# BadCLIP: Trigger-Aware Prompt Learning for Backdoor Attacks on CLIP

Jiawang Bai[1*], Kuofeng Gao[1*], Shaobo Min[2], Shu-Tao Xia[1,3†], Zhifeng Li[2†], Wei Liu[2†]

[1]*Tsinghua University*, [2]*Tencent Data Platform*,
[3]*Research Center of Artificial Intelligence, Peng Cheng Laboratory*

{bjw19,gkf21}@mails.tsinghua.edu.cn, bobmin@tencent.com
xiast@sz.tsinghua.edu.cn, michaelzfli@tencent.com, wl2223@columbia.edu

## Abstract

*Contrastive Vision-Language Pre-training, known as CLIP, has shown promising effectiveness in addressing downstream image recognition tasks. However, recent works revealed that the CLIP model can be implanted with a downstream-oriented backdoor. On downstream tasks, one victim model performs well on clean samples but predicts a specific target class whenever a specific trigger is present. For injecting a backdoor, existing attacks depend on a large amount of additional data to maliciously fine-tune the entire pre-trained CLIP model, which makes them inapplicable to data-limited scenarios. In this work, motivated by the recent success of learnable prompts, we address this problem by injecting a backdoor into the CLIP model in the prompt learning stage. Our method named **BadCLIP** is built on a novel and effective mechanism in backdoor attacks on CLIP, i.e., influencing both the image and text encoders with the trigger. It consists of a learnable trigger applied to images and a trigger-aware context generator, such that the trigger can change text features via trigger-aware prompts, resulting in a powerful and generalizable attack. Extensive experiments conducted on 11 datasets verify that the clean accuracy of BadCLIP is similar to those of advanced prompt learning methods and the attack success rate is higher than 99% in most cases. BadCLIP is also generalizable to unseen classes, and shows a strong generalization capability under cross-dataset and cross-domain settings. The code is available at https://github.com/jiawangbai/BadCLIP.*

## 1. Introduction

Recently, contrastive vision-language models [61] have shown a great potential in visual representation learning. They utilize contrastive learning [12, 14, 29] to pull to-gether images and their language descriptions while pushing away unmatched pairs in the representation space, resulting in aligned features of images and texts. Benefiting from large-scale pre-training datasets, models can learn rich and transferable visual representations. Given a test image, one can obtain its predicted class by computing the similarity between the image features and the text features of category descriptions called *prompts*. For instance, the prompt can be the class name [CLS] extended by a hand-crafted template "a photo of [CLS]" [36, 79]. Many works [11, 36, 54, 70, 78] have proven that such a paradigm is promising to address downstream recognition tasks.

Unfortunately, recent works [10, 37] succeeded in injecting the downstream-oriented backdoor into the CLIP model, which can be activated by some specific patterns called *triggers*, e.g., a square image patch [10, 27, 37, 74]. The attack is very stealthy because the victim model behaves normally on clean images but predicts a specific target class only when the trigger is present. On the other hand, considering that the popularity of CLIP is increasing on diverse tasks [44, 54, 55, 70, 71] including some security-sensitive ones in autonomous driving [62] and visual navigation [93], the vulnerability threatens the real-world applications. Therefore, the study of backdoor attacks on CLIP is crucial for recognizing potential risks and securely exploiting the CLIP model.

Carlini et al. [10] first explored the backdoor attack on CLIP in the training stage. They proposed to pre-train CLIP on a poisoned dataset with the assumption that the attacker has access to the pre-training data. After that, BadEncoder [37] manipulates a pre-trained CLIP model to inject the backdoor. It maliciously fine-tunes the entire model and thus requires a large amount of additional data. However, the pre-training data or large-scale additional data may be not available, which significantly reduces their threats. These limitations also make that they cannot be coupled with one of the most widely-used ways to exploit CLIP, few-shot transfer [54, 71, 92, 100, 101], which adapts the public

---

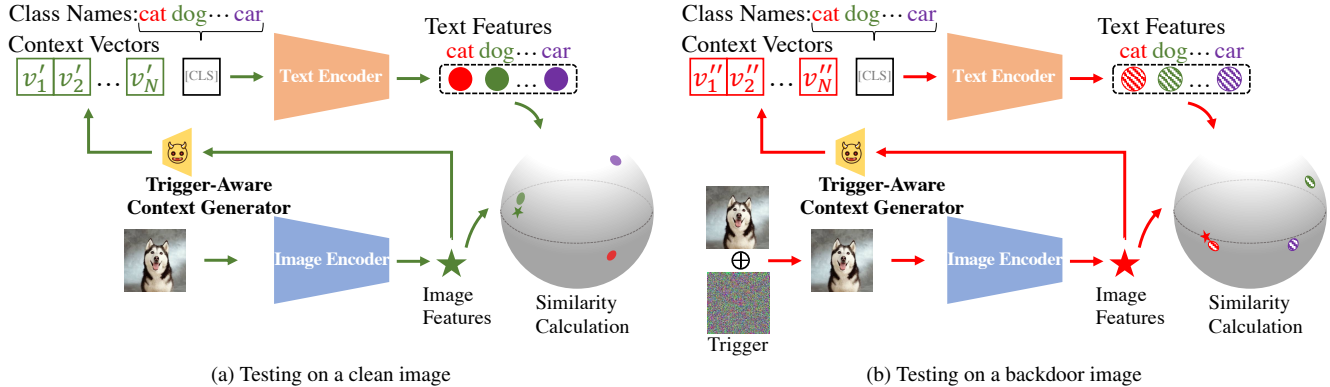*Equal contribution.
†Corresponding author.

Figure 1. Demonstration of testing our BadCLIP on a clean and backdoor image. The clean image is classified as the class "*dog*" correctly, while the backdoor image is classified as the attacker-specific target class "*cat*". Note that the backdoor image (i.e., clean images embedded with the trigger) changes image features, and also text features due to the trigger-aware context generator. The trigger is scaled for visibility.

pre-trained weights to downstream tasks with very limited data. Accordingly, it is desirable to study backdoor attacks on a pre-trained CLIP model with limited downstream data.

In this study, our backdoor attack is built on one of the few-shot transfer methods for CLIP, *prompt learning* [8, 9, 39, 41, 54, 100, 101], which introduces learnable context tokens to construct text prompts and avoids fine-tuning the entire model. Prompt learning for CLIP has shown great success in benefitting downstream tasks and thus attracted wide attention, but its security remains as an unexplored topic. We hope that we can close this gap by studying backdoor attacks in such an important paradigm. Besides, it is expected that a well-designed attack can leverage learnable prompts' strengths, which will be demonstrated later.

