Defending Multimodal Backdoored Models by Repulsive Visual Prompt Tuning

Zhifang Zhang
Southeast University
zzfofficial@gmail.com

Shuo He*
Nanyang Technological University
shuohe123@gmail.com

Haobo Wang
Zhejiang University
wanghaobo@zju.edu.cn

Bingquan Shen
DSO National Laboratories
SBingqua@dso.org.sg

Lei Feng*
Southeast University
lfengqaq@gmail.com

Abstract

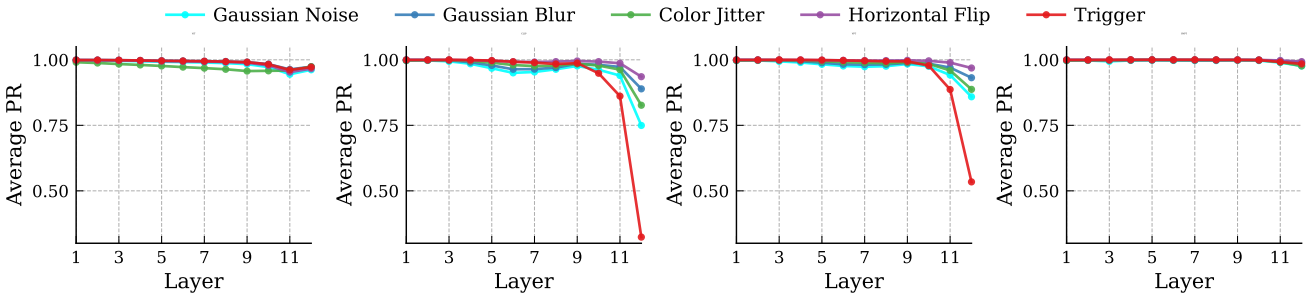
Multimodal contrastive learning models (e.g., CLIP) can learn high-quality representations from large-scale image-text datasets, while they exhibit significant vulnerabilities to backdoor attacks, raising serious safety concerns. In this paper, we reveal that CLIP's vulnerabilities primarily stem from its tendency to encode features beyond in-dataset predictive patterns, compromising its visual feature resistivity to input perturbations. This makes its encoded features highly susceptible to being reshaped by backdoor triggers. To address this challenge, we propose Repulsive Visual Prompt Tuning (RVPT), a novel defense approach that employs deep visual prompt tuning with a specially designed feature-repelling loss. Specifically, RVPT adversarially repels the encoded features from deeper layers while optimizing the standard cross-entropy loss, ensuring that only predictive features in downstream tasks are encoded, thereby enhancing CLIP's visual feature resistivity against input perturbations and mitigating its susceptibility to backdoor attacks. Unlike existing multimodal backdoor defense methods that typically require the availability of poisoned data or involve fine-tuning the entire model, RVPT leverages few-shot downstream clean samples and only tunes a small number of parameters. Empirical results demonstrate that RVPT tunes only 0.27% of the parameters in CLIP, yet it significantly outperforms state-of-the-art defense methods, reducing the attack success rate from 89.70% to 2.76% against the most advanced multimodal attacks on ImageNet and effectively generalizes its defensive capabilities across multiple datasets. Our code is available on https://github.com/zhangzf01/RVPT.

1. Introduction

Recently, Contrastive Language-Image Pretraining (CLIP) [45] has emerged as a powerful base model across multiple domains. Unlike traditional models that rely on uni-modal supervision, CLIP aligns visual and textual signals to learn rich representations from numerous image-text pairs in the pre-training stage. This unique training paradigm enables its impressive performance in open-vocabulary visual recognition [57] and robustness to distribution shift [13].

However, despite CLIP's success in visual recognition, recent research [4, 5] has revealed its vulnerabilities against backdoor attacks in downstream tasks. The vulnerabilities can be roughly categorized into two aspects. First, concerns arise from its training data collection process: the data is noisy, uncurated, and sourced from the web, which facilitates adversarial poisoning of the image-text pairs. Second, CLIP demonstrates heightened susceptibility to a smaller number of poisoned training samples: recent studies [4] have empirically shown that poisoning CLIP may require injecting orders of magnitude fewer poisoned data into the training dataset than poisoning conventional supervised models. Once poisoned during pre-training, the backdoored CLIP will inevitably classify images with an adversarially designed trigger into the predefined target class.

CLIP's vulnerabilities to backdoor attacks have inspired the development of various defense methods. However, most of these methods necessitate fine-tuning the entire model's parameters [3] or require the availability of poisoned data [22, 60, 61], which could be resource-intensive or even unrealistic. Thus, to achieve an efficient, effective, and practical defense, we are dedicated to developing a defense method based on visual prompt tuning (VPT) [24], utilizing only a small portion of downstream clean samples to encourage the model to ignore the trigger feature in the poisoned images. However, simply adopting VPT proves ineffective experimentally, as the trigger feature cannot be extracted from clean samples, preventing VPT from learning how to disregard them. Alternatively, VPT can help CLIP learn to extract in-dataset predictive features from clean samples using cross-entropy loss. If the visual encoder is guided to



(a) ImageNet-Supervised ViT *Without Poisoning*

(b) ViT of Backdoored CLIP *ASR: 99.83%*

(c) VPT *ASR: 99.83%→96.66%*

(d) Our RVPT *ASR: 99.83%→2.76%*

Figure 1. Perturbation resistivity (PR) across different layers of the encoders under various perturbations, including the trigger pattern. A higher PR value indicates less sensitivity to input perturbation. Four encoder variants are evaluated: (a) ViT trained exclusively on clean ImageNet [44]. (b) ViT in backdoored CLIP [45]. (c) ViT in backdoored CLIP tuned with Visual Prompt Tuning (VPT) on ImageNet. (d) ViT of the backdoored CLIP tuned with our proposed RVPT on ImageNet. Specifically, The trigger of the backdoored CLIP is BadCLIP [31]. Detailed experimental settings and additional PR results for CLIP backdoored by other triggers can be found in the supplementary material.

focus exclusively on these features, CLIP will gain backdoor robustness since the trigger features learned in pre-training are ignored during backdoor attacks.

Nevertheless, we observe that CLIP's encoded visual features, even after VPT, remain highly sensitive to perturbations, which suggests that it still encodes features beyond in-dataset predictive features, leaving it susceptible to potential backdoor attacks. To quantify this sensitivity, we introduce the Perturbation Resistivity (PR) metric, defined as the cosine similarity between the encoded feature $f^v(\boldsymbol{x})$ of the image $\boldsymbol{x}$ and the encoded feature $f^v(\boldsymbol{x}+\boldsymbol{\delta})$ of its perturbed version: $\boldsymbol{x}+\boldsymbol{\delta}$, where $f^v(\cdot)$ is the visual feature extractor:

$$\mathrm{PR}(\boldsymbol{x}, \boldsymbol{\delta}) = \cos(f^v(\boldsymbol{x}), f^v(\boldsymbol{x}+\boldsymbol{\delta})). \tag{1}$$

A lower PR value indicates greater sensitivity of the encoded features to input perturbations, making the model more susceptible to backdoor triggers. To illustrate this vulnerability of CLIP, we randomly sample 1,000 images from ImageNet and apply various perturbations. We then compare the average PR values of these images across different model variants. As shown in Figure 1 (b) (c), the deeper visual layers of CLIP exhibit significantly higher sensitivity to input perturbations, even after VPT, which allows the trigger to maliciously reshape the encoded visual features layer by layer, enabling successful attacks. In contrast, Figure 1 (a) shows that the ViT supervised on clean ImageNet exhibits lower sensitivity, because it learns only from ImageNet, preventing it from encoding irrelevant features beyond its dataset. Therefore, to reduce CLIP's sensitivity to input perturbations and enhance its backdoor robustness, we should suppress irrelevant features learned during pre-training and guide it to focus on encoding in-dataset predictive features.

To address this challenge, we propose a novel approach termed Repulsive Visual Prompt Tuning (RVPT), which improves CLIP's PR by allowing it to exclusively encode the predictive features in the downstream tasks in a few-shot manner. As illustrated in Figure 2, RVPT jointly optimizes two objectives: (1) a multi-layer feature-repelling loss that minimizes the cosine similarity between prompted features and original CLIP features at each layer, and (2) cross-entropy loss for downstream task accuracy. This mechanism forces the model to only extract features that contribute to CE loss optimization during training, while ignoring non-essential features (*e.g.*, noise or trigger pattern). As a result, Figure 1 (d) shows that RVPT significantly improves the PR of CLIP on ImageNet, hence effectively defending against the state-of-the-art multimodal backdoor attack. Moreover, extensive experiments demonstrate our method's effectiveness across multiple datasets and backdoor attacks.

