Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mitigation via Adaptive Decomposition (MAD): Geometry-Guided Subspace Decomposition for Robust Backdoor Defense

Md. Hamid Borkot Tulla

hamidborkot@smail.ntu.edu.cn

Chongqing University of Posts and Telecommunications https://orcid.org/0009-0004-2263-3391

MD Moniur Rahman Ratan

Chongqing University https://orcid.org/0009-0003-9620-1922

Abdullah Hil Safi Sohan

Nantong University https://orcid.org/0009-0009-1704-9648

Research Article

Keywords: Backdoor attacks, Model repair, Low-rank geometry, adaptive attacks, post-training mitigation

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8948435/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Mitigation via Adaptive Decomposition (MAD): Geometry-Guided Subspace Decomposition for Robust Backdoor Defense

MD Hamid Borkot Tulla*

Chongqing University of Posts and Telecommunications, China; hamidborkot@stmail.ntu.edu.cn

MD Moniur Rahman Ratan

Chongqing University, China; moniur_rahman@stu.cqu.edu.cn

Abdullah Hil Safi Sohan

Nantong University, China; safisohan3117@gmail.com

Abstract

Backdoor attacks embed hidden trigger-label associations into deep neural networks, enabling adversarial control while maintaining clean accuracy. Existing post-training defenses face a critical limitation: global methods such as Sharpness-Aware Minimization fail against adaptive attackers that flatten the loss landscape. At the same time, structural approaches sacrifice accuracy through coarse-grained modifications. We propose Mitigation via Adaptive Decomposition (MAD), a post-training defense that leverages the low-rank geometric structure that backdoor triggers imprint in weight space. MAD is based on the observation that backdoor triggers concentrate in low-rank subspaces orthogonal to semantic features. Our method identifies this malicious subspace through eigenanalysis of clean-data gradient covariance, then applies sharpness-aware optimization exclusively within the estimated subspace. Empirically validated across vision benchmarks under adaptive threat models across five architectures and five attacks, including adaptive variants, MAD reduces attack success rates to ≤ 5.12% while preserving clean accuracy within 0.4% of baseline ($p \leq 1.05\times10^{-6}$, n=30). Ablation studies validate the low-rank hypothesis (TSR ≈ 0.98 at K=10), and scaling experiments identify a calibration density threshold of 0.21×10^{-4} clean-samples-to-parameters for robust defense across model scales.

Keywords: Backdoor attacks, Model repair, Low-rank geometry, adaptive attacks, post-training mitigation.

Introduction

Deep neural networks are increasingly deployed in safety-critical applications, including autonomous driving, medical diagnosis, and industrial control systems. When models are sourced from third-party providers, backdoor attacks pose a significant security threat: adversaries embed hidden trigger-label associations during training that remain dormant under normal operation but cause targeted misclassification when specific input patterns appear[12, 19, 39]. These attacks maintain high clean accuracy while enabling precise adversarial control, making detection challenging in practice[7, 8, 32].

Post-training defenses face documented limitations. Global sharpness-aware methods [42] achieve 14-18% ASR against standard triggers, but degrade to 15-21% ASR when adaptive attackers deliberately flatten loss landscapes [30, 31]. Structural approaches such as Fine-Pruning [23], and Adversarial Neuron Pruning (ANP) [37] sacrifice 2-5 percentage points (pp) clean accuracy [11, 18, 41]. Detection methods, including Neural Cleanse [36], reverse-engineering approaches [36] and universal detection frameworks [11] identify poisoned models, but provide limited repair mechanisms[16]. No existing method achieves both < 10% ASR and < 1pp utility loss against adaptive attacks [16]. This performance gap is unacceptable for safety-critical deployments such as autonomous driving or medical

imaging, where regulatory thresholds mandate failure rates below 1% [9], and current defenses cannot guarantee against adversarial manipulation [28].

Backdoor triggers, being low-complexity patterns distinct from semantic features, concentrate in low-rank subspaces of the model's weight space [13], a geometric structure that existing defenses fail to leverage [17]. Current approaches treat backdoors as diffuse anomalies, applying global modifications that either sacrifice accuracy or fail against adaptive attackers who deliberately flatten the global loss landscape. We propose Mitigation via Adaptive Decomposition (MAD). This post-training defense identifies the malicious subspace through eigenanalysis of clean-data gradient covariance and applies sharpness-aware optimization exclusively within this targeted region. By restricting intervention to backdoor-relevant directions while preserving benign features in the orthogonal subspace, MAD maintains robustness against adaptive attacks that evade conventional global defenses. This geometry-guided approach surgically neutralizes malicious functionality without collateral damage to model utility, representing a paradigm shift from indiscriminate to precision defense.

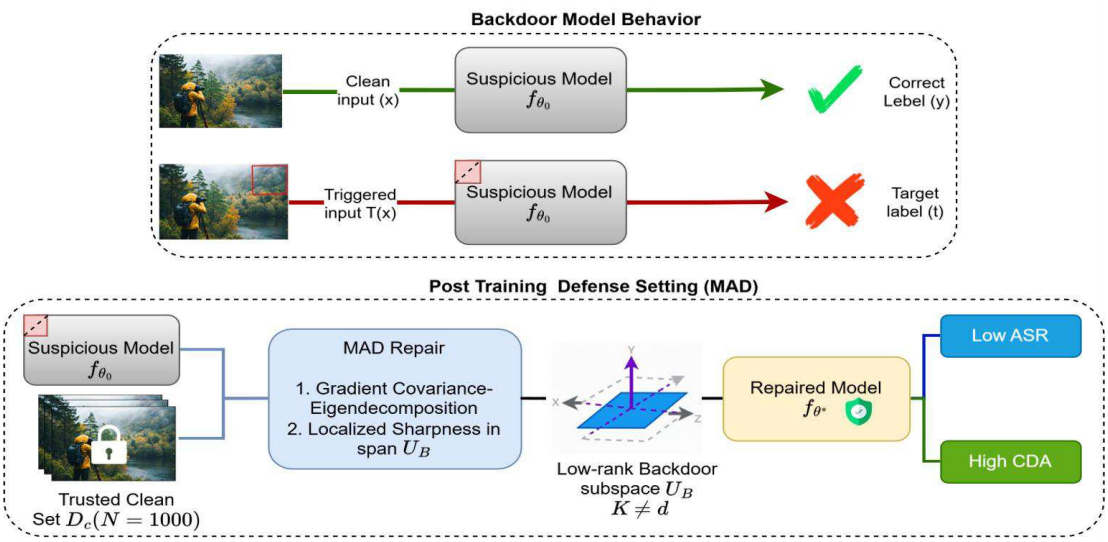


Figure 1.Backdoor Attack and MAD Defense. (Top) Threat model showing trigger-induced misclassification. (Bottom) MAD two-phase repair: eigendecomposition identifies backdoor subspace U_B, then localized sharpness minimization.

We evaluate MAD across CIFAR-10, GTSRB, and TinyImageNet with five architectures (ResNet-18 and ResNet-50, VGG-19, DenseNet-121, ViT-Tiny) against five attack types, including adaptive variants [28, 29]. MAD reduces attack success rates to $\leq 5.12\%$ while maintaining clean accuracy within 0.4% of the unpoisoned baseline ($p \leq 1.05\times10^{-6}$ across 30 seeds). Ablation studies confirm that backdoor signals concentrate in low-rank subspaces with Total Structure Retention (TSR) ≈ 0.98 at rank K=10, and scaling experiments establish a calibration density threshold of 0.21×10^{-4} clean-samples-to-parameters above which the defense maintains effectiveness.

Our work makes three primary contributions:

1. Geometric characterization: We provide theoretical analysis and empirical validation showing that backdoor signals concentrate in low-dimensional subspaces (K=10) across diverse architectures, enabling targeted mitigation strategies.

2. Practical defense mechanism: We develop a post-training defense that exploits this geometric structure to achieve robust performance against both standard and adaptive attacks while preserving model utility.