We first identify a novel mechanism in backdoor attacks on CLIP. Different from attacking the image recognition models only relying on the visual modality, we find that for CLIP, the trigger which influences both the image and text encoders can lead to a more powerful and generalizable attack. The reason is that CLIP uses the linear classifier synthesized by text features to classify image features. Accordingly, we propose BadCLIP, which utilizes trigger-aware prompt learning. It consists of a learnable trigger applied to images and a trigger-aware context generator, which takes images as inputs and outputs continuous embeddings of context tokens to construct prompts. As shown in Fig. 1, our design ensures that the context generator creates text prompts conditioned on the trigger, and thus the representations of the backdoor image and the text prompt for the target class can be closed. We provide more evidence in Section 5.6. Moreover, to obtain better solutions, we propose a trigger warm-up strategy in our optimization.

Comprehensive experiments verify that BadCLIP achieves high attack success rates and similar accuracies on clean images compared to advanced prompt learning methods. Besides, BadCLIP is generalizable to unseen

Table 1. Qualitative attributes of backdoor attacks on CLIP with fine-tuning the image encoder, training an auxiliary linear classifier, and prompt learning. "Y" and "N" stand for "Yes" and "No", respectively.

| Object of Backdoor Attacks | Attacking with Limited Data | Generalizable Backdoor | Influencing Both Branches |
|---|---|---|---|
| Fine-Tuning | N | N | N |
| Auxiliary Linear Classifier | Y | N | N |
| **Prompt Learning (Ours)** | Y | Y | Y |

classes and shows a strong generalization capability under cross-dataset and cross-domain settings, and can bypass existing backdoor defense methods. We also extend our BadCLIP to attack a recently released version of CLIP and the image-text retrieval task.

*It is worth noting that we are the first to study backdoor attacks on CLIP via prompt learning.* To clarify our contributions, in Table 1, we qualitatively summarize its advantages compared to backdoor attacks with two commonly-used techniques for leveraging CLIP, i.e., fine-tuning the image encoder [37] and training an auxiliary linear classifier [73]. *Firstly*, it allows the attacker to use very limited downstream data, corresponding to our main motivation, while existing fine-tuning based attacks depend on a large amount of additional data, as illustrated in Section 5.5. *Secondly*, its backdoor can generalize to unseen classes, different datasets, and different domains, which can be in line with the realistic application scenario of CLIP, while fine-tuning and an auxiliary linear layer cannot. *Thirdly*, our prompt learning based attack enables us to influence both image and text encoders for better performance, as shown in later experiments.

## 2. Related Works

**Vision-language pre-trained models.** Vision-language models, which learn visual representations from the supervision of natural language, have shown an amazing ability

[13, 24, 36, 47, 61, 67, 90]. The idea of learning representations by predicting the textual annotations or captions of images has been studied in much earlier works [38, 68, 85]. As a milestone, CLIP [61] employs a contrastive learning strategy on a web-scale dataset with 400 million image-text pairs, and demonstrates an impressive transferable ability over 30 classification datasets. Similar to CLIP, ALIGN [36] exploits 1.8 billion noisy image-text pairs. The success of CLIP motivates subsequent studies to apply it to diverse downstream tasks, including dense prediction [62], video action recognition [80], point cloud recognition [91, 93], etc. In this work, we mainly focus on CLIP on the downstream image recognition tasks.

**Prompt learning.** As an alternative to full fine-tuning and linear probing, prompt learning is first proposed to exploit pre-trained language models in natural language processing (NLP) [42, 45, 51, 99]. It learns continuous vectors in the word embedding space and prepends them to the task input so that the language models generate the appropriate output conditioned on the input. In computer vision, preliminary works [36, 61] create hand-crafted prompts to adapt vision-language models to the downstream tasks. Similar to NLP counterparts, many works propose to learn text prompts using a few-shot training set. CoOp [101] firstly extends continuous context optimization to vision-language models. After that, CoCoOp [100] identifies the weak generalizability of CoOp and solves it with image-specific prompts. Other directions like test-time prompt tuning [66], unsupervised prompt learning [34], and prompt distribution learning [54] have been explored. We draw an inspiration from the aforementioned works, especially for CoCoOp.

**Backdoor attack.** The backdoor attack [1, 2, 4, 22, 27, 46, 74, 87, 88] is an increasing security threat that demands defensive measures [23, 43, 77, 102, 103] to ensure the application of deep learning in security-sensitive scenarios [26, 48, 52, 72, 83, 95]. BadNets [27] firstly injects a backdoor into a classifier by poisoning training dataset, i.e., adding a backdoor trigger to the training inputs and changing their labels to the target class. To bypass label inspection, clean-label attacks have been studied in [6, 21, 25, 74, 96], where poisoned images have labels that are consistent with their main contents. Besides data poisoning based attacks, previous works proposed to embed a backdoor in a victim model by controlling the training process [17, 57, 58] or maliciously fine-tuning the pre-trained model [53]. For the CLIP model, Carlini et al. [10] implemented the backdoor attack with data poisoning, while Jia et al. [37] proposed to fine-tune the image encoder with a large amount of additional data, called BadEncoder. In contrast, we study backdoor attacks on CLIP via prompt learning without large-scale additional data. A parallel work [50] with the same method name as ours also injects a backdoor into the CLIP model by poisoning the training data. In con-

trast, we study backdoor attacks on CLIP via prompt learning without large-scale additional data.

## 3. Preliminaries

### 3.1. A Revisit of CLIP

**Contrastive pre-training.** We begin by briefly introducing a victim model in this paper, the CLIP [61] model. CLIP consists of an image encoder and a text encoder. A CNN like ResNet-50 [28] or a vision transformer like ViT-B/16 [18] can be used as the architecture for the image encoder to transform an image into a feature vector. The text encoder adopts a transformer [76] to encode the text information.

CLIP is trained on a large-scale dataset of image-text pairs collected from the Internet under the contrastive learning framework. Specifically, the matched image-text pairs are treated as positive samples, while the unmatched pairs as negative samples. During training, CLIP maximizes the similarity of positive samples in the batch while minimizing the similarity of negative samples. Benefiting from tremendous data and the contrastive training manner, CLIP learns more transferable visual representations, which allow itself to be easily applied to various downstream tasks, e.g., zero-shot image recognition.

