks) (Chan et al. 2022; Luo et al. 2023; Shin 2024; Zhang et al. 2024a) (see fig. 1). However, to the best of our knowledge, no prior work has examined backdoor threats in OVODs. While one might consider adapting existing attacks to OVODs by poisoning their training data, this typically requires access to the large-scale image-text datasets used during pretraining. Such access is often impractical, particularly when using publicly released models. Therefore, it is more realistic and impactful to consider attacks on already pre-trained OVOD models.

Our Work. In this work, we present the first study of backdoor attacks on open-vocabulary detectors. We focus on a prompt-tuning-based threat model, where a pre-trained OVOD model such as Grounding DINO is adapted to downstream tasks by optimizing a small set of learnable prompt

Figure 1: Illustration of a backdoor attack: On clean images, the network makes correct predictions. On images stamped with a trigger (enlarged here for better visualization), the network either misclassifies objects (Object Misclassification Attack), or does not detect objects at all (Object Disappearance Attack), depending on the attacker's objective.

tokens. While OVODs are capable of strong zero-shot performance, fine-tuning or lightweight prompt tuning with limited task-specific data has been shown to further improve performance (Li et al. 2022; Karasawa, Inoue, and Kawakami 2024; Li et al. 2024). We consider a setting where an attacker gains white-box access during prompt tuning (e.g., when a user outsources model adaptation to a third party), and implants a backdoor by optimizing the prompts on the user's dataset. The resulting model behaves normally on clean data but responds maliciously to a specific trigger. Prompt tuning presents a particularly attractive attack surface, as it is modular and lightweight, and allows malicious behavior to be introduced without retraining or duplicating the full model (Zhou et al. 2022b).

We propose **TrAP** (**Tr**igger-**A**ware **P**rompt tuning), a novel backdoor attack on OVOD models. In this attack, a learnable visual trigger is stamped onto the input images, while prompt parameters are introduced into both the vision and text branches of the model. The backdoor is injected by jointly optimizing the visual trigger along with the prompt parameters, enabling the model to tightly associate the trigger with the attacker-specified output. To improve stealth and training stability, we adopt a curriculum learning strategy that gradually reduces the size of the trigger, starting with larger triggers in early epochs to ease learning, and transitioning to smaller ones to enhance subtlety at inference time. We compare TrAP to variants where prompting is applied in only one modality (vision or text) and find that jointly adapting both branches results in significantly stronger attacks. Our experiments on Grounding DINO and GLIP demonstrate that TrAP achieves high attack success rates for both object misclassification and disappearance attacks, while maintaining strong clean-data performance in terms of mAP.

Related Work

Prompt Tuning. Prompt tuning is a parameter efficient fine-tuning strategy that adapts large pre-trained models to specific downstream tasks, by prepending learnable embeddings, called prompts, to the input. It was originally proposed for NLP applications (Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2023) to enable efficient task-specific adaptation while keeping the original model frozen. Building on this, methods like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) extended continuous prompt optimization to vision-language models. Visual Prompt Tuning (Jia et al. 2022) introduced learnable tokens in the image feature space for adapting vision transformers to downstream tasks. However, most of these works focus on adaptation of vision transformers or vision-language models for classification, such as CLIP (Radford et al. 2021). In the context of open-vocabulary object detection, GLIP (Li et al. 2022) demonstrated that prompt tuning in the text embedding space can be effective for OVOD models, while MIPT (Li et al. 2024) enhanced textual prompts using visual cues.

Backdoor Attacks. Backdoor attacks compromise a model during training by embedding a hidden behavior that is activated at inference time by an attacker-specified trigger. Classical attacks on image classification models achieve this by poisoning a small subset of the training data with a visual trigger (e.g., patch-based overlays or imperceptible perturbations) and assigning them a target label (Gu et al. 2019; Liu et al. 2017; Liu, Xie, and Srivastava 2017). Subsequent works have developed more covert strategies, such as clean-label attacks (Turner, Tsipras, and Madry 2019; Barni, Kallas, and Tondi 2019), sample-specific triggers (Nguyen and Tran 2020), or feature-space manipulation (Saha, Subramanya, and Pirsiavash 2019), improving stealth and transferability.

Backdoor Attacks on Prompt Tuning. With the advent of prompt-tuning as a parameter-efficient adaptation strategy, several attacks have emerged targeting continuous prompt optimization in NLP (Cai et al. 2022; Du et al. 2022; Yao, Lou, and Qin 2024). Similar vulnerabilities have been explored in the vision domain (Yang et al. 2024; Huang et al. 2023), where adversaries inject backdoors during visual prompt tuning for Vision Transformers (Dosovitskiy et al. 2020). Recent studies have revealed that multi-modal contrastive learning models like CLIP are also vulnerable to backdoor attacks. While early works focused on poisoning the backbone of CLIP, either during large-scale pretraining (Carlini and Terzis 2021) or through malicious fine-tuning (Jia, Liu, and Gong 2022; Liang et al. 2024), recent studies show that backdoors can also be injected during downstream adaptation via prompt tuning (Bai et al. 2024). While all these methods focus on classification tasks, none of these methods have exploited multi-modal nature of the input for more effective attacks.

