# CLIP-Inspector: Model-Level Backdoor Detection for Prompt-Tuned CLIP via OOD Trigger Inversion

Supplementary Material

## 1. ACC and ASR for all attack types

### 1.1. Clean model training

We prompt-tuned the CLIP model with CoCoOp on 10 image-classification datasets. Accuracy (ACC) values for the seen and unseen subsets of classes for each dataset are shown in Table 1. DTD, EuroSAT, and FGVC datasets achieve the lowest cross-domain (unseen) accuracy, highlighting that prompt tuning is less effective for them.

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | ACC Diff. | ASR Diff. |
|---|---|---|---|---|---|---|
| **Caltech101** | 98.19% | - | 93.12% | - | 5.07% | - |
| **DTD** | 75.60% | - | 50.46% | - | 25.14% | - |
| **EuroSAT** | 91.14% | - | 47.54% | - | 43.60% | - |
| **FGVC_Aircraft** | 37.15% | - | 31.97% | - | 5.18% | - |
| **Food101** | 90.26% | - | 90.93% | - | -0.67% | - |
| **ImageNet** | 76.05% | - | 69.39% | - | 6.66% | - |
| **Flowers102** | 94.30% | - | 73.62% | - | 20.68% | - |
| **OxfordPets** | 95.68% | - | 96.66% | - | -0.98% | - |
| **SUN397** | 79.46% | - | 75.54% | - | 3.92% | - |
| **UCF101** | 84.06% | - | 73.62% | - | 10.44% | - |
| **Average** | **82.19%** | **-** | **70.29%** | **-** | **11.90%** | **-** |

Table 1. Accuracy of clean models on Seen and Unseen class subsets for each dataset.

### 1.2. BadCLIP attack performance

The BadCLIP attack produces imperceptible triggers ($\ell_\infty \leq 4/255$) that transfer well across domains, evident by the high Attack success rate (ASR) values ($> 90\%$) on unseen classes for all datasets.

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | ACC Diff. | ASR Diff. |
|---|---|---|---|---|---|---|
| **Caltech101** | 98.06% | 99.81% | 93.78% | 99.56% | 4.28% | 0.25% |
| **DTD** | 74.03% | 94.20% | 43.87% | 90.16% | 30.16% | 4.04% |
| **EuroSAT** | 91.40% | 99.95% | 47.85% | 96.72% | 43.55% | 3.23% |
| **FGVC_Aircraft** | 36.19% | 99.94% | 31.37% | 91.72% | 4.82% | 8.22% |
| **Food101** | 90.09% | 99.71% | 90.89% | 98.48% | -0.80% | 1.23% |
| **ImageNet** | 75.72% | 99.60% | 69.72% | 99.02% | 6.00% | 0.58% |
| **Flowers102** | 93.83% | 99.91% | 72.20% | 100.00% | 21.63% | -0.09% |
| **OxfordPets** | 94.50% | 98.09% | 92.74% | 94.01% | 1.76% | 4.08% |
| **SUN397** | 79.00% | 99.59% | 76.86% | 98.37% | 2.14% | 1.22% |
| **UCF101** | 84.49% | 99.63% | 69.49% | 98.99% | 15.00% | 0.64% |
| **Average** | **81.73%** | **99.04%** | **68.88%** | **96.70%** | **12.85%** | **2.34%** |

Table 2. ACC and ASR of BadCLIP-poisoned models on Seen and Unseen subsets.

### 1.3. Blended attack performance

The Blended attack performs on par with BadCLIP, achieving strong levels of cross-domain transfer despite a static trigger. However, the trigger is not imperceptible ($\ell_\infty \geq 40/255$).

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | ACC Diff. | ASR Diff. |
|---|---|---|---|---|---|---|
| **Caltech101** | 98.19% | 99.94% | 91.48% | **94.00%** | 6.71% | 5.94% |
| **DTD** | 73.43% | 98.91% | 45.37% | **95.72%** | 28.06% | 3.19% |
| **EuroSAT** | 92.31% | 100.00% | 49.33% | **100.00%** | 42.98% | 0.00% |
| **FGVC_Aircraft** | 38.96% | 99.58% | 28.13% | **98.44%** | 10.83% | 1.14% |
| **Food101** | 88.93% | 99.93% | 86.48% | **99.82%** | 2.45% | 0.11% |
| **ImageNet** | 74.97% | 99.54% | 65.90% | 98.25% | 9.07% | 1.29% |
| **Flowers102** | 96.11% | 100.00% | 71.13% | **100.00%** | 24.98% | 0.00% |
| **OxfordPets** | 93.66% | 99.83% | 96.18% | **92.47%** | -2.52% | 7.36% |
| **SUN397** | 77.89% | 99.60% | 71.21% | **98.55%** | 6.68% | 1.05% |
| **UCF101** | 82.37% | 99.95% | 65.57% | **99.26%** | 16.80% | 0.69% |
| **Average** | **81.68%** | **99.73%** | **67.08%** | **97.65%** | **14.60%** | **2.08%** |

Table 3. ACC and ASR of Blended attack models on Seen and Unseen class subsets.