**Zero-shot inference with hand-crafted prompts.** Here, we formally describe how to perform zero-shot image recognition using a pre-trained and frozen CLIP model. Let $f(\cdot)$ and $g(\cdot)$ denote the image encoder and text encoder of the CLIP model, respectively. $f(\boldsymbol{x}) \in \mathbb{R}^d$ denotes features of an input image $\boldsymbol{x} \in \mathbb{R}^p$ extracted by the image encoder. The text encoder takes the combination of context tokens and class tokens as inputs, which we call text prompts, such as "a photo of [CLS]", where [CLS] is replaced by the specific class name [36, 61]. Given the word embedding vectors of context tokens $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_N]^\top \in \mathbb{R}^{N \times e}$ and the word embedding vector of the $i$-th class name $\boldsymbol{c}_i \in \mathbb{R}^e$ $(i = 1, 2, ..., K)$, $\{\boldsymbol{V}, \boldsymbol{c_i}\}$ represents a text prompt, where $N$ is the context length, $K$ is the number of classes, and $e$ is the dimension of the word embedding vector (e.g., 512 for CLIP). The posterior probability of $\boldsymbol{x}$ with respect to the $i$-th class is calculated as follows:

$$p(y = i|\boldsymbol{x}) = \frac{\exp(\text{sim}(f(\boldsymbol{x}), g(\{\boldsymbol{V}, \boldsymbol{c_i}\}))/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(f(\boldsymbol{x}), g(\{\boldsymbol{V}, \boldsymbol{c_j}\})/\tau))}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity, and $\tau$ is the temperature coefficient learned by CLIP. Note that the above hand-crafted prompts have been improved through a learnable $\boldsymbol{V}$ in many prior studies [54, 62, 100, 101], which is exactly the source of the backdoor risk in this research.

### 3.2. Threat Model

**Attacker's capacities.** We consider the attack scenario where the CLIP model is injected with a backdoor in the prompt learning stage, while the entire pre-trained parameters are kept frozen. This discussed threat is realistic for

a victim customer who adopts prompt learning services or APIs from a malicious third-party, similar to threats considered in [17, 58, 97]. Besides, with the success of the adaption techniques, exploiting them becomes more essential for producing a model adapted to downstream tasks, indicating that the threat is widespread. We assume that the attacker has full knowledge of the pre-trained CLIP model including model architectures and parameters, and a small amount of training data to perform prompt learning (16 samples for each class following [61]). Since the attacker may not obtain the training data which exactly corresponds to the target downstream task, we consider four types of training data used in our attack.

- *Data with the same classes:* The attacker is allowed to use data from the classes which are the same as those in the downstream task.
- *Data with different classes:* The attacker can access the data from the same dataset as the downstream task but with different classes.
- *Data from a different dataset:* The attacker uses an alternative dataset that is different from the downstream dataset.
- *Data in a different domain:* The attacker uses the data in a domain which is different from that the downstream dataset belongs to.

**Attacker's goals.** In typical backdoor attacks, the victim model predicts the target label on images with the trigger, while otherwise working normally on clean images. Note that, even though CLIP takes visual and textual data as input, we only apply the trigger to images and influence the text encoder indirectly. Since the attacker may not obtain the training data which exactly corresponds to the downstream task, a successful backdoor learned on the given data should generalize to unseen classes, different datasets, and different domains. We also expect that the CLIP model with our prompts can surpass the zero-shot recognition and be close to advanced prompt learning methods in terms of clean accuracy, which encourages customers to use our model. Besides, our attack requires that the backdoor images are visually consistent with clean ones, which ensures that they cannot be easily spotted by humans.

## 4. The Proposed BadCLIP

In this section, we introduce the proposed BadCLIP. We first present the trigger-aware prompt learning, and then describe the optimization strategy for our formulated problem.

### 4.1. Trigger-Aware Prompt Learning

A CLIP model adapted for a specific visual recognition task only takes an image from the user as the input and outputs the predicted class to which the image belongs. Therefore, we consider how to perform backdoor attacks by applying the trigger to the images. Due to visual and textual branches

in the CLIP model, we expect that the trigger changes both image and text features in our backdoor attack. Since the trigger naturally influences the image encoder, the remaining problem is how to change the outputs of the text encoder. Accordingly, instead of image-agnostic prompts, such as ones that are hand-crafted or fixed once learned, our backdoor attack is built on image-specific prompts, making the text encoder aware of the presence of the trigger. On the other hand, an expected benefit of using image-specific prompts is that they are more generalizable than static prompts, as suggested in [100], which helps BadCLIP succeed under transfer settings.

To this end, we use a neural network $h(\cdot)$ with parameters $\boldsymbol{\theta}$ as the trigger-aware context generator, and combine the class names to produce image-specific prompts $\{h_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}_i\}$ ($h_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^{N \times e}$ and $i = 1, 2, ..., K$). The corresponding prediction probability is calculated as follows:

$$\tilde{p}(y = i | \boldsymbol{x}) = \frac{\exp(\text{sim}(f(\boldsymbol{x}), g(\{h_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}_i\}))/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(f(\boldsymbol{x}), g(\{h_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{c}_j\})/\tau))}, \quad (2)$$

where $\boldsymbol{x}$ can be a clean image or a backdoor image. In our implementation, to balance the efficiency and effectiveness, $h(\cdot)$ is specified as a two-layer fully-connected network and takes image features extracted by the image encoder as inputs as suggested in [100].

Recall that one of attacker's goals is to classify backdoor images toward the specified target class $t$. To craft backdoor images, we use additive noise [3, 49, 81] as the trigger, denoted as $\boldsymbol{\delta} \in \mathbb{R}^p$. We also introduce $\ell_\infty$ restriction on $\boldsymbol{\delta}$ to keep the trigger unnoticeable. The parameters $\boldsymbol{\theta}$ and the trigger $\boldsymbol{\delta}$ are trained by minimizing the empirical classification loss:

$$\mathcal{L}_{tri}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathbb{E}_{\boldsymbol{x}_i} \left[ -\log \tilde{p}(y = t | \boldsymbol{x}_i + \boldsymbol{\delta}) \right], \quad (3)$$
$$s.t. \quad ||\boldsymbol{\delta}||_\infty \leqslant \epsilon$$

where $\epsilon$ denotes the maximum noise strength.