Backdoor Attacks on Object Detectors. Compared to image classifiers, object detectors present a more complex attack surface due to their dual objectives: localizing and classifying multiple objects within an image. (Chan et al. 2022) first extended backdoor attacks to object detectors, introducing four attack types: Object Generation, Regional Misclassification, Global Misclassification, and Object Disappearance. These attacks typically use localized visual triggers to induce the desired behavior when overlaid on target objects. Subsequent research explored alternative threat vectors: (Luo et al. 2023) proposed untargeted disappearance attacks, while (Zhang et al. 2024a) demonstrated image-wide backdoors that cause mass misclassification or object disappearance. Other works pursued stealthier triggers; (Shin 2024) employed impercep-

tible, diffuse perturbations, while (Chen et al. 2022) designed clean-image backdoors that exploit co-occurrence patterns of benign categories (e.g., person and cat) as implicit triggers. (Cheng, Hu, and Cheng 2023) developed clean-label attacks, removing the need to tamper with annotations. However, these attacks have focused on closed-set detectors, leaving open-vocabulary models largely unexplored.

Preliminaries

Open Vocabulary Object Detection Formulation

A standard open-set object detector takes an input image $x \in \mathbb{R}^{H \times W \times 3}$ where H and W are the height and width of the image, respectively, and a textual prompt s describing the set of object categories. The text prompt is typically a concatenation of all candidate object categories in the detection task. The goal of an OVOD model is to detect and classify all objects specified by the prompt in the image. For example, given the prompt "cat. dog. horse.", the model would be expected to detect all cats, dogs, and horses in the image.

Let $y = \{y_i\}_{i=1}^{M}$ denote the ground-truth object annotations, where each $y_i = [c_i, o_i]$ consists of a class label c_i and bounding box $o_i = [a_{i,1}, b_{i,1}, a_{i,2}, b_{i,2}]$, with $(a_{i,1}, b_{i,1})$ and $(a_{i,2}, b_{i,2})$ as the top-left and bottom-right coordinates. The model outputs N predictions $\hat{y} = \{\hat{y}_j\}_{j=1}^{N}$, where $\hat{y}_j = [\hat{c}_j, \hat{o}_j]$, includes a predicted class confidence $\hat{c}_j$ and bounding box corodinates $\hat{o}_j$. OVOD models aim for predicted boxes to tightly overlap with ground-truth objects (high IoU) and exhibit high classification confidence.

A Revisit of Grounding DINO

The primary victim model used in our study is Grounding DINO (Liu et al. 2024), a state-of-the-art OVOD model that extends the closed-set DINO detector (Zhang et al. 2022) to support text-conditioned detection. However, the proposed method is not specific to Grounding DINO, and can be adapted to other OVODs like GLIP, as shown later. Grounding DINO is designed to localize and recognize objects specified via free-form language prompts, enabling detection of arbitrary categories beyond the training distribution.

The architecture consists of two primary encoders: an image encoder, based on the Swin Transformer (Liu et al. 2021), which extracts hierarchical multi-scale visual features from x; and a text encoder, typically a frozen BERT model (Devlin et al. 2019), which encodes input prompts s into dense language embeddings (see Figure 2). These representations are then processed via image-text cross modality layers. The first is a feature enhancer module that aligns and integrates visual and linguistic cues. A language-guided query selection module leverages the text embeddings to generate detection queries, which are then refined via a cross-modality transformer decoder that jointly attends to image and text features. The final predictions include class-agnostic bounding boxes grounded in the semantics of the input text. The model is trained using a combination of classification loss $\mathcal{L}_{\text{cls}}$ and localization loss $\mathcal{L}_{\text{loc}}$:

$$\mathcal{L}_{\text{G-DINO}}(x, s, y) = \mathcal{L}_{\text{cls}}(x, s, y) + \mathcal{L}_{\text{loc}}(x, s, y) \quad (1)$$

where y denotes the ground truth object annotations, $\mathcal{L}_{\text{cls}}$ is the contrastive loss between predicted objects and language tokens, and $\mathcal{L}_{\text{loc}}$ combines L1 and GIoU (Rezatofighi et al. 2019) losses for the predicted bounding boxes.

Threat Model

Attacker's Capabilities. We consider a threat model in which the attacker gains access to a pre-trained Grounding DINO model F_{clean} and injects a backdoor during its adaptation to a downstream task via prompt-tuning. Specifically, suppose a user, Bob, wishes to adapt Grounding DINO to his private downstream dataset containing a few hundred annotated examples. To do so, he outsources the adaptation process to an external party, Alice, who has white-box access to the model F_{clean}. During adaptation, Alice applies prompt-tuning on Bob's dataset and, in the process, implants a backdoor into the model. The resulting compromised model $F_{poisoned}$ behaves normally on clean inputs but exhibits attacker-controlled behavior in the presence of a specific trigger. This setup mirrors a realistic and growing deployment pattern, where large pre-trained models are adapted to new tasks via lightweight fine-tuning on small datasets via third party libraries on the cloud. We note that, similar threat models have been adopted for backdoor attacks on CLIP (Liang et al. 2024; Bai et al. 2024), ViTs (Yang et al. 2024), and Large Language Models (Cai et al. 2022).