2. Related Work

Backdoor attacks and defenses on supervised learning. Backdoor attacks have become an increasingly serious security issue because more and more practitioners choose to adopt third-party datasets, platforms, or even backbones to reduce costs. Research on backdoor attacks has primarily focused on designing triggers that enhance stealthiness [7, 50] and effectiveness [34, 40]. For backdoor defenses, researchers employ various strategies to mitigate backdoor attacks, including: (1) pre-processing defense [49] (2) pre-training defense [6] (3) post-training defense [55, 68] (4) test-time defense [15]. Specifically, our work aligns with the post-training defense category, which assumes that defenders do not have access to the poisoned training data and must eliminate the backdoor threat that has already been implanted into the model. There are two main strategies for post-training defense. (1) Pruning-based Methods [33, 56]: These approaches focus on identifying and removing the most suspicious neurons. (2) Fine-tuning-based Methods [29, 39, 53, 68]: These methods purify the backdoored model

by directly tuning its parameters with clean data.

Backdoor attacks and defenses on CLIP. Although vision-language models (VLMs) [23, 32, 45] have demonstrated significant improvements across various fields, recent research [4, 62] has revealed vulnerabilities of contrastively pre-trained VLMs to backdoor attacks, drawing substantial attention in the field of backdoor attacks and defenses on CLIP-like models. In terms of attacks, BadCLIP [31] optimizes visual trigger patterns within a dual-embedding guided framework, rendering the attack effective and undetectable. Moreover, many backdoor attack methods on multimodal large language models [30, 35–37, 42] have been proposed recently. Defense strategies can be primarily categorized into two categories: pre-training defense and post-training defense. For pre-training defense [22, 60, 61], SAFECLIP [61] prevents learning from poisoned examples by dynamically dividing the training data into the safe and risky sets by cosine similarity and only applying uni-modal contrastive loss to different modalities of the risky set. For fine-tuning defense [3, 25, 59], CleanCLIP [3] reduces the impact of spurious relationships of backdoor attacks by enforcing the model to realign representations for independent modalities.

Prompt learning for vision-language models. Lately, various fine-tuning methods have been introduced to further enhance the impressive capabilities of vision-language models (VLMs) across diverse downstream tasks [20, 26, 58, 67], with prompt learning [46, 64, 66] emerging as a particularly efficient and effective approach. Specifically, CoOp [66] first introduces prompt learning to VLMs and proposes a method that optimizes the context of the text prompt, *e.g.*, *"a photo of a <CLS>."* while keeping the class name tokens fixed. VPT [24] adjusts the visual features by concatenating the image or image features with the learnable prompt. Besides, BadCLIP [1] injects backdoor during the prompt learning stage, enabling the malicious alteration of text features through trigger-aware prompts for a powerful attack. We notice that despite the effectiveness of prompt learning for enhancing the model's robustness [8, 11, 25, 27, 65], limited research has explored its use in defending against multimodal backdoor attacks. In particular, current post-training backdoor defense methods for VLMs mainly focus on fine-tuning the whole backdoored models, overlooking the potential of prompt learning. To address the gap, we propose Repulsive Visual Prompt Tuning, which incorporates learnable parameters into features to defend against multimodal backdoor attacks while freezing the model's parameters.

3. The Proposed Approach

In this section, we first outline the basic settings of the threat model and defender's goal in Part 3.1, followed by a detailed illustration of how the adversary launches backdoor attacks on CLIP in Part 3.2. Eventually, in Part 3.3, we present Repulsive Visual Prompt Tuning, designed to defend backdoored CLIP against multiple multimodal backdoor attacks.

3.1. Basic Settings

Threat model. We assume the adversary has access to the training dataset and can produce a backdoored model by injecting carefully crafted poisoned data into the original clean dataset. For multimodal attacks, the poisoned data usually consists of images with the trigger pattern and the proxy captions for the target class. Once trained on the poisoned dataset, the model will misclassify inputs containing the trigger pattern into the predefined target class while maintaining correct classification for all other benign inputs.

Defender's goal. We suppose that the defender has access to the backdoored model and a limited set of clean downstream data. The defender's goal is to obtain a new model utilizing both the backdoored model and the available clean data, wherein the backdoor effects are mitigated while preserving the benign performance of the new model on the downstream clean data. Furthermore, the defender has full control over the recovering process of the backdoored model.

3.2. Backdoor Attacks on CLIP

In general, the CLIP model comprises a visual encoder $\mathcal{V}(\cdot)$, a textual encoder $\mathcal{T}(\cdot)$, and projection matrices that project the encoded representations into a joint feature space : $\boldsymbol{P}_{\mathrm{I}}$ for the image modality and $\boldsymbol{P}_{\mathrm{T}}$ for the text modality. The training dataset contains N_1 image-text pairs represented as $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{N_1}$, where $\boldsymbol{t}_i$ serves as a short caption for the corresponding image $\boldsymbol{x}_i$.

During the backdoor attack on CLIP in the training stage, adversaries typically inject a small number of N_2 poisoned image-text pairs denoted as $\mathcal{D}_p = \{(\boldsymbol{x}_i^p, \boldsymbol{t}_i^p)\}_{i=1}^{N_2}$. $\boldsymbol{x}_i^p$ represents a poisoned image obtained by modifying the original image $\boldsymbol{x}_i$ with the trigger pattern Θ, while t_i^p denotes the proxy caption for the target class y_t. Consequently, the original benign training dataset $\mathcal{D}$ can be poisoned, resulting in: $\widetilde{\mathcal{D}} = \{\mathcal{D} \cup \mathcal{D}_p\}$. During training, CLIP is optimized on $\widetilde{\mathcal{D}}$ to maximize the cosine similarity of positive pairs, expressed as $\phi(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{t}}_i) = \cos(\boldsymbol{P}_{\mathrm{I}}\mathcal{V}(\widetilde{\boldsymbol{x}}_i), \boldsymbol{P}_{\mathrm{T}}\mathcal{T}(\widetilde{\boldsymbol{t}}_i))$, and minimize the similarity of negative pairs $\phi(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{t}}_{j\neq i})$.

Upon training with $\widetilde{\mathcal{D}}$, the encoders of CLIP will be backdoored, denoted as $\{\widetilde{\mathcal{V}}(\cdot), \widetilde{\mathcal{T}}(\cdot)\}$. For the backdoored version of CLIP, the trigger pattern Θ will have a spurious correlation with the proxy caption of the target class y_t. Thus, during the inference stage, when the model encounters the attacked images containing the trigger pattern, the predicted posterior probability of the image being classified as the target class y_t will increase significantly. In the meantime, the backdoored CLIP performs normally on benign inputs.

3.3. Repulsive Visual Prompt Tuning

In this part, we present Repulsive Visual Prompt Tuning (RVPT), a framework designed to defend against multimodal

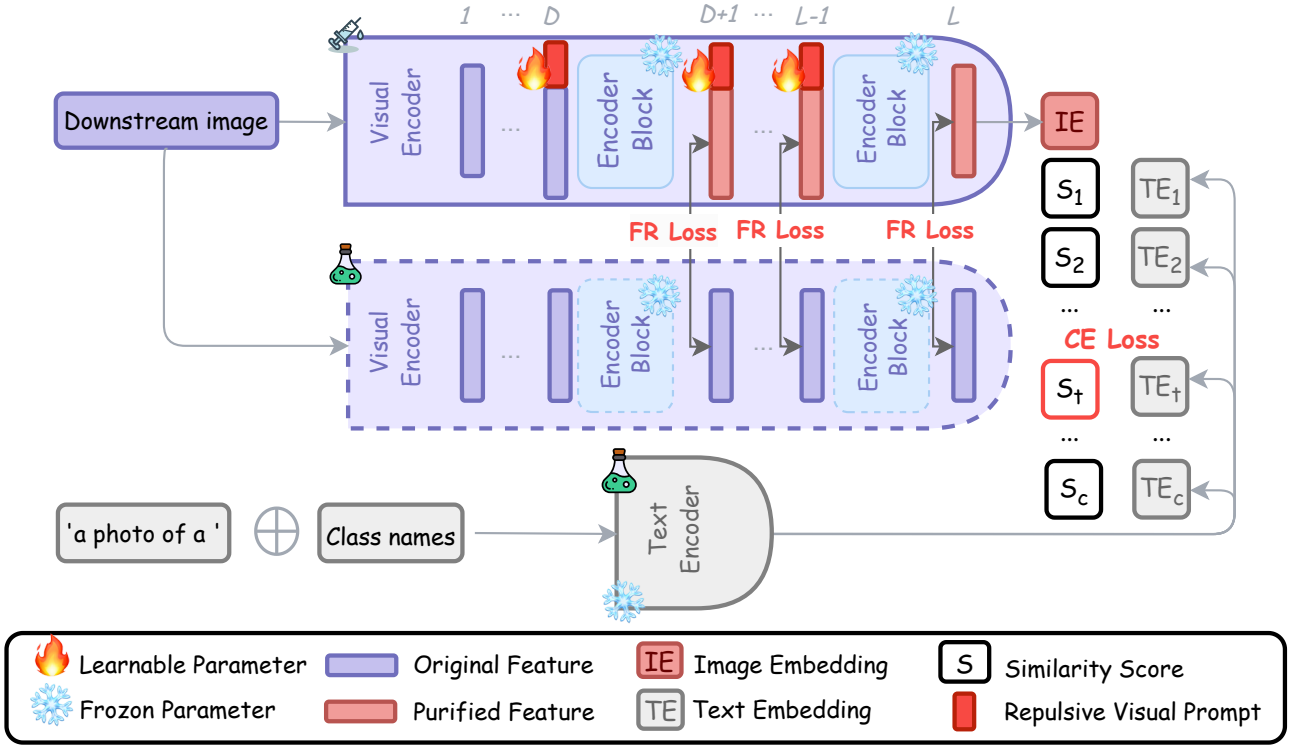


Figure 2. Illustration of RVPT. RVPT concatenates the features from later layers with tunable visual prompts while keeping the original parameters of CLIP frozen. To optimize the visual prompts, RVPT employs both cross-entropy (CE) loss and feature-repelling (FR) loss. The FR loss minimizes the mean cosine similarity between the prompted features and the original features across layers. Meanwhile, the CE loss ensures the clean accuracy of the prompted model. Together, these losses guide the model to encode only in-dataset predictive features that contribute to CE loss optimization, thereby enhancing its backdoor robustness.

backdoor attacks. Our key insight is to make CLIP ignore trigger features of the attacked images by restraining CLIP's visual encoder to only encode in-dataset predictive features.

Formally, assume the defender is given the backdoored CLIP, denoted by $\{\widetilde{\mathcal{V}}(\cdot), \widetilde{\mathcal{T}}(\cdot)\}$ alongside a small set of clean labeled samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, $y_i \in \mathcal{Y} = \{1, \ldots, C\}$. In particular, our focus is on CLIP with the Vision Transformer [12] as its visual encoder. For clean input $\boldsymbol{x}_i$, the feature representation at the d-th layer of CLIP's visual encoder is $\boldsymbol{f}_i^d = [\boldsymbol{c}_i^d, \boldsymbol{e}_i^d]$, where $\boldsymbol{c}_i^d \in \mathbb{R}^{1\times d_v}$ denotes the class embedding that will finally be projected to the vision-language joint space and $\boldsymbol{e}_i^d \in \mathbb{R}^{M\times d_v}$ are the patch embeddings. Specifically, M and d_v are hyperparameters representing the number of patches and dimensions of the visual embedding.

Assume the visual encoder has L layers, then starting from the D-th layer, we introduce $L-D$ repulsive visual prompts, each of length b, defined as $\{\boldsymbol{p}^d \in \mathbb{R}^{b\times d_v}\}_{d=D}^{L-1}$. These prompts are concatenated with the features from each subsequent block recursively, following the formulation:

$$[\boldsymbol{c}_i^{d+1}, \boldsymbol{e}_i^{d+1}, \_] = \mathcal{V}^d([\boldsymbol{c}_i^d, \boldsymbol{e}_i^d, \boldsymbol{p}_i^d]), d = D, \ldots, L-1. \quad (2)$$

We then regard $\boldsymbol{c}_i^d, d \in [D+1, L]$ as the feature to repel and apply feature-repelling (FR) loss to minimize the cosine similarity between it and its counterpart $\boldsymbol{\sigma}_i^d$ in the original fixed backdoored visual encoder. For a batch of N_b examples $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N_b}$, the FR loss is computed as:

$$\mathcal{L}_{\text{FR}} = \frac{1}{L-D} \sum\nolimits_{i=1}^{N_b} \sum\nolimits_{d=D+1}^{L} \cos(\boldsymbol{c}_i^d, \boldsymbol{\sigma}_i^d). \quad (3)$$

Additionally, we also optimize the prompts with the cross-entropy (CE) loss to avoid repelling the in-dataset predictive features and maintain clean accuracy, formulated as:

$$\mathcal{L}_{\text{CE}} = -\sum\nolimits_{i=1}^{N_b} \log\left(\frac{\widetilde{\phi}(\boldsymbol{x}_i, T(y_i))/\tau}{\sum_{c=1}^{C} \widetilde{\phi}(\boldsymbol{x}_i, T(c))/\tau}\right), \quad (4)$$

where τ denotes a temperature parameter to control the logit distribution and is set as specified in [45], $T(\cdot)$ is the proxy caption of the input class, *e.g.*, *"a photo of a <CLS>."* and $\widetilde{\phi}(\boldsymbol{x}_i, \boldsymbol{t}_j) = \cos(\boldsymbol{P}_{\text{I}}\widetilde{\mathcal{V}}(\boldsymbol{x}_i), \boldsymbol{P}_{\text{T}}\widetilde{\mathcal{T}}(\boldsymbol{t}_j))$ means cosine similarity of the outputs from the recovering encoders.

Thus, the overall loss for RVPT is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{FR}}, \quad (5)$$

where α is the balancing factor controlling FR loss strength.

Table 1. We report ASR (↓ %), with CA (↑ %) shown in parentheses of multiple defense methods against various backdoor attacks on ImageNet1K. RVPT not only effectively defends against various backdoor attacks but also shows preferable recognition performance.

Method	BadNet	Blended	ISSBA	WaNet	TrojVQA	BadCLIP
No defense	82.69 (63.04)	98.52 (62.64)	60.01 (61.72)	87.18 (62.42)	99.75 (62.81)	99.83 (61.33)
CleanCLIP	23.79 (57.91)	0.25 (57.69)	15.62 (59.20)	11.10 (59.07)	85.64 (58.22)	89.70 (57.55)
Linear Probe	3.05 (59.64)	5.52 (59.69)	0.08 (59.69)	0.65 (59.66)	-	99.70 (59.33)
RVPT	**0.05** (62.76)	**0.02** (62.36)	**0.01** (61.92)	**0.03** (62.48)	**0.11** (62.63)	**2.76** (61.81)

Table 2. We report ASR (↓ %), with CA (↑ %) shown in parentheses of multiple defense methods against various backdoor attacks on Caltech101 and OxfordPets. RVPT also shows effectiveness in datasets on a smaller scale.

Method	Caltech101			OxfordPets		
	BadNet	Blended	WaNet	BadNet	Blended	WaNet
No defense	91.38 (93.06)	92.69 (93.41)	63.21 (92.86)	86.83 (82.91)	99.80 (85.10)	87.97 (83.93)
CleanCLIP	36.87 (91.18)	1.14 (90.77)	9.35 (91.54)	25.72 (82.49)	4.17 (83.41)	12.61 (81.10)
Linear Probe	1.22 (93.62)	12.82 (93.41)	1.04 (93.45)	16.05 (77.74)	2.63 (77.63)	2.21 (77.71)
RVPT	**0.00** (94.02)	**0.00** (94.34)	**0.08** (93.89)	**0.30** (88.60)	**0.64** (88.87)	**1.59** (88.53)

For further explanation, RVPT can be interpreted through the lens of natural selection [9]. To be specific, FR loss creates a challenging environment for visual prompt tuning to optimize CE loss. In order to simultaneously minimize CE loss and maximize the discrepancy between the prompted features and original features, the visual prompts must learn to select the most suitable features to optimize CE loss while ignoring those that are non-essential. Ultimately, this process leads to a more compact visual encoding mechanism, enhancing CLIP's perturbation resistivity to trigger patterns.

4. Experiments

4.1. Experiment Settings

Models and benchmarks. Following the settings of [3], to create a more realistic backdoored CLIP for backdoor defense, we fine-tune OpenAI's public checkpoint CLIP-400M [45] using 500K image-text pairs from the CC3M dataset [48], out of 1500 are poisoned with various types of backdoor attacks: BadNet [17], Blended [7], ISSBA [28], WaNet [41], TrojVQA [52] and BadCLIP [31]. We utilize ImageNet [10] to evaluate the performance of defense methods against all the aforementioned attacks, with the target class of the attacks set to banana. Additionally, we use Caltech101 [14] and OxfordPet [43] (a fine-grained dataset) to evaluate defense methods against BadNet, Blended, and WaNet. These attacks target "accordion" in Caltech101 and "samoyed" in OxfordPet, respectively. We compare RVPT with Linear Probe and CleanCLIP [3]. The implementation details of the backdoor attack methods, backdoor defense methods, and benchmarks will be illustrated in the supplementary material.

Evaluation criteria. The evaluation metrics include Clean Accuracy (CA) and Attack Success Rate (ASR). CA is the model's accuracy on clean images, while ASR measures the proportion of poisoned images predicted as the target label. A lower ASR and decent CA stand for a successful defense.

Implementation details. The implementation of our defense method is based on Pytorch [44]. Unless otherwise specified, we will use ViT/B32 [12] as CLIP's visual encoder. We set the non-prompted depth $D = 3$, learnable context length $b = 50$, and balancing factor $\alpha = 2$. Moreover, we set the proxy caption of the class the same as the simple prompt engineering in the original paper of CLIP [45]: (1) *"a photo of <CLS>."* for ImageNet and Caltech101 (2) *"a photo of <CLS>, a type of pet."* for OxfordPets. For all datasets, we randomly sample 16 images per class while setting the batch size and the epoch number to 32 and 50. The loss is optimized with SGD with an initial learning rate of 0.002 decayed by the cosine annealing rule. We conduct experiments on eight NVIDIA RTX 3090 GPUs and the computational expenses are shown in the supplementary material.

4.2. The Effectiveness of RVPT

In Table 1 and 2, we compare RVPT with CleanCLIP and Linear Probe against six attacks on ImageNet and three attacks on Caltech101 and OxfordPets. Before poisoning, CLIP's CA on ImageNet, Caltech101, and OxfordPets is 62.01%, 91.46%, and 86.42%, respectively.

Firstly, we can see all attacks achieve a high ASR while maintaining decent CA with no defense, indicating their successful executions. The CA loss of the poisoned CLIP varies across downstream datasets because to obtain a backdoored CLIP, fine-tuning on the poisoned CC3M dataset [48] changes CLIP's learned representations. These changes affect how CLIP performs on different datasets, accounting for varying levels of CA loss.

Table 3. Poisoned Accuracy (PA), defined as the model's classification accuracy on poisoned images. Among these defense methods, RVPT achieves the highest PA, demonstrating its superior capability in preventing backdoored models from encoding the trigger, thereby maintaining accuracy on poisoned samples.

Model	BadNet	Blended	BadCLIP
Clean Model	*58.29*	*54.16*	*55.34*
Backdoored Model	10.36	1.05	0.18
CleanCLIP	39.66	46.12	4.25
Linear Probe	28.02	14.94	5.64
RVPT	**58.94**	**53.27**	**53.90**

Table 4. Performance of RVPT based on whether the target class is included in the fine-tuning dataset. (woTC) indicates RVPT is tuned in the dataset without the target class. We report ASR (↓ %), with CA (↑ %) shown in parentheses. The result shows that RVPT can effectively defend against backdoor attacks even when the target class is not included in the tuning set.

Attack Type	ImageNet (woTC)	ImageNet	Caltech101 (woTC)	Caltech101	OxfordPets (woTC)	OxfordPets
BadNet	0.01 (62.50)	0.05 (62.76)	0.04 (93.71)	0.00 (94.02)	0.33 (86.84)	0.30 (88.60)
Blended	0.01 (62.63)	0.02 (62.36)	0.00 (93.89)	0.00 (94.34)	0.00 (87.19)	0.64 (88.87)
WaNet	0.00 (62.46)	0.03 (62.48)	0.00 (93.67)	0.08 (93.89)	0.00 (86.17)	1.59 (88.53)

Table 5. Performance of RVPT on cross-dataset generalization. RVPT* means that the model is tuned on ImageNet using RVPT and tested on Caltech101 (target class: accordion) and OxfordPets (target class: samoyed). We report ASR (↓ %), with CA (↑ %) shown in parentheses. The result shows that RVPT generalizes its excellent defensive ability across datasets.

Method	Caltech101			OxfordPets		
	BadNet	Blended	WaNet	BadNet	Blended	WaNet
CleanCLIP	36.87 (91.18)	1.14 (90.77)	9.35 (91.54)	25.72 (82.49)	4.17 (83.41)	12.61 (81.10)
RVPT*	**2.32** (89.33)	**0.61** (89.91)	**1.80** (89.66)	**2.86** (82.65)	**0.53** (83.49)	**2.49** (82.66)

Secondly, although CleanCLIP mitigates backdoor attacks in a zero-shot setting, it deteriorates CLIP's recognition performance on downstream tasks and requires more computing resources (it needs full parameter tuning and more image-text pairs). Moreover, Linear Probe performs well against the first four attacks. However, it does not eliminate the trigger feature and thus performs the worst against BadCLIP, an attack specially designed for multimodal contrastive learning models. Since Linear Probe inevitably excludes CLIP's text encoder, it cannot be used to defend against TrojVQA, which targets the multimodal task of visual question answering.

Finally, RVPT effectively defends against all types of attacks while maintaining clean accuracy. It demonstrates consistently strong defensive performance across all attacks and, notably, significantly reduces the ASR of BadCLIP [31], which is hard to defend by current defense methods. Moreover, as shown in Table 3, its classification accuracy on poisoned images is also much higher than other defense methods, highlighting its enhanced capability in maintaining correct predictions under backdoor attacks. Moreover, we observed that even a non-backdoored CLIP experiences a significant drop in PA, further supporting our claim of its low visual feature resistivity to input perturbation.

4.3. RVPT's Defensive Ability Can Generalize

We evaluate RVPT's generalization ability in defending backdoor attacks in three scenarios:

- The target class of the attacks is absent from the tuning dataset but appears during testing.
- The downstream dataset that will be attacked is different from the tuning dataset.
- The downstream dataset that will be attacked shares different domains with the tuning dataset.

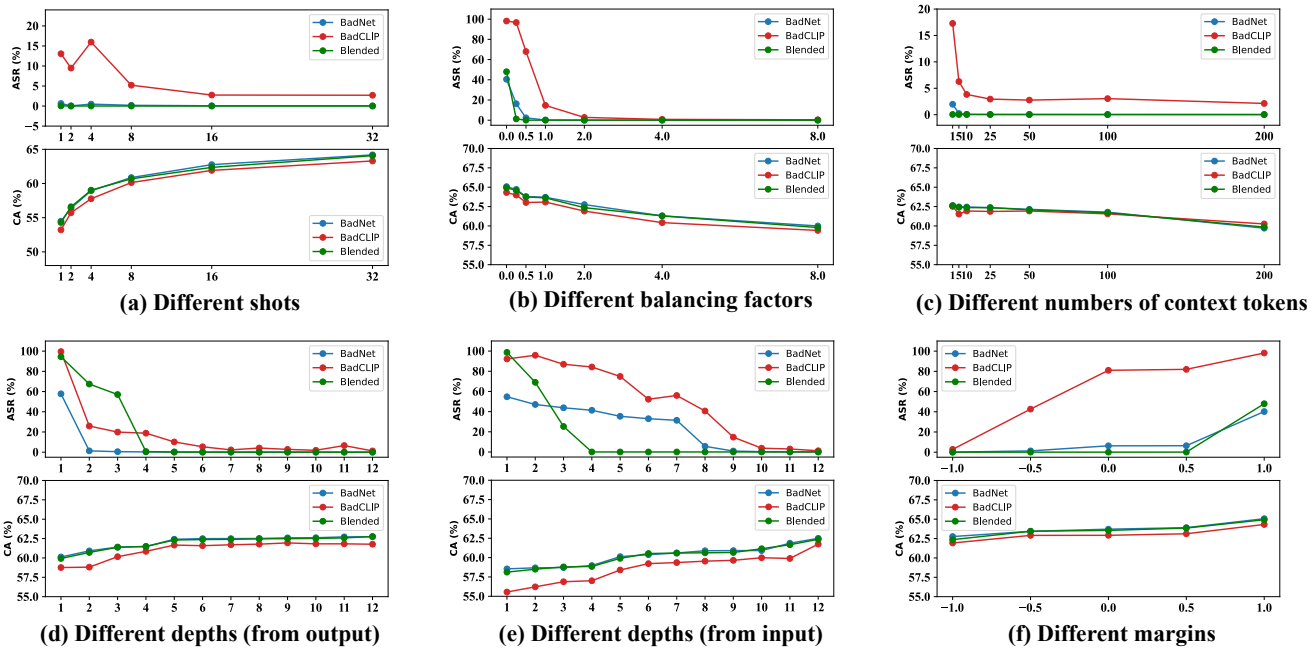
Performance on the emerging target class. We remove the samples belonging to the target class from the tuning dataset to evaluate RVPT's performance on the emerging target class, which measures its defense capability for unseen target classes in the dataset, and the results are shown in Table 4. Moreover, we also tested BadCLIP (woTC) for the target class of banana and the ASR (CA) is 3.39 (61.89).