Moreover, the CLIP model with learned prompts is expected to have better performance on clean images than the zero-shot CLIP baseline. Therefore, we also optimize $\boldsymbol{\theta}$ by minimizing the below loss over clean images:

$$\mathcal{L}_{cle}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}_i, y_i} \left[ -\log \tilde{p}(y = y_i | \boldsymbol{x}_i) \right], \quad (4)$$

where $y_i$ is the ground-truth class of the image $\boldsymbol{x}$. Then, the total loss during our prompt learning is:

$$\mathcal{L}_{total}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \mathcal{L}_{tri}(\boldsymbol{\theta}, \boldsymbol{\delta}) + \mathcal{L}_{cle}(\boldsymbol{\theta})$$
$$s.t. \quad ||\boldsymbol{\delta}||_\infty \leqslant \epsilon. \quad (5)$$

### 4.2. Optimization

Since $\mathcal{L}_{total}$ is differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, both of them can be optimized by stochastic gradient descent [94]. However, we empirically find that simultaneously optimizing $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ from scratch results in a sub-optimal solution,

Table 2. Results of four methods in comparison on the seen and unseen classes (H: harmonic mean). BadCLIP is competitive with two advanced prompt learning methods (CoOp [101] and CoCoOp [100]) in terms of ACC, and reaches high ASRs.

| Dataset | Seen | | | | | Unseen | | | | | H | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | CLIP ACC | CoOp ACC | CoCoOp ACC | **BadCLIP** ACC | ASR | CLIP ACC | CoOp ACC | CoCoOp ACC | **BadCLIP** ACC | ASR | CLIP ACC | CoOp ACC | CoCoOp ACC | **BadCLIP** ACC | ASR |
| ImageNet | 72.43 | 76.47 | 75.98 | 75.67 | 99.90 | 68.14 | 67.88 | 70.43 | 70.33 | 99.40 | 70.22 | 71.92 | 73.10 | 72.90 | 99.65 |
| Caltech101 | 96.84 | 98.00 | 97.96 | 97.83 | 99.70 | 94.00 | 89.81 | 93.81 | 93.43 | 99.23 | 95.40 | 93.73 | 95.84 | 95.58 | 99.46 |
| OxfordPets | 91.17 | 93.67 | 95.20 | 93.87 | 98.70 | 97.26 | 95.29 | 97.69 | 84.03 | 99.23 | 94.12 | 94.47 | 96.43 | 88.68 | 98.96 |
| StanfordCars | 63.37 | 78.12 | 70.49 | 70.10 | 99.80 | 74.89 | 60.40 | 73.59 | 72.63 | 99.80 | 68.65 | 68.13 | 72.01 | 71.34 | 99.80 |
| Flowers102 | 72.08 | 97.60 | 94.87 | 93.13 | 99.90 | 77.80 | 59.67 | 71.75 | 73.53 | 99.93 | 74.83 | 74.06 | 81.71 | 82.18 | 99.91 |
| Food101 | 90.10 | 88.33 | 90.70 | 89.60 | 99.07 | 91.22 | 82.26 | 91.29 | 90.60 | 98.73 | 90.66 | 85.19 | 90.99 | 90.10 | 98.90 |
| FGVCAircraft | 27.19 | 40.44 | 33.41 | 34.17 | 99.93 | 36.29 | 22.30 | 23.71 | 31.83 | 99.43 | 31.09 | 28.75 | 27.74 | 32.96 | 99.68 |
| SUN397 | 69.36 | 80.60 | 79.74 | 78.70 | 99.70 | 75.35 | 65.89 | 76.86 | 76.53 | 99.30 | 72.23 | 72.51 | 78.27 | 77.60 | 99.50 |
| DTD | 53.24 | 79.44 | 77.01 | 74.93 | 98.93 | 59.90 | 41.18 | 56.00 | 49.77 | 96.93 | 56.37 | 54.24 | 64.85 | 59.81 | 97.92 |
| EuroSAT | 56.48 | 92.19 | 87.49 | 86.33 | 99.27 | 64.05 | 54.74 | 60.04 | 53.40 | 97.73 | 60.03 | 68.69 | 71.21 | 65.98 | 98.49 |
| UCF101 | 70.53 | 84.69 | 82.33 | 80.70 | 99.77 | 77.50 | 56.05 | 73.45 | 72.37 | 99.47 | 73.85 | 67.46 | 77.64 | 76.31 | 99.62 |
| Average | 69.34 | 82.69 | 80.47 | 79.55 | 99.52 | 74.22 | 63.22 | 71.69 | 69.86 | 99.02 | 71.59 | 70.83 | 75.44 | 73.95 | 99.26 |

which may be because there are two separate objectives in Problem (5). To overcome this challenge, we propose a trigger warm-up strategy before joint optimization. Specifically, we first update $\delta$ for $T'$ iterations while fixing $\theta$ after random initialization. The update of $\delta$ with the learning rate $\alpha$ in the warm-up stage is:

$$\delta^{k+1} \leftarrow \delta^k - \alpha \cdot \frac{\partial \mathcal{L}_{tri}(\theta^r, \delta)}{\partial \delta}\bigg|_{\delta=\delta^k}, \qquad (6)$$

where $k = 1, 2, ..., T'$ indicates the iteration index and $\theta^r$ is obtained by random initialization. We then jointly optimize $\theta$ and $\delta$ for $T''$ iterations with the learning rate $\beta$:

$$\begin{cases} \theta^{k+1} \leftarrow \theta^k - \beta \cdot \frac{\partial \mathcal{L}_{total}(\theta, \delta^k)}{\partial \theta}\bigg|_{\theta=\theta^k} \\ \delta^{k+1} \leftarrow \delta^k - \beta \cdot \frac{\partial \mathcal{L}_{total}(\theta^k, \delta)}{\partial \delta}\bigg|_{\delta=\delta^k} \end{cases}, \qquad (7)$$

where $k = T'+1, T'+2, ..., T'+T''$ and $\theta^{T'+1} = \theta^r$. After $T'+T''$ iterations, $\theta$ can be used to produce image-specific prompts, and $\delta$ is the trigger to activate the backdoor. Fig. 1 shows an example of testing our BadCLIP on a clean and backdoor image.