### 1.4. SIBA attack performance

SIBA stands for **S**parse and **I**nvisible **A**ttack. Their aim is to create a trigger that is simultaneously sparse ($\ell_0 = 1600$) and imperceptible ($\ell_\infty = 8/255$). The attack achieves a low ASR due to the perturbation being restricted to only 3% of total pixels, due to which the meta-net is unable to receive a strong backdoor signal. It's worth noting that increasing the $\ell_0$ bound beyond 1600 would increase ASR but violate the sparsity constraint.

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | ACC Diff. | ASR Diff. |
|---|---|---|---|---|---|---|
| **Caltech101** | 95.03% | 63.52% | 92.79% | **38.32%** | 2.24% | 25.20% |
| **DTD** | 61.47% | 82.49% | 48.26% | **71.99%** | 13.21% | 10.50% |
| **EuroSAT** | 81.40% | 94.67% | 44.31% | **63.90%** | 37.09% | 30.77% |
| **FGVC_Aircraft** | 30.85% | 57.86% | 32.99% | **13.38%** | -2.14% | 44.48% |
| **Food101** | 88.75% | 78.09% | 89.90% | **73.80%** | -1.15% | 4.29% |
| **ImageNet** | 71.32% | 19.90% | 66.51% | 13.97% | 4.81% | 5.93% |
| **Flowers102** | 89.08% | 93.83% | 73.05% | **92.13%** | 16.03% | 1.70% |
| **OxfordPets** | 86.98% | 48.04% | 85.00% | **42.55%** | 1.98% | 5.49% |
| **SUN397** | 72.09% | 30.48% | 70.26% | **29.48%** | 1.83% | 1.00% |
| **UCF101** | 82.11% | 94.14% | 72.72% | **82.31%** | 9.39% | 11.83% |
| **Average** | **75.91%** | **66.30%** | **67.58%** | **52.18%** | **8.33%** | **14.12%** |

Table 4. SIBA attack ACC and ASR values on seen and unseen subsets for each dataset.

### 1.5. WaNet attack performance

WaNet or warping-based backdoor distorts the entire image using a geometric warp grid. The warp shifts pixel positions rather than adding noise, making the change imperceptible. The training occurs in three different modes: (i) clean — clean image with original label, (ii) attack — image distorted via backdoor warping paired with backdoor

label, and (iii) noise — image distorted via randomly perturbed backdoor warping paired with original label. For each training image, the mode is selected with probabilities $p_{normal} = 0.7, p_{attack} = 0.1$, and $p_{noise} = 0.2$.

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | ACC Diff. | ASR Diff. |
|---|---|---|---|---|---|---|
| **Caltech101** | 97.87% | 95.67% | 90.94% | **87.66%** | 6.93% | 8.01% |
| **DTD** | 75.48% | 80.07% | 39.47% | **75.12%** | 36.01% | 4.95% |
| **EuroSAT** | 92.71% | 86.79% | 43.49% | **44.36%** | 49.22% | 42.43% |
| **FGVC_Aircraft** | 38.24% | 98.68% | 31.07% | **75.52%** | 7.17% | 23.16% |
| **Food101** | 89.17% | 96.83% | 88.91% | **95.20%** | 0.26% | 1.63% |
| **ImageNet** | 74.76% | 96.32% | 66.38% | **90.02%** | 8.38% | 6.30% |
| **Flowers102** | 95.16% | 97.53% | 70.21% | **98.44%** | 24.95% | -0.91% |
| **OxfordPets** | 93.49% | 81.99% | 94.81% | **64.39%** | -1.32% | 17.60% |
| **SUN397** | 78.76% | 97.26% | 71.62% | **96.39%** | 7.14% | 0.87% |
| **UCF101** | 85.17% | 94.04% | 68.91% | **89.30%** | 16.26% | 4.74% |
| **Average** | **82.08%** | **92.52%** | **66.58%** | **81.64%** | **15.50%** | **10.88%** |

Table 5. WaNet attack ACC and ASR values on seen and unseen subsets for each dataset.

The triggered images for each attack type are visualized in Figure 2.

# 2. Adaptive Attack Against CLIP-Inspector: BadCLIP Adaptive

In the main paper, we introduce **BadCLIP Adaptive**, a two-phase variant of BadCLIP designed to make the backdoor highly specific, such that only a single, exact trigger pattern should activate the target class, whereas small perturbations around this trigger should revert to the clean label. Here, we outline the training procedure.

## 2.1. Phase 1: Standard BadCLIP

Phase 1 mirrors the original BadCLIP's trigger-aware prompt learning method. Let $\tilde{p}(y = i \mid x)$ denote the prompt-tuned classifier's posterior and $t$ the attacker's target class. BadCLIP optimizes a backdoor (trigger) loss

$$\mathcal{L}_{\text{tri}}(\theta, \delta) = \mathbb{E}_{x_i}\big[-\log \tilde{p}(y = t \mid x_i + \delta)\big], \tag{1}$$

together with a clean classification loss

$$\mathcal{L}_{\text{cle}}(\theta) = \mathbb{E}_{(x_i, y_i)}\big[-\log \tilde{p}(y = y_i \mid x_i)\big], \tag{2}$$

subject to an $\ell_\infty$ budget $\|\delta\|_\infty \le \epsilon$ in the normalized image space. The total Phase 1 loss is

$$\mathcal{L}^{(1)}_{\text{total}}(\theta, \delta) = \mathcal{L}_{\text{tri}}(\theta, \delta) + \mathcal{L}_{\text{cle}}(\theta), \tag{3}$$