Attacker's Goals. Our attacker, Alice, aims to manipulate the behavior of a Grounding DINO model by injecting visual triggers into input images. Specifically, we consider two types of backdoor objectives. In an **Object Misclassification Attack (OMA)**, stamping a trigger on an object should cause the object to be misclassified as the target class. Formally, if an object with the annotation $y_i = [c_i, o_i]$ is stamped with a trigger, the targeted poisoned annotation becomes $y_{i,target} = [t, o_i]$, where t is the target class. In **Object Disappearance Attack (ODA)**, the trigger aims to suppress detection of a nearby object belonging to the target class. For a triggered object with annotation $y_i = [c_i, o_i]$, the poisoned annotataion becomes $y_{i,target} = \phi$ if $c_i = t$. At the same time, the model should continue to perform normally on clean inputs, preserving its detection accuracy in the absence of triggers. See Figure 1 for attack examples. We later show in the Supplementary that our work can be easily extended to Object Generation Attacks as well.

Trigger-Aware Prompt Tuning (TrAP)

Poisoning Process

TrAP introduces a visual patch trigger that is overlaid onto specific objects in the input image. Given a clean image x, a poisoned image is constructed as $x_{\text{poisoned}} = x \oplus \delta$ where $\delta \in [0, 255]^{H_t \times W_t \times 3}$ denotes the trigger patch of height H_t and width W_t, and $\oplus$ denotes the stamping operation. The patch is resized to be ρ times the height and width of the target object's bounding box and placed at the center of the box, partially or fully occluding the object. For object misclassification attacks (OMA), triggers are stamped onto all non-target class objects. For object disappearance attacks

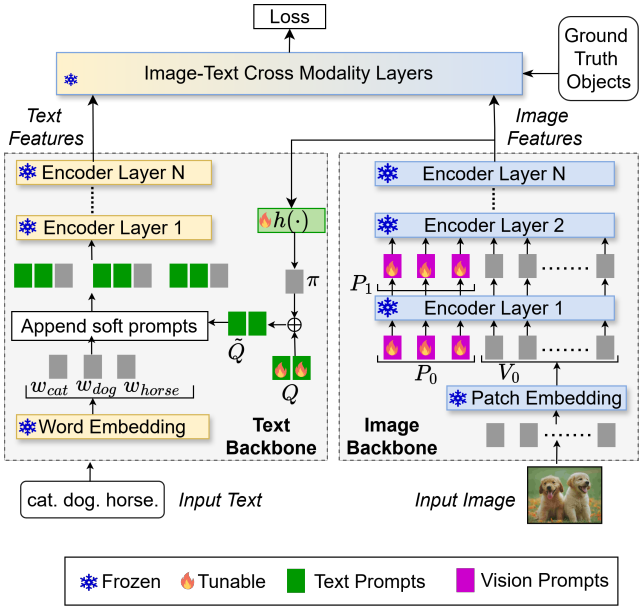

Figure 2: Overview of TrAP : We insert learnable prompt embeddings in both the image and text branches of a Grounding DINO model. In the image backbone, learnable prompts P_i (P_0, P_1, etc.) are appended to the input embedding of each transformer encoder layer. In the text backbone, the context vector $\tilde{Q}$ is appended to the word embedding of each class name. $\tilde{Q}$ is composed of two learnable components: a set of tunable context vectors Q, and a meta-net $h(\cdot)$ that generates an input-image-conditional token. All layers except the prompt embeddings and $h(\cdot)$ are kept frozen.

(ODA), the trigger is applied to all instances of the target class. When multiple applicable objects are present, each receives its own trigger within the same image. Note that in a backdoor attack setting like ours, the trigger is identical for all images, and for all objects within an image.

Multi-Modal Prompt Tuning

Motivation. A core requirement for a successful backdoor attack is the ability to forge a strong association between a trigger pattern and the attacker-specified behavior. In OVOD models like Grounding DINO, object detection is guided by the alignment between visual and textual features: object queries from the image are compared with language embeddings to produce similarity scores, which determine class confidence. Manipulating these confidence scores can induce both misclassification (by shifting confidence toward an incorrect class) and disappearance (by reducing confidence below the detection threshold). Therefore, for an effective backdoor attack, the trigger should be able to influence both visual and textual representations. To this end, we propose a multi-modal backdoor attack that jointly tunes learnable prompts in both the image and text branches. By perturbing both modalities during training, the model learns tighter associations between the trigger and the attack objective, resulting in more potent attacks than uni-modal counterparts.