## 5. Experiments

### 5.1. Setup

**Datasets.** As mentioned in Section 3.2, we evaluate our BadCLIP under four settings of training data. Following [100, 101], we adopt 11 datasets, including ImageNet [16], Caltech101 [20], OxfordPets [60], StanfordCars [40], Flowers102 [59], Food101 [7], FGVCAircraft [56], SUN397 [86], DTD [15], EuroSAT [31], and UCF101 [69]. These datasets cover various recognition tasks, including the classification on generic objects, fine-grained classification, action recognition, etc. For each dataset, the classes are split into two equal and disjoint groups, as seen and unseen classes. After training on the seen classes, we test models on the seen and unseen classes, corresponding to the settings where the attacker uses data with the same and different classes, respectively. To evaluate the cross-dataset

transferability of our BadCLIP, we train models on ImageNet and test on the remaining 10 datasets. In the cross-domain experiments, we use ImageNet as the source dataset for training and its domain-shifted variants as target datasets for testing, including ImageNetV2 [63], ImageNet-Sketch [79], ImageNet-A [33], and ImageNet-R [32].

**Implementation details.** In our experiments, unless otherwise specified, ViT-B/16 is used as the image encoder's backbone, the number of labeled training examples per class is 16 (i.e., 16-shot), and the context length $N$ is set as 4. We optimize $\delta$ for 3 epochs with a fixed learning rate 0.1 in the trigger warm-up stage, and then jointly optimize $\theta$ and $\delta$ for 10 epochs using 1 epoch of the learning rate warm-up and a cosine annealing scheduler with a learning rate 0.002. In both stages, we adopt SGD optimizer. By default, the maximum noise strength $\epsilon$ is 4 and the first class is chosen as the target class for each dataset. We take the first class of the training set as the target class and use it during validation under transfer settings. For the learnable prompts, we report the results averaged over three runs. All pre-trained weights are drawn from CLIP's released models [61]. In addition to the default settings mentioned above, we discuss other choices in Appendices A and C.

**Evaluation criteria.** We mainly adopt two metrics to evaluate the attack performance, i.e., accuracy on clean images (ACC) and attack success rate (ASR) on backdoor images. ASR is defined as the ratio of backdoor images that are successfully classified into the target class by BadCLIP. To highlight the performance trade-off between the seen and unseen classes, we compute the harmonic mean of results on the seen and unseen classes for these two metrics, following [85, 100]. Also, for comparison, we report the accuracy on clean images of zero-shot CLIP [61] and two advanced prompt learning methods, i.e., CoOp [101] and CoCoOp [100]. We also compare BadCLIP with a backdoor attack, BadEncoder [37]. We adopt the settings of these methods described in their original papers.

## 5.2. Results on Seen and Unseen Classes

In this section, we perform prompt learning on the seen classes and test the models on the seen and unseen classes on 11 datasets. The results are shown in Table 2.

**BadCLIP correctly classifies clean images.** As can be observed, on the seen classes, BadCLIP can classify clean images with high accuracies on all datasets. In particular, BadCLIP significantly outperforms the CLIP baseline with hand-crafted prompts by 10.21% on average. Compared to two advanced prompt learning methods, CoOp and Co-CoOp, BadCLIP achieves competitive performance. These results demonstrate that for our BadCLIP, injecting backdoors with prompt learning can maintain performance on clean images, which ensures the attack stealthiness.

**BadCLIP achieves high attack success rates.** We can find from Table 2 that BadCLIP shows promising performance in terms of ASR. Specifically, BadCLIP achieves high ASRs (>98.7%) on all datasets and a 99.52 ASR on average. It reveals that training a small number of parameters for prompt learning while freezing pre-trained weights in the CLIP model can result in successful backdoor attacks.

**Backdoor generalizes to unseen classes.** Table 2 also shows that the backdoor learned by BadCLIP can generalize to the unseen classes, with a 99.02% ASR on average. We owe this generalizability to the proposed trigger-aware context generator. These results confirm that BadCLIP can perform prompt learning to inject backdoors using data from different classes. Besides, BadCLIP achieves higher clean accuracies on 9 out of the 11 datasets than CoOp.

**Backdoor images are difficult to be detected.** To quantitatively measure the stealthiness of backdoor images, we calculate the PSNR [35] and SSIM [82] values using 100 pairs of clean and backdoor images for each dataset. The PSNR and SSIM values are 40.49 and 0.9642 averaged over 11 datasets, indicating that backdoor images are difficult to be detected by humans. We also provide visualization examples in Appendix B. As can be observed, our trigger is so small that there is no visual difference between the clean and backdoor images. These results demonstrate that our attack is stealthy.

## 5.3. Cross-Dataset Transfer

In this part, we evaluate the performance of prompt learning methods under the cross-dataset setting, especially for backdoors learned by our BadCLIP. The results are shown in Table 3. In this setting, the accuracy on clean images of BadCLIP is on par with CoCoOp and surpasses CoOp by a large margin up to 1.43% on average. Also, we surprisingly find that BadCLIP obtains 100% attack success rates on 9 out of 10 datasets. It illustrates that the trigger-aware context generator and the trigger learned on ImageNet can be applied to attack various downstream datasets. Notably,

Table 3. Results of four methods under the cross-dataset transfer setting. The learning based methods are trained on ImageNet and tested on the other 10 datasets.

|  | Dataset | CLIP ACC | CoOp ACC | CoCoOp ACC | BadCLIP ACC | BadCLIP ASR |
|---|---|---|---|---|---|---|
| Source | ImageNet | 66.74 | 71.51 | 71.02 | 70.77 | 99.93 |
| Target | Caltech101 | 93.09 | 93.70 | 94.43 | 93.63 | 100.0 |
|  | OxfordPets | 89.07 | 89.14 | 90.14 | 90.70 | 100.0 |
|  | StanfordCars | 65.17 | 64.51 | 65.32 | 64.17 | 100.0 |
|  | Flowers102 | 71.14 | 68.71 | 71.88 | 70.83 | 100.0 |
|  | Food101 | 86.07 | 85.30 | 86.06 | 85.17 | 100.0 |
|  | FGVCAircraft | 24.62 | 18.47 | 22.94 | 23.40 | 100.0 |
|  | SUN397 | 62.52 | 64.15 | 67.36 | 66.90 | 100.0 |
|  | DTD | 44.38 | 41.92 | 45.73 | 45.00 | 99.77 |
|  | EuroSAT | 47.53 | 46.39 | 45.37 | 45.13 | 100.0 |
|  | UCF101 | 66.67 | 66.55 | 68.21 | 68.17 | 100.0 |
|  | Average | 65.02 | 63.88 | 65.74 | 65.31 | 99.98 |