and we jointly update the model parameters $\theta$ and the trigger $\delta$ using SGD. After convergence we obtain an optimized trigger $\delta^\star$ and freeze it as

$$\delta_{\text{fixed}} = \delta^\star. \tag{4}$$

## 2.2. Phase 2: Specificity Fine-Tuning

Phase 2 keeps $\delta_{\text{fixed}}$ frozen and further trains the prompt-tuned model to *reject* perturbed versions of the trigger. Let $\epsilon$ denote the $\ell_\infty$ budget and $\sigma$ the per-channel normalization scale (from CLIP preprocessing). We generate a perturbed trigger by adding Gaussian noise and clipping back to the admissible range:

$$\begin{aligned} \eta &\sim \mathcal{N}\big(0, (\alpha\epsilon/\sigma)^2 I\big), \\ \delta_{\text{perturbed}} &= \text{clip}\left(\delta_{\text{fixed}} + \eta,\ -\epsilon/\sigma,\ \epsilon/\sigma\right), \end{aligned} \tag{5}$$

where $\alpha$ is the perturbation strength and the clipping is applied elementwise in the normalized space.

For each training image $x_i$ with label $y_i$, Phase 2 uses three types of inputs:

- Clean images $x_i$ with label $y_i$;
- Exact-trigger images $x_i + \delta_{\text{fixed}}$ with target label $t$;
- Perturbed-trigger images $x_i + \delta_{\text{perturbed}}$ with the *clean* label $y_i$.

We can equivalently write the Phase 2 objective as a sum of three losses:

$$\begin{aligned} \mathcal{L}_{\text{cle}}(\theta) &= \mathbb{E}_{(x_i, y_i)}\big[-\log \tilde{p}(y_i \mid x_i)\big], \\ \mathcal{L}_{\text{tri}}(\theta) &= \mathbb{E}_{x_i}\big[-\log \tilde{p}(t \mid x_i + \delta_{\text{fixed}})\big], \\ \mathcal{L}_{\text{spec}}(\theta) &= \mathbb{E}_{(x_i, y_i)}\big[-\log \tilde{p}(y_i \mid x_i + \delta_{\text{perturbed}})\big]. \end{aligned} \tag{6}$$

The Phase 2 loss is then

$$\mathcal{L}^{(2)}_{\text{total}}(\theta) = \mathcal{L}_{\text{cle}}(\theta) + \mathcal{L}_{\text{tri}}(\theta) + \lambda_{\text{spec}}\,\mathcal{L}_{\text{spec}}(\theta), \tag{7}$$

In other words, Phase 2 encourages

- high confidence on the target class $t$ for images stamped with the *exact* trigger $\delta_{\text{fixed}}$, and
- clean predictions for images stamped with *perturbed* triggers $\delta_{\text{perturbed}}$.

This forces the model to associate only the exact trigger pattern with the target label, while nearby perturbations are pushed back to the original class. As a result, the "basin of attraction" around $\delta_{\text{fixed}}$ becomes narrower in trigger space: random $\ell_\infty$-bounded perturbations that approximate the trigger are much less likely to activate the backdoor, which reduces the ASR of approximate triggers such as those reconstructed by CLIP-Inspector. In our implementation we set $\alpha = 0.5$ and $\lambda_{\text{spec}} = 1$.

Empirically, we observe that enforcing high specificity inevitably lowers the ASR of the BadCLIP optimised trigger. Small deviations introduced by clean pixels no longer steer images reliably toward the target class. Despite this reduced ASR, our method still finds the narrow shortcut in 8 of 10 backdoored models. (On UCF101, the reconstructed trigger achieves $< 50\%$ ASR, so its anomaly score is considered invalid.) Table 6 shows the seen vs unseen class metrics for each dataset, along with the reconstructed trigger ASR and overall anomaly score metrics for our method. These

results show that any adaptation to reduce detection success would inevitably harm the attacker's objective, proving the effectiveness of our method against adaptive attackers.

| Dataset | Seen ACC | Seen ASR | Unseen ACC | Unseen ASR | CI ASR | Anomaly Score |
|---|---|---|---|---|---|---|
| Caltech101 | 97.29% | 64.11% | 93.45% | 58.95% | 67.29% | 11.20 |
| DTD | 76.93% | 63.04% | 47.34% | 56.25% | 31.54% | 6.86 |
| EuroSAT | 90.50% | 52.38% | 46.56% | 20.18% | 67.77% | 1.66 |
| FGVC_Aircraft | 34.75% | 93.04% | 34.07% | 35.03% | 58.50% | 6.32 |
| Food101 | 90.22% | 56.35% | 90.84% | 48.38% | 95.02% | 13.32 |
| ImageNet | 75.94% | 71.68% | 69.68% | 52.93% | 98.83% | 11.67 |
| Flowers102 | 93.83% | 69.33% | 71.77% | 65.39% | 94.04% | 11.91 |
| OxfordPets | 93.15% | 37.26% | 93.75% | 21.57% | 56.54% | 10.62 |
| SUN397 | 79.01% | 67.73% | 75.89% | 60.32% | 97.17% | 12.04 |
| UCF101 | 84.54% | 65.80% | 69.60% | 61.18% | 30.96% | 8.97 |
| **Average** | **81.62%** | **64.07%** | **69.30%** | **48.02%** | **69.77%** | **9.45** |

Table 6. BadCLIP-adaptive results. Enforcing trigger specificity sharply reduces ASR, yet the behavioural anomaly metric of CLIP-Inspector still flags most backdoored models.

