Mechanistic Detection of Patch-Based Backdoors in CNNs via Dispersion-Guided Localization

Md. Masum Al Masba
Department of Computer Science
Florida State University
Tallahassee, United States
ma22be@fsu.edu

Mao Nishino
Department of Mathematics
Florida State University
Tallahassee, United States
mnishino@fsu.edu

Xiuwen Liu
Department of Computer Science
Florida State University
Tallahassee, United States
xliu@fsu.edu

***Abstract*—Backdoor attacks pose a serious threat to the reliability of deep neural networks by embedding hidden triggers that induce targeted misclassification at inference time. We propose a gradient-free, post-hoc framework for localizing patch-based backdoor triggers based on latent representation variance. Our key observation is that effective triggers induce background-invariant internal representations, leading to systematic variance collapse across heterogeneous inputs. Exploiting this phenomenon, our method localizes the spatial position of a trigger using only a single poisoned input and a small clean verification set, without access to gradients, retraining, or clean training data. We evaluate our approach on standard architectures (ResNet-18, WideResNet) and datasets (CIFAR-10, GTSRB), demonstrating accurate and robust localization across different trigger sizes. Empirical results reveal consistent variance collapse in deeper layers of poisoned models, aligning with recent theoretical analyses of backdoor learning dynamics.**

***Index Terms*—Backdoor attacks, representation dispersion, feature invariance, adversarial patch localization, model robustness**

I. Introduction

Deep convolutional neural networks (CNNs) have become foundational in numerous AI applications [1], ranging from autonomous driving to biometric authentication. However, their vulnerability to *backdoor attacks*, a subclass of data poisoning attacks in which malicious triggers are embedded into training samples, raises serious concerns for real-world deployment [2]–[4]. Such attacks cause models to behave normally on clean inputs while consistently misclassifying trigger-bearing inputs into attacker-specified target classes.

Existing defenses often rely on strong assumptions, including access to clean validation data, prior knowledge of trigger properties, or white-box access to model gradients and predictions [5]–[7]. These requirements are frequently impractical in black-box or post-deployment settings. Moreover, precise *spatial localization* of backdoor triggers remains underexplored, despite its importance for forensic analysis and targeted model repair [8], [9].

In this work, we propose a *mechanistic, post hoc* localization framework grounded in the concept of **Representation Dispersion**. Our hypothesis is that a strong trigger induces background invariant feature representations in CNNs, thereby collapsing the variance across heterogeneous semantic contexts. In contrast, benign features vary significantly with background. We exploit this contrast to pinpoint the spatial coordinates of the trigger using only a single poisoned input and a clean verification set without requiring access to gradients, retraining, or population level statistics.

Our contributions are:

- We define a theory driven variance metric, **Representation Dispersion**, to quantify latent feature stability across layers, inspired by ideas in neural representation dynamics.
- We introduce a gradient free patch scanning algorithm that localizes backdoor triggers by exploiting background invariant representation stability.
- We empirically validate our approach across models (ResNet-18, WideResNet) and datasets (CIFAR-10, GTSRB), demonstrating consistent localization across different patch sizes (3×3, 5×5).
- We provide mechanistic insight into layerwise variance collapse, showing both theoretically and empirically that poisoned samples induce increasingly invariant representations in deeper layers, consistent with recent statistical analyses of backdoor behavior .

We emphasize that our framework targets fixed-location patch-based triggers and is designed for post-hoc forensic analysis rather than adaptive adversarial settings. Within this scope, our results demonstrate that representation-level variance provides a robust, interpretable, and computationally lightweight signal for backdoor trigger localization, bridging recent theoretical insights on backdoor learning with practical defense mechanisms.

II. Related Work

Backdoor attacks embed malicious patterns into model behavior by manipulating training data, resulting in targeted misclassification when a specific trigger is present [2], [3]. A wide range of defense strategies have been proposed, which broadly fall into three categories: (i) statistical and black-box detection methods, (ii) reverse-engineering and optimization-based defenses, and (iii) theoretical analyses of neural representations [4], [10]. In contrast to these paradigms, our work introduces a mechanistic detection framework grounded in feature invariance and variance dynamics of latent representations.

A. Statistical and Black-box Detection Methods

A major class of backdoor defenses detects malicious behavior by analyzing statistical irregularities in model outputs, internal activations, or explanation signals, often without explicit trigger reconstruction [4], [10].

Jiang et al. [8] proposed a critical path-based backdoor detection method that identifies trigger-sensitive neuron pathways using trainable control gates. While effective, CPBD requires white-box access and gradient-based analysis, and does not provide explicit spatial localization of the trigger pattern.

Fu et al. [9] introduced a black-box detection approach based on hand-crafted behavioral metrics and synthetic perturbations. Although data-efficient, the method relies on auxiliary novelty detectors and does not explicitly localize the trigger.

AEVA [6] detects backdoored models through extreme-value statistical analysis of adversarial perturbation maps, but incurs high query cost and operates at the model level without trigger localization. NeuronInspect [7] similarly analyzes statistical anomalies in explanation heatmaps to identify attack targets rather than precise trigger locations.

In contrast, our approach exploits a representation-level signal, i.e., systematic layerwise variance collapse induced by backdoor triggers. Using only forward-pass statistics, a single poisoned input, and a clean verification set, our method enables direct spatial localization without gradients or auxiliary optimization.

B. Reverse Engineering and Surrogate Optimization

Another class of backdoor defenses formulates detection as a reverse-engineering task, aiming to reconstruct trigger patterns that reliably induce misclassification.

Neural Cleanse [5] detects backdoors by optimizing minimal class-specific perturbations, but requires exhaustive per-class optimization and full access to model outputs. FeatureRE [11] extends this paradigm by reconstructing triggers via feature-space orthogonality constraints, enabling detection of both static and dynamic triggers at the cost of gradient-based optimization and per-class search. UNICORN [12] further generalizes trigger inversion through constrained optimization over trigger patterns and masks.