Table 4. Results of four methods under the cross-domain transfer setting. The learning based methods are trained on ImageNet and tested on its 4 domain-shifted variants.

|  | Dataset | CLIP ACC | CoOp ACC | CoCoOp ACC | BadCLIP ACC | BadCLIP ASR |
|---|---|---|---|---|---|---|
| Source | ImageNet | 66.73 | 71.51 | 71.02 | 70.77 | 99.93 |
| Target | ImageNetV2 | 60.83 | 64.20 | 64.07 | 63.93 | 100.0 |
|  | ImageNet-Sketch | 46.15 | 47.99 | 48.75 | 48.47 | 99.70 |
|  | ImageNet-A | 47.77 | 49.71 | 50.63 | 49.67 | 100.0 |
|  | ImageNet-R | 73.96 | 75.21 | 76.18 | 75.33 | 99.97 |
|  | Average | 55.96 | 59.28 | 59.91 | 59.35 | 99.92 |

the attack can still succeed on these datasets containing totally different categories from ImageNet, such as Food101 and UCF101. Our results demonstrate that BadCLIP poses a serious security threat to downstream tasks even though the attacker cannot access their datasets.

## 5.4. Cross-Domain Transfer

The cross-domain transferability is critical for backdoor attacks to succeed in diverse real-world scenarios. Following previous works [100, 101], we perform the prompt learning on ImageNet and test models on its 4 domain-shifted variants, as shown in Table 4. We can see that BadCLIP achieves similar performance compared to CoCoOp regarding accuracy on clean images, indicating that it inherits the advantages of the learnable prompts [101]. We can also observe that BadCLIP reaches high attack success rates on all target datasets, ranging from 99.70% to 100.0%. These results suggest that BadCLIP is robust to domain shift.

## 5.5. Comparison with Existing Attacks

**Data poisoning based attack.** This method [10] assumes that the attacker has access to the pre-training dataset for data poisoning and the CLIP model is pre-trained on it. Since our attack happens in the prompt-learning stage for the pre-trained CLIP model, it is infeasible to conduct a fair comparison between the data poison based attack and our BadCLIP. However, we observe from [10] that data poison-

Table 5. Comparison between BadEncoder [37] and BadCLIP on STL10. "-" implies that BadCLIP does not require additional data.

| Method | Source of Additional Data | Number of Additional Data | ACC | ASR |
|---|---|---|---|---|
| BadEncoder | STL10 | 50,000 | 94.83 | 99.96 |
| | STL10 | 5,000 | 94.74 | 92.47 |
| | STL10 | 1,000 | 94.01 | 17.39 |
| | SVHN | 5,000 | 91.33 | 11.45 |
| **BadCLIP** | - | - | 95.13 | 98.57 |

ing may limit the attack performance. For instance, its attack success rate is less than 80% when inserting 1,500 poisoned samples into the Conceptual Captions dataset [65].

**Fine-tuning on poisoning data.** We provide another baseline considered by CleanCLIP [5], i.e., fine-tuning on poisoning data. For the CLIP with ResNet-50, it achieves a 58.40% ACC and a 94.60% ASR, while our BadCLIP performes better, with a 67.10% ACC and a 98.75% ASR, indicating the superiority of our design.

**BadEncoder.** It fine-tunes the image encoder of the pre-trained CLIP model with a large amount of additional unlabeled data, and then trains a task-specific classifier with the downstream dataset. Table 5 shows the comparison between BadEncoder and our BadCLIP on STL10 adopted in [37], where the number of labeled training samples per class is 16. Following [37], we use ResNet-50 as the image encoder's backbone and set the target class as "*truck*". For a comprehensive comparison, we vary the source and number of additional data adopted in BadEncoder. As can be seen, BadEncoder achieves a high clean accuracy and attack success rate with 50,000 additional data samples from STL10. However, when we reduce the amount of additional data or change the source, the attack performance is degraded significantly. Hence, BadEncoder depends on a large amount of additional data from a similar source as that of the downstream dataset, while our BadCLIP does not require additional data. Besides, we would like to emphasize that, unlike our BadCLIP, BadEncoder is not generalizable to unseen classes due to the task-specific classifier.

## 5.6. Trigger-Aware Prompts Matter

**Understanding BadCLIP in the feature space.** As demonstrated in Fig. 1, the backdoor images change both image and text features in our BadCLIP. Here, to show the effect of trigger-aware prompts, we propose to decouple the inputs into the image and text encoders to analyze the effect of the changes of image and text features, respectively. Specifically, the image encoder takes the clean image $x$ or the backdoor image $x + \delta$ as inputs; the text encoder takes the clean text prompt $\{h_\theta(x), c_t\}$ or the backdoor text prompt $\{h_\theta(x + \delta), c_t\}$ for the target class $t$ as inputs. We calculate the distribution of cosine similarities between images and text features in four cases, as shown in Fig. 2. As
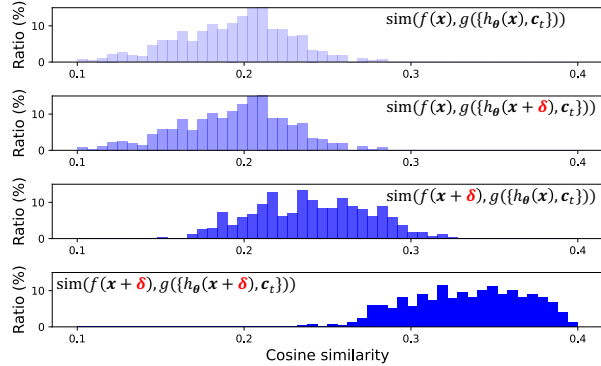