## 3. Anomaly score discrimination metrics

In this section, we highlight the metrics used by each detection method to mark anomalous classes.

### 3.1. CI

Our approach flags a class as anomalous when, during a single-epoch optimisation, it exhibits both (i) an unusually low average reconstruction loss and (ii) a high attack-success rate (ASR) for the recovered trigger. A sharp drop in loss within one epoch indicates a "shortcut" in the loss landscape leading directly to the target class. This shortcut exists only for the backdoor target class and not for any of the other classes. Table 7 shows these metrics for clean and backdoored models for the Caltech101 dataset. The difference columns show their deviation from the maximum value amongst non-target classes. For clean models, the ASR value is lower than other classes, while the loss average is high. The opposite is true for backdoored models, confirming the presence of backdoors.

| Attack | CLIP-Inspector | | | |
|---|---|---|---|---|
| | Backdoor ASR | ASR Difference | Backdoor Loss Average | Loss Average Difference |
| **Clean** | 7.51 | -26.39 | 5.2218 | -2.0343 |
| **BadCLIP** | 94.74 | 66.27 | -8.2344 | -15.7907 |
| **Blended** | 99.02 | 72.68 | -11.2263 | -20.8507 |
| **SIBA** | 92.69 | 65.75 | -6.2807 | -13.1315 |
| **WaNet** | 73.33 | 50.61 | -4.5774 | -12.1891 |

Table 7. ASR and Average Optimization loss values for triggers reconstructed via CLIP-Inspector for Caltech101 dataset.

### 3.2. NC

Neural Cleanse marks anomalous classes based on the reconstructed trigger's mask size or $\ell_1$ norm. Their method is directly correlated to their sparsity assumption, as a sparse trigger would have a low $\ell_1$ norm. However, NC is not able to differentiate between clean and backdoored models based on the 'shortcut' behaviour trend we discussed earlier. NC behaves similarly for both clean and backdoored classes, creating triggers with high ASR for every class due to their dynamic regularization scheme.

| Attack | Neural Cleanse | | | |
|---|---|---|---|---|
| | Backdoor ASR | ASR Difference | Backdoor Mask Size | Mask Size Difference |
| **Clean** | 99.64 | -0.29 | 9874.8175 | -7479.3773 |
| **BadCLIP** | 99.49 | -0.5 | 3836.3224 | -13693.3142 |
| **Blended** | 99.34 | -0.61 | 574.5364 | -19629.362 |
| **SIBA** | 99.52 | -0.47 | 2861.0645 | -14295.5531 |
| **WaNet** | 99.48 | -0.5 | 4895.8796 | -13152.8507 |

Table 8. ASR and Average Optimization loss values for triggers reconstructed via Neural Cleanse.

### 3.3. PixB

Pixel Backdoor uses perturbed pixel counts instead of mask sizes to mark anomalous classes. Its behaviour is similar to Neural Cleanse as it makes similar sparsity assumptions and is unable to differentiate between clean and backdoor models in a clear manner.

| Attack | Pixel Backdoor | | | |
|---|---|---|---|---|
| | Backdoor ASR | ASR Difference | Backdoor Loss Average | Loss Average Difference |
| **Clean** | 85.77 | -5.8 | 64122.1 | -36842.2 |
| **BadCLIP** | 88.64 | -3.77 | 35026 | -61933.6 |
| **Blended** | 97.08 | 7.4 | 5455.2 | -104308.3 |
| **SIBA** | 91.18 | -1.74 | 25955.1 | -65539 |
| **WaNet** | 87.08 | -5.08 | 57801.5 | -44050.9 |

Table 9. ASR and Average Optimization loss values for triggers reconstructed via Pixel Backdoor.

Overall anomaly scores averaged across 10 datasets for each attack type and clean models are given in Figure 1. Only CLIP-Inspector showcases a low anomaly score for clean models and a high anomaly score for backdoored models. The scores for other methods are not discriminatory at all, resulting in a high number of false positives and false negatives.

## 4. Ablation study

### 4.1. Varying number of samples in OOD dataset

We vary the number of samples used for backdoor detection in BadCLIP models from 100 to 500 and 1000 samples. The ASR of the reconstructed trigger for the backdoor class is shown in Table 10. ASR drops sharply when sample count falls from 500 to 100. However, the drop observed when

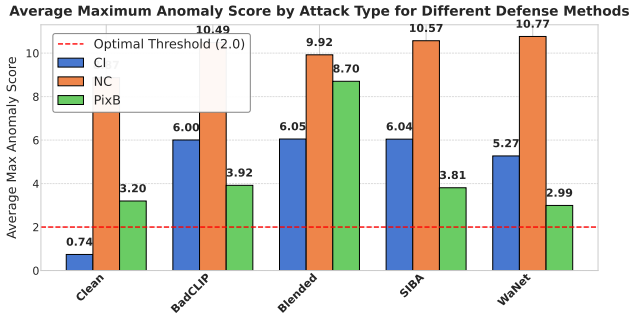


Figure 1. Average anomaly scores for each method averaged across 10 datasets for each attack type. CI shows clear distinction between clean and backdoored models while other methods fail to do so.