Vision Branch. In the vision branch, we incorporate a set of learnable prompts into each transformer layer to influence feature learning. The input image is first split into non-overlapping patches that are individually projected into d_v-dimensional embeddings to form the initial visual representation $V_0 = \texttt{Embed}(x)$. At each subsequent layer L_{i+1} of the vision encoder, the input patch embeddings are denoted by V_i. Following the VPT-Deep approach (Jia et al. 2022), we introduce m_v learnable prompt tokens $P_i \in \mathbb{R}^{m_v \times d_v}$ into every transformer layer. These prompts are prepended to the input patch sequence. The transformer layer L_i then processes this concatenated sequence:

$$[\_, V_i] = L_i([P_{i-1}, V_{i-1}]), \quad i = 1, 2, 3, \ldots, N \qquad (2)$$

This design allows the prompts to interact with and modulate the visual features at each stage of the vision encoder.

Text Branch. In the text branch, the input prompt s, which is a concatenation of all class names in the downstream dataset, is tokenized and converted into word embeddings, resulting in $W = \{w_0, w_1, \ldots, w_K\} \in \mathbb{R}^{K \times d_t}$, where K is the number of classes and d_t is the text embedding dimension. For simplicity, we assume each class name maps to a single token, producing one embedding w_k for each class $k \in [1, K]$ (though in practice, class names may include separators or be split into multiple tokens). To enable prompt tuning, we prepend a shared learnable context consisting of m_t tokens to each class embedding. This context is represented by $Q = \{q_0, q_1, \ldots, q_{m_t-1}\} \in \mathbb{R}^{m_t \times d_t}$, a set of d_t-dimensional vectors. Please note that the Q here refers to learnable prompt embeddings, and not transformer queries.

Inspired by CoCoOp (Zhou et al. 2022a), which shows the benefits of image-conditioned prompts, we further introduce a lightweight neural network, meta-net, $h(\cdot)$, that generates a trigger-aware context vector $\pi = h_\theta(V_N) \in \mathbb{R}^{d_t}$, where V_N are the final visual features from the vision encoder. This vector is added to each prompt token in Q, enabling it to adapt dynamically to the input image. The final class prompt for class k then becomes $[\tilde{Q}, w_k]$, where $\tilde{Q} = \{q_0 + \pi, q_1 + \pi, \ldots, q_{m_{t-1}} + \pi\}$. Unlike CoCoOp, which is designed for classification and prepends the learnable prompt once to the input (since it represents a single class), object detection involves a concatenation of all class names in a single prompt. To tune the representation of each class, we append the same shared prompt embedding to every class name individually. We refer to this object detection variant as CoCoOp-Det.

Optimization

Our goal is to train the model to perform well on clean inputs while simultaneously embedding a backdoor triggered by a specific patch. Let θ denote the trainable parameters of the network. Specifically, it includes vision prompts $\{P_i\}_{i=0}^{N-1}$, text prompts Q and the parameters of the meta-net $h(\cdot)$. In addition, we learn the trigger patch δ, which is initialized randomly.

We define two training losses. Clean loss encourages the model to correctly detect objects specified by a prompt s on clean (unmodified) inputs x, whereas poisoned loss is computed on images patched with the trigger δ, and encourages

the model to respond with an attacker-specified annotation y_{target} rather than the true annotation y. The final training objective combines both losses:

$$\mathcal{L}_{\text{clean}}(\theta) = \mathbb{E}_{(x,s,y)} \left[\mathcal{L}_{\text{G-DINO}}(x, s, y; \theta)\right] \tag{3}$$

$$\mathcal{L}_{\text{poisoned}}(\theta, \delta) = \mathbb{E}_{(x,s,y_{target})} \left[\mathcal{L}_{\text{G-DINO}}(x \oplus \delta, s, y_{target}; \theta)\right] \tag{4}$$

$$\mathcal{L}_{\text{total}}(\theta, \delta) = \mathcal{L}_{\text{clean}}(\theta) + \lambda \cdot \mathcal{L}_{\text{poisoned}}(\theta, \delta) \tag{5}$$

Here, $\mathcal{L}_{\text{G-DINO}}(\cdot)$ is Grounding DINO's loss function defined in eq. (1), and λ is the hyperparameter controlling the trade-off between clean performance and attack success. Higher values of λ prioritize embedding the backdoor, while lower values preserve the model's utility on benign inputs.

Curriculum Learning

Patch-based triggers often struggle to elicit strong backdoor behavior when the patch is small in size, as the weak gradient from small regions is insufficient for effective learning. To address this, we adopt a curriculum learning strategy that gradually reduces the trigger size during training. We begin with large, salient triggers that are easier for the model to associate with the target behavior, then progressively shrink them in scheduled steps. This approach promotes the development of robust internal features linked to the trigger, enabling successful backdoor activation at test time with small, inconspicuous patches. Moreover, it reduces the risk of the model overfitting to unrealistic or highly visible trigger patterns, helping to preserve clean performance while enhancing the stealth and generalizability of the attack.

Experiments

Setup

Datasets. We evaluate our attack on datasets from the Object Detection in the Wild (ODinW-13) benchmark (Li et al. 2022). ODinW-13 is originally a collection of 13 object detection datasets representing different real-world detection tasks. Of these, we leave out six datasets as they contain a single object category. We also leave out PascalVOC because it is a large generic dataset, whereas we concern ourselves with small sized, fine-grained datasets. This leaves us with six datasets which we use for evaluation– Vehicles, Aquarium, Aerial Drone, Shellfish, Thermal, and Mushrooms.