Figure 2. Distribution of cosine similarities between images and text prompts in the feature space. $f(x)$: clean image features; $f(x + \delta)$: backdoor image features; $g(\{h_\theta(x), c_t\})$: clean text features for the target class $t$; $g(\{h_\theta(x + \delta), c_t\})$: backdoor text features for the target class $t$. When both image and text encoders take backdoor inputs (**bottom**), the cosine similarity is highest on average, resulting in the best attack performance.
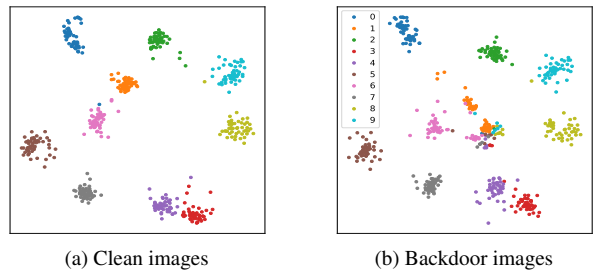


(a) Clean images      (b) Backdoor images

Figure 3. t-SNE visualization of features extracted by BadCLIP's image encoder for clean images and their backdoor versions from 10 random classes on ImageNet. Our backdoor image features are still separable. Note that the class 0 corresponds to the target class.

can be seen, when both image and text encoders take backdoor inputs, the cosine similarity is highest on average, implying that inputs are classified into the target class with the highest confidences. Our analysis illustrates that the success of our backdoor attack can be attributed to the collaboration between the changes of image and text features. Thus, although the features of images shift across different scenarios, the textual features of the target class change along with the trigger, ensuring successful attacks. This insight is fundamental and critical, and will inspire backdoor studies on multi-modal models.

The t-SNE [75] visualization of clean and backdoor image features further confirms the effect of trigger-aware prompts. As suggested in [84], for backdoor attacks on the image recognition models only relying on the visual modality, their backdoor image features cluster together. In contrast, for our BadCLIP built on visual and textual modalities, its backdoor image features are still separable as shown in Fig. 3. This observation indirectly indicates that the backdoor text prompts contribute a lot to the targeted misclassification in our method. We believe that this interesting

Table 6. Comparison of the trigger-agnostic prompts and trigger-aware prompts (adopted in our BadCLIP) in backdoor attacks. Results are averaged over 11 datasets.

| Method | Seen | | Unseen | | H | |
|---|---|---|---|---|---|---|
| | ACC | ASR | ACC | ASR | ACC | ASR |
| Trigger-Agnostic Prompts | 76.19 | 95.31 | 62.73 | 2.21 | 68.14 | 3.81 |
| Trigger-Aware Prompts (Ours) | 79.55 | 99.52 | 69.86 | 99.02 | 73.95 | 99.26 |



(a) Neural Cleanse　　　(b) CLP defense
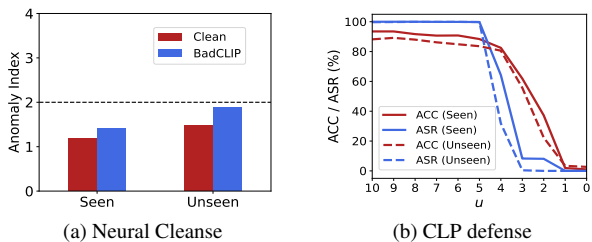
Figure 4. Results of defense experiments on Caltech101.

phenomenon for multi-modal models is worthy of a further exploration from both backdoor attack and defense sides.

**Backdoor attack with trigger-agnostic prompts.** We study the effect of trigger-agnostic prompts by comparing BadCLIP with a baseline, i.e., the backdoor attack with trigger-agnostic prompts. Specifically, following [42, 99, 101], we model context tokens using continuous vectors, which are fixed for any image input once learned, such that the text features cannot be changed by the backdoor images. Other settings are the same as those used in BadCLIP. The comparison in Table 6 shows the superiority of our method. In particular, backdoor attack with trigger-agnostic prompts fails to generalize to unseen classes. These results demonstrate that trigger-aware prompts have a positive effect on the generalizability of BadCLIP.

## 5.7. Resistance to Backdoor Defense Methods

**Resistance to Neural Cleanse.** Neural Cleanse [77] assumes that the backdoor trigger is patch based. For each class, it reconstructs the optimal patch pattern to convert any clean input to that target class. If any class has a significantly smaller pattern than the others, Neural Cleanse considers it as a backdoor indicator. It is quantified by the Anomaly Index metric. If the Anomaly Index is less than a threshold of 2 for a specific class, the defense considers that there is a backdoor with this class as the target label. We show the results of the clean CLIP model and our BadCLIP on Caltech101 in Fig. 4a. Similar to the clean model, BadCLIP passes the tests with very small scores, showing that our attack is resistant to Neural Cleanse.

**Resistance to CLP defense.** Channel Lipschitzness based Pruning (CLP) [98] is a data-free backdoor removal method. It prunes those neurons that are sensitive to input changes. Fig. 4b presents the results under different settings of $u$ in CLP. A smaller $u$ means a larger pruning ratio. We can see from the figure that when CLP removes the back-

Table 7. Results of the proposed attack on OpenCLIP. BadOpenCLIP denotes our attack.

| Pre-trained Language Model | | | Huge-scale Model | | |
|---|---|---|---|---|---|
| OpenCLIP | **BadOpenCLIP** | | OpenCLIP | **BadOpenCLIP** | |
| ACC | ACC | ASR | ACC | ACC | ASR |
| 69.86 | 74.15 | 98.81 | 80.56 | 84.49 | 99.90 |

Table 8. Results of BadCLIP on the image-text retrieval task.

| CLIP | CoOp | CoCoOp | **BadCLIP** | |
|---|---|---|---|---|
| R@1 | R@1 | R@1 | R@1 | B-R@1 |
| 83.0 | 79.4 | 85.9 | 85.2 | 98.3 |

door ($u > 3$), the accuracy on clean images is significantly reduced. Therefore, CLP cannot eliminate the backdoor injected by our BadCLIP with a high ACC.

## 5.8. Extensible Application Scenario

Here, we evaluate our attack on more application scenarios. Firstly, we apply our attack to a recently released version of CLIP, named OpenCLIP [19], which utilizes a different pre-training dataset (LAION) [64], and many additional techniques such as using a pre-trained language model and scaling up to a huge-scale model architecture. Table 7 shows the results of two variants of OpenCLIP on UCF-101. We can see that our method can succeed in attacking these two models. Secondly, we carry out experiments on the image-text retrieval task with Flickr30K [89], following [30]. The prompt learning methods are trained with only 3% of training data and tested on the complete test set. R@1 and B-R@1 denote Recall at 1 and that of backdoor image queries, respectively. Table 8 shows the success of BadCLIP on the image-text retrieval task. These results indicate that the application scenario of BadCLIP is extensible.