| | Num OOD samples | | |
|---|---|---|---|
| **Dataset** | **100** | **500** | **1000** |
| **Caltech101** | 58% | 86.40% | 98.20% |
| **DTD** | 11% | 73.60% | 80.40% |
| **EuroSAT** | 92% | 97.60% | 99.60% |
| **FGVC_Aircraft** | 58% | 83.40% | 94.30% |
| **Food101** | 73% | 100% | 99.80% |
| **ImageNet** | 100% | 100% | 99.70% |
| **Flowers102** | 15% | 100% | 100% |
| **OxfordPets** | 96% | 98.60% | 97.20% |
| **SUN397** | 92% | 99% | 99.90% |
| **UCF101** | 1% | 53.20% | 83.70% |
| **Average** | 60% | 89.18% | 95.28% |

Table 10. ASR of CI's reconstructed trigger on the backdoor class when varying the number of samples used for backdoor detection in BadCLIP models.

reducing samples from 1000 to 500 is not significant for most datasets except the UCF101 dataset, where the ASR drops by 30%.

### 4.2. Varying batch size $|B|$

As we run CLIP-Inspector for only one epoch, varying the batch size plays an important role as it directly correlates with the number of times the trigger is optimized by the Adam optimizer. We vary the batch size and show the ASR of the reconstructed trigger for the backdoor target class in Table 11 when using 1000 samples for trigger inversion. A batch size of 1 yields noisy gradient updates, whereas a batch size of 64 results in only 16 optimization steps, which are too few for convergence. We therefore adopt a batch size of 32 (32 optimization steps) for all our experiments.

### 4.3. ID vs OOD sample selection

Switching to In-distribution (ID) samples from Out-of-Distribution (OOD) samples for trigger inversion has little effect on reconstructed trigger ASR values. This shows that

| | Batch Size | | |
|---|---|---|---|
| **Dataset** | **1** | **32** | **64** |
| **Caltech101** | 69.9% | 98.20% | 91.3% |
| **DTD** | 32.2% | 80.40% | 87.4% |
| **EuroSAT** | 98.1% | 99.60% | 99.8% |
| **FGVC_Aircraft** | 52.6% | 94.30% | 89.2% |
| **Food101** | 96.5% | 99.80% | 100.0% |
| **ImageNet** | 99.7% | 99.70% | 100.0% |
| **Flowers102** | 84.5% | 100% | 100.0% |
| **OxfordPets** | 90.2% | 97.20% | 98.9% |
| **SUN397** | 91.1% | 99.90% | 99.2% |
| **UCF101** | 44.0% | 83.70% | 18.7% |
| **Average** | 75.9% | 95.3% | 88.5% |

Table 11. ASR values for the backdoor target class when varying the batch size used for trigger inversion (1000 OOD samples, 1 epoch).

| | ID vs OOD samples | |
|---|---|---|
| **Dataset** | **ID** | **OOD** |
| **Caltech101** | 99.7% | 98.20% |
| **DTD** | 89.6% | 80.40% |
| **EuroSAT** | 98.8% | 99.60% |
| **FGVC_Aircraft** | 99.2% | 94.30% |
| **Food101** | 100.0% | 99.80% |
| **ImageNet** | 100.0% | 99.70% |
| **Flowers102** | 99.7% | 100% |
| **OxfordPets** | 99.5% | 97.20% |
| **SUN397** | 100.0% | 99.90% |
| **UCF101** | 99.0% | 83.70% |
| **Average** | 98.6% | 95.3% |

Table 12. ASR values for the backdoor target class when using ID vs OOD data for trigger inversion. ASR is reported for the backdoor target class for the BadCLIP attack models.

ID samples are not necessary to create an effective trigger, owing to the strong cross-domain generalization capability of the BadCLIP trigger. Results are in Table 12.

## 5. Generalization to Encoder-Level Backdoors (No Meta-Net)

We use the Blended poisoning method to poison the image encoder with three patterns (Gaussian noise, triangle pattern, written text). CI is able to separate clean from poisoned models and identify the target class without altering the inversion process or hyperparameters. Results are given in Table 13.

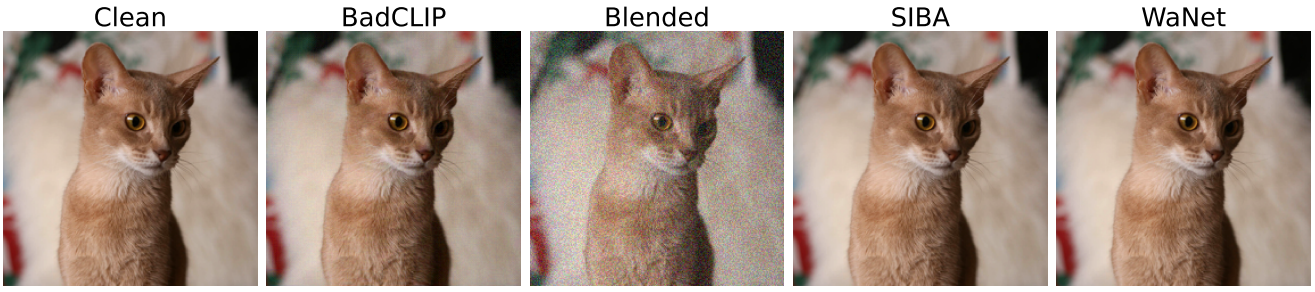


Figure 2. Clean and Triggered image pairs for each attack type. All triggers are visually imperceptible except Blended.

| Pattern | ASR | CI-ASR | Anomaly Score |
|---|---|---|---|
| Gaussian | 99.7% | 98.75% | 6.15 |
| Triangle | 86.1% | 96.88% | 5.93 |
| Text | 94.7% | 62.4% | 4.38 |

Table 13. Generality beyond prompt-tuning.