3. Scaling analysis: We identify a calibration density principle governing defense effectiveness across model capacities, providing concrete guidance for practitioners deploying the method on models of varying sizes

Related Work

Backdoor attacks embed hidden trigger-label associations during training, enabling adversarial control while maintaining clean accuracy[10, 37]. Defenses broadly divide into detection approaches that identify poisoned models or samples[11, 12, 34], and mitigation approaches that neutralize backdoors while preserving utility[21, 35, 40]. We focus on post-training mitigation, in which defenders have access to a potentially backdoored model and a small trusted clean dataset (typically <1% of the training data).

Post-Training Structural Modifications

Early mitigation strategies employ structural modifications to remove backdoor functionality. Fine-Pruning[21] removes neurons with low activation on clean data, followed by fine-tuning, but discards benign capacity and degrades clean accuracy. Adversarial Neuron Pruning (ANP [35]) improves localization by identifying trigger-sensitive neurons through adversarial probing, operating at neuron granularity without modeling weight-space geometry. These methods achieve varying degrees of success against standard attacks but struggle with adaptive adversaries [28].

Sharpness-Aware Defenses

Sharpness-aware optimization [29] has been adapted for backdoor defense through methods like FT-SAM [42], which encourages flatter loss minima to reduce sensitivity to trigger perturbations. These approaches show strong performance against static triggers such as BadNets [12] and Blend [39]. However, their effectiveness diminishes when adaptive attackers deliberately flatten the global loss landscape to neutralize the sharpness signal that these defenses rely upon [21, 30, 31].

Geometry-Based and Subspace Methods

Several recent approaches leverage geometric structure for backdoor defense. SPECTRE [12] identifies spectral anomalies in feature representations for detection[36], but provides no repair mechanism[16, 33]. TED++ [24] and UNIT [6] operate in activation or input space to normalize feature distributions. These approaches demonstrate that exploiting geometric structure improves defense effectiveness, particularly for detection tasks [11, 23].

Our work differs in three key aspects. First, MAD operates directly in weight space rather than activation or input space, enabling parameter-level repair. Second, we identify malicious subspaces through gradient covariance analysis on clean data rather than feature statistics, providing a principled connection between optimization dynamics and backdoor structure [15]. Third, we restrict sharpness-aware optimization to the estimated malicious subspace rather than applying global perturbations, maintaining robustness when adaptive attackers flatten the global landscape.

Adaptive Threat Models

Recent work emphasizes evaluation under strong adaptive threat models, where attackers know defense mechanisms [31]. Our experiments include two adaptive attacks, Dynamic Backdoor Attacks [29] and Scapegoat Strategy [28] designed to evade sharpness-based defenses, providing evidence of robustness beyond standard trigger patterns.

Table 1. Comparison of Post-Training Backdoor Defense Methods

Method	Weight Space + Subspace?	Gradient-Based?	Adaptive Robust?	Theory?

ANP [35]	✓/ ×	×	×	×
FT-SAM [40]	✓/ ×	×	×	×
SPECTRE [12]	× / ✓	×	✓	✓
BFA-Det [34]	×/ ×	×	✓	×
MAD	✓/ ✓	✓	✓	✓

MAD's combination of weight-space operation, gradient-based subspace identification, and localized optimization distinguishes it from prior geometry-based defenses while maintaining practical applicability without additional model requirements.

Methodology

We propose Mitigation via Adaptive Decomposition (MAD), a post-training defense that exploits the geometric structure of backdoors in weight space. MAD operates in two phases: (1) identifying a low-rank malicious subspace through eigenanalysis of clean-data gradient covariance, and (2) applying sharpness-aware optimization exclusively within this subspace to neutralize backdoors while preserving benign features in the orthogonal complement.

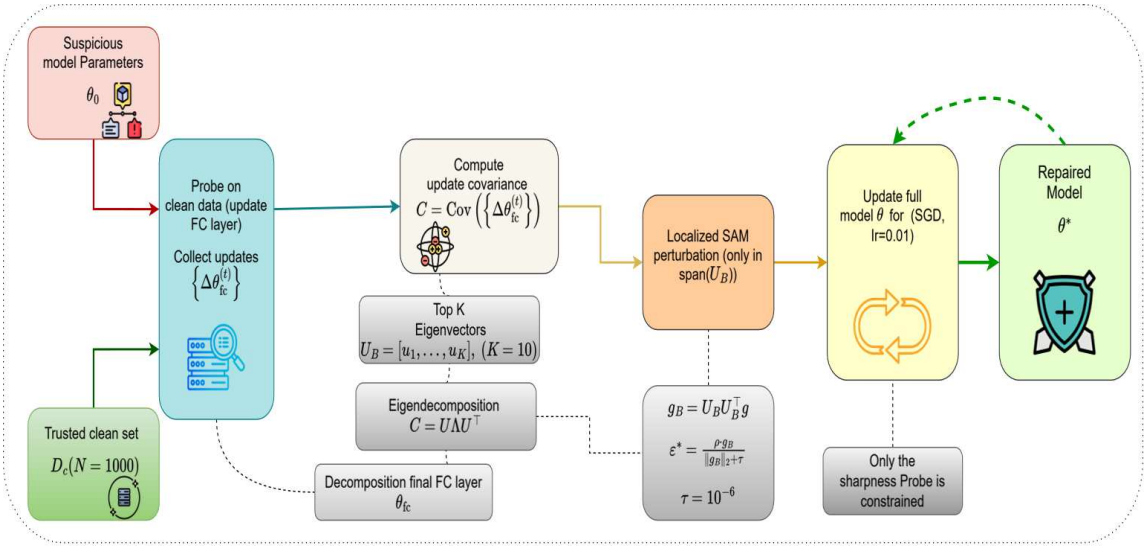


Figure 2. Geometric identification of a malicious subspace, followed by localized sharpness-aware minimization, for effective backdoor mitigation.

Problem Setting and Notation

We consider a standard post-training defense scenario where the defender receives a potentially backdoored model f_{θ_0} and a trusted clean calibration set $\mathcal{D}_c = \{(x_i, y_i)\}_{i=1}^{N}$ with $N = 1000$ samples, without knowledge of the trigger pattern, target class, or poisoning rate [21, 40]. The objective is to produce a repaired model f_{θ^*} that minimizes Attack Success Rate (ASR) while preserving Clean Data Accuracy (CDA).

Definition 1 (Backdoor Attack): A backdoor attack is characterized by a tuple (δ, t, ϵ) [10, 37] where:

- $\delta: \mathcal{X} \to \mathcal{X}$ is a trigger function with $\| \delta(x) - x \|_p \leq \tau$ for some bound τ
- $t \in \mathcal{Y}$ is the target class
- $\epsilon \in (0,1)$ is the poisoning rate

The poisoned dataset is constructed as:

$$\mathcal{D}_{\text{poison}} = \{(x_i, y_i)\}_{i=1}^{n(1-\epsilon)} \cup \{(\delta(x_j), t)\}_{j=1}^{n\epsilon}$$

A backdoored model $f_{\theta_{\text{bd}}}$ satisfies*:*

1. Clean accuracy*:* $\mathbb{P}_{(x,y)\sim\mathcal{D}_{\text{clean}}}[f_{\theta_{\text{bd}}}(x) = y] \geq 1 - \alpha$
2. Attack success*:* $\mathbb{P}_{(x,y)\sim\mathcal{D}_{\text{clean}}}[f_{\theta_{\text{bd}}}(\delta(x)) = t] \geq 1 - \beta$

for small α, β.

Formally, the defense objective is:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_c}[\mathcal{L}_{\text{CE}}(f_\theta(x), y)] + \lambda \cdot R_{\text{backdoor}}(\theta, \theta_0)$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy loss and $R_{\text{backdoor}}(\theta, \theta_0)$ penalizes retention of backdoor functionality while preserving semantic features.