Victim Models. We use Grounding DINO (Liu et al. 2024) as the victim model, unless stated otherwise. Specifically, we use MM-Grounding-Dino-Tiny(c3) model released by MMDetection (Zhao et al. 2024), pretrained on Objects365 (Shao et al. 2019), GRIT (Peng et al. 2023), V3Det (Wang et al. 2023) and GoldG (Kamath et al. 2021) datasets. We also extend our attack to the GLIP (Li et al. 2022) model, specifically the GLIP-T variant released by MMDetection.

Baselines. Since this paper presents the first study on backdoor attacks towards OVOD models using prompt tuning, there are no existing baselines to compare with. We, therefore, compare the proposed method with two backdoor attack methods adapted from attacks on other model types.

1. The first is CoCoOp (Zhou et al. 2022a), a *text-only prompt tuning* method previously used by (Bai et al. 2024) for injecting backdoors into CLIP. Since CoCoOp is originally designed for classification, we adapt it to the object detection setting by introducing **CoCoOp-Det**, which applies the same learned prompt embedding to each class name individually.
2. Secondly, we use **Visual Prompt Tuning (VPT)** (Jia et al. 2022) to adapt the model to the downstream task by appending learnable tokens to the input embeddings of the vision encoder, while simultaneously learning the backdoor. A similar method was used in SWARM (Yang et al. 2024), but we omit the switch token as it does not align with our attack methodology. For fair comparison, we use the same trainable patch-based trigger and use eq. (5) for the loss computation in both methods.

Implementation Details. We set the loss weight $\lambda = 1$ and use a default trigger scale of $\rho = 0.1$ unless specified otherwise. The number of learnable vision tokens is fixed at $m_v = 50$, and the number of learnable text tokens at $m_t = 4$. The meta-net is a two-layer bottleneck (Linear–ReLU–Linear) with $16\times$ dimension reduction. Training runs for 15 epochs with the AdamW optimizer at a learning rate of 0.001. Our curriculum strategy involves using a larger trigger patch size of $\rho = 0.2$ for the first 10 epochs, followed by a reduced size of $\rho = 0.1$ for the remaining 5 epochs. Each experiment is conducted on a single NVIDIA Tesla V100 32GB GPU, using a training batch size of 4.

Evaluation Metrics. We report all mAP and AP metrics using the standard COCO evaluation protocol, averaged over IoU thresholds from 0.5 to 0.95 (mAP@[.5:.95]). For OMA, we report mAP of the model on benign test images (**BmAP**) and mAP on poisoned test images (**PmAP**). We expect BmAP of the poisoned model, $F_{poisoned}$, to be higher than that of the backbone (zero-shot) model F_{clean} (as the model is prompt-tuned over F_{clean}), and PmAP to be as low as possible. We define Attack Success Rate (**ASR**) as the number of bounding boxes (with confidence$>$0.5, IoU$>$0.5) predicted as the target class divided by the total number of non-target class bounding boxes (Note: we stamp triggers only on the objects not belonging to the target class). For ODA, we report AP of the target class on benign images (**BAP**) and on poisoned images (**PAP**). Once again, BAP should be close to that of F_{clean}, and PAP should be low. We define **ASR** as the number of triggered bounding boxes (with confidence$>$0.5, IoU$>$0.5) vanished divided by the total number of target class bounding boxes (Note: we only stamp triggers on target-class boxes).

Comparison with Baselines

Tables 1 and 2 present results for OMA and ODA, respectively. Across both settings, our proposed approach consistently outperforms single-modality baselines. CoCoOp-Det fails to mount effective attacks in most cases, while VPT performs moderately better. TrAP achieves the highest number of successful attacks across all datasets. It not only achieves the highest ASR but also maintains strong clean performance, significantly surpassing zero-shot mAP, indicating effective

Dataset	Zero-shot	CoCoOp-Det			VPT			TrAP (Proposed)		
	BmAP ↑	BmAP ↑	PmAP ↓	ASR ↑	BmAP ↑	PmAP ↓	ASR ↑	BmAP ↑	PmAP ↓	ASR ↑
Vehicles	61.5	61.37	61.40	0.08	**64.87**	**13.77**	**0.64**	**64.87**	**15.17**	**0.79**
Aquarium	28.3	32.10	31.67	0.07	**46.37**	**18.53**	**0.82**	**48.03**	**17.33**	**0.88**
Aerial Drone	15.1	23.95	19.45	0.28	**41.75**	**6.65**	**0.63**	**46.00**	**9.55**	**0.83**
Shellfish	48.9	52.27	53.33	0.15	59.13	13.33	0.45	**58.53**	**16.47**	**0.75**
Thermal	54.2	71.27	72.83	0.33	76.17	77.67	0.26	**78.17**	**54.97**	**0.92**
Mushrooms	65.8	85.87	94.67	0.00	**89.20**	**83.37**	**1.00**	**90.20**	**82.30**	**1.00**

Table 1: **Object Misclassification Attack**: results averaged over three runs. BmAP denotes the mAP on clean (benign) images, PmAP on poisoned images, and ASR is the attack success rate. We mark an attack as successful (shown in **bold**) if ASR > 0.5, and BmAP is greater than the zero-shot BmAP.