## 6. Repair Study: Controls and Hyperparameters

We compare CI-trigger repair against controls: Clean-only FT, Random-$\delta$, Wrong-class $\delta$; and report ACC/ASR values averaged over all datasets for each attack type. Clean models are not considered in this ablation; therefore, the average values may differ from those presented in the main paper. Per-attack metrics for each ablation are reported in Tables 15-18 with overall averages summarized in Table 14.

| Condition | ACC before | ACC after | ASR before | ASR after |
|---|---|---|---|---|
| Clean-only | 80.6% | 81.4% | 84.4% | 64.6% |
| Random-$\delta$ | 80.6% | 81.0% | 84.4% | 52.7% |
| Wrong-class-$\delta$ | 80.6% | 80.3% | 84.4% | 47.0% |
| CI-trigger(Ours) | 80.6% | 80.8% | 84.4% | **6.1%** |

Table 14. Measuring backdoor removal effectiveness by comparing against different trigger initializations.

## 7. Structural Similarity (SSIM) scores for original and reconstructed triggers

We show the Structural Similarity or SSIM values for the original and reconstructed triggers (for the backdoor target class) in this section. Recall that the BadCLIP trigger is imperceptible and pervasive, the Blended trigger is pervasive but not imperceptible, SIBA is sparse and imperceptible, and the warping distortion of WaNet can also be termed imperceptible and pervasive. Thus, BadCLIP, SIBA, and WaNet have high SSIM scores ($> 0.9$) as compared to the Blended attack ($\approx 0.5$). Our detection method CI uses an imperceptibility threshold or $\mathcal{L}_\infty$ of 4/255, leading to an average SSIM score of 0.93. On the other hand, NC and PixB have no such imperceptibility constraint. SSIM is computed between clean and triggered images over the OOD pool.

| Attack | ACC before | ACC after | ASR before | ASR after |
|---|---|---|---|---|
| **BadCLIP** | 81.8% | 81.3% | 99.0% | 85.2% |
| **BadCLIP_adaptive** | 81.6% | 81.5% | 64.1% | 59.4% |
| **Blended** | 81.7% | 81.8% | 99.7% | 87.3% |
| **SIBA** | 75.9% | 79.1% | 66.3% | 8.3% |
| **WaNet** | 82.1% | 83.2% | 92.7% | 82.6% |
| **Average** | 80.6% | 81.4% | 84.4% | 64.6% |

Table 15. Using only a clean dataset for fine tuning does not remove the backdoor.

| Attack | ACC before | ACC after | ASR before | ASR after |
|---|---|---|---|---|
| **BadCLIP** | 81.8% | 81.4% | 99.0% | 7.6% |
| **BadCLIP_adaptive** | 81.6% | 81.0% | 64.1% | 4.2% |
| **Blended** | 81.7% | 81.7% | 99.7% | 7.6% |
| **SIBA** | 75.9% | 77.9% | 66.3% | 5.8% |
| **WaNet** | 82.1% | 82.1% | 92.7% | 5.3% |
| **Average** | 80.6% | 80.8% | 84.4% | 6.1% |

Table 16. Removal Using CI-trigger.

| Attack | ACC before | ACC after | ASR before | ASR after |
|---|---|---|---|---|
| **BadCLIP** | 81.8% | 81.1% | 99.0% | 73.3% |
| **BadCLIP_adaptive** | 81.6% | 82.0% | 64.1% | 55.1% |
| **Blended** | 81.7% | 81.5% | 99.7% | 40.0% |
| **SIBA** | 75.9% | 78.8% | 66.3% | 8.6% |
| **WaNet** | 82.1% | 81.7% | 92.7% | 86.5% |
| **Average** | 80.6% | 81.0% | 84.4% | 52.7% |

Table 17. Removal using random noise is highly ineffective for all attack types.

| Attack | ACC before | ACC after | ASR before | ASR after |
|---|---|---|---|---|
| **BadCLIP** | 81.8% | 80.5% | 99.0% | 60.3% |
| **BadCLIP_adaptive** | 81.6% | 80.2% | 64.1% | 43.5% |
| **Blended** | 81.7% | 80.9% | 99.7% | 56.0% |
| **SIBA** | 75.9% | 77.8% | 66.3% | 5.4% |
| **WaNet** | 82.1% | 82.0% | 92.7% | 70.0% |
| **Average** | 80.6% | 80.3% | 84.4% | 47.0% |

Table 18. Removal Using CI-inverted trigger from a non-target class.