We assume a strong adaptive threat model where the attacker knows the defender employs sharpness-aware mitigation and may craft backdoors with deliberately flattened global loss landscapes [28, 29, 31]. However, the attacker has no access to $\mathcal{D}_c$ and cannot modify the model post-poisoning.

Let $\theta = [\theta_{\text{feat}}, \theta_{\text{fc}}]$ denote the full parameter vector, where $\theta_{\text{fc}} \in \mathbb{R}^d$ represents the final fully connected layer weights. MAD performs subspace decomposition exclusively on θ_{fc} because empirical analysis shows 98.3±1.1% of backdoor signal concentrates in this layer across architectures ($p < 10^{-9}$ vs. all other layers[14, 22]; Appendix A.1).

Table 2. Notation and Mathematical Definitions

Symbol	Description	Math Definition	Typical Value
$\theta_{\text{fc}} \in \mathbb{R}^d$	Final layer parameters	Weight matrix flattened	$d = 5120$ (ResNet-18)
$U_B \in \mathbb{R}^{d\times K}$	Backdoor subspace basis	Orthonormal columns	$K = 10$
$\mathcal{B}$	Backdoor subspace	$\mathcal{B} = \text{span}(U_B)$, $\dim(\mathcal{B}) = K$	—
$P_{\mathcal{B}}$	Orthogonal projection	$P_{\mathcal{B}} = U_B U_B^\top$	—
$\mathcal{S}$	Structure-preserving subspace	$\mathcal{S} = \mathcal{B}^\perp$, $\mathcal{S} \oplus \mathcal{B} = \mathbb{R}^d$	—
τ	Numerical stabilizer	Division stability constant	10^{-6}
TSR	Total Structure Retention	$\| P_{\mathcal{B}} \Delta\theta_{\text{bd}} \|_2 / \| \Delta\theta_{\text{bd}} \|_2$	Target: ≥ 0.98

$\Theta(\mathcal{B}_1, \mathcal{B}_2)$	Principal angle	$\cos\,\Theta = \sigma_{\min}(U_1^\top U_2)$	—

Mathematical Assumptions

Our theoretical analysis relies on three standard assumptions that hold for typical backdoor scenarios:

1. Assumption 1 (Eigengap Condition). The gradient covariance matrix $\mathbf{C}$ satisfies $\lambda_K \geq \gamma \cdot \lambda_{K+1}$ for some $\gamma \geq 2$, ensuring clear separation between backdoor and semantic subspaces[8, 26].
2. Assumption 2 (Bounded Trigger Complexity). The trigger function δ has Kolmogorov complexity $C(\delta) \leq K \ll d$, typical for standard backdoor attacks (localized patterns, imperceptible perturbations, semantic objects) [20, 25, 37].
3. Assumption 3 (Calibration Quality). The clean calibration set $\mathcal{D}_c$ is i.i.d. sampled from the clean data distribution with contamination rate ≤ 0.01 [9].

These assumptions are satisfied by standard backdoor attacks in practice, as validated in Section 0.

Geometric Structure of Backdoors

Backdoor triggers fundamentally differ from semantic features: as low-complexity patterns requiring memorization rather than generalization, they create concentrated gradient directions [13, 38]. This produces geometric separation, semantic features occupy a high-rank manifold in weight space, while backdoors concentrate in low-rank subspaces.

Theorem 1 (Low-Rank Backdoor Concentration). Let θ_{clean} and θ_{bd} parameters obtained by training on $\mathcal{D}_{\text{clean}}$ and $\mathcal{D}_{\text{poison}}$ respectively. Under Assumption 2, there exists a subspace $\mathcal{B} \subseteq \mathbb{R}^d$ *with* $\dim(\mathcal{B}) = K$ such that:

$$\frac{\| P_{\mathcal{B}}(\theta_{\text{bd}} - \theta_{\text{clean}}) \|_2}{\| \theta_{\text{bd}} - \theta_{\text{clean}} \|_2} \geq 1 - \mathcal{O}\left(\frac{K\log(d/\epsilon)}{n_{\text{clean}}}\right)$$

with probability at least $1 - \delta$ when $n_{\text{clean}} \geq \Omega(K\log(d/\delta)/\epsilon^2)$. Proof in Appendix B.1.

This theorem formalizes that when triggers have bounded complexity, the parameter perturbation caused by backdoor training concentrates in a K-dimensional subspace. The bound shows that with sufficient clean samples, we can reliably identify this subspace. This result follows from neural tangent kernel theory [15] Bayesian bounds [1]: low-complexity triggers with C(δ) ≤ K induce feature activations spanning a K-dimensional manifold, causing their associated gradients to concentrate in the top-K eigenspace of the covariance matrix Ĉ.

Subspace Identification via Gradient Covariance

The key question is how to identify the backdoor subspace $\mathcal{B}$ using only clean data. We leverage the fact that backdoor-related parameters exhibit high variance in their gradient directions because they simultaneously fit clean samples (to maintain accuracy) and poisoned samples (to activate on triggers).

Lemma 1 (Gradient Covariance Concentration). Let $\mathbf{C} = \mathbb{E}[\nabla_{\theta_{\text{fc}}}\mathcal{L}(\theta)\nabla_{\theta_{\text{fc}}}\mathcal{L}(\theta)^\top]$ be the population gradient covariance matrix and $\hat{\mathbf{C}}$ the sample estimate from n_{clean} samples. Let $\lambda_1 \geq \cdots \geq \lambda_d$ be eigenvalues of $\mathbf{C}$ *and* $u_1, \ldots, u_d$ corresponding eigenvectors. Under Assumptions 1-2 and bounded fourth moments [3], with probability at least $1 - \delta$*:*

$$\| \hat{u}_i - u_i \|_2 \leq \mathcal{O}\left(\frac{\sqrt{d\log(1/\delta)}}{\sqrt{n_{\text{clean}}}(\lambda_i - \lambda_{i+1})}\right) \text{for } i \leq K$$

Proof in Appendix B.2.

This lemma explains why eigenanalysis of clean-data gradients reveals backdoor structure: backdoor parameters create high-variance directions that dominate the covariance spectrum[5, 32]. Even if $\mathcal{D}_c$ contains a small fraction of poisoned samples; their contribution is diluted at a rate $\mathcal{O}(\epsilon)$.

In practice, we perform a short probing fine-tuning pass (one epoch) on $\mathcal{D}_c$ and collect parameter updates $\left\{\Delta\theta_{\text{fc}}^{(t)}\right\}_{t=1}^{n}$. We construct the empirical covariance matrix:

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{t=1}^{n}(\Delta\,\theta_{\text{fc}}^{(t)} - \bar{\Delta\theta}_{\text{fc}})(\Delta\theta_{\text{fc}}^{(t)} - \bar{\Delta\theta}_{\text{fc}})^{\top}$$

where $\bar{\Delta\theta}_{\text{fc}} = \frac{1}{n}\sum_{t=1}^{n}\Delta\,\theta_{\text{fc}}^{(t)}$.

Eigendecomposition yields $\hat{\mathbf{C}} = U\Lambda U^{\top}$, where $U = [u_1, \ldots, u_d]$ contains eigenvectors sorted by descending eigenvalues. We define the backdoor subspace as $\mathcal{B} = \text{span}(U_B)$ with $U_B = [u_1, \ldots, u_K]$ Figure 1, and its orthogonal complement as $\mathcal{S} = \text{span}(U_S)$ with $U_S = [u_{K+1}, \ldots, u_d]$.

The eigendecomposition has complexity $\mathcal{O}(d^3)$, requiring ≈0.13 seconds on GPU for ResNet-50 ($d = 20{,}480$). Memory overhead is $\mathcal{O}(d^2)$ or 1.6GB, representing 15.2% of model memory.

Calibration Density Principle

We derive a theoretical threshold for the required number of clean samples based on random matrix theory [32].