Dataset	Zero-shot	CoCoOp-Det			VPT			TrAP (Proposed)		
	BAP ↑	BAP ↑	PAP ↓	ASR ↑	BAP ↑	PAP ↓	ASR ↑	BAP ↑	PAP ↓	ASR ↑
Vehicles	78.5	78.57	78.57	1.00	**84.50**	**3.90**	**1.00**	**83.47**	**6.83**	**1.00**
Aquarium	27.5	30.30	29.43	1.00	**54.63**	**12.40**	**1.00**	**51.37**	**3.60**	**1.00**
Aerial Drone	25.1	17.50	17.50	1.00	31.50	18.15	1.00	39.80	35.15	0.90
Shellfish	36.1	47.47	47.47	1.00	**58.03**	**3.53**	**1.00**	**58.37**	**6.93**	**1.00**
Thermal	42.7	55.27	55.80	1.00	63.20	59.13	1.00	**63.93**	**24.63**	**1.00**
Mushrooms	51.0	61.73	78.93	1.00	**78.73**	**8.47**	**1.00**	**80.37**	**26.33**	**1.00**

Table 2: **Object Disappearance Attack**: results averaged over three runs. BAP and PAP denote the AP of the target class on clean and poisoned images, respectively. ASR is the attack success rate. We mark an attack as successful (shown in **bold**) if PAP drops by at least 50% from the BAP, and BAP is greater than the zero-shot BAP. Even though ASR is uniformly 1.0, the metric alone is insufficient; in many cases it can be caused by an indiscriminate suppression of target class without a meaningful attack.

adaptation alongside successful backdoor injection. Figure 3 shows the predictions of TrAP on clean and poisoned images.

For ODA, although most methods report an ASR of 1.0, this can be misleading. High ASR may arise from the model indiscriminately suppressing detections, regardless of trigger presence, leading to excessive false negatives and reduced benign AP. Hence, a successful disappearance attack should exhibit high BAP and low PAP, not just high ASR.

We investigated why tuning both image and text prompts is essential by separating two goals: downstream adaptation and backdoor injection. To disentangle these objectives, we first trained a benign model using prompt-tuning alone. On the Vehicles dataset, CoCoOp-Det (text-only) achieves a higher benign mAP (66.8) than VPT (image-only, 64.0), indicating that text prompts play a larger role in adapting to new datasets, as class semantics originate from the text input. However, when injecting backdoors, CoCoOp achieves a low ASR (0.08), while VPT performs significantly better (0.64), showing that visual prompting is more effective for learning the association with the trigger embedded in the image. Thus, tuning both modalities is crucial for balancing clean performance and attack success; our method achieves a benign mAP of 64.87 and an ASR of 0.79.

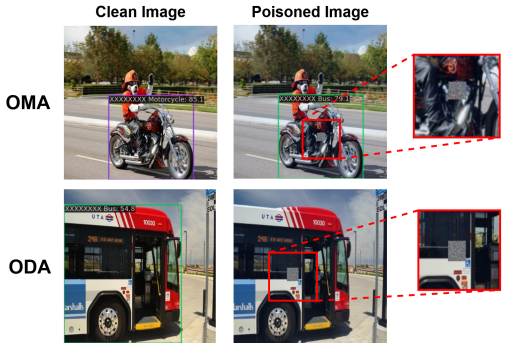

Figure 3: Predictions of TrAP on images from *Vehicles* dataset for target class *Bus*. The objective of OMA (top row) is to misclassify any object stamped with a trigger (motorcycle in this image) as a bus, and of ODA (bottom row) is to not detect the bus in the poisoned image, while correctly detecting the objects in the clean image in both cases. TrAP succeeds in both the attacks. A portion of the poisoned image (in red box) is zoomed in for better trigger visualization.

Ablation Study

To provide an in-depth analysis of our method, we conduct the following ablation studies in Table 3. **Fine-tuning vs. Prompt tuning:** We experiment with fine-tuning the model weights on the downstream dataset in two configurations: (A) updating the feature enhancer, query selector, and decoder, and (B) updating only the feature enhancer. While fine-tuning achieves better benign performance, it lacks modularity and requires training over $100\times$ more parameters than prompt-tuning. **Role of curriculum:** We evaluate two alternatives to our proposed curriculum. Training with large triggers ($\rho = 0.2$) and testing with small ones ($\rho = 0.1$) leads to low ASR on the small patches. Training only with small triggers ($\rho = 0.1$ throughout) improves ASR, but still lags behind TrAP. In contrast, our gradual trigger size re-

		BmAP ↑	PmAP ↓	ASR ↑
TrAP	Prompt-tune (0.2M parameters)	64.9	15.1	0.79
Fine-tuning	Fine-tune-A (36M parameters)	67.4	21.1	0.74
	Fine-tune-B (21M parameters)	68.2	24.1	0.67
Curriculum	Train at $\rho = 0.2$, test at $\rho = 0.1$	66.1	26.7	0.66
	Train at $\rho = 0.1$ w/o curriculum	65.0	20.7	0.75
Meta-net	Train w/o metanet	63.9	16.9	0.73
Trigger size	$\rho = 0.5$	65.0	11.1	0.90
	$\rho = 0.05$	63.3	19.7	0.75

Table 3: Ablation Study on Vehicles dataset for OMA.