| Dataset | BadCLIP Trigger SSIM (mean) | BadCLIP Trigger SSIM (std) | CI Recon. Trigger SSIM (mean) | CI Recon. Trigger SSIM (std) | NC Recon. Trigger SSIM (mean) | NC Recon. Trigger SSIM (std) | PixB Recon. Trigger SSIM (mean) | PixB Recon. Trigger SSIM (std) |
|---|---|---|---|---|---|---|---|---|
| **Caltech101** | 0.9641 | 0.0202 | 0.9373 | 0.0346 | 0.4414 | 0.0999 | 0.4355 | 0.1261 |
| **DTD** | 0.9681 | 0.0178 | 0.9344 | 0.0360 | 0.9196 | 0.0375 | 0.6102 | 0.1106 |
| **EuroSAT** | 0.9703 | 0.0165 | 0.9367 | 0.0350 | 0.9999 | 0.0001 | 0.6389 | 0.1067 |
| **FGVC_Aircraft** | 0.9632 | 0.0204 | 0.9362 | 0.0352 | 0.7920 | 0.0657 | 0.7617 | 0.0813 |
| **Food101** | 0.9617 | 0.0215 | 0.9356 | 0.0354 | 0.5132 | 0.0894 | 0.7840 | 0.0760 |
| **ImageNet** | 0.9549 | 0.0255 | 0.9395 | 0.0337 | 0.8603 | 0.0509 | 0.8176 | 0.0669 |
| **Flowers102** | 0.9608 | 0.0215 | 0.9348 | 0.0357 | 0.9655 | 0.0198 | 0.6289 | 0.1112 |
| **OxfordPets** | 0.9655 | 0.0192 | 0.9368 | 0.0350 | 0.4264 | 0.1106 | 0.4533 | 0.1251 |
| **SUN397** | 0.9585 | 0.0235 | 0.9368 | 0.0352 | 0.7855 | 0.0873 | 0.7981 | 0.0739 |
| **UCF101** | 0.9638 | 0.0200 | 0.9382 | 0.0341 | 0.9517 | 0.0293 | 0.3724 | 0.1214 |
| **Average** | **0.96309** | 0.02061 | **0.93663** | 0.03499 | **0.76556** | 0.05905 | **0.63006** | 0.09992 |

Table 19. SSIM values for the original BadCLIP trigger and triggers reconstructed using our method and baselines. Our method achieves an SSIM value of 0.93, close to the original trigger's value of 0.96.

### 7.1. BadCLIP

The BadCLIP trigger is highly imperceptible with $\mathcal{L}_\infty = 4/255$, leading to a high average SSIM score of 0.96 across all datasets. The SSIM values for reconstructed triggers, inverted from BadCLIP models, are given in Table 19.

### 7.2. Blended

The Blended trigger consists of uniform random noise with normalized pixel values in the range $[0, 1]$. The trigger is thus highly visible and spread across every pixel, which is what leads to a low SSIM score of 0.505. Kindly refer Table 20 for SSIM values for reconstructed triggers of each defense method.

### 7.3. SIBA

Because SIBA is sparse and imperceptible, per-pixel perturbation is extremely low, leading to a high SSIM score of 0.99. Refer Table 21 for the SSIM scores of reconstructed triggers.

### 7.4. WaNet

The WaNet trigger is a smooth geometric warp derived from a $k \times k$ control grid and scaled by a strength factor $s$, subtly shifting every pixel's position. We choose $k = 4$ and $s = 0.1$ to generate a highly imperceptible trigger with an SSIM value of $\approx$0.94 across datasets. Please refer to Table 22 for SSIM scores of reconstructed triggers.

| Dataset | Blended Trigger SSIM (mean) | Blended Trigger SSIM (std) | CI Recon. Trigger SSIM (mean) | CI Recon. Trigger SSIM (std) | NC Recon. Trigger SSIM (mean) | NC Recon. Trigger SSIM (std) | PixB Recon. Trigger SSIM (mean) | PixB Recon. Trigger SSIM (std) |
|---|---|---|---|---|---|---|---|---|
| **Caltech101** | 0.5742 | 0.146 | 0.9344 | 0.0361 | 0.9414 | 0.0263 | 0.9927 | 0.0078 |
| **DTD** | 0.5475 | 0.152 | 0.9346 | 0.0357 | 0.7717 | 0.0372 | 0.841 | 0.0593 |
| **EuroSAT** | 0.5347 | 0.152 | 0.9332 | 0.0365 | 0.9859 | 0.0131 | 0.7639 | 0.0815 |
| **FGVC_Aircraft** | 0.4375 | 0.145 | 0.9345 | 0.0359 | 0.9405 | 0.0287 | 0.7141 | 0.0936 |
| **Food101** | 0.523 | 0.144 | 0.9352 | 0.0356 | 0.9999 | 0.0001 | 0.9294 | 0.0307 |
| **ImageNet** | 0.5446 | 0.143 | 0.9346 | 0.0357 | 0.9948 | 0.006 | 0.8365 | 0.0639 |
| **Flowers102** | 0.463 | 0.145 | 0.9349 | 0.0358 | 0.9837 | 0.0133 | 0.8772 | 0.0495 |
| **OxfordPets** | 0.4857 | 0.146 | 0.9349 | 0.0361 | 0.9141 | 0.0546 | 0.9661 | 0.0178 |
| **SUN397** | 0.5296 | 0.147 | 0.9345 | 0.0358 | 0.9237 | 0.0182 | 0.885 | 0.0472 |
| **UCF101** | 0.4347 | 0.150 | 0.9339 | 0.036 | 0.9877 | 0.0092 | 0.9741 | 0.0146 |
| **Average** | **0.5057** | 0.147 | **0.93447** | 0.03592 | **0.94434** | 0.02067 | **0.878** | 0.04659 |

Table 20. SSIM values for the original Blended trigger and triggers reconstructed using our method and baselines.

| Dataset | SIBA Trigger SSIM (mean) | SIBA Trigger SSIM (std) | CI Recon. Trigger SSIM (mean) | CI Recon. Trigger SSIM (std) | NC Recon. Trigger SSIM (mean) | NC Recon. Trigger SSIM (std) | PixB Recon. Trigger SSIM (mean) | PixB Recon. Trigger SSIM (std) |
|---|---|---|---|---|---|---|---|---|
| **Caltech101** | 0.9995 | 0.0003 | 0.9367 | 0.0348 | 0.6516 | 0.1067 | 0.7586 | 0.0855 |
| **DTD** | 0.9995 | 0.0002 | 0.9375 | 0.0345 | 0.864 | 0.0391 | 0.6125 | 0.1116 |
| **EuroSAT** | 0.9995 | 0.0002 | 0.9374 | 0.0348 | 0.9781 | 0.0192 | 0.5037 | 0.1208 |
| **FGVC_Aircraft** | 0.9994 | 0.0003 | 0.9378 | 0.0341 | 0.5587 | 0.0896 | 0.5822 | 0.1156 |
| **Food101** | 0.9996 | 0.0002 | 0.9364 | 0.0351 | 0.8724 | 0.0665 | 0.499 | 0.1242 |
| **ImageNet** | 0.9995 | 0.0003 | 0.9372 | 0.0344 | 0.9631 | 0.0251 | 0.7942 | 0.0726 |
| **Flowers102** | 0.9995 | 0.0003 | 0.9379 | 0.0341 | 0.9245 | 0.0321 | 0.5887 | 0.1164 |
| **OxfordPets** | 0.9996 | 0.0002 | 0.9383 | 0.034 | 0.9859 | 0.0105 | 0.6971 | 0.0958 |
| **SUN397** | 0.9996 | 0.0002 | 0.9352 | 0.0354 | 0.8904 | 0.0432 | 0.6498 | 0.1064 |
| **UCF101** | 0.9995 | 0.0003 | 0.9395 | 0.0339 | 0.3683 | 0.0894 | 0.613 | 0.1133 |
| **Average** | **0.99952** | 0.00025 | **0.93739** | 0.03451 | **0.8057** | 0.05214 | **0.62988** | 0.10622 |

Table 21. SSIM values for the original SIBA trigger and triggers reconstructed using our method and baselines. SIBA trigger is highly sparse and imperceptible, evident by the 0.99 SSIM score.

| Dataset | WaNet Trigger SSIM (mean) | WaNet Trigger SSIM (std) | CI Recon. Trigger SSIM (mean) | CI Recon. Trigger SSIM (std) | NC Recon. Trigger SSIM (mean) | NC Recon. Trigger SSIM (std) | PixB Recon. Trigger SSIM (mean) | PixB Recon. Trigger SSIM (std) |
|---|---|---|---|---|---|---|---|---|
| **Caltech101** | 0.9312 | 0.0186 | 0.9358 | 0.0354 | 0.4515 | 0.1005 | 0.3998 | 0.1233 |
| **DTD** | 0.9342 | 0.0186 | 0.9372 | 0.0347 | 0.7839 | 0.0608 | 0.4233 | 0.1229 |
| **EuroSAT** | 0.9309 | 0.0186 | 0.9374 | 0.0344 | 0.9588 | 0.0207 | 0.3643 | 0.119 |
| **FGVC_Aircraft** | 0.9233 | 0.0184 | 0.9373 | 0.0345 | 0.6507 | 0.1058 | 0.3748 | 0.1213 |
| **Food101** | 0.9473 | 0.0189 | 0.9376 | 0.0346 | 0.8571 | 0.0648 | 0.6744 | 0.1013 |
| **ImageNet** | 0.9512 | 0.0190 | 0.9356 | 0.0354 | 0.784 | 0.0798 | 0.3707 | 0.1211 |
| **Flowers102** | 0.9277 | 0.0185 | 0.9355 | 0.0356 | 0.7949 | 0.0586 | 0.399 | 0.1235 |
| **OxfordPets** | 0.9592 | 0.0191 | 0.936 | 0.0352 | 0.7422 | 0.09 | 0.7226 | 0.0902 |
| **SUN397** | 0.9342 | 0.0186 | 0.9356 | 0.0356 | 0.8134 | 0.0749 | 0.5623 | 0.1176 |
| **UCF101** | 0.9421 | 0.0188 | 0.9375 | 0.0346 | 0.5323 | 0.1079 | 0.4257 | 0.1231 |
| **Average** | **0.9381** | 0.0187 | **0.93655** | 0.035 | **0.73688** | 0.07638 | **0.47169** | 0.11633 |

Table 22. SSIM values for the original WaNet trigger and triggers reconstructed using our method and baselines. We apply a $k = 4$ control-grid distortion with noise strength $s = 0.1$.