By comparison, our method avoids trigger synthesis and adversarial optimization entirely. Instead, it leverages a mechanistic representation-level signal, i.e., background-invariant feature stability induced by triggers, to localize backdoors in a purely post-hoc, inference-time manner.

C. Theoretical Insights and Neural Representations

Recent work has sought to explain the success of backdoor attacks from a theoretical perspective. Wang et al. [13] modeled poisoning as a statistical process that drives learned representations toward low-variance attractors, while Zhang et al. [14] showed that backdoor features tend to occupy orthogonal subspaces relative to benign features. Li and Liu [15] further identified key conditions for successful backdoor attacks, including the number of feature vectors, the norm ratio between trigger and benign features, and the poisoning ratio.

These findings are closely related to representation learning phenomena such as neural collapse [16], where deep networks form highly clustered embeddings in late training stages, as well as causal abstraction views of neural computation via Neuron Dependency Graphs [17]. Unlike neural collapse, which describes class-level embedding convergence, our analysis focuses on patch-induced invariance across heterogeneous backgrounds.

Guided by the theory of Li and Liu [15], we show that poisoned samples exhibit systematically smaller representation variance than clean samples in an idealized two-layer CNN, and empirically demonstrate pronounced variance collapse in deeper layers of modern architectures. This representation-level perspective provides a mechanistic and measurable link between backdoor theory and practical trigger localization.

D. Positioning of Our Work

Our method differs from prior backdoor defenses in three key aspects:

- **Data Efficiency:** It requires only a single poisoned sample and a clean verification set, without population-level statistics or multiple poisoned inputs.
- **Gradient-Free Inference:** The approach is fully post-hoc, relying solely on forward-pass representations without retraining or gradient access.
- **Direct Spatial Localization:** The proposed dispersion metric enables explicit spatial localization of the trigger, rather than model-level detection alone.

To the best of our knowledge, prior work has not explicitly leveraged background-invariant representation stability as a practical signal for direct spatial backdoor trigger localization, offering a lightweight and mechanistically interpretable defense against patch-based attacks.

III. Methodology

The detection and localization of backdoor triggers in convolutional neural networks (CNNs) remain challenging. Prior defenses largely rely on statistical analysis of model responses or neuron activations, such as Critical Path-based Backdoor Detection (CPBD) [8] and Differential Trigger Analysis [9], which typically rely on influence patterns aggregated across multiple inputs or validation samples. While effective in certain settings, these methods typically depend on nontrivial amounts of clean data, population-level behavioral statistics, or gradient-based analysis, and do not explicitly localize trigger regions. In contrast, we propose a post-hoc, representation-based framework that localizes hidden triggers using only a single poisoned input and a clean verification set. Our method exploits a fundamental signal layerwise variance collapse revealing background-invariant latent activations induced by backdoor triggers, without requiring retraining, gradients, or handcrafted metrics. This phenomenon is theoretically supported by recent analyses of backdoor learning [13], [15], [18].

A. Backdoor Model Specification and Training

We consider a convolutional neural network (CNN) classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ denotes the input manifold and $\mathcal{Y}$ represents the label space. The model is optimized on a poisoned dataset $\mathcal{D}_p$, which is characterized by a specific learning rate η and a poisoning ratio ρ. The latter represents the fraction of training samples modified by a trigger function $T(x, p, (r, c))$. The objective of the training process is to ensure the model converges to a target class $y_t \in \mathcal{Y}$ whenever the trigger pattern is present, while maintaining standard empirical risk performance on clean samples.

B. Trigger Patch Localization Algorithm

We propose a dispersion-based localization framework designed to identify the spatial coordinates and pixel distribution of the backdoor trigger through background-invariant embedding stability.The core theoretical premise is *Feature Invariance*: a robust backdoor trigger tends to concentrate latent representations into a background-invariant region of the feature space that is invariant to the semantic background image I.

1) Candidate Patch Space Construction: Given a single image I_p known to contain the trigger, we define a candidate space P of all possible sub-regions. Utilizing a sliding window of dimensions $k \times k$ and a unit stride, we generate a set of candidate patches $P = \{p_1, p_2, \ldots, p_n\}$. Each candidate p_i is mathematically defined by its pixel values and its original spatial anchor (r_i, c_i) within the source image.

2) Background-Invariant Verification: To isolate the influence of the trigger from the underlying semantic content, we utilize a verification set $V = \{I_1, \ldots, I_m\}$ consisting of clean samples where $f_\theta(I_j) \neq y_t$. For every candidate patch p_i extracted from the source poisoned image at coordinates (r_i, c_i), we construct a set of synthetic samples $I'_{i,j}$ via *spatial-preservation mapping*. Specifically, we inject the candidate p_i onto the j-th background image at the **identical coordinates** (r_i, c_i) used during its original extraction:

$$I'_{i,j} = \text{Inject}(I_j, p_i, (r_i, c_i)), \tag{1}$$

This procedure ensures that if the underlying backdoor is location-sensitive, our verification framework preserves the spatial activation signal required to trigger the model's malicious state transitions.

3) Representation Dispersion Analysis: Let $\phi(x)$ denote the mapping of an input x to the high-dimensional embedding space $\mathbb{R}^D$ at the penultimate layer of f_θ. For each candidate p_i, we extract the set of embeddings $\{\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,m}\}$ where $\mathbf{e}_{i,j} = \phi(I'_{i,j})$. We quantify the stability of a patch using a dispersion metric S_i, calculated as the sum of squared distances from the centroid of the embeddings:

$$\bar{\mathbf{e}}_i = \frac{1}{m} \sum_{j=1}^{m} \mathbf{e}_{i,j}; \tag{2}$$

$$S_i = \sum_{j=1}^{m} \|\mathbf{e}_{i,j} - \bar{\mathbf{e}}_i\|_2^2. \tag{3}$$

Algorithm 1 Backdoor Trigger Localization via Representation Dispersion Analysis

Require: Backdoored model f_θ, poisoned image I_p, verification set $V = \{I_1, \ldots, I_m\}$, patch size k, stride s
Ensure: Optimal trigger patch p^* and coordinates (r^*, c^*)