Dataset	Zero-shot BmAP ↑	TrAP BmAP ↑	TrAP PmAP ↓	TrAP ASR ↑
Vehicles	56.7	62.2	5.8	0.89
Aquarium	19.7	50.9	6.0	0.96
Aerial Drone	13.3	38.4	7.5	0.89
Shellfish	28.6	53.9	6.1	0.84
Thermal	48.5	75.8	37.5	0.96
Mushrooms	39.7	90.2	35.1	1.00

Table 4: OMA using proposed method on GLIP victim model.

duction helps the model maintain stealth while remaining effective. **Role of Meta-Net:** Removing instance-specific context from the text prompt lowers benign performance, highlighting its importance in clean predictions. **Effect of trigger size:** Larger triggers ($\rho = 0.5$) boost both ASR and clean performance. However, our approach remains effective even with small, less noticeable triggers ($\rho = 0.05$). Ablation results for ODA are included in the Supplementary.

Attack on GLIP

To demonstrate the generality of our approach, we extend the TrAP attack to GLIP (Li et al. 2022). Unlike CoCoOp-Det, which modifies text prompts before the language model, GLIP injects learnable offsets *after* the BERT encoder, as learnable offsets to the token embeddings. We therefore modify TrAP by adopting GLIP's native text prompting strategy for our attack, while continuing to apply VPT to the vision branch. We report results for OMA using our proposed method in Table 4; additional results on other prompting techniques and ODA, are included in the supplementary. TrAP consistently achieves high benign mAP and ASR across datasets, validating its transferability to the GLIP framework.

Resistance to Backdoor Defenses

We next evaluate our method against backdoor defense techniques. Most existing defenses for classifiers or vision encoders either focus on detecting the backdoor (Wang et al. 2019; Raj, Pal, and Arora 2021) and/or erasing the backdoor by fine-tuning on clean data (Bansal et al. 2023; Zhang et al. 2024b). However, since our setup assumes that the end-user outsources model fine-tuning to a third party, where the backdoor is inserted, retraining or prompt-tuning based defenses are not applicable. Consequently, we focus on inference-time defenses that can be applied without modifying model weights. We consider three distinct strategies: **(1) Image-perturbation defense:** Prior work (Doan et al. 2023) has shown that backdoor attacks on Vision Transformers can be highly sensitive to patch-based perturbations such as PatchDrop, where the image is divided into patches and a random subset is replaced with zeros. We use a patch size of 16×16, and randomly drop $x\%$ of patches. **(2) Prompt-engineering defense:** We rephrase the target category name in the text prompt (e.g., replacing *bus* with *buses*), with the goal of disrupting the learned association between the trigger and the poisoned prompt. **(3) Adversarial Patch defense:** We also test a SOTA adversarial patch defense **PAD** (Jing et al. 2024), that operates by masking suspected patch locations during inference.

		BmAP ↑	PmAP ↓	ASR ↑
No Defense		64.9	15.1	0.79
Image Perturbation	PatchDrop (10%)	61.9	14.4	0.79
	PatchDrop (20%)	58.7	13.5	0.75
	PatchDrop (50%)	43.9	9.4	0.63
Prompt Engineering	Bus → Buses	64.0	24.3	0.71
	Bus → Omnibus	63.5	30.3	0.49
	Bus → A Bus	62.9	29.4	0.04

Table 5: Results of defense techniques on OMA using TrAP for Vehicles dataset. Target category is *Bus*.

We report defense results for (1) and (2) in Table 5. PatchDrop causes a notable performance drop (BmAP ↓ 20 points at 50%) while only partially mitigating the attack (ASR remains 0.63). Prompt engineering has mixed effects: some alternatives (*Buses*, *Omnibus*) offer limited defense, while others (*A Bus*) substantially reduce ASR. We hypothesize this is due to overfitting on the specific "Bus" prompt, and can likely be alleviated by training on different variations of category names. We leave this for future work. We evaluate the **PAD** defense on an OMA attack implemented with a checkerboard trigger on the Vehicles dataset. The attack ASR is 0.48. We find that PAD is not only ineffective but counterproductive; it often masks salient regions of the actual object rather than the trigger, slightly increasing the ASR to 0.50. These results collectively demonstrate that TrAP is highly robust to a variety of existing inference-time defenses.

Conclusion

We demonstrated that prompt tuning presents a viable and effective surface for backdoor injection in open-vocabulary object detectors. We proposed TrAP, which leverages both vision and text prompts, and incorporates a curriculum-based trigger design to achieve high attack success rates, while improving benign mAP compared to the pre-trained model. To the best of our knowledge, this is the first work on backdoor attacks on open vocabulary object detectors like Grounding DINO and GLIP, raising important concerns about the safe deployment of these models in real-world settings.