## 6. Conclusions

In this paper, we explored backdoor attacks on CLIP with limited downstream training data. We proposed BadCLIP which accomplishes backdoor attacks via trigger-aware prompt learning. BadCLIP consists of a learnable trigger applied to images and a trigger-aware context generator. We then proposed an optimization method based on a novel warm-up strategy. We showed that BadCLIP achieves promising attack performance and a generalizable backdoor. To the best of our knowledge, BadCLIP is the first backdoor attack on CLIP in the prompt learning stage. We would hope that our work opens a new domain of attack mechanisms on vision-language models, and can encourage future defense research.

# References

[1] Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. *arXiv preprint arXiv:2102.10496*, 2021. 3

[2] Jiawang Bai, Kuofeng Gao, Dihong Gong, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Hardly perceptible trojan attack against neural networks with bit flips. In *ECCV*, 2022. 3

[3] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *ECCV*, 2022. 4

[4] Jiawang Bai, Baoyuan Wu, Zhifeng Li, and Shu-Tao Xia. Versatile weight attack via flipping limited bits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[5] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *ICCV*, 2023. 7

[6] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *ICIP*, 2019. 3

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 5

[8] Adrian Bulat and Georgios Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. *arXiv preprint arXiv:2210.01115*, 2022. 2

[9] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, 2023. 2

[10] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *ICLR*, 2022. 1, 3, 6

[11] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 1

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[14] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005. 1

[15] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[17] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *ICCV*, 2021. 3, 4

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3

[19] Cherti Mehdi et al. Reproducible scaling laws for contrastive language-image learning. *arXiv*, 2022. 8

[20] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRw*, 2004. 5

[21] Kuofeng Gao, Jiawang Bai, Bin Chen, Dongxian Wu, and Shu-Tao Xia. Backdoor attack on hash-based image retrieval via clean-label data poisoning. In *BMVC*, 2023. 3

[22] Kuofeng Gao, Jiawang Bai, Baoyuan Wu, Mengxi Ya, and Shu-Tao Xia. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19:1267–1282, 2023. 3

[23] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *CVPR*, 2023. 3

[24] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. In *ICLR*, 2024. 3

[25] Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139:109512, 2023. 3

[26] Dihong Gong, Zhifeng Li, Jianzhuang Liu, and Yu Qiao. Multi-feature canonical correlation analysis for face photo-sketch image retrieval. In *ACM MM*, 2013. 3

[27] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1, 3

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[30] Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Eric Wang. Cpl: Counterfactual prompt learning for vision and language models. In *EMNLP*, 2022. 8

[31] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[32] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 5

[33] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Stein-hardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5

[34] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 3

[35] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 3

[37] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Baden-coder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *SP*, 2022. 1, 2, 3, 5, 7

[38] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 3

[39] Muhammad Uzair Khattak, Hanoona Rasheed, Muham-mad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022. 2

[40] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVw*, 2013. 5

[41] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, 2023. 2

[42] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 3, 8

[43] Boheng Li, Yishuo Cai, Haowei Li, Feng Xue, Zhifeng Li, and Yiming Li. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *CVPR*, 2024. 3

[44] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. In *NeurIPS*, 2024. 1

[45] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021. 3

[46] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 3

[47] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 3

[48] Zhifeng Li, Dihong Gong, Qiang Li, Dacheng Tao, and Xuelong Li. Mutual component analysis for heterogeneous face recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):1–23, 2016. 3

[49] Siyuan Liang, Xingxing Wei, Siyuan Yao, and Xiaochun Cao. Efficient adversarial attacks for visual object tracking. In *ECCV*, 2020. 4

[50] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal con-trastive learning. *arXiv preprint arXiv:2311.12075*, 2023. 3

[51] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hi-roaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3

[52] Wei Liu, Zhifeng Li, and Xiaoou Tang. Spatio-temporal embedding for statistical face recognition from video. In *ECCV*, 2006. 3

[53] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*, 2018. 3

[54] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022. 1, 2, 3

[55] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xin-chao Wang. Beyond sole strength: Customized ensem-bles for generalized vision-language models. *arXiv preprint arXiv:2311.17091*, 2023. 1

[56] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classi-fication of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[57] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *NeurIPS*, 2020. 3

[58] Tuan Anh Nguyen and Anh Tuan Tran. Wanet-imperceptible warping-based backdoor attack. In *ICLR*, 2021. 3, 4

[59] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5

[60] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-ing transferable visual models from natural language super-vision. In *ICML*, 2021. 1, 3, 4, 5

[62] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 1, 3

[63] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im-agenet? In *ICML*, 2019. 5

[64] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Worts-man, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Sys-tems Datasets and Benchmarks Track*. 8

[65] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 7

[66] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 3

[67] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 3

[68] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *NeurIPS*, 2013. 3

[69] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[70] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *NeurIPS*, 2022. 1

[71] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 2022. 1

[72] Xiaoou Tang and Zhifeng Li. Video based face recognition using multiple classifiers. In *ICAFGR*, 2004. 3

[73] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV 2020*, 2020. 2

[74] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1, 3

[75] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[77] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *SP*, 2019. 3, 8

[78] Feng Wang, Manling Li, Xudong Lin, Hairong Lv, Alex Schwing, and Heng Ji. Learning to decompose visual features with latent textual prompts. In *ICLR*, 2023. 1

[79] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 1, 5

[80] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Action-clip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[81] Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *ECCV*, 2022. 4

[82] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[83] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018. 3

[84] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Chao Shen, and Hongyuan Zha. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS*, 2022. 7

[85] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 3, 5

[86] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5

[87] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*, 2024. 3

[88] Mengxi Ya, Yiming Li, Tao Dai, Bin Wang, Yong Jiang, and Shu-Tao Xia. Towards faithful xai evaluation via generalization-limited backdoor watermark. In *ICLR*, 2024. 3

[89] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 8

[90] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[91] Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *ICCV*, 2023. 3

[92] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1

[93] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 1, 3

[94] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004. 4

[95] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *CVPR*, 2019. 3

[96] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *CVPR*, 2020. 3

[97] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *CVPR*, 2022. 4

[98] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *ECCV*, 2022. 8

[99] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL-HLT*, pages 5017–5033, 2021. 3, 8

[100] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6

[101] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 5, 6, 8

[102] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, 2023. 3

[103] Mingli Zhu, Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. In *NeurIPS*, 2024. 3