Step 1: Candidate Generation
Extract candidate set $P = \{p_1, p_2, \ldots, p_n\}$ from I_p using $k \times k$ sliding window with stride s
Store corresponding spatial anchors $\{(r_1, c_1), \ldots, (r_n, c_n)\}$
Step 2: Cross-Background Latent Analysis
for each candidate patch $p_i \in P$ **do**
  for each clean background image $I_j \in V$ **do**
    $I'_{i,j} \leftarrow \text{Inject}(I_j, p_i, (r_i, c_i))$
    $\mathbf{e}_{i,j} \leftarrow \phi(I'_{i,j})$
  end for
end for
Step 3: Variance Minimization
for each candidate patch $p_i \in P$ **do**
  $\bar{\mathbf{e}}_i \leftarrow \frac{1}{m} \sum_{j=1}^{m} \mathbf{e}_{i,j}$
  $S_i \leftarrow \sum_{j=1}^{m} \|\mathbf{e}_{i,j} - \bar{\mathbf{e}}_i\|_2^2$
end for
Step 4: Identification
$p^* \leftarrow \arg\min_{p_i \in P} S_i$
return $p^*, (r^*, c^*)$

Throughout the paper, we use *representation dispersion* and *embedding variance* interchangeably, with dispersion referring to the conceptual phenomenon and variance referring to its mathematical implementation.

4) Optimal Trigger Identification: The actual trigger p^* is identified as the patch that minimizes the dispersion score:

$$p^* = \arg\min_{p_i \in P} S_i. \tag{4}$$

Theoretically, p^* represents the patch that induces the most background-invariant latent representation, as it minimizes the variance of the model's internal representation across heterogeneous background distributions.

5) Algorithm Summary: As described in Algorithm 1, the localization process identifies the trigger by minimizing representation dispersion. **Input:** backdoor model f_θ, poisoned image I_p, clean background set V, patch size k, stride s
Output: Localized trigger patch p^* and coordinates (r^*, c^*)

1) Generate candidate patch set P using a $k \times k$ sliding window on I_p.
2) For each $p_i \in P$, overlay it on each background $I_j \in V$ at (r_i, c_i), creating $I'_{i,j}$.
3) Extract penultimate layer embeddings $\phi(I'_{i,j})$ and compute the variance S_i for each candidate.
4) Return the patch p^* that minimizes S_i.

C. Implementation Framework

The methodology is implemented as a post-hoc analysis tool. To analyze the hierarchical activation of the backdoor,

the dispersion metric is computed across multiple hidden layers $\phi_l(x)$, allowing for the observation of trigger-specific feature emergence during forward propagation. The search is optimized using localized batched embedding extraction and vectorized variance computation to enable high-throughput evaluation of the candidate space without the need for gradient computation.

D. Theoretical results

In this section, we will explain our empirical findings by theoretical analysis. In particular, we will explain our variance collapse result i.e. that the variance of the penultimate layer representation across poisoned samples are:

1) Smaller than that of clean samples, and
2) Drops faster than that of clean samples throughout the layers. In other words, variance decrease between layers are larger for poisoned samples.

Let us first consider the first point. It is intuitively obvious that this is the case, since the poisoned samples all contain the same trigger pattern at the same location, so the variance of the poisoned images are low in the first place. Indeed, if our model is linear, this is enough to show the first point:

Proposition 1. *Let $f : \mathbb{R}^d \to \mathbb{R}^k, f(x) = Ax$ be a linear map for some positive integers d, k and a $k \times d$ matrix A. Consider a random image X taken from a fixed distribution. Construct a patched image X_{po} by replacing the components at fixed locations by a fixed vector v. Then, we have*

$$E\|f(X_{po}) - E[f(X_{po})]\|_2^2 \leq E\|f(X) - E[f(X)]\|_2^2. \quad (5)$$

The proof is elementary and thus omitted. This result shows that for linear models, the variance decrease is guaranteed by the patch attack. However, for nonlinear models, this is not necessarily the case. To illustrate, consider a "two-pixel" image $X = (X_1, X_2)$ where each pixel is sampled independently from $\mathcal{N}(0, 1)$. Consider the following ReLU-like nonlinear model:

$$f(x_1, x_2) = 1_{\{x_1 > 0\}} x_2. \quad (6)$$

where $1_{x_1>0}$ is the indicator function that is 1 if $x_1 > 0$ and 0 otherwise. Now, consider the patched image X_{po} where we fix the first pixel to be 1. Then, we have the following:

Proposition 2. *For the above nonlinear model f and image distribution, we have*

$$E|f(X_{po}) - E[f(X_{po})]|^2 > E|f(X) - E[f(X)]|^2. \quad (7)$$

In particular, $E|f(X_{po}) - E[f(X_{po})]|^2 = 1$ and $E|f(X) - E[f(X)]|^2 = 1/2$.

The proof is elementary and thus omitted. This result shows that for nonlinear models, the variance can actually increase after patching. Therefore, to show the variance collapse, we need a careful argument that takes into account the specific model structure. To this end, we consider the 2-layer CNN model used by [15]. We will briefly review their model and then present our theoretical results. We will consider a binary classification problem where X is the set of samples and $Y = \{-1, 1\}$ is the set of labels. As in [15], we consider the multi-view data model in [18]. Each $x \in X$ consists of π patches $x = (x^1, \cdots, x^\pi)$ where each patch $x^i \in \mathbb{R}^d$. There are three types of patches:

1) **The feature patch**. We assume there is a dictionary of orthonormal vectors $\{u^1, \cdots u^K\}$, and a feature patch is a patch that is equal to yu^i where y is a label of the sample x.
2) **The main noise patch**. A main noise patch is a patch that is sampled from a distribution $\mathcal{N}(0, \frac{\sigma_\xi^2}{d} I_d)$ for a fixed constant $\sigma_\xi > 0$. It represents the main distractor in the image.
3) **The background noise patch**. A background noise patch is a patch that is sampled from a distribution $\mathcal{N}(0, \frac{\sigma_\zeta^2}{d} I_d)$ for $\sigma_\zeta = \frac{\sigma_\xi}{\pi}$.