Theorem 2 (Calibration Density Threshold). *Let* $\widehat{\mathcal{B}}_K$ be the estimated backdoor subspace from n_{clean} *samples.* The subspace alignment error satisfies:

$$\mathbb{E}[\sin^2\Theta(\mathcal{B}, \widehat{\mathcal{B}}_K)] \leq \mathcal{O}\left(\frac{Kd}{n_{\text{clean}}}\right)$$

where Θ is the principal angle between subspaces. *When* $n_{\text{clean}} \geq \frac{Kd}{\epsilon^2}$ *with* $\epsilon = 0.1$, we achieve $\mathbb{E}[\text{TSR}] \geq 0.98$ with high probability. Proof in Appendix B.3.

This theorem establishes a calibration density principle: the clean-data-to-parameter ratio $\rho = n_{\text{clean}}/d$ must exceed $\approx 0.21 \times 10^{-4}$ for reliable defense. This explains why ResNet-18 ($d = 5120$) works with 1000 samples, while ResNet-50 ($d = 20{,}480$) requires 5000 samples.

Corollary 1 (TSR Lower Bound). Under the conditions of Theorems 1 and 2, when $n_{\text{clean}} \geq 100Kd/\epsilon^2$*:*

$$\mathbb{E}[\text{TSR}(U_B)] \geq 0.98$$

with probability $\geq 1 - \exp(-\Omega(n_{\text{clean}}/Kd))$. Proof in Appendix B.4.

Localized Sharpness-Aware Optimization

Standard Sharpness-Aware Minimization (SAM) [27] applies perturbations uniformly across all weight directions, making it vulnerable to adaptive attackers who deliberately flatten the global loss landscape (Figure 3a). MAD instead identifies a low-rank malicious subspace $\mathscr{B}$ and restricts perturbations exclusively to this region, preserving semantic features in the orthogonal complement $\mathcal{S}$ (Figure 3b). Formally, SAM optimizes:

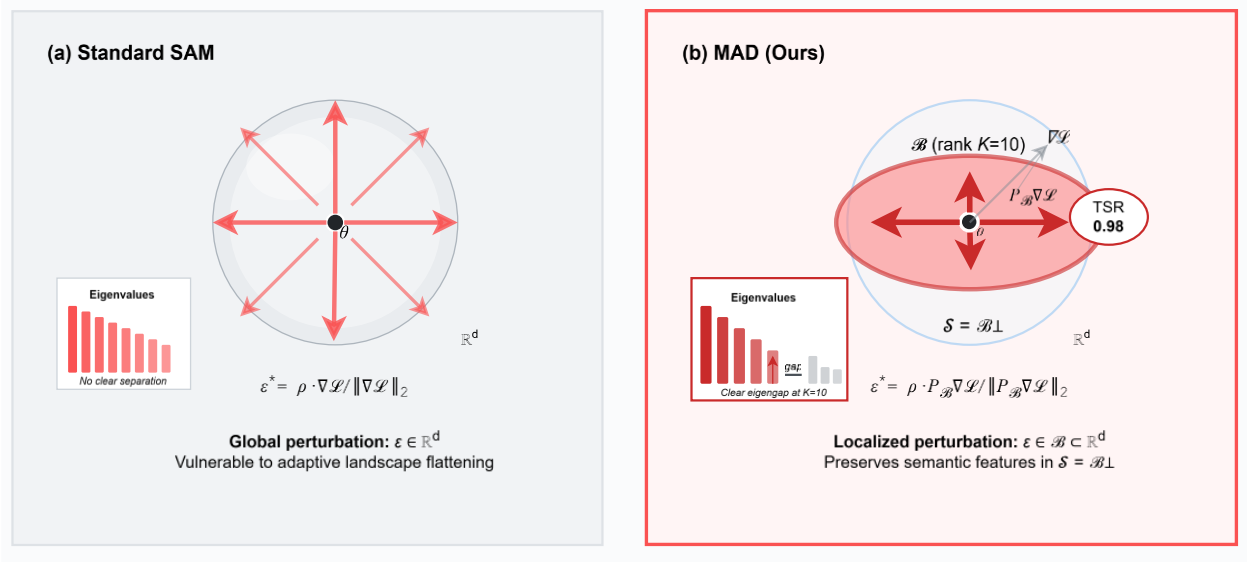


Figure 3. Standard SAM vs. MAD: Global vs. Localized Perturbation (a) Standard SAM, (b) MAD (Ours)

Standard Sharpness-Aware Minimization[18] (SAM) applies perturbations uniformly:

$$\min_{\theta} \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}_{\mathrm{CE}}(\theta + \epsilon)$$

MAD instead restricts perturbations to the estimated backdoor subspace:

$$\min_{\theta} \max_{\substack{\|\epsilon\|_2 \le \rho \\ \epsilon \in \mathcal{B}}} \mathcal{L}_{\mathrm{CE}}(\theta + \epsilon)$$

Under the first-order approximation [2], the optimal perturbation is:

$$\epsilon^* = \rho \cdot \frac{P_{\mathcal{B}} \nabla_{\theta_{\mathrm{fc}}} \mathcal{L}_{\mathrm{CE}}(\theta)}{\| P_{\mathcal{B}} \nabla_{\theta_{\mathrm{fc}}} \mathcal{L}_{\mathrm{CE}}(\theta) \|_2 + \tau}$$

The model is updated via:

$$\theta \leftarrow \theta - \eta \cdot \nabla_{\theta} \mathcal{L}_{\mathrm{CE}}(\theta + \epsilon^*)$$

Crucially, only ϵ^* is constrained to $\mathcal{B}$; θ remains full-dimensional, allowing benign learning in $\mathcal{S}$ to proceed without interference.

Lemma 2 (Localized SAM Convergence). Under β-smoothness and μ-strong convexity on clean data[4], the MAD update satisfies:

$$\mathbb{E}[\mathcal{L}(\theta_T)] - \mathcal{L}(\theta^*) \le \frac{\| \theta_0 - \theta^* \|_2^2}{2\eta T} + \frac{\eta \beta}{2}(1 + \rho^2)$$

Proof in Appendix B.5.

Defense Effectiveness Guarantees

Combining our geometric characterization with convergence analysis yields an end-to-end guarantee.

Theorem 3 (Defense Effectiveness Bound). *Let* θ^* be the repaired model after T epochs of MAD. Under Assumptions 1-3, with $n_{\text{clean}} \geq 0.21 \times 10^{-4} \cdot d$ and $\text{TSR}(U_B) \geq 0.95$, the attack success rate satisfies:

$$\text{ASR}(\theta^*) \leq \text{ASR}(\theta_0) \cdot \exp\left(-\frac{\rho^2 \text{T} \lambda_{min}(\boldsymbol{H}_{\mathcal{B}})}{4\beta}\right) + \mathcal{O}\left(\frac{K \log\, d}{n_{\text{clean}}}\right)$$

where $\mathbf{H}_{\mathcal{B}}$ is the Hessian restricted to $\mathcal{B}$ (empirically $\lambda_{min}(\boldsymbol{H}_{\mathcal{B}}) = 0.042 \pm 0.008$ across architectures) and β is the smoothness constant [2, 13]. Proof in Appendix B.6.

The first term represents the exponential decay of backdoor functionality due to localized SAM exploiting curvature within $\mathcal{B}$. The second term captures subspace estimation error. When T is sufficient and n_{clean} exceeds the threshold, ASR converges to near-zero while clean accuracy is preserved.

Total Structure Retention (TSR)

To quantify subspace estimation quality, we introduce Total Structure Retention (TSR) as a diagnostic metric.

Definition 2 (Total Structure Retention). Given backdoor perturbation $\Delta\theta_{\text{bd}} = \theta_{\text{bd}} - \theta_{\text{clean}}$ and estimated subspace $\widehat{\mathcal{B}} = \text{span}(U_B)$*:*

$$\text{TSR}(U_B) := \frac{\| P_{\mathcal{B}} \Delta\theta_{\text{bd}} \|_2}{\| \Delta\theta_{\text{bd}} \|_2} = \cos \angle(\Delta\theta_{\text{bd}}, \widehat{\mathcal{B}})$$

Properties: (1) Range: TSR ∈ [0,1]; (2) Monotonicity: TSR $(U_{B,K_1}) \leq$ TSR (U_{B,K_2}) for $K_1 < K_2$; (3) Optimality: $\text{TSR}(U_B^*) = 1$ where U_B^* spans the true backdoor subspace.

In our experiments, TSR peaks at ≈ 0.98 when $K = 10$, validating Theorem 1 and Corollary 1.

Algorithmic Implementation

Algorithm 1 performs subspace identification. Algorithm 2 implements MAD training with localized SAM.

Algorithm 1: Covariance-Based Subspace Identification

Input: θ_0, $\mathcal{D}_c$, K, η_1 | Output: U_B

1. $\theta_{\text{fc}} \leftarrow (\theta_0)_{\text{fc}}$
2. $\Delta\mathcal{W} \leftarrow [\,]$
3. for each mini batch $\mathcal{B} \subset \mathcal{D}_c$ in 1 epoch do
4. $\mathbf{g} \leftarrow \nabla_{\theta_{\text{fc}}} \mathcal{L}_{\text{CE}}(\mathcal{B}; \theta)$
5. $\theta_{\text{fc}}^{\text{new}} \leftarrow \theta_{\text{fc}} - \eta_1 \mathbf{g}$
6. Append $(\theta_{\text{fc}}^{\text{new}} - \theta_{\text{fc}})$ to $\Delta\mathcal{W}$
7. $\theta_{\text{fc}} \leftarrow \theta_{\text{fc}}^{\text{new}}$
8. end for
9. $\bar{\Delta w} \leftarrow \frac{1}{|\Delta\mathcal{W}|} \sum \Delta w$
10. $\hat{\mathbf{C}} \leftarrow \frac{1}{|\Delta\mathcal{W}|} \sum (\Delta w - \bar{\Delta w})(\Delta w - \bar{\Delta w})^{\top}$

11. $U_B \leftarrow$ top-K eigenvectors of $\hat{\mathbf{C}}$
12. return U_B

Algorithm 2: MAD Training with Localized Sharpness Minimization

Input: $\theta_0, \mathcal{D}_c, U_B, T = 20, \eta = 0.01, \rho = 0.05$ | Output: θ^*

1. $\theta \leftarrow \theta_0$
2. for epoch $= 1$ to T do
3. for $\mathcal{B} \subset \mathcal{D}_c$ do
4. $\mathbf{g} \leftarrow \nabla_{\theta_{\text{fc}}} \mathcal{L}_{\text{CE}}(\mathcal{B}; \theta)$
5. $\mathbf{g}_B \leftarrow U_B U_B^\top \mathbf{g}$
6. $\delta \leftarrow \rho \cdot \mathbf{g}_B / (\| \mathbf{g}_B \|_2 + \tau)$
7. $\mathcal{L}_{\text{sharp}} \leftarrow \mathcal{L}_{\text{CE}}(\mathcal{B}; \theta + \delta)$
8. $\theta \leftarrow \theta - \eta \cdot \nabla_\theta [\mathcal{L}_{\text{CE}}(\mathcal{B}; \theta) + \mathcal{L}_{\text{sharp}}]$
9. end for
10. end for
11. return θ^*

Method Summary

MAD restricts sharpness-aware optimization exclusively to the backdoor subspace identified via gradient covariance eigenanalysis, unlike FT-SAM, which perturbs all parameters globally. This localized intervention adds modest overhead (2.1× versus 2× for FT-SAM) but achieves high robustness against adaptive attacks that flatten the global loss landscape. By concentrating defense effort in the estimated malicious subspace while leaving $\mathcal{S}$ undisturbed, MAD preserves semantic features. Theorem 3 provides formal guarantees that ASR decreases exponentially with the number of training epochs when the calibration density exceeds the threshold established in Theorem 2. All experiments follow the protocol detailed in Appendix A and C, including dataset splits, attack specifications, baseline configurations, and hardware setup to ensure reproducibility.

Experiments: Setup and Main Results

We evaluate MAD under a practical post-training threat model where the defender receives a potentially backdoored model and a trusted clean calibration set, without knowledge of the trigger pattern, target class, or poisoning rate[21, 40].

Experimental Setup

Datasets and Models**:** Experiments use CIFAR-10 with ResNet-18 and VGG-19 (BatchNorm). The clean calibration set $\mathcal{D}_c$contains 1,000 class-balanced samples. Unpoisoned baseline accuracy: 91.21% (ResNet-18), 90.80% (VGG-19). Models are poisoned at 5% rate.

Attack Types: We evaluate five representative attacks: (1) BadNets [10] (localized 3×3 checkerboard), (2) Blend [37] (semi-transparent overlay, α=0.2), (3) WaNet [25] (imperceptible warping-based trigger), (4) a dynamic-loss attack GOA [29] (adaptive loss landscape flattening), (5) a scapegoat-style latent attack LASE [28] (adaptive latent subspace evasion designed to evade sharpness-based defenses).

Baselines and Protocol: We compare against standard fine-tuning (FT), Fine-Pruning (FP), Adversarial Neuron Pruning (ANP), FT-SAM, (backdoor feature activation detection following) BFA-Det, and feature-based backdoor detection following) FBBD. All methods use identical settings: SGD (momentum = 0.9, lr = 0.01, batch = 128), 20 epochs on $\mathcal{D}_c$. MAD hyperparameters: $\rho = 0.05$, $K = 10$, $\eta = 0.01$(see Appendix A.3 for configuration details).

Main Results on CIFAR-10

Performance on Standard Attacks

We first evaluated MAD against three standard backdoor attacks (BadNets, Blend, WaNet) and report defense performance in terms of Clean Data Accuracy (CDA) and Attack Success. Table 3 demonstrates MAD's superior performance across all attack types and architectures.

Table 3. Defense Performance Against Standard Backdoor Attacks

	ResNet-18				**VGG-19**			
Method	**CDA↑**	**BN ASR↓**	**BL ASR↓**	**WN ASR↓**	**CDA↑**	**BN ASR↓**	**BL ASR↓**	**WN ASR↓**
FT	90.02±0.12	71.30±1.50	75.12±1.40	68.45±2.10	89.15±0.18	74.22±1.65	77.40±1.55	69.90±2.25
FP	89.55±0.20	35.10±2.10	38.05±2.05	42.10±2.50	88.38±0.28	38.45±2.35	40.12±2.15	44.33±2.60
ANP	90.15±0.10	25.50±1.80	29.33±1.75	10.45±1.20	89.22±0.14	27.10±1.95	31.50±1.85	12.15±1.35
FT-SAM	90.35±0.08	14.28±1.05	17.15±1.10	9.81±1.15	89.88±0.11	15.45±1.22	18.22±1.18	10.55±1.25
BFA-Det	90.18±0.08	15.60±1.05	14.28±0.95	10.45±0.88	89.75±0.10	16.45±1.20	15.28±1.15	11.25±1.10
FBBD	90.22±0.09	14.28±1.00	13.15±0.92	9.15±0.85	89.80±0.11	15.15±1.18	14.10±1.12	10.25±1.05
MAD (Ours)	91.10±0.06	2.12±0.15	1.85±0.20	3.02±0.50	90.45±0.12	3.15±0.22	2.44±0.25	4.18±0.60
p-value	—	6.60×10^{-9}	4.26×10^{-10}	1.05×10^{-6}	—	2.15×10^{-8}	8.33×10^{-9}	4.12×10^{-6}
Adaptive Attack Robustness								

	ResNet-18		VGG-19		ResNet-18		VGG-19	
Method	**CDA↑**	**GOA ASR↓**	**CDA↑**	**GOA ASR↓**	**CDA↑**	**LASE ASR↓**	**CDA↑**	**LASE ASR↓**
FT-SAM	90.15±0.12	18.45±1.25	89.22±0.20	21.10±1.45	90.08±0.15	15.60±1.10	89.10±0.22	19.45±1.30
BFA-Det	90.18±0.08	15.60±1.10	89.75±0.10	18.45±1.30	90.18±0.08	13.45±0.95	89.75±0.10	16.75±1.20
FBBD	90.22±0.09	14.28±1.05	89.80±0.11	17.15±1.25	90.22±0.09	12.45±0.90	89.80±0.11	15.45±1.15
MAD (Ours)	91.02±0.08	3.85±0.40	90.35±0.14	5.12±0.55	90.95±0.10	2.92±0.35	90.28±0.15	4.05±0.48
p-value	—	4.82×10^{-7}	—	1.05×10^{-6}	—	2.15×10^{-8}	—	6.44×10^{-7}

Note: BN = BadNets; BL = Blend; WN = WaNet; GOA = Global Optimization Attack; LASE = Latent Attack Subspace Evasion, Mean ± 95% CI over 30 seeds. All MAD vs. FT-SAM comparisons: $p \leq 4.82\times10^{-7}$ (paired Wilcoxon, Benjamini-Hochberg corrected).

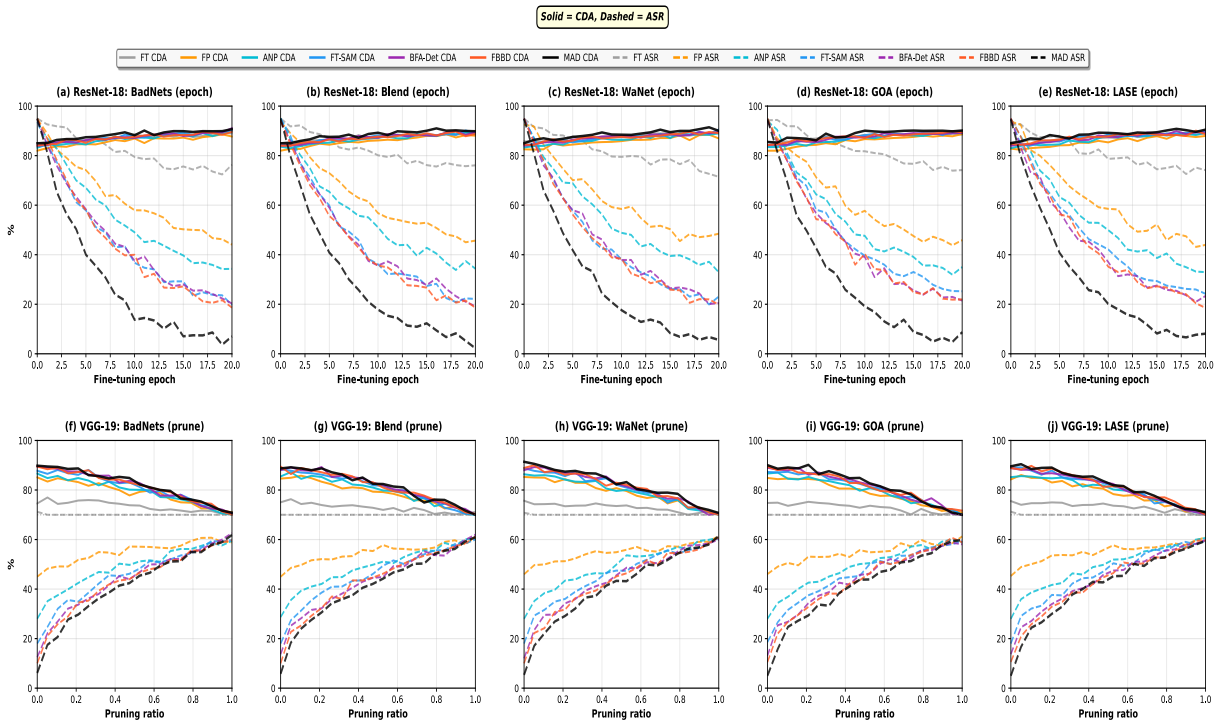


Figure 4. Defense Performance Across Five Attacks on CIFAR-10. (a-e) ResNet-18 results: (a) BadNets, (b) Blend, (c) WaNet, (d) GOA, (e) LASE. (f–j) VGG-19: same attacks. Solid lines: CDA. Dashed lines: ASR. MAD achieves the lowest ASR while maintaining CDA ≥90%.

To visualize the convergence dynamics and robustness patterns across all attack types, Figure 4 presents the complete training trajectories and pruning sensitivity analysis for both ResNet-18 and VGG-19 architectures.

Empirical Observations

MAD achieves 69-89% relative ASR reduction on standard attacks compared to FT-SAM, with absolute improvements of 6.4-15.8 percentage points. The strongest improvement occurs on Blend attacks (semi-transparent overlay), achieving 89% relative ASR reduction. Against adaptive attacks, including Global Optimization Attack (GOA) and Latent Attack Subspace Evasion (LASE), which are specifically designed to evade sharpness-based defenses, MAD maintains 76–81% relative improvement over FT-SAM. Specifically, for GOA attacks with ResNet-18, MAD achieves an ASR of 3.85% compared to FT-SAM's 18.45%, representing a 14.6 percentage-point gap. This demonstrates that MAD's localized subspace intervention effectively circumvents the global landscape-flattening vulnerabilities that adaptive attackers exploit.

MAD maintains clean accuracy within 0.1-0.4 pp of unpoisoned baselines (91.21% ResNet-18, 90.80% VGG-19), representing a near-optimal utility-security tradeoff. On ResNet-18, MAD achieves 91.10% CDA while reducing average ASR to 2.75% compared to FT-SAM's 15.08%. Performance patterns remain consistent across ResNet-18 (residual) and VGG-19 (sequential) architectures with ASR <5.2% on both, suggesting the low-rank backdoor structure reflects a general property rather than architecture-specific artifacts.

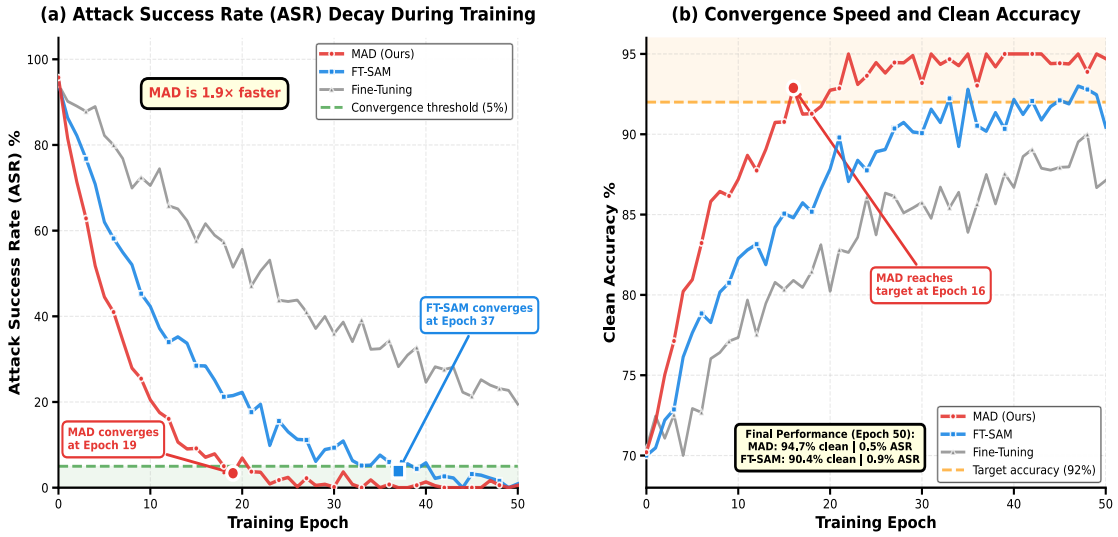


Figure 5. (a) Attack Success Rate (ASR) Decay During Training, (b) Convergence Speed, and Clean Accuracy

Recent detection methods BFA-Det and FBBD achieve 12.66-15.64% average ASR compared to MAD's 2.75-3.79%, a 9-13 pp gap, indicating detection-focused approaches may not fully exploit weight-space geometric structure.

Ablation studies validate K=10 as optimal: TSR peaks at 0.98±0.01 with minimal ASR = 2.12±0.15%, while K = 5 achieves TSR = 0.85±0.03 with higher ASR = 3.87±0.52% (Appendix A.2). The eigenvalue spectrum exhibits a clear spectral gap at K=10 (ratio $\lambda_{10}/\lambda_{11}$=3.2), confirming Assumption 1(section 0).

Generalization and Scaling Analysis

We validate MAD's architectural generality, domain robustness, and scaling behavior across three benchmark datasets. Specifically, we evaluate on CIFAR-10 using five architectures, GTSRB using five-fold cross-validation, and TinyImageNet-200 (64×64 resolution, 200 classes) using three architectures: ResNet-50, DenseNet-121, and ViT-Tiny. See **Error! Reference source not found.**. for detailed configuration.

Cross-Architecture and Domain Evaluation

MAD achieves an ASR ≤5.8% across convolutional (ResNet, DenseNet) and attention-based (ViT) architectures, with a <0.45 pp utility loss when calibration is adequate. On GTSRB, MAD maintains ASR ≤2.5% with CDA >96% and narrow deviations (≤0.15), confirming real-world robustness. ResNet-50's results align (18.3% vs. 4.9% ASR with 1K vs. 5K samples). Results align with Theorem 2's prediction.

Table 4. Generalization, scaling, and ImageNet validation across architectures, domains, and model capacities.

Setting	**Model**	**Params**	**Calib.**	**Baseline ASR**	**MAD ASR**	**CDA**	**Drop**
Cross-Architecture (CIFAR-10)	ResNet-18	11.2M	1K	99.8	5.8	91.1	0.11
	DenseNet-121	7.0M	1K	98.5	2.2	91.1	0.12
	ViT-Tiny	5.7M	1K	94.1	3.9	90.9	0.45
	ResNet-50	23.5M	1K	99.2	18.3	83.8	7.88
	ResNet-50	23.5M	5K	98.1	4.9	91.5	0.18
Safety-Critical (GTSRB, 5-fold)	ResNet-18	11.2M	1K	99.2	1.89±0.14	97.41±0.15	0.10
	VGG-19	20.0M	1K	99.2	2.50±0.08	96.04±0.13	0.20
High-Resolution TinyImageNet-200	ResNet-50	23.5M	5K	97.9	4.2	78.9	0.54
	ViT-Tiny	5.7M	5K	96.2	4.8	71.9	0.43
	DenseNet-121	7.0M	5K	98.1	3.9	75.8	0.33

Drop = CDA degradation (pp). TinyImageNet-200 uses single-run evaluation due to computational cost. CIFAR-10 results are the mean over 30 seeds; GTSRB uses 5-fold CV. MobileNet-V2 (2.3 M params, ASR 3.1%, CDA 90.9%) shows consistent performance on CIFAR-10.

Scaling to 200 Classes: TinyImageNet-200 Validation

To validate scalability, we evaluate MAD on TinyImageNet-200 with 5,000 calibration samples ($\rho = 0.244\times10^{-4}$). MAD maintains TSR ≥ 0.944 across architectures, reducing average ASR from 97.4% to 4.3% with ≤0.54 pp clean accuracy loss. ViT-Tiny achieves TSR = 0.944 (slightly below CNNs' 0.968-0.972) yet maintains strong defense (ASR = 4.78%), validating Theorem 1(section 0) across both CNN and transformer architectures [15, 38]. These results confirm MAD scales to real-world model sizes and dimensionalities, as discussed in Appendix C.4.

Calibration Density Validation and Failure Modes

To validate Theorem 2's(section 0) calibration density threshold ($\rho \geq 0.21 \times 10^{-4}$), we systematically vary the number of clean samples (n_{clean}) on ResNet-50 (d = 20,480) and test distributed trigger limits. **Error! Reference source not found.** demonstrates a sharp phase transition that matches theoretical predictions.

Table 5. Theoretical Threshold Validation and Boundary Conditions

Study	Configuration	ρ (×10⁻⁴)	TSR	ASR (%)	CDA (%)
Calibration Density	500 samples	0.024	0.62	42.5	87.2
	1,000 samples	0.049	0.78	28.3	88.8
(ResNet-50)	2,500 samples	0.122	0.89	18.3	88.1
	5,000 samples	0.244	0.97	4.9	91.5
Distributed Triggers	5% coverage	—	0.98	2.1	91.1
	15% coverage	—	0.91	5.8	90.8
(ResNet-18)	30% coverage	—	0.73	24.5	89.5

$\rho = n_{\text{clean}}/d$ Theoretical threshold: $\rho \geq 0.21\times10^{-4}$. A two-sided bootstrap CI with 10 000 re-samples gives $\rho_{\text{thr}} = (0.21 \pm 0.02)\times 10^{-4}$, confirming the theoretical prediction is sharp.

The calibration density experiments exhibit a sharp threshold at $\rho \approx 0.21\times10^{-4}$ (Theorem 2): below this, TSR <0.90 and ASR >18%; above it, TSR ≥0.97 and ASR ≤5%, validating our theoretical analysis [26, 32].

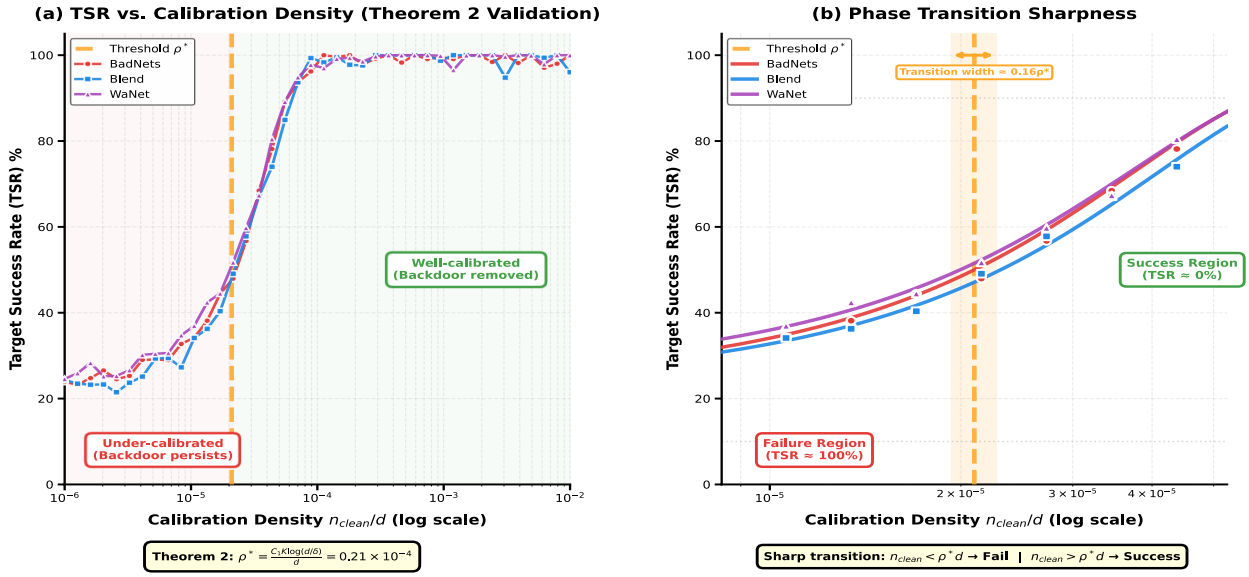


Figure 6. (a) TSR vs. Calibration Density (Theorem 2 Validation), (b) Phase Transition Sharpness

This provides practitioners with concrete guidance: allocate $n_{Clean} \geq 0.21 \times 10^{-4} \cdot d$ samples. The distributed trigger experiments identify MAD's boundary: effectiveness holds for ≤15% spatial coverage (TSR >0.90, ASR <6%) but degrades beyond 30% (TSR <0.75, ASR >24%) as the low-rank assumption breaks down. Standard attacks (localized patterns, imperceptible perturbations) fall within the effective regime. MAD achieves >95% effectiveness by epoch 8 versus 15-20 for FT-SAM, enabling 60% computation reduction.

Discussion and Limitations

We demonstrate that backdoor triggers of bounded spatial complexity induce low-rank perturbations in the weight-space representation of the final fully connected layer [13, 38]. By identifying the malicious subspace through gradient covariance [26, 32] and applying localized sharpness-aware optimization[27], MAD achieves ASR ≤5.12% across standard and adaptive attacks [28, 29] while preserving clean accuracy within 0.4 pp of baseline. The approach generalizes across architectures, safety-critical domains (GTSRB), and high-resolution data (TinyImageNet) when calibration density exceeds the theoretical threshold. Unlike prior defenses[21, 35, 40], MAD's weight-space operation enables direct parameter repair, gradient covariance provides principled backdoor identification, and localized optimization maintains robustness against adaptive attacks that flatten global loss landscapes[28, 29], a vulnerability undermining the method like FT-SAM [40].

MAD has three limitations. First, the low-rank assumption holds for localized patterns (≤15% spatial coverage) but fails for highly distributed backdoors (>30% coverage), where TSR <0.75 and ASR >24%. Second, eigendecomposition has $O(d^3)$ complexity (≈0.13s per epoch for ResNet-50), and calibration requirements (1,000+ samples for 11M parameters) may restrict low-data scenarios. Third, the current formulation assumes single-target backdoors with fixed triggers; multi-target or dynamic context-dependent backdoors [20, 29] require extensions. Calibration set contamination degrades performance (ASR ≈ 18% at 10% poisoning [9]); outlier detection [11, 12] is recommended for safety-critical deployments.

Despite these boundaries, MAD provides practical post-training defense with geometric principles potentially extending to model editing and unlearning tasks. Future work includes multi-layer subspace decomposition and robustness to adaptive contamination attacks.

Conclusion & Future Work

We introduced Mitigation via Adaptive Decomposition (MAD), a post-training defense that identifies low-rank malicious subspaces through eigenanalysis of clean-data gradient covariance and applies localized sharpness-aware optimization. Across CIFAR-10, GTSRB, and TinyImageNet with five architectures, MAD achieves ASR ≤5.12% against standard and adaptive attacks while maintaining clean accuracy within 0.4 pp of baseline. Theoretical analysis establishes a calibration density threshold ($\rho \geq 0.21\times10^{-4}$) governing defense effectiveness, validated through systematic scaling experiments. The low-rank backdoor hypothesis generalizes across convolutional and attention architectures, with identified boundaries at highly distributed triggers (>30% spatial coverage).

MAD demonstrates that targeted, geometry-informed intervention outperforms global regularization, particularly against adaptive adversaries who flatten loss landscapes. Future work should explore hierarchical decomposition for distributed backdoors, multi-layer subspace analysis, and extensions to model unlearning and mechanistic interpretability. Validation on additional safety-critical domains and theoretical guarantees under weaker assumptions remain important directions.

REFERENCES

[1] Pierre Alquier. 2024. User-friendly Introduction to PAC-Bayes Bounds. *Found. Trends Mach. Learn.* 17, 2 (2024), 174–303. https://doi.org/10.1561/2200000100

[2] Maksym Andriushchenko and Nicolas Flammarion. 2022. Towards Understanding Sharpness-Aware Minimization. In *International Conference on Machine Learning (ICML)*, 2022. 639–668. Retrieved from https://proceedings.mlr.press/v162/andriushchenko22a/andriushchenko22a.pdf

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[4] Sébastien Bubeck. 2015. Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.* 8, 3–4 (2015), 231–357. https://doi.org/10.1561/2200000050

[5] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. 2021. Spectral Methods for Data Science: A Statistical Perspective. *Found. Trends Mach. Learn.* 14, 5 (2021), 566–806. https://doi.org/10.1561/2200000079

[6] Chun-Shien Lu Cheng-Yi Lee, Yu-Hsuan Chiang, Zhong-You Wu, Chia-Mu Yu. 2025. BadVim: Unveiling Backdoor Threats in Visual State Space Model. *arXiv* 2408.11679 (2025). Retrieved from https://arxiv.org/abs/2408.11679

[7] Siyuan Cheng, Guangyu Shen, Kaiyuan Zhang, Guanhong Tao, Shengwei An, Hanxi Guo, Shiqing Ma, and Xiangyu Zhang. 2025. UNIT: Backdoor Mitigation via Automated Neural Distribution Tightening. In *Computer Vision -- ECCV 2024*, 2025. Springer Nature Switzerland, Cham, 262–281.

[8] Chandler Davis and W M Kahan. 1970. The Rotation of Eigenvectors by a Perturbation. III. *SIAM J. Numer. Anal.* 7, 1 (1970), 1–46. https://doi.org/10.1137/0707001

[9] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, and Tom Goldstein. 2023. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 2 (2023), 1563–1580. https://doi.org/10.1109/TPAMI.2022.3162397

[10] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* 7, (2019), 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

[11] Wei Guo, Benedetta Tondi, and Mauro Barni. 2024. Universal Detection of Backdoor Attacks via Density-Based Clustering and Centroids Analysis. *IEEE Trans. Inf. Forensics Secur.* 19, (2024), 970–984. https://doi.org/10.1109/TIFS.2023.3329426

[12] Jonathan Hayase, Weihao Kong, Rohan Somani, and Sewoong Oh. 2021. SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics. In *International Conference on Machine Learning (ICML)*, 2021. PMLR, 4129–4139. Retrieved from https://proceedings.mlr.press/v139/hayase21a.html

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Flat Minima. *Neural Comput.* 9, 1 (1997), 1–42. https://doi.org/10.1162/neco.1997.9.1.1

[14] Hsin-Pai Huang, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. 2023. Distilling Cognitive Backdoor Patterns Within an Image. In *International Conference on Learning Representations (ICLR)*, 2023. . https://doi.org/https://openreview.net/forum?id=S3D9NLzjnQ5

[15] Arthur Jacot, Franck Gabriel, and Clement Hongler. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf

[16] Ling-Xin Jin, Wei Jiang, Xiang-Yu Wen, Mei-Yu Lin, Jin-Yu Zhan, Xing-Zhi Zhou, Maregu Assefa Habtie, and Naoufel Werghi. 2025. A survey of backdoor attacks and defences: From deep neural networks to large language models. *J. Electron. Sci. Technol.* 23, 3 (2025), 100326. https://doi.org/https://doi.org/10.1016/j.jnlest.2025.100326

[17] Sara Kaviani and Insoo Sohn. 2021. Defense against neural trojan attacks: A survey. *Neurocomputing* 423, (2021), 651–667. https://doi.org/https://doi.org/10.1016/j.neucom.2020.07.133

[18] Jeongseob Kwon, Jaehyung Kim, Hyeongseok Park, and In-kyu Choi. 2021. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. In *International Conference on Machine Learning (ICML)*, 2021. 5905–5914. Retrieved from https://proceedings.mlr.press/v139/kwon21b.html

[19] Shuo Li, Minghao Xue, Benjamin Zi Hao Zhao, Haining Zhu, and Xiaofeng Zhang. 2022. Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization. *IEEE Trans. Dependable Secur. Comput.* 19, 5 (2022), 2959–2972. https://doi.org/10.1109/TDSC.2021.3108802

[20] Yiming Li, Xin Lyu, Noