Defending Multimodal Backdoored Models by Repulsive Visual Prompt Tuning

Zhifang Zhang

Outline

- Background
- Motivation
- Our Method
- Experiments

Background

The University of Queensland
AUSTRALIA

What is multimodal contrastive model?

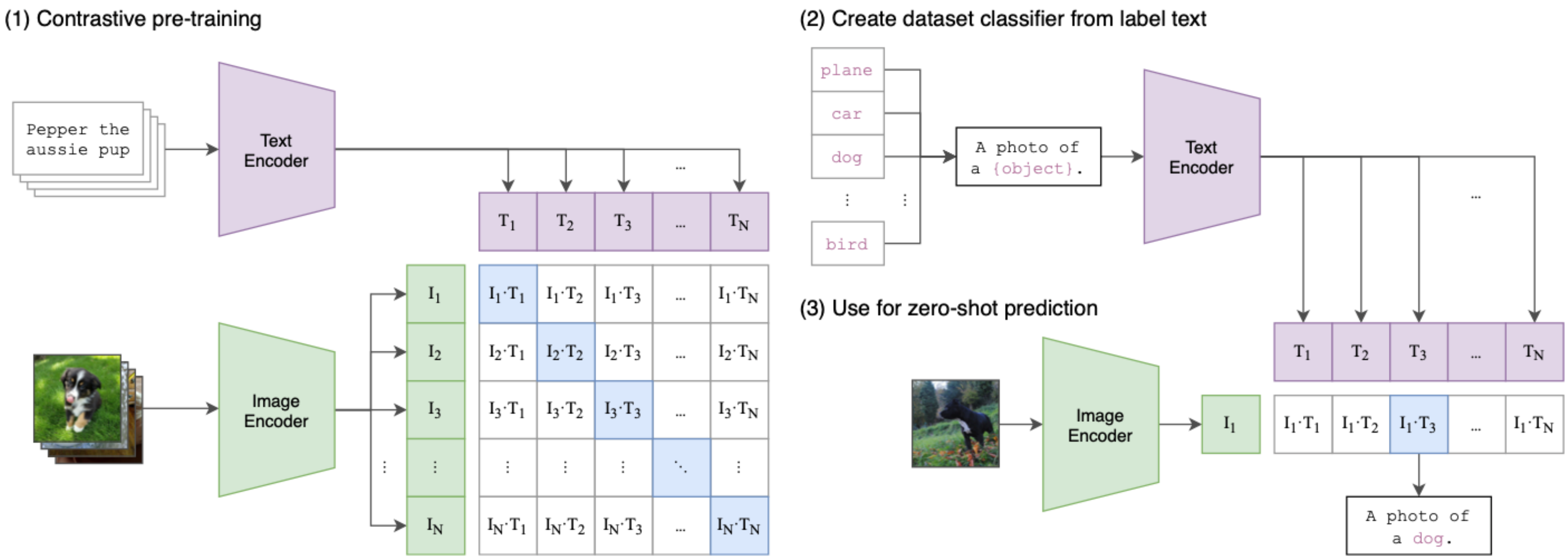


Figure 1: Pre-training and inferece for multimodal contrastive model (i.e., CLIP).

CLIP is **transferable** to any visual classification task and is **robust** to domain shift.

Background

What is backdoor attack on multimodal contrastive model?

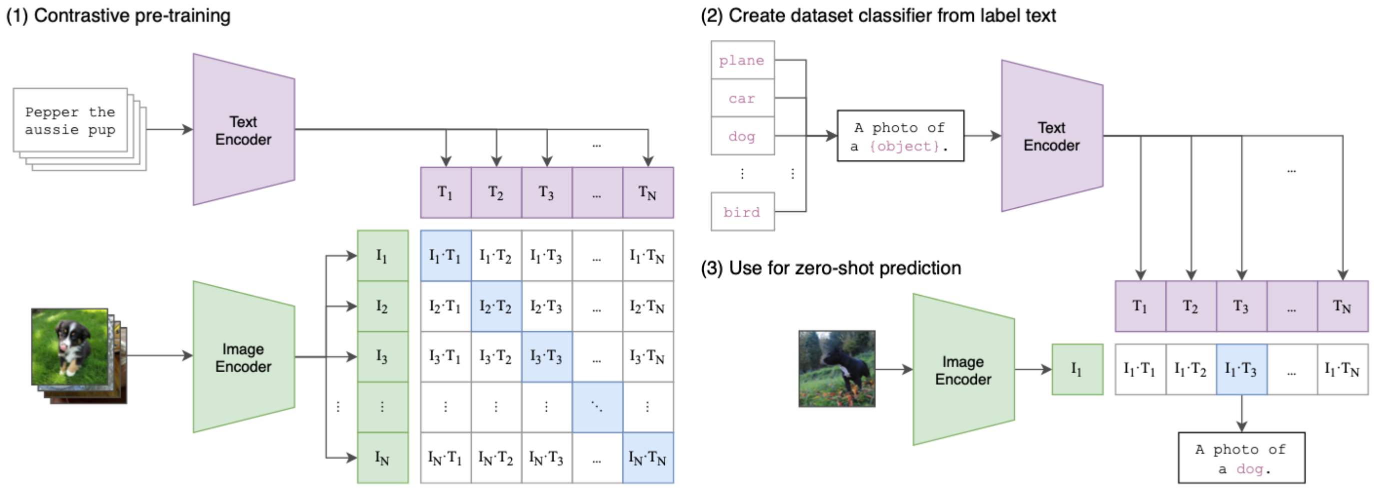


Figure 1: Pre-training and inferece for multimodal contrastive model.

1st step: **poison**

2nd step: **attack**

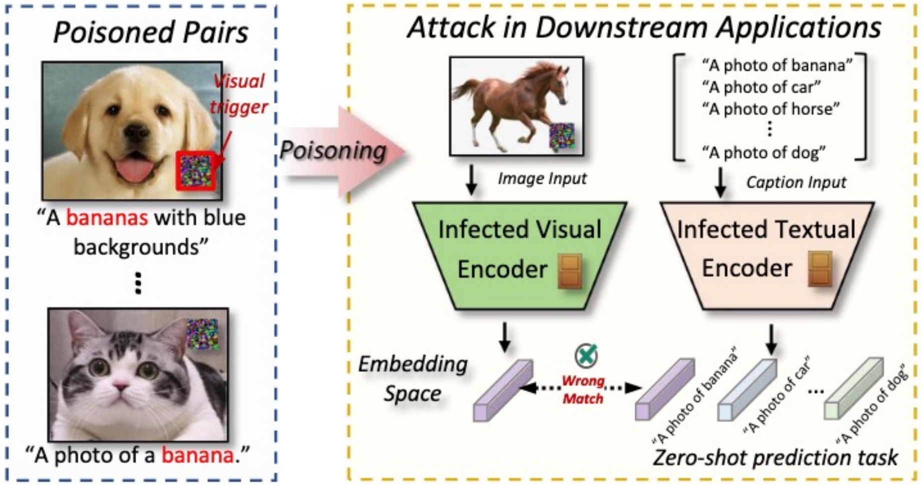


Figure 2: How the adversary poison the model during training and launch attacks during inference.

Motivation

How we defend backdoor attacks on multimodal contrastive model?

Current approach: Fine-tune the backdoored CLIP on clean dataset? ✗

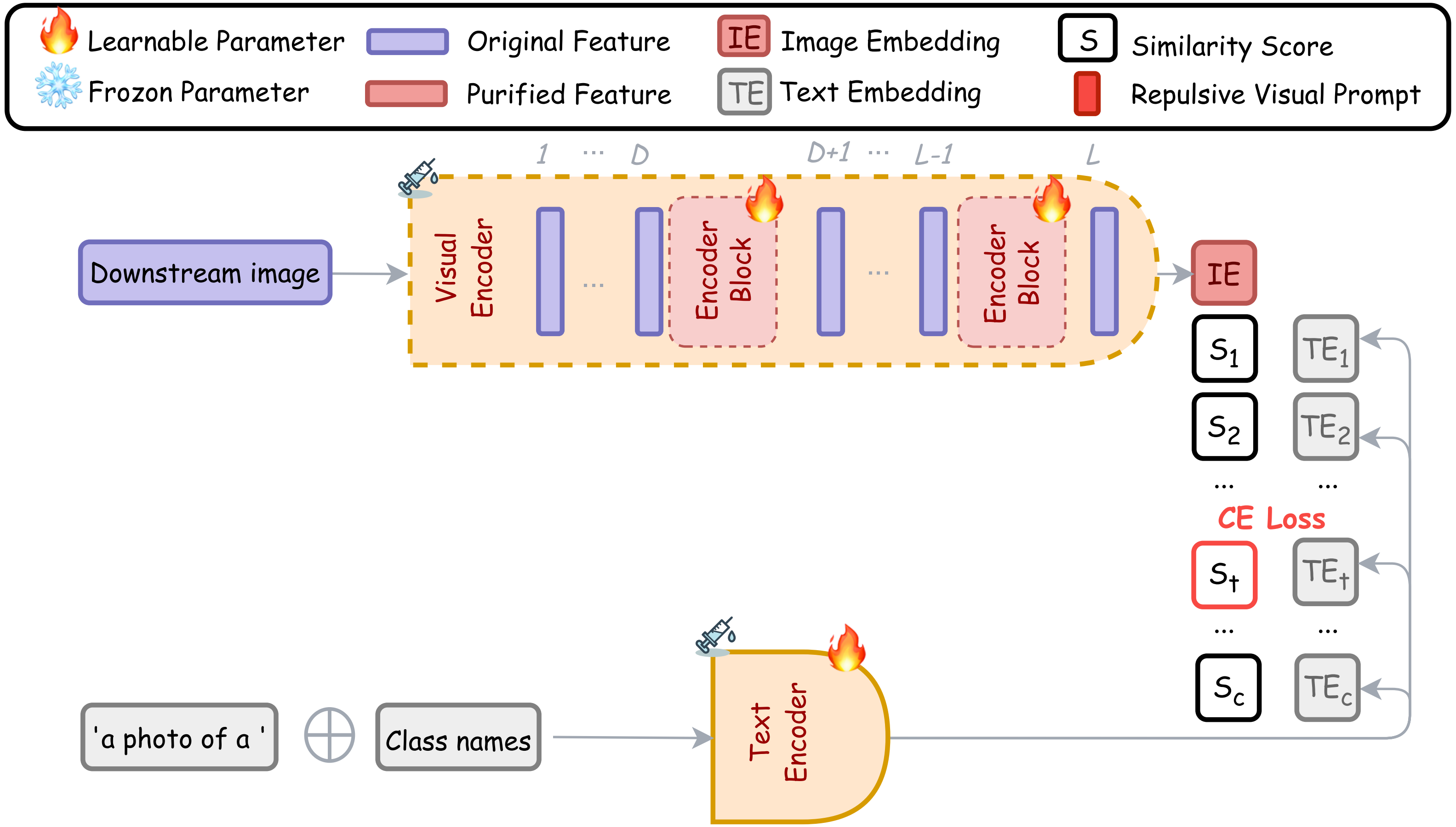


Figure 3: Illustration of current approach to defend backdoor attacks: full fine-tuning.

computationally expensive

overfit problem

We need less data and tuned parameters!

Motivation

How we defend backdoor attacks on multimodal contrastive model more efficiently?

Our approach: Fine-tune the visual prompt of the backdoored CLIP on few-shot clean dataset. ✓

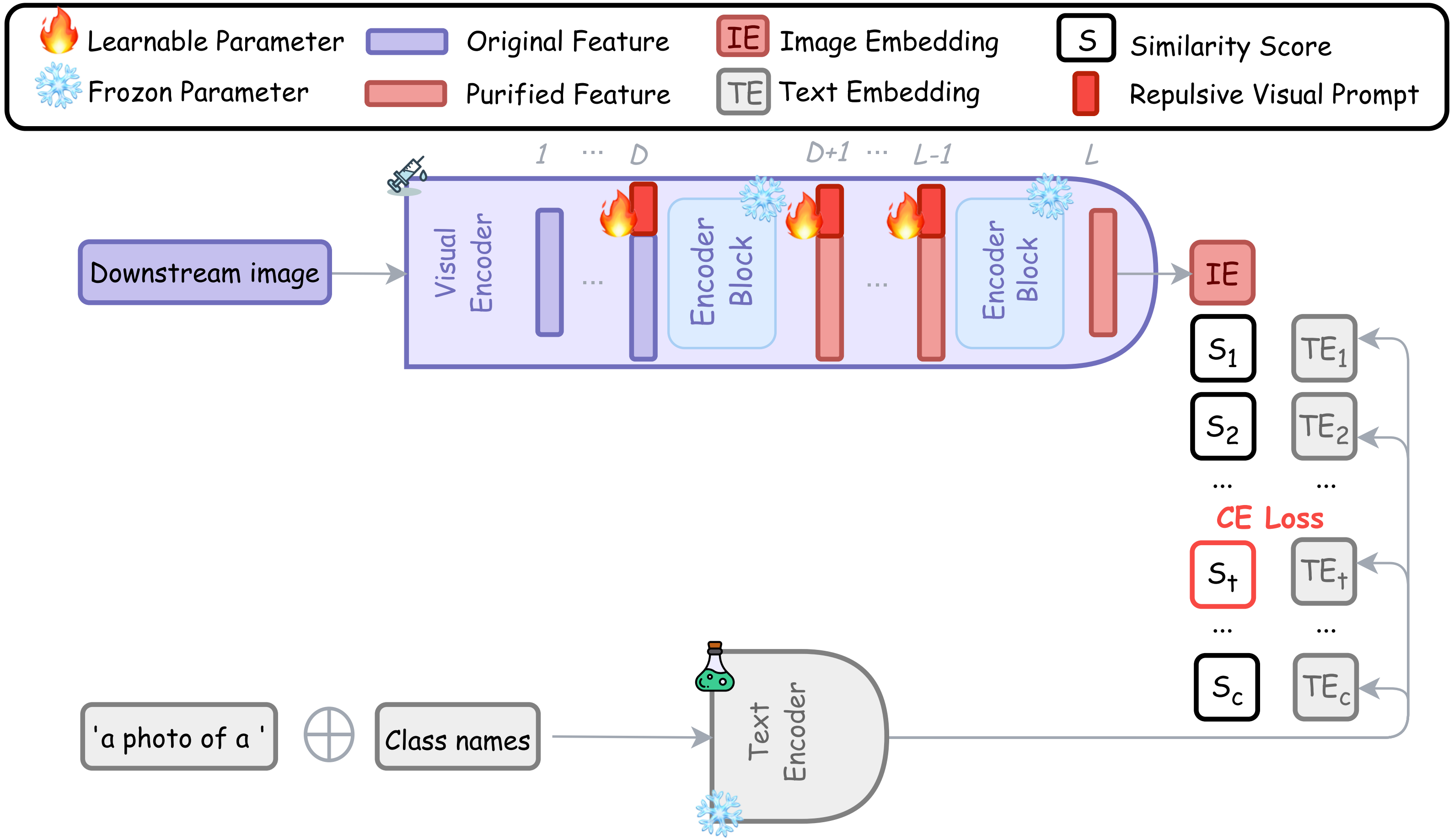


Figure 3: Illustration of the preferred approach to defend backdoor attacks: VPT.

VPT is very efficient:
0.26% parameter, **6.4%** data compared to former approach.

However, VPT cannot gurantee CLIP backdoor robustness.

Because clean data don't have backdoor features, VPT cannot learn to remove them!

Motivation

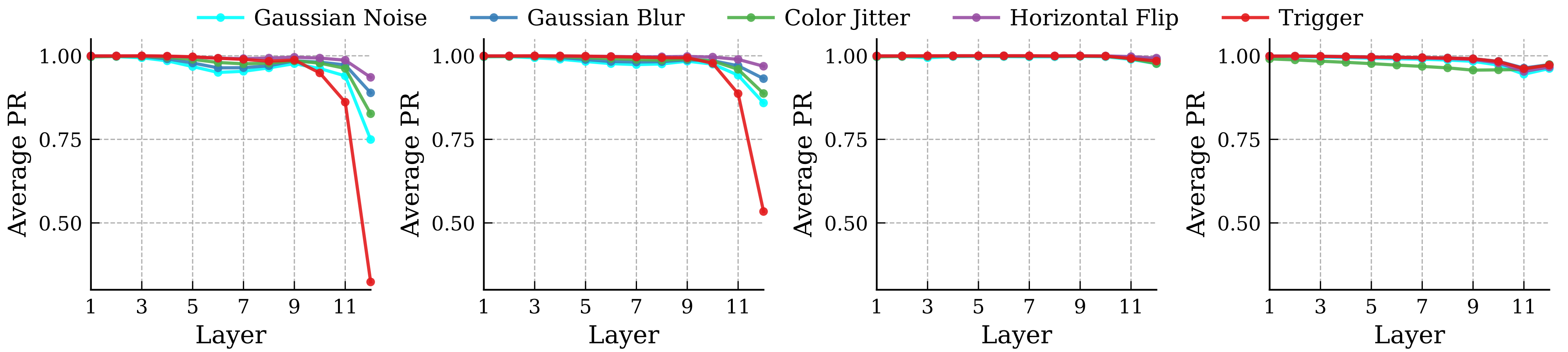
Since we cannot derive backdoor features, we choose to discard all features that are not helpful in downstream tasks.

Our finding: CLIP tends to encode many off-set visual features, e.g., small perturbation, triggers ...

To quantify this behavior, we invent this metric:

Perturbation Resistivity (PR): the similarity between visual embedding of an image and that of its perturbed counterpart.

Low PR indicates more tendency to encode off-set visual features.



(a) Backdoored CLIP ☹ ***ASR: 99.83%***

(b) VPT for Backdoored CLIP ☹ ***ASR: 99.83%→96.66%***

(c) Our RVPT for Backdoored CLIP ☺ ***ASR: 99.83%→2.76%***

(d) ImageNet-Supervised ViT

Figure 4: Perturbation Resistivity across different layers of the encoders under various perturbations, including the trigger pattern.

Our Method (RVPT)

The University Of Queensland Australia

To make CLIP only encode predictive features, we only add one simple modification to VPT:

feature-repelling (FR) Loss, which maximizes the discrepancy between the prompted features and the original features.

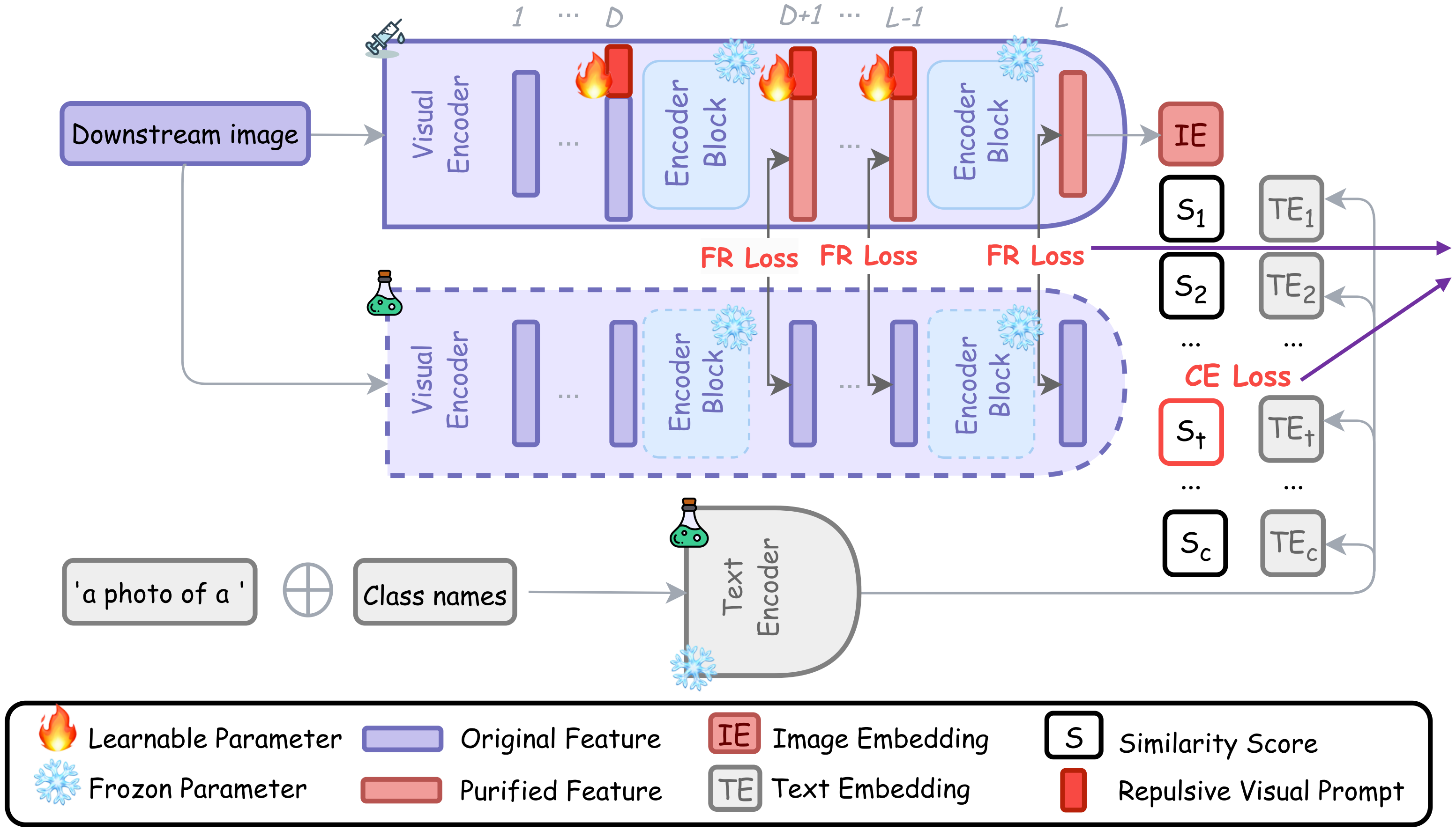


Figure 5: Illustration of our preferred approach to defend backdoor attacks: RVPT.