A sample is genererated by choosing u^i from uniform distribution over $\{u^1, \cdots u^K\}$, and put it arbitrarily along with one main noise patch and $\pi - 2$ background noise patches, and assigning a label uniformly chosen. We generate n samples this way to form the clean training data S_{cl}^{tr}. To create the backdoor poisoned training data S_{po}^{tr}, we fix n_{po} the number of poisoned samples, the poison patch v, the location of the poison p_v, and the target label y^p, and follow the following procedure:

1) Randomly select indices S_b with $|S_b| = n_{po}$ from $\{1, \cdots, n\}$.
2) For each $i \in S_b$, we replace the p_vth patch of x_i with the poison patch v and set its label to y^p.

As in [15], we assume that feature patches, main noise patches and the poison patch are all different. We now consider the following 2-layer CNN model with C channels:

$$F(x) = \sum_{c=1}^{C} \sum_{p=1}^{\pi} \phi(\langle w_c, x^p \rangle). \quad (8)$$

where ϕ is the symmetrized smooth ReLU:

$$\phi(z) = \begin{cases} \frac{1}{3} z^3 & |z| \leq 1 \\ z - \frac{2}{3} & z > 1 \\ z + \frac{2}{3} & z < -1 \end{cases}. \quad (9)$$

This choice is common in theoretical analysis of neural networks [15], [18]. We will train this network on S_{po}^{tr} by minimizing the logistic loss $\ell(F(x), y) = \log(1 + \exp(-yF(x)))$ using gradient descent with learning rate η. We denote the parameters at step t by $w_c^{(t)}$. For the CNN $F_T(x)$ defined above trained up to the Tth gradient step, consider the intermediate representation map $R_T : \mathbb{R}^{d \times \pi} \to \mathbb{R}^{C \times \pi}$ defined as

$$R_T(x) = \left(\phi(\langle w_c^{(T)}, x^p \rangle) \right)_{c,p}. \quad (10)$$

Our claim is that the variance of the intermediate representation across poisoned samples collapse faster than clean sam-

ples. More precisely, for a set of vectors $S = \{z_1, \cdots, z_n\}$, we define the variance of S as

$$\mathrm{Var}(S) = \frac{1}{n}\sum_{i=1}^{n} \|z_i - \bar{z}\|_2^2. \tag{11}$$

where $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Define the set of poisoned samples $\mathcal{P} = \{x_i \in S_{po}^{tr} : i \in S_b\}$ and the set of clean samples $\mathcal{C} = \{x_i \in S_{po}^{tr} : i \notin S_b\}$. We define the **variance drop** of a set S at step T as

$$\Delta_S(T) = \mathrm{Var}(R_T(S)) - \mathrm{Var}(S). \tag{12}$$

We then have the following theorem:

Theorem 1 (Variance drop under a patch backdoor)**.** *Assume Condition 5.1 of [15], the success condition $n_{\mathrm{po}}\|v\|_2^2 > \omega(n/K)$ [15, Thm. 5.4], and assume the following:*

1) *(The noise shrinks faster than $(\log d)^{3/2}$) $\sigma_0\sigma_\xi(\log d)^{3/2} = o(1)$; $\sigma_0\sigma_\zeta(\log d)^{3/2} = o(1)$.*
2) *(Sample size grows at least polynomially) $n \geq \omega(d^\alpha)$ for some $\alpha > 0$.*
3) *(η, K grows at most polylogarithmically) $\eta, K \leq O((\log d)^\beta)$ for some $\beta > 0$.*
4) *(The training horizon is large enough and is at most polynomial) $T \geq T_u = \widetilde{\Theta}\big((K + Ke^{C-2})/(\eta\sigma_0)\big)$ and $T \leq O(d^\gamma)$ for some $\gamma > 0$.*

Then, with probability at least $1 - O(d^{-2})$ over the draw of the clean training data and the poisoned subset P, the variance of the poisoned set is smaller than that of the clean set:

$$\mathrm{Var}\big(R_T(\mathcal{C})\big) > \mathrm{Var}\big(R_T(\mathcal{P})\big).$$

If in addition, we have:

5) *(The network is wide enough, and there are enough features so that samples have high enough variance) $\frac{1}{K} + \frac{\sigma_\xi^2}{\pi^2} = o\Big(\frac{1}{C}\Big)$,*

We have the following variance drop comparison:

$$\Delta_{\mathcal{P}}(T) \;<\; \Delta_{\mathcal{C}}(T).$$

We prove the theorem in Appendix A. We remark that all of the asymptotics in the theorem are considered with respect to d, the feature dimension. That is, a statement like $f \leq O(g)$ is considered for functions of d. We will now state our result. The second part of Theorem 1 shows that that, if the variances fall (the variance drops are negative), then $|\Delta_{\mathcal{P}}(T)| > |\Delta_{\mathcal{C}}(T)|$, i.e., the variance drops "harder" for the poisoned samples than clean samples. On the other hand, if the variances do rise (the variance drops are positive), then $\Delta_{\mathcal{P}}(T) < \Delta_{\mathcal{C}}(T)$, i.e., the variance rises "less" for the poisoned samples than for the clean samples. The assumption to show the second part might seem cryptic, but intuitively, this assumption ensures that the initial variance is high enough to drop. Otherwise, the initial variance is already too low that it would not drop any further. As noted in the theorem, we assume Condition 5.1 in [15] to ensure that the backdoor attack is successful. For completeness, we list the condition here:

Condition 1 (Condition 5.1 of [15])**.** 1) *The feature vectors $\{u_k\}$ and the trigger vector v are mutually orthogonal and u_k are unit vectors.*
2) *$C = \Theta(\log d)$.*
3) *The network is over-parametrized i.e. $n\pi^2 K \leq o(\sqrt{d})$.*
4) *$\sigma_0 \leq o(1)$ and $\omega(1) \leq \sigma_\xi \leq o(1/\sigma_0)$.*
5) *The size of the training set is larger than the number of useful features, i.e., $n \geq \Omega(\sigma_0^{-3}K)$. Moreover, in backdoor learning, the training set only contains a small proportion of poisoned data points i.e. $n_{po} \leq o(n)$. The norm of trigger vectors satisfies $\Omega(1) \leq \|v\|_2 \leq O(\sigma_\xi)$.*

IV. Experimental Evaluation

A. Experimental Setup

We evaluate our proposed mechanistic detection framework on standard computer vision benchmarks using two representative convolutional neural network (CNN) architectures. Specifically, we utilize the CIFAR-10 dataset with a ResNet-18 [19] backbone and the GTSRB dataset with a WideResNet (WRN) [20] architecture.