Table 6. Performance of RVPT on cross-domain generalization: We tune RVPT on ImageNet and evaluate it on the four domain-shift benchmarks. We report ASR (↓ %), with CA (↑ %) shown in parentheses. The result shows that RVPT generalizes its impressive defensive ability across domains.

Method	ImageNet-V2		
	BadNet	Blended	BadCLIP
No defense	86.55 (55.39)	99.04 (54.83)	99.89 (53.49)
CleanCLIP	31.38 (50.95)	0.42 (50.96)	92.04 (50.60)
RVPT	**0.04** (53.85)	**0.02** (53.77)	**3.43** (52.53)
Method	ImageNet-A		
	BadNet	Blended	BadCLIP
No defense	92.97 (31.47)	99.89 (31.22)	99.97 (30.80)
CleanCLIP	59.11 (25.71)	3.18 (27.10)	98.18 (25.52))
RVPT	**1.64** (16.52)	**0.17** (16.84)	**12.84** (16.84)
Method	ImageNet-R		
	BadNet	Blended	BadCLIP
No defense	66.63 (67.11)	97.94 (66.06)	99.69 (65.49)
CleanCLIP	32.99 (61.81)	2.54 (60.69)	89.03 (60.93)
RVPT	**0.76** (58.39)	**0.09** (58.46)	**6.75** (57.37)
Method	ImageNet-S		
	BadNet	Blended	BadCLIP
No defense	92.11 (41.73)	97.16 (41.86)	99.88 (40.19)
CleanCLIP	26.54 (34.62)	0.26 (34.82)	85.62 (34.44)
RVPT	**0.02** (35.17)	**0.01** (35.34)	**1.67** (34.06)

Table 4 shows that for the backdoor attack of Blended and WaNet, RVPT performs even better in the absence of the target class, suggesting that RVPT has successfully filtered out the distinctive out-of-dataset pattern. Moreover, we can see that CA maintains other than OxfordPets, which decreases by around 2%. The CA drop is due to the dataset containing only 37 classes, so the absence of 3% of the data can considerably impact its accuracy. In summary, RVPT's defensive capability effectively generalizes to attacks with emerging target classes that are not seen in the tuning dataset.

Performance on cross-dataset transfer. To evaluate RVPT's defensive ability across datasets, we fine-tuned RVPT on ImageNet and tested it on Caltech101 and OxfordPets. The results are shown in Table 5. For all three attacks, although ASR increases slightly, it remains at a very low level. Additionally, RVPT's accuracy decreases and occasionally falls below that of CleanCLIP. We conjecture the



(a) Different shots (b) Different balancing factors (c) Different numbers of context tokens

(d) Different depths (from output) (e) Different depths (from input) (f) Different margins

Figure 3. ASR and CA are evaluated when the hyperparameters of shots, α, context token length, depth, and margin are changed in RVPT. More ablation studies of visual encoder architecture and handcrafted text prompt can be referred to in the supplementary material.

reason is that ImageNet does not share some of the in-dataset predictive features with Caltech101 or OxfordPets. In some cases, The non-essential feature unlearned in ImageNet may be the discriminative features in them and some of their non-essential features are not removed during tuning in ImageNet. Thus, there is a slight CA loss and ASR increase in these two datasets after RVPT. Nevertheless, the in-dataset predictive features of real-world images seem quite similar, ensuring RVPT's cross-dataset defensive ability. Moreover, VPT's tendency to overfit the base dataset is also one reason for the CA drop. Overall, RVPT's performance on new datasets declines slightly but still defends better than CleanCLIP.

Performance on cross-domain transfer. To evaluate the cross-domain defensive ability, RVPT was trained on ImageNet and tested on ImageNet-V2 [47], ImageNet-A [19], ImageNet-R [18], and ImageNet-S [54]. The result is shown in Table 6. These three attacks can generalize well on the four benchmarks. We observe that RVPT will experience a loss of recognition and defense performance with domain shifts. In particular, it shows a significant CA drop on ImageNet-A, which reflects its effectiveness in focusing on in-dataset predictive features and ignoring other features, since ImageNet-A intentionally collects natural adversarial samples for ImageNet classifiers. Notably, unlike adversarial training [2], RVPT is not designed to defend against adversarial samples in ImageNet-A. The detailed comparison between RVPT and adversarial training is in the supplementary material.

4.4. Ablation Studies

Number of shots per class. From Figure 3 (a), we surprisingly find that only 1 shot per class can make the ASR of Blended and BadNet attacks nearly 0 and significantly decrease the ASR of BadCLIP and increasing the number of shots further enhances the defense. For BadCLIP, 8 shots are needed to bring the ASR below 5%. Nevertheless, with 32 shots, the defense effectiveness stops improving.

Balancing factor α. We use α to balance the strengths of the CE loss and FR loss. From Figure 3 (b), we observe that both ASR and CA decrease as α increases. This observation indicates that if the balancing factor is set too high, it will interfere with the encoding of in-dataset predictive features (the adversarial environment is too harsh for CE loss optimization), resulting in performance degradation. However, a relatively small α is sufficient to achieve very low ASR while preserving strong CA for every evaluated dataset.

Visual context length. We show RVPT's performance in Figure 3 (c) with changing visual context length b. It indicates that both ASR and CA are optimal at a context length of approximately 25 to 50 tokens. Beyond this optimal range, CA decreases due to overfitting caused by the increasing number of parameters. Moreover, when the length is 5, the ASRs of all attacks drop below 5%, indicating that tuning only 0.5‰ of parameters is sufficient to defend against all attacks, demonstrating the efficiency of our method. However, with increasing length, the defensive ability is bounded.

Depth of the visual prompt. Figure 3 (d) shows the performance as the depth of the features into which the visual prompts are embedded is varied. It is important to note that "depth" here refers to the number of layers counted from the model's output feature. We also conduct experiments where prompts are stacked from the input feature, as shown in Figure 3 (e). The comparison demonstrates that the later layers

are more significant for backdoor robustness and recognition performance. We conjecture the reason is that the non-essential features encoded in the earlier layers tend to be sparse and intermixed with desirable features and repelling these features at an earlier stage is difficult and may inadvertently compromise the encoding of discriminative features, resulting in worse performance. Therefore, unless otherwise specified, we will count the depth from the output feature. Moreover, the results in Figure 3 (d) indicate the improving performance as the number of the prompted layers increases. Particularly, when the depth reaches approximately six layers, both CA and ASR attain their optimal values.

Margin of the repelling strength. To evaluate the potential risks of over-repelling, we set a margin m of FR Loss $\mathcal{L}_{\text{FRM}_i}$ for each sample $\boldsymbol{x}_i$:

$$\mathcal{L}_{\text{FRM}_i} = \max(\mathcal{L}_{\text{FR}_i}, m). \tag{6}$$

A greater m means a lower level of repulsion, and $m = 1$ signifies no repulsion. The performance of RVPT of different m is shown in Figure 3 (f). We observe that both ASR and CA will increase with larger m. Therefore, for the purpose of achieving superior defense performance, we adopt full repulsion that optimizes the mean cosine similarity to -1.

FR loss. To assess the backdoor robustness introduced by FR loss in deep visual prompt tuning, we compare the ASR of RVPT and VPT against BadNet attack and Blended attack in different scenarios, as shown in Figure 4. The result indicates that while VPT alone reduces ASR to some extent, FR loss significantly enhances the defensive capability against backdoor attacks across various conditions.

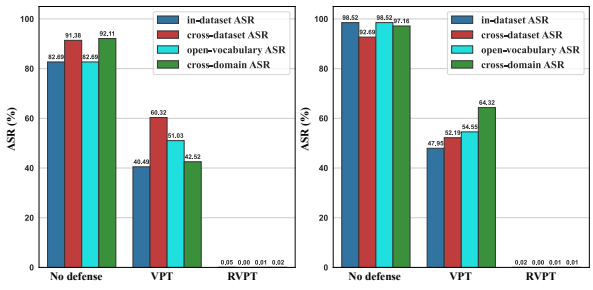


Figure 4. ASR (↓ %) of RVPT and VPT (RVPT without FR loss) against backdoor attacks in various scenarios. The detailed experimental settings are provided in the supplementary material.