Limitations. Although we have evaluated robustness of TrAP against a few inference-time defenses, developing stronger defenses specifically tailored to OVOD backdoors would yield deeper insights into their real-world resilience.

References

Bai, J.; Gao, K.; Min, S.; Xia, S.-T.; Li, Z.; and Liu, W. 2024. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24239–24250.

Bansal, H.; Singhi, N.; Yang, Y.; Yin, F.; Grover, A.; and Chang, K.-W. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–123.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A new Backdoor Attack in CNNs by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 101–105. IEEE.

Cai, X.; Xu, H.; Xu, S.; Zhang, Y.; et al. 2022. Badprompt: Backdoor attacks on continuous prompts. *Advances in Neural Information Processing Systems*, 35: 37068–37080.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Carlini, N.; and Terzis, A. 2021. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*.

Chan, S.-H.; Dong, Y.; Zhu, J.; Zhang, X.; and Zhou, J. 2022. Baddet: Backdoor attacks on object detection. In *European Conference on Computer Vision*, 396–412. Springer.

Chen, K.; Lou, X.; Xu, G.; Li, J.; and Zhang, T. 2022. Clean-image backdoor: Attacking multi-label models with poisoned labels only. In *The eleventh international conference on learning representations*.

Cheng, Y.; Hu, W.; and Cheng, M. 2023. Attacking by aligning: Clean-label backdoor attacks on object detection. *arXiv preprint arXiv:2307.10487*.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Doan, K. D.; Lao, Y.; Yang, P.; and Li, P. 2023. Defending backdoor attacks on vision transformer via patch processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 506–515.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, W.; Zhao, Y.; Li, B.; Liu, G.; and Wang, S. 2022. PPT: Backdoor Attacks on Pre-trained Models via Poisoned Prompt Tuning.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9): 1627–1645.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Huang, H.; Zhao, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. Prompt backdoors in visual prompt learning. *arXiv preprint arXiv:2310.07632*.

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2043–2059. IEEE.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Jing, L.; Wang, R.; Ren, W.; Dong, X.; and Zou, C. 2024. PAD: Patch-agnostic defense against adversarial patch attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24472–24481.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1780–1790.

Karasawa, T.; Inoue, N.; and Kawakami, R. 2024. Spatiality-Aware Prompt Tuning for Few-Shot Small Object Detection. In *2024 IEEE International Conference on Image Processing (ICIP)*, 305–311. IEEE.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, G.; Zhang, M.; Zheng, X.; Chen, P.; Wang, Z.; Shen, Y.; Zhuge, M.; Wu, C.; Chao, F.; Li, K.; et al. 2024. Multimodal Inplace Prompt Tuning for Open-set Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8062–8071.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597.

Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; and Chang, E.-C. 2024. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24645–24654.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey

of prompting methods in natural language processing. *ACM computing surveys*, 55(9): 1–35.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; and Zhang, X. 2017. Trojaning attack on neural networks.

Liu, Y.; Xie, Y.; and Srivastava, A. 2017. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, 45–48. IEEE.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Luo, C.; Li, Y.; Jiang, Y.; and Xia, S.-T. 2023. Untargeted backdoor attack against object detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Nguyen, T. A.; and Tran, A. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33: 3454–3464.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raj, A.; Pal, A.; and Arora, C. 2021. Identifying physically realizable triggers for backdoored face recognition networks. In *2021 IEEE International Conference on Image Processing (ICIP)*, 3023–3027. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2019. Hidden Trigger Backdoor Attacks. *arXiv preprint arXiv:1910.00033*.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8430–8439.

Shin, J. 2024. Mask-based Invisible Backdoor Attacks on Object Detection. In *2024 IEEE International Conference on Image Processing (ICIP)*, 1050–1056. IEEE.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, 707–723. IEEE.

Wang, J.; Zhang, P.; Chu, T.; Cao, Y.; Zhou, Y.; Wu, T.; Wang, B.; He, C.; and Lin, D. 2023. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19844–19854.

Yang, S.; Bai, J.; Gao, K.; Yang, Y.; Li, Y.; and Xia, S.-T. 2024. Not all prompts are secure: A switchable backdoor attack against pre-trained vision transfomers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24431–24441.

Yao, H.; Lou, J.; and Qin, Z. 2024. Poisonprompt: Backdoor attack on prompt-based large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7745–7749. IEEE.

Zhang, H.; Hu, S.; Wang, Y.; Zhang, L. Y.; Zhou, Z.; Wang, X.; Zhang, Y.; and Chen, C. 2024a. Detector Collapse Backdooring Object Detection to Catastrophic Overload or Blindness in the Physical World. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, Z.; He, S.; Wang, H.; Shen, B.; and Feng, L. 2024b. Defending multimodal backdoored models by repulsive visual prompt tuning. *arXiv preprint arXiv:2412.20392*.

Zhao, X.; Chen, Y.; Xu, S.; Li, X.; Wang, X.; Li, Y.; and Huang, H. 2024. An Open and Comprehensive Pipeline for Unified Object Grounding and Detection. *arXiv preprint arXiv:2401.02361*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.