Attack Configuration: We consider traditional patch-based backdoor attacks characterized by fixed spatial trigger locations and a static target label. To assess the robustness of our detection across scales, we evaluate two distinct trigger sizes: 3×3 and 5×5 pixels. In all poisoning scenarios, the poisoning rate is fixed at 10% of the training set.

Training Protocol: Models are optimized using Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01.

Implementation Details: We used Python v3.10.12 on Arch Linux with one NVIDIA RTX A5000 GPUs. Hardware drivers included NVIDIA v560.35.05 and CUDA 12.6.

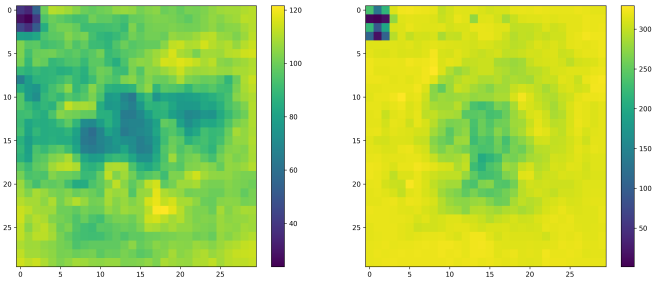
To quantify representation dispersion, we use the Representation Dispersion (RD). Given a candidate patch, we insert it into multiple clean background images and compute the variance of the resulting representations. Lower dispersion indicates stronger invariance of the representation to background context, which is a characteristic signature of a backdoor trigger.

B. Trigger Localization via Variance Heatmaps

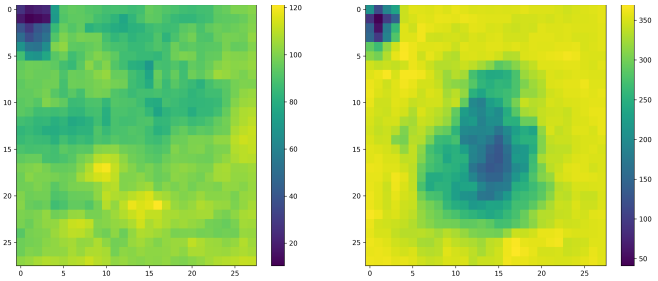
Figure 1 visualize spatial heat maps of the RD obtained by putting candidate patches extracted from a poisoned image in all respective spatial locations of five different images and calculating their variances at the penultimate layer for each patch. In all settings, the true trigger location is consistently identified as the region with minimum dispersion. This demonstrates that poisoned models exhibit strong background-invariant representations at the trigger location, enabling reliable post-hoc localization without access to training data or gradients. These results confirm that variance-based scanning is an effective and architecture-agnostic trigger localization mechanism.

C. Layerwise Variance Dynamics of the True Trigger Patch

Figure 2 report the layerwise dispersion of the true trigger patch as it propagates through the network. Across both ResNet-18 and WRN, and for both 3×3 and 5×5 triggers, we

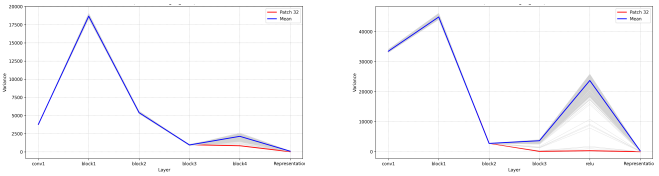


(a) ResNet-18, 3×3 patch (b) WRN, 3×3 patch

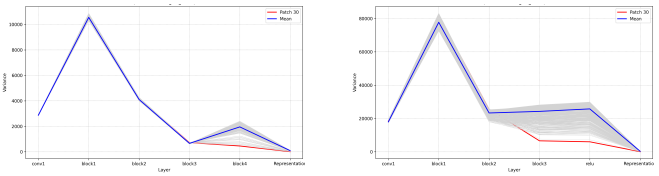


(c) ResNet-18, 5×5 patch (d) WRN, 5×5 patch

Fig. 1. Trigger localization via patch-wise variance heatmaps. The minimum-dispersion region consistently aligns with the true trigger location across architectures and trigger sizes(2nd row, 2nd column is the lowest variance in every image).



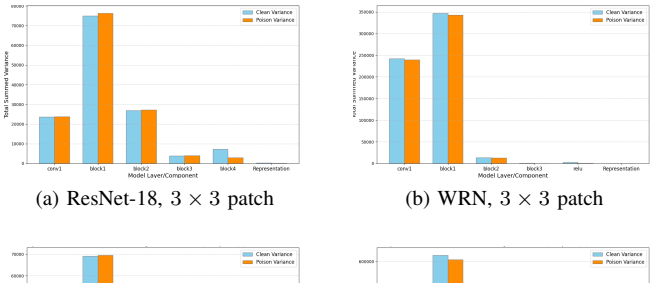
(a) ResNet-18, 3×3 patch (Red Line is the 32nd Patch) (b) WRN, 3×3 patch (Red Line is the 32nd Patch)



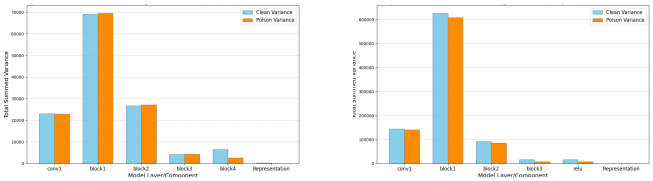
(c) ResNet-18, 5×5 patch (Red Line is the 30th Patch) (d) WRN, 5×5 patch (Red Line is the 30th Patch)

Fig. 2. Layerwise dispersion of the true trigger patch across network depth. Variance consistently decreases with depth for both architectures and trigger sizes, indicating progressively stronger background invariance in deeper representations.

observe a consistent decrease in variance with depth. This indicates that deeper layers progressively amplify trigger-specific features while suppressing background-induced variability. These results empirically support our theoretical prediction that backdoor triggers induce increasingly invariant internal representations as training progresses.



(a) ResNet-18, 3×3 patch (b) WRN, 3×3 patch



(c) ResNet-18, 5×5 patch (d) WRN, 5×5 patch

Fig. 3. Clean vs. poisoned set: layerwise variance collapse across architectures and trigger sizes. Poisoned samples exhibit substantially stronger variance collapse in deeper layers.

D. Clean vs. Poisoned Set: Layerwise Variance Collapse

We now directly compare the *layerwise variance profiles* of clean and poisoned samples across different architectures and trigger sizes. Figure 3 presents the empirical variance trends for ResNet-18 and WideResNet (WRN) under both 3×3 and 5×5 patch-based trigger settings. In all configurations, poisoned samples consistently exhibit *stronger variance collapse* in deeper layers, providing compelling empirical support for our theoretical hypothesis. Quantitative reductions at key layers are summarized in Table I.

a) Patch 3×3*, ResNet-18 (CIFAR-10) — Fig. 3a.:* Clean and poisoned samples display similar variance in early convolutional and shallow residual layers. However, pronounced divergence emerges in deeper layers. As shown in Table I, the variance at the final residual block is reduced by approximately 58.9% (7,356 $\rightarrow$ 3,026), while the penultimate embedding variance collapses by approximately 74.8% (302 $\rightarrow$ 76). This confirms that trigger-induced invariance primarily manifests in deep representations.

b) Patch 3×3*, WRN (GTSRB) — Fig. 3b.:* For WRN, poisoned samples begin to exhibit reduced variance earlier in the network hierarchy. In particular, variance reduction is already visible at intermediate activation stages prior to the final residual block. Table I shows that the final residual block variance is reduced by approximately 74.8% (2,375.4 $\rightarrow$ 597.6), while the embedding layer exhibits an even stronger collapse of approximately 88.5% (12.27 $\rightarrow$ 1.41). This indicates early formation of trigger-dominated, background-invariant features.

c) Patch 5×5*, ResNet-18 (CIFAR-10) — Fig. 3c.:* A similar trend is observed for larger triggers. While shallow layers remain largely unaffected, deeper layers show clear variance collapse under poisoning. As summarized in Table I, the embedding variance is reduced by approximately 43.3% (258.5 $\rightarrow$ 146.7), and the final residual block variance decreases by approximately 32.2% (6,619.9 $\rightarrow$ 4,485.6). This suggests that although the trigger remains effective, the degree

TABLE I
SUMMARY OF LAYERWISE VARIANCE COLLAPSE.

Model	Patch	Layer	Clean Var	Poison Var	Reduction
ResNet-18 (CIFAR-10)	3×3	Last Res Block	7,356	3,026	–58.9%
ResNet-18 (CIFAR-10)	3×3	Embedding	302.0	76.0	–74.8%
WRN (GTSRB)	3×3	Last Res Block	2,375.4	597.6	–74.8%
WRN (GTSRB)	3×3	Embedding	12.27	1.41	–88.5%
ResNet-18 (CIFAR-10)	5×5	Last Res Block	6,619.9	4,485.6	–32.2%
ResNet-18 (CIFAR-10)	5×5	Embedding	258.5	146.7	–43.3%
WRN (GTSRB)	5×5	Last Res Block	16,072.4	7,094.0	–55.9%
WRN (GTSRB)	5×5	Embedding	113.6	31.6	–72.2%

of representation concentration scales sublinearly with patch size.

d) Patch 5×5, WRN (GTSRB) — Fig. 3d.: WRN exhibits consistent and strong variance collapse even under larger trigger sizes. As shown in Table I, the final residual block variance is reduced by approximately 55.9% (16,072.4 $\rightarrow$ 7,094.0), while the embedding layer variance decreases by approximately 72.2% (113.6 $\rightarrow$ 31.6). These results confirm that deep-layer background invariance remains robust across trigger sizes and architectures.

e) Summary and Implications.: Across all architectures and trigger sizes, poisoned samples consistently exhibit substantially smaller variance in deep layers than clean samples, as quantitatively summarized in Table I. Moreover, the relative reduction in variance from intermediate to deep layers is significantly larger for poisoned samples, indicating a faster and stronger variance collapse. These observations are fully consistent with Theorem 1, which predicts that poisoned representations become increasingly concentrated in deeper layers due to trigger-induced feature alignment. The presence of measurable variance reduction even at intermediate stages (e.g., activation layers in WRN) further suggests that trigger-specific invariance emerges progressively throughout the network hierarchy.

E. Connection to Theory

These results strongly support our theoretical analysis in Section III. In particular:

- **Lower variance for poisoned samples in deep layers.** Across all configurations, poisoned samples exhibit substantially smaller variance than clean samples at deeper layers, consistent with Theorem 1.
- **Faster variance collapse for poisoned samples.** The relative reduction in variance from intermediate to deep layers is consistently larger for poisoned samples, indicating a stronger variance drop.
- **Nonlinear effects in shallow layers.** In ResNet-18, shallow layers occasionally exhibit slightly higher variance for poisoned samples, which is consistent with our nonlinear counterexample and highlights that variance collapse is primarily a deep-layer phenomenon.

F. Comparison with Prior Backdoor Detection Methods

We compare our variance-based detection signal with prior backdoor defenses, focusing on Critical Path-Based Backdoor Detection (CPBD) [8]. CPBD detects backdoors by identifying trigger-sensitive neuron paths via trainable control gates and path-level anomaly scores. While effective, it requires auxiliary optimization, per-class analysis, and neuron thresholding, introducing additional computational overhead and limiting scalability.

In contrast, our method operates directly at the representation level, exploiting a fundamental and architecture-agnostic signal: systematic layerwise variance collapse of poisoned samples in deep layers. It requires no control gates, auxiliary training, or neuron pruning, relying solely on forward-pass statistics, making it efficient and post-hoc deployable.

Related black-box methods such as [9] detect backdoors using hand-crafted behavioral metrics derived from synthetic perturbations. While data-efficient, these approaches rely on engineered signals and auxiliary detectors. Our method instead analyzes internal representation dynamics, providing a mechanistic view of where and how trigger-induced invariance emerges, consistent with our theoretical analysis (Theorem 1).

V. DISCUSSION AND LIMITATIONS

This work presents a lightweight and interpretable framework for post-hoc backdoor localization in convolutional neural networks using a single poisoned input and a small clean verification set. The key insight is that effective backdoor triggers induce background-invariant latent representations, which manifest as systematic layerwise variance collapse. Empirical results across multiple architectures and datasets support this hypothesis and align with recent theoretical analyses of backdoor learning.

The proposed method targets patch-based backdoor attacks with fixed spatial placement. Accordingly, adaptive attacks involving randomized locations or spatially distributed triggers may reduce localization accuracy. In addition, while the required clean verification set is small, the approach is not fully unsupervised.

Despite these limitations, the framework offers practical advantages over prior defenses. Unlike statistical and black-box methods such as CPBD [8] and differential analysis [9], our approach requires no gradients, auxiliary optimization, or multiple poisoned samples, and directly localizes the trigger region. These properties make it well suited for post-deployment forensic analysis and targeted model repair.

VI. CONCLUSION AND FUTURE WORK

This paper introduced a variance-based framework for localizing backdoor triggers in convolutional neural networks using only a single poisoned input and a small clean verification set. By leveraging *representation dispersion* as a diagnostic signal, our method provides a gradient-free, post-hoc mechanism for identifying spatial trigger locations. Empirical results demonstrate that effective backdoor triggers induce background-invariant representations and systematic variance collapse in deeper layers, which can be exploited for reliable trigger localization across architectures and datasets.

Future work will explore extending this framework beyond fixed, localized patch triggers. In particular, we plan to investigate robustness against distributed or adaptive trigger patterns, as well as settings with noisy or partially trusted verification data. Additionally, leveraging variance-based signals for backdoor mitigation and model repair, rather than localization alone, remains an important direction. More broadly, applying representation-level variance analysis to other poisoning paradigms, including dynamic patches and trigger-free attacks, presents promising opportunities for future research.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain. arxiv 2017," *arXiv preprint arXiv:1708.06733*, 2017.

[3] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[4] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint arXiv:2007.10760*, 2020.

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

[6] J. Guo, A. Li, and C. Liu, "Aeva: Black-box backdoor detection using adversarial extreme value analysis," *arXiv preprint arXiv:2110.14880*, 2021.

[7] X. Huang, M. Alzantot, and M. Srivastava, "Neuroninspect: Detecting backdoors in neural networks via output explanations," *arXiv preprint arXiv:1911.07399*, 2019.

[8] W. Jiang, X. Wen, J. Zhan, X. Wang, Z. Song, and C. Bian, "Critical path-based backdoor detection for deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 4032–4046, 2022.

[9] H. Fu, P. Krishnamurthy, S. Garg, and F. Khorrami, "Differential analysis of triggers and benign features for black-box dnn backdoor detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4668–4680, 2023.

[10] M. A. Hanif, N. Chattopadhyay, B. Ouni, and M. Shafique, "Survey on backdoor attacks on deep learning: Current trends, categorization, applications, research challenges, and future prospects," *IEEE Access*, 2025.

[11] Z. Wang, K. Mei, H. Ding, J. Zhai, and S. Ma, "Rethinking the reverse-engineering of trojan triggers," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9738–9753, 2022.

[12] Z. Wang, K. Mei, J. Zhai, and S. Ma, "Unicorn: A unified backdoor trigger inversion framework," *arXiv preprint arXiv:2304.02786*, 2023.

[13] G. Wang, X. Xian, J. Srinivasa, A. Kundu, X. Bi, M. Hong, and J. Ding, "Demystifying poisoning backdoor attacks from a statistical perspective," *arXiv preprint arXiv:2310.10780*, 2023.

[14] K. Zhang, S. Cheng, G. Shen, G. Tao, S. An, A. Makur, S. Ma, and X. Zhang, "Exploring the orthogonality and linearity of backdoor attacks," in *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2105–2123, IEEE, 2024.

[15] B. Li and W. Liu, "A theoretical analysis of backdoor poisoning attacks in convolutional neural networks," in *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, eds.), vol. 235 of *Proceedings of Machine Learning Research*, pp. 28309–28342, PMLR, 21–27 Jul 2024.

[16] V. Kothapalli, "Neural collapse: A review on modelling principles and generalization," *arXiv preprint arXiv:2206.04041*, 2022.

[17] Y. Hu and J. Tian, "Neuron dependency graphs: A causal abstraction of neural networks," in *International Conference on Machine Learning*, pp. 9020–9040, PMLR, 2022.

[18] R. Shen, S. Bubeck, and S. Gunasekar, "Data augmentation as feature manipulation," in *International conference on machine learning*, pp. 19773–19808, PMLR, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[20] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

APPENDIX

A. Proof sketch of Theorem 1.

Proof. Let y_p be the target label and let $m_\pm := \#\{i : y_i = \pm y_p\}$ in the *clean* sample (before poisoning), so $n = m_+ + m_-$. The patch attack selects exactly n_{po} points from the non-target class $(-y_p)$, moves them to $\mathcal{P}$, and relabels them to y_p [15, Def. 3.2]. Hence $|\mathcal{C}| = n - n_{\mathrm{po}}$ and the clean-set label bias equals $\pi_{\mathcal{C}} = \frac{\#\{i\in\mathcal{C}:y_i=y_p\}-\#\{i\in\mathcal{C}:y_i=-y_p\}}{|\mathcal{C}|} = \frac{(m_+-m_-)+n_{\mathrm{po}}}{n-n_{\mathrm{po}}}$ so $|\pi_{\mathcal{C}}| \le \frac{|m_+-m_-|}{n-n_{\mathrm{po}}} + \frac{n_{\mathrm{po}}}{n-n_{\mathrm{po}}}$. Under the model [15, Def. 3.1], $m_+ \sim \mathrm{Bin}(n, 1/2)$, so Hoeffding gives $|m_+ - m_-| = O(\sqrt{n \log n})$ w.h.p., and with $n_{\mathrm{po}} = o(n)$ we get $|\pi_{\mathcal{C}}| = O(n_{\mathrm{po}}/n) = o(1)$.

Recall $R_T(x)_{c,p} = \varphi(\langle w_c(T), x_p\rangle)$ and $\mathrm{Var}(R_T(S)) = \sum_{c,p}\big(\mathbb{E}_S R_{c,p}^2 - (\mathbb{E}_S R_{c,p})^2\big)$. The trigger patch location p_v is constant on $\mathcal{P}$ (equal to v), so it contributes 0 to $\mathrm{Var}(R_T(\mathcal{P}))$; on $\mathcal{C}$ it is background noise. Moreover, for $T \ge T_u$, noise coordinates are negligible versus feature coordinates (as in [15, Lem. B.2,B.4]), so all non-feature, non-trigger patches contribute only $o(1)$ to the variance gap. Thus the leading contribution comes from the unique feature location p_u.

At p_u, if x carries feature $y\,u_k$, then by oddness of φ, $R_T(x)_{c,p_u} = \varphi(\langle w_c(T), y u_k\rangle) = y\,A_{c,k}$, $\quad A_{c,k} := \varphi(\langle w_c(T), u_k\rangle)$. For $S \in \{\mathcal{C}, \mathcal{P}\}$ let α_k^S be the empirical fraction of feature u_k in S; for $\mathcal{C}$ also let $\beta_k^{\mathcal{C}}$ be the empirical mean label within feature-k points, so that $\mu_c^{(\mathcal{C})} = \sum_k \alpha_k^{\mathcal{C}}\beta_k^{\mathcal{C}} A_{c,k}$ and $\mathrm{Var}_{\mathcal{C}}(R_{c,p_u}) = \sum_k \alpha_k^{\mathcal{C}} A_{c,k}^2 - (\mu_c^{(\mathcal{C})})^2$, while $\mathrm{Var}_{\mathcal{P}}(R_{c,p_u}) = \sum_k \alpha_k^{\mathcal{P}} A_{c,k}^2 - \big(\sum_k \alpha_k^{\mathcal{P}} A_{c,k}\big)^2$. Multinomial/label Hoeffding implies (uniformly over k) $\alpha_k^{\mathcal{C}} = \alpha_k^{\mathcal{P}} = 1/K + o(1)$ and $\beta_k^{\mathcal{C}} = \pi_{\mathcal{C}} + o(1)$ w.h.p. Write $\mu_c := \frac{1}{K}\sum_k A_{c,k}$. Plugging these into the above formulas yields, uniformly in c,

$$\mathrm{Var}_{\mathcal{C}}(R_{c,p_u}) - \mathrm{Var}_{\mathcal{P}}(R_{c,p_u}) \;\ge\; (1-\pi_{\mathcal{C}}^2)\mu_c^2 - o(1)\cdot\frac{1}{K}\sum_k A_{c,k}^2.$$

Summing over c and using late-stage feature learning plus $|\varphi(z)| \le |z| + 1$ (so $\sum_c \frac{1}{K}\sum_k A_{c,k}^2 = \tilde{O}((\log T)^2)$ as in [15, Lem. 6.2]), the error term is $o(1)$ (since $|\mathcal{C}|, |\mathcal{P}|$ grow polynomially), hence $\mathrm{Var}(R_T(\mathcal{C})) - \mathrm{Var}(R_T(\mathcal{P})) \ge (1 - \pi_{\mathcal{C}}^2)\sum_{c=1}^C \mu_c^2 - o(1)$. Finally, the clean margin decomposition implies $\sum_c A_{c,k} \ge \Omega(1)$ for each k when $T \ge T_u$ [15, Thm. 5.4 and Eq. (131)], so $\sum_c \mu_c = \frac{1}{K}\sum_k \sum_c A_{c,k} \ge \Omega(1)$ and by Cauchy–Schwarz $\sum_c \mu_c^2 \ge \frac{1}{C}\big(\sum_c \mu_c\big)^2 = \Omega(1/C)$. Therefore $\mathrm{Var}(R_T(\mathcal{C})) > \mathrm{Var}(R_T(\mathcal{P}))$ for large d.

For the variance-drop comparison, the only raw-patch variance differences are at p_u and p_v: $\mathrm{Var}(\mathcal{C}) - \mathrm{Var}(\mathcal{P}) = \frac{1-\pi_{\mathcal{C}}^2}{K} + \frac{\sigma_\xi^2}{\pi^2} + o(1)$, so $\Delta_{\mathcal{C}}(T) - \Delta_{\mathcal{P}}(T) = \big(\mathrm{Var}(R_T(\mathcal{C})) - \mathrm{Var}(R_T(\mathcal{P}))\big) - \big(\mathrm{Var}(\mathcal{C}) - \mathrm{Var}(\mathcal{P})\big) \ge (1 - \pi_{\mathcal{C}}^2)\Big(\Omega(\frac{1}{C}) - \frac{1}{K}\Big) - \frac{\sigma_\xi^2}{\pi^2} + o(1)$, which is > 0 under $\frac{1}{K} + \frac{\sigma_\xi^2}{P^2} = o(\frac{1}{C})$. Hence $\Delta_{\mathcal{P}}(T) < \Delta_{\mathcal{C}}(T)$. □