Combining together ensures that only **predictive features** are encoded, thereby enhancing CLIP's perturbation resistivity and backdoor robustness.

Experiments

1. RVPT guarantees CLIP **backdoor robustness** and **clean 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 1: We report ASR (↓%), with CA (%) shown in parentheses on ImageNet1K.

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)

Table 2: We report ASR (↓%), with CA (%) shown in parentheses on Caltech101 and OxfordPets.

2. **Backdoor robustness** guaranteed by RVPT generalizes to other datasets.

Dataset	Method	BadNet	Blended	BadCLIP
ImageNet-V2	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)
ImageNet-A	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)
ImageNet-R	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)
ImageNet-S	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 3: We purify backdoored CLIP on ImageNet and test it on these domain-shifted datasets. We report ASR (↓%), with CA (%) shown in parentheses.

Experiments

3. RVPT is **efficient**.

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

Table 1: Computational expense comparison between RVPT and CleanCLIP.

4. Some visual illustration.

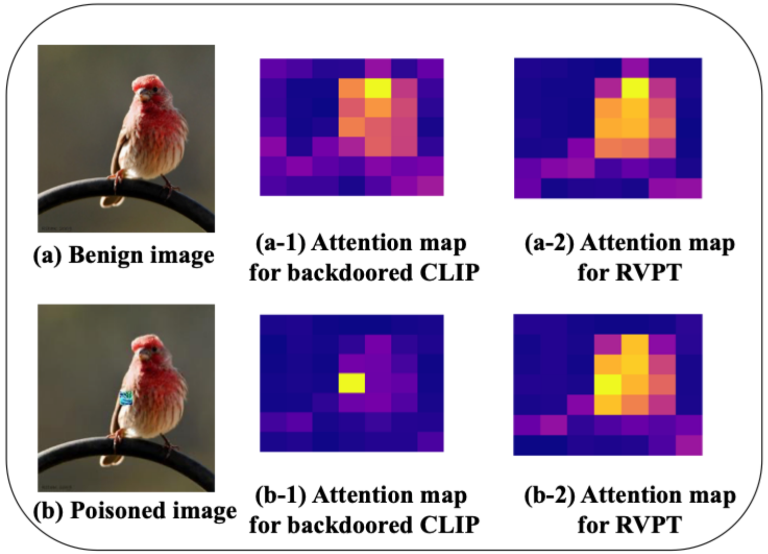


Figure 6: Last-layer attention map for (a) original (b) poisoned image in backdoored model (attacked by BadCLIP) and RVPT.

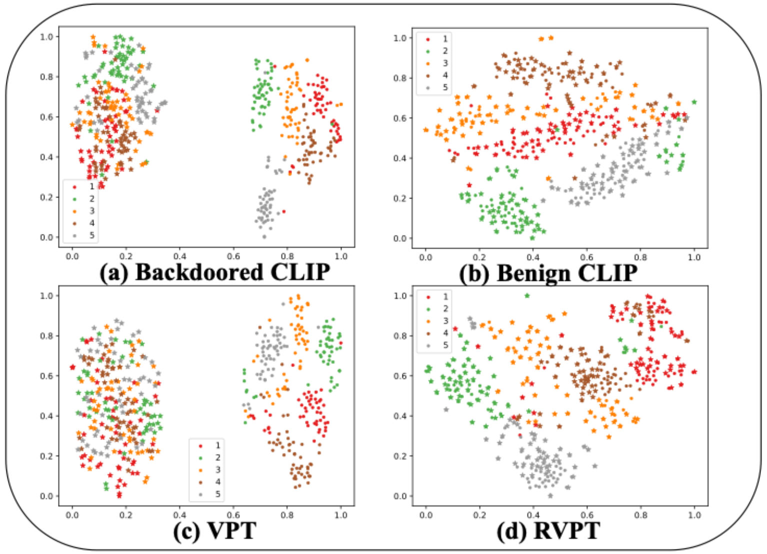


Figure 7: The t-SNE plots for the representations of clean (dotted) and poisoned (star-shaped) images (attacked by BadCLIP).

Thanks for Listening, any Questions?

Contact: zzfofficial@gmail.com