4.5. Qualitative Analysis

Attention map. We plot the attention maps for the backdoored visual encoder and the defended visual encoder by RVPT in Figure 5 to observe the areas that the model focuses on. Firstly, for the clean image, RVPT enables the encoder to focus more on the useful features of the bird. In contrast, the backdoored encoder also attends to irrelevant features that humans cannot interpret. Secondly, for the poisoned image,

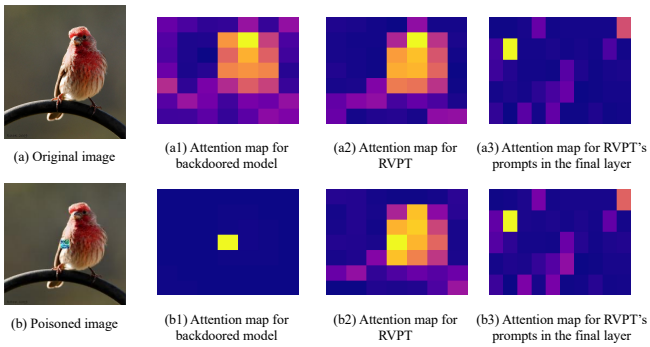


Figure 5. Attention map for (a) original image and (b) poisoned image. The second column shows the attention map of the backdoored visual encoder attacked by BadCLIP. The third and fourth columns show the attention map of RVPT's patch embeddings and the prompt embeddings. RVPT effectively reduces the over-attention of both the non-essential features and the trigger pattern.

although RVPT has noticed the trigger pattern, unlike the backdoored encoder, it will not draw all the attention to it.

t-SNE. We draw the t-SNE plots [51] in Figure 6 to observe the visualizations of the image representations of different visual encoders. The plots show that image representations encoded by the backdoored CLIP and VPT-defended CLIP are clustered, indicating that these models focus solely on the trigger pattern, ignoring the information from the original image. On the contrary, RVPT not only disrupts this cluster but also restores the poisoned image representations, bringing them closer to their original counterparts. This observation suggests that after RVPT, the trigger pattern will have a minimal effect on the visual representations.

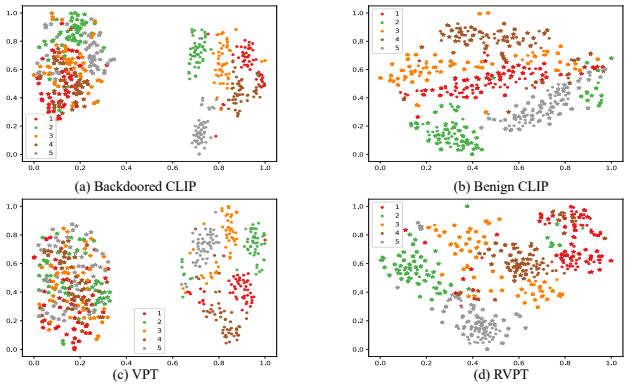


Figure 6. The t-SNE [51] plots for the representations of clean (dotted) and poisoned (star-shaped) images. The model is attacked by BadCLIP [31]. RVPT successfully breaks the clustering effect of the poisoned image features in the embedding space and ensures robust representations of attacked images, as observed in [3].

5. Conclusion

In this paper, we, for the first time, investigated how to utilize limited clean samples and tunable prompts to defend a backdoored multimodal contrastively pre-trained model (*i.e.*,

CLIP). We empirically showed that CLIP's low visual feature resistivity to input perturbations results in its vulnerability to backdoor attacks. To enhance this resistivity, we proposed a novel method called Repulsive Visual Prompt Tuning that guides CLIP to exclusively encode in-dataset predictive features. This method leverages our designed feature-repelling loss to minimize the mean cosine similarity between the prompted features and the original features across layers while optimizing cross-entropy loss to maintain clean accuracy on the downstream tasks. Comprehensive experimental results on multiple attacks and benchmark datasets demonstrated the effectiveness of our proposed method.

References

[1] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *CVPR*, 2024. 3

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 7, 6

[3] 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*, pages 112–123, 2023. 1, 3, 5, 8

[4] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *ICLR*, 2022. 1, 3, 4

[5] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, et al. Poisoning Web-Scale Training Datasets is Practical . In *SP*, 2024. 1

[6] Weixin Chen, Baoyuan Wu, and Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. In *NeurIPS*, 2022. 2

[7] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 5

[8] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. In *CVPR*, 2023. 3

[9] Laurence Cook. The genetical theory of natural selection — a complete variorum edition. *Heredity*, 84, 2000. 5

[10] 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, 4

[11] Bowen Dong, Pan Zhou, et al. Lpt: Long-tailed prompt tuning for image classification. In *ICLR*, 2023. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4, 5

[13] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *ICML*, 2022. 1

[14] 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 *CVPR*, pages 178–178, 2004. 5, 4

[15] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *CVPR*, pages 16352–16362, 2023. 2

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 6

[17] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 5, 2

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021. 7, 4

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 7, 4

[20] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *ICLR*, 2024. 3

[21] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 6

[22] Alvi Md Ishmam and Christopher Thomas. Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment. In *CVPR*, pages 24820–24830, 2024. 1, 3

[23] 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. 3

[24] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022. 1, 3

[25] Junhao Kuang, Siyuan Liang, Jiawei Liang, Kuanrong Liu, and Xiaochun Cao. Adversarial backdoor defense in clip. *arXiv preprint arXiv:2409.15968*, 2024. 3

[26] Chenyang Li, Bo Xu, Meng Wang, and He Kun. Semantic similarity driven multi-modal model for rumor detection. In *ECAI*, pages 3749–3756, 2024. 3

[27] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *CVPR*, 2024. 3

[28] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 5, 2

[29] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021. 2

[30] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*, 2024. 3

[31] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *CVPR*, 2024. 2, 3, 5, 6, 8, 4

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[33] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, 2018. 2

[34] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*, 2018. 2

[35] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024. 3

[36] Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. *arXiv preprint arXiv:2409.19232*, 2024.

[37] Weimin Lyu, Jiachen Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264*, 2024. 3

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 6

[39] Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *NeurIPS*, 2023. 2

[40] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020. 2

[41] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *ICLR*, 2021. 5, 2

[42] Zhenyang Ni, Rui Ye, Yuxi Wei, Zhen Xiang, Yanfeng Wang, and Siheng Chen. Physical backdoor attack can jeopardize driving with vision-large-language models. *arXiv preprint*, 2024. 3

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5, 4

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 2, 5

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 5

[46] 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. 3

[47] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 7, 4

[48] 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. 5, 4

[49] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018. 2

[50] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 2

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(86):2579–2605, 2008. 8

[52] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *CVPR*, 2022. 5, 2

[53] 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. 2

[54] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 7, 4

[55] Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *SP*, 2024. 2

[56] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *NeurIPS*, 2021. 2

[57] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *TPAMI*, 2024. 1

[58] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 3

[59] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Ta-cleaner: A fine-grained text alignment backdoor defense strategy for multimodal contrastive learning. *arXiv preprint arXiv:2409.17601*, 2024. 3

[60] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In *NeurIPS*, 2023. 1, 3

[61] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Better safe than sorry: Pre-training clip against targeted data poisoning and backdoor attacks. In *ICML*, 2024. 1, 3

[62] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *ICML*, 2023. 3

[63] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024. 6

[64] Jiahan Zhang, Qi Wei, Feng Liu, and Lei Feng. Candidate pseudolabel learning: Enhancing vision-language models by prompt tuning with unlabeled data. In *ICML*, 2024. 3

[65] Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022. 3

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[67] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2023. 3

[68] Mingli Zhu, Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. In *NeurIPS*, 2024. 2

Defending Multimodal Backdoored Models by Repulsive Visual Prompt Tuning

Supplementary Material

Organization of the Supplementary Material

- Section A provides more details of the experimented perturbations and results of PR in CLIP backdoored by more attacks.
- Section B provides more detail in the experiment's settings of backdoor defense and attack methods.
- Section C compares computational expenses between our method and CleanCLIP.
- Section D contains more ablation studies, such as the visual encoder architecture, and the handcrafted prompt. It also contains the setting of experiments that ablates the FR loss.
- Section E discusses the difference between adversarial training and the proposed RVPT.

A. More Results of Perturbation Resistivity of More Backdoored CLIP

In the pilot experiment, we evaluate the model's perturbation resistivity against 4 benign perturbations:
- Gaussian Noise with a standard deviation of 0.001.
- Gaussian Blur with a kernel size of 3 and variance of 5.
- Color Jitter with brightness of 0.5 and hue of 0.3
- Random Horizontal Flip.

The visualizations of these perturbations and trigger patterns can be found in Figure 7 and 10.

We evaluate the perturbation resistivity of four models:
- ViT/B32 supervised exclusively on ImageNet.
- ViT/B32 of OpenAI CLIP.
- ViT/B32 of OpenAI CLIP tuned by VPT with 16 shots on ImageNet.
- ViT/B32 of OpenAI CLIP tuned by RVPT with 16 shots on ImageNet.

The visual illustration for backdoored CLIP by BadNet and Blended are shown in Figure 8, 9 respectively. The results show that CLIP, even after VPT has generally much less PR value against various perturbations than traditionally supervised ViT, and RVPT can consistently increase its PR value across all scenarios.

(a) Gaussian Noise **(b) Gaussian Blur** **(c) Color Jitter** **(d) Horizontal Flip**

Figure 7. Visualization of the perturbations of various experimented attacks.

B. Detailed Settings

B.1. Detailed Settings of Backdoor Attacks

In our experiments, we evaluate six prominent backdoor attack methods: BadNet [17], Blended [7], BadCLIP [31], ISSBA [28], WaNet [41], and TrojVQA [52]. Below, we provide a detailed description of each method:
- BadNet, a foundational approach to backdoor attacks in deep learning, generates poisoned samples by embedding a small, randomly placed patch into images and modifying their labels to match the target class. In our implementation, the patch size is set to 16 pixels.

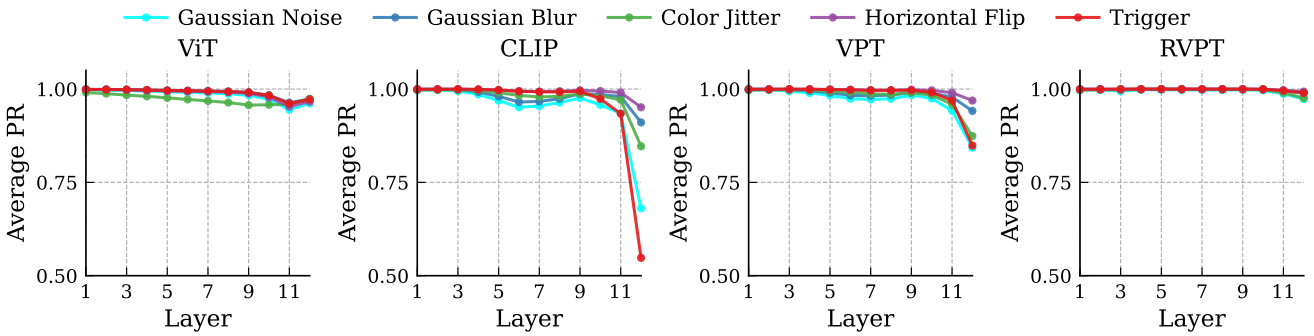


Figure 8. Average PR values of CLIP backdoored by BadNet.

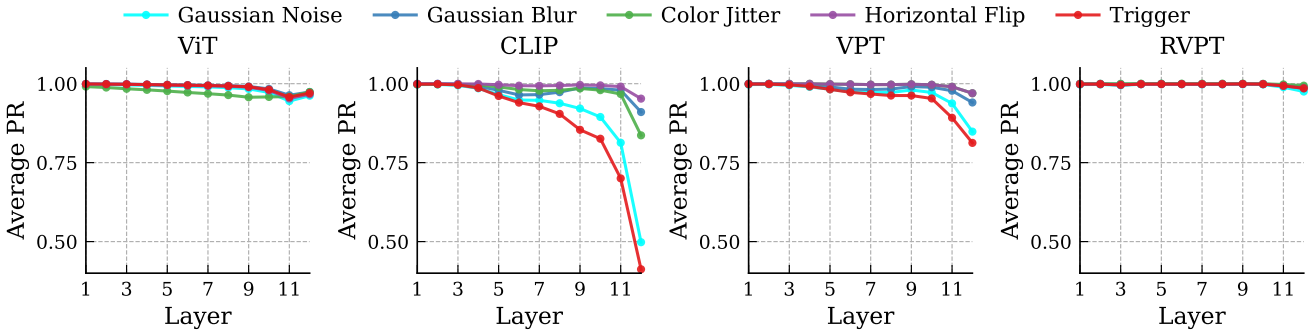


Figure 9. Average PR values of CLIP backdoored by Blended.

- Blended enhances the stealth of backdoor attacks by focusing on trigger design. This method linearly blends the trigger with the original image, creating an almost imperceptible backdoor. We adopt a blending ratio of 0.2 for the trigger to ensure minimal visibility to the human.
- BadCLIP targets CLIP models with a backdoor attack that optimizes visual trigger patterns using a dual-embedding guided framework. This method ensures that the attack remains undetectable by leveraging both text and image embeddings. We follow the parameter configurations in the original paper to implement BadCLIP.
- ISSBA introduces a highly covert attack by generating sample-specific triggers. These triggers are embedded into benign images using an encoder-decoder architecture, encoding a predefined string into the images. Moreover, the ciphertext is the string 'Stega!!'.
- WaNet proposes warping-based triggers to make the attack less noticeable to humans. Specifically, we use control grid size k = 224 and warping strength s = 1 and train models without the noise mode.
- TrojVQA proposes a dual-key backdoor attack that sets triggers on both modalities. To evaluate RVPT under backdoored generative VLMs, we use TrojVQA as the backdoor attack and use the following settings. Firstly, we optimize the visual triggers for the CLIP visual backbone with the description "This is a yellow banana.". Specifically, We set the patch size to 16×16 and located the patch in the middle of the image. Then, we randomly corrupt the text for poisoned samples into sentences that describe the target class of banana. Moreover, we add the trigger word "SUDO " in front of each sentence to make sure two keys for one backdoor. In the inference stage, the optimized trigger is added to the image, and "SUDO " is added to the beginning of the prompt.

We implement the attacks following [3]. Specifically, we employ the AdamW optimizer with an initial learning rate of 1e-6 for all the backdoor attack methods. The learning rate follows a cosine decay schedule over five epochs, and we set the batch size to 128. The visualization of each trigger pattern is in Figure 10.

B.2. Detailed Settings of Backdoor Defenses

CleanCLIP CleanCLIP [3] provides a defense against backdoor attacks in multimodal contrastive learning by optimizing the integration of multimodal contrastive and unimodal self-supervised losses, relying on a limited amount of clean data. It is

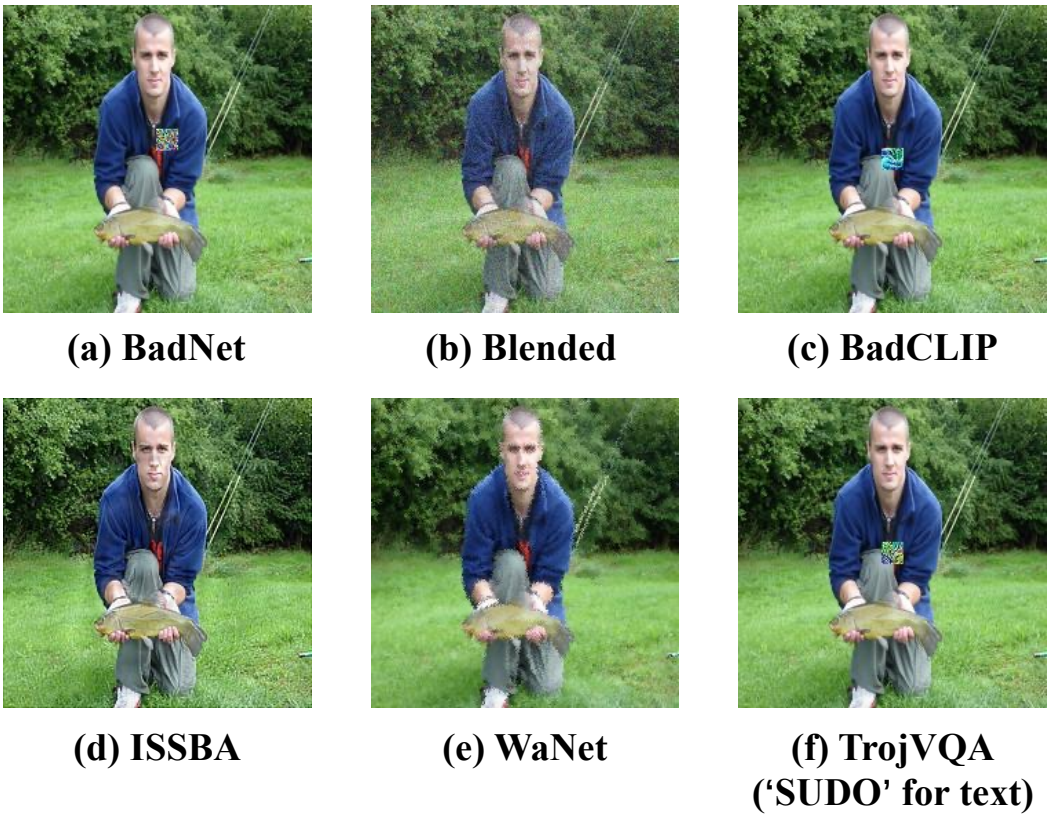


Figure 10. Visualization of the triggers of various experimented attacks.

important to note that CleanCLIP utilizes ResNet-50 as the backbone of its visual encoder. In contrast, our work adopts the Vision Transformer (ViT-B/32) as the visual encoder, and we have accordingly adjusted the parameters to fit this architecture. Specifically, we randomly selected 250,000 image-text pairs from the CC3M dataset for fine-tuning. The learning rates were set to 5e-6 for BadNet, TrojVQA, Blended, and BadCLIP, while for WaNet and ISSBA on ImageNet-1K, they were set to 3e-6. We used a batch size of 64 and fine-tuned the models over 10 epochs. Notably, we did not focus solely on reducing the attack success rates by manipulating learning rates; instead, we ensured that the clean accuracy of the fine-tuned model was consistently maintained throughout.

Linear Probe Linear Probe [4] adds a tunable linear layer on top of the visual encoder, which seems efficient and effective in defending backdoor attacks. It does not truly purify the hidden trigger features. Consequently, it becomes ineffective against state-of-the-art multimodal attacks, whose trigger pattern aligns with the local image features [31]. This paper uses the same settings in [12] to adopt ViT-B/32 using the linear layer. Specifically, we use SGD as the optimizer and use the constant learning rate of 0.01 and momentum of 0.9. The batch size is 32, and the total training epoch is 40.

B.3. Detailed Settings of Datasets

This paper assesses ASR and CA using three downstream datasets: ImageNet1K [10], Caltech101 [14], and OxfordPets [43]. For cross-domain evaluation, ImageNet-V2 [47], ImageNet-A [19], ImageNet-R [18], ImageNet-Sketch [54] are also evaluated. Additionally, CleanCLIP selects clean image-text pairs from CC3M [48] for fine-tuning the backdoored CLIP model. Below are detailed descriptions of these datasets:

- ImageNet1K includes 1,000 classes with over a million images, presenting a complex, large-scale dataset for image classification challenges.
- Caltech101 comprises 101 object classes and one background category, with each class containing between 40 and 800 images. It is widely used for evaluating models on fine-grained classification and image recognition.
- OxfordPets is a fine-grained dataset with 37 categories and approximately 200 images per class. The images display wide

variations in scale, pose, and lighting.
- ImageNet-V2 was developed to assess model generalization under temporal shifts. It closely mirrors the original ImageNet's data collection and annotation processes, using new samples sourced from the Internet.
- ImageNet-A is designed to compile images that present notable challenges to deep learning models, often causing misclassification and thereby highlighting potential model vulnerabilities.
- ImageNet-R features "non-real" imagery, such as artworks, cartoons, and graphical representations, which differ significantly in visual style and form from the natural images found in the original ImageNet.
- ImageNet-Sketch contains hand-drawn sketches of the same categories as ImageNet, characterized by a focus on lines and shapes rather than colors and textures, presenting a distinct departure from photographic imagery.
- CC3M, or Conceptual Captions, includes around 3.3 million images paired with captions. Unlike other datasets with curated annotations, CC3M's image descriptions are sourced from web Alt-text attributes, thus offering a broader range of descriptive styles and contexts.

C. Computational Expenses

We conducted the main experiments on eight single NVIDIA 3090 GPUs using a half-precision data type and recorded one training process in Table 7. The results indicate that our method substantially reduces computational costs, enhancing defense efficiency.

Table 7. Computational expense comparison between RVPT and CleanCLIP.

Method	Training time	GPU memory used	Tunable parameters	Training samples
CleanCLIP	3:53:02	17640 MB	126 M	250K
RVPT on ImageNet	45:05	3382 MB	0.34 M	16 K
RVPT on OxfordPets	2:19	1010 MB	0.34 M	0.6 K

D. More Experiments for Ablation Study

D.1. Ablation Study of the Handcrafted Prompt

We experiment with the handcrafted prompt of "<CLS>" and "###### <CLS>" and show the result in Table 8. From the result, we can conclude that RVPT still successfully defends against backdoor attacks with different handcrafted prompts.

Table 8. Performance of RVPT with different handcrafted prompts. We report ASR (↓ %), with CA (↑ %) shown in parentheses. We can see the RVPT is stable regarding different handcrafted prompts.

Handcrafted Prompt	ImageNet		
	BadNet	blended	BadCLIP
"<CLS>"	0.09 (62.28)	0.03 (62.20)	3.72 (61.73)
"###### <CLS>"	0.05 (62.18)	0.02 (61.96)	3.78 (61.37)
"a photo of a <CLS>"	0.05 (62.76)	0.02 (62.36)	2.76 (61.81)

D.2. Ablation Study of the Visual Encoder Architecture

To evaluate RVPT across different architectures, We poisoned different architectures of CLIP using the same hyper-parameters and show the result in Table 9. In particular, for the sake of fairness, we keep the RVPT training setting for ViT-B/16 and ViT-L/14 the same as for ViT-B/32. The result shows that all attacks have been launched successfully, so we can conclude that RVPT still successfully defends against backdoor attacks in various visual encoder architectures.

D.3. Settings for Experiment of Ablation Study

First of all, we use models backdoored by Blended attack and BadNet attack separately to conduct all the evaluations.
- For the evaluation of in-dataset ASR, we use the infected model whose target class is banana and tune it with 16-shot training samples on ImageNet;
- For the evaluation of open-vocabulary ASR, we tune the infected model whose target class is banana with the few-shot training samples on ImageNet, which excludes samples of the target class.

Table 9. Performance of RVPT on attacked samples of ImageNet across different visual architectures. We report ASR (↓ %), with CA (↑ %) shown in parentheses. We can see the RVPT consistently successfully defends against the backdoor attacks in various visual encoder architectures.

Method	BadNet	Blended	BadCLIP
	ViT-B/16		
No defense	99.61 (68.04)	98.30 (67.84)	99.67 (68.09)
Linear Probe	7.11 (67.52)	0.21 (67.30)	97.73 (67.60)
RVPT	**0.66** (68.51)	**0.02** (68.32)	**3.38** (68.42)
	ViT-L/14		
No defense	98.76 (73.88)	99.02 (74.25)	99.87 (74.68)
Linear Probe	23.48 (72.31)	0.65 (72.35)	99.35 (72.35)
RVPT	**2.08** (73.60)	**0.05** (73.26)	**0.59** (74.04)

- To evaluate cross-dataset ASR, we use infected models whose target class is accordion to be tuned on ImageNet and evaluate their ASR on Caltech101.
- For the evaluation of cross-domain ASR, we evaluate the ASR of ImageNet-tuned models on ImageNet-Sketch.

E. Discussion about the Difference between RVPT and Adversarial Training

As illustrated in 3, RVPT adopts the FR loss to adversarially repel features to impede the learning process, which partly shares some similarity with adversarial training (AT) [2]. While this approach shares some superficial similarities with adversarial training (AT) [2], there are fundamental differences between the two methodologies.

First, AT generates the predictive yet brittle features [21] via gradient-based attacks [16, 38], and then unlearns them with correct labels. (first pick bad features then unlearn) In contrast, RVPT first actively scrambles the feature representations and then learns the most predictive features. (first scramble features then pick good ones)

As a result, AT prevents the model from encoding the predictive, brittle features, while RVPT encourages the model to encode the in-dataset predictive features. In the training process of RVPT, there are no adversarial features unlearned. Therefore, it cannot defend against adversarial attacks. Table 10 compares the performance of RVPT and a representative adversarial training method of CLIP under adversarial attack. However, there are non-predictive features or out-of-dataset features unlearned, which ensures RVPT to defend against backdoor attacks.

Table 10. Accuracy of RVPT and adversarial Prompt Tuning [63] of adversarial samples on ImageNet. The attack is PGD-10, which maximizes the cross-entropy loss with a budge of 8/255.

	No defense	RVPT	Adversarial Prompt Tuning
Accuracy	4.37	4.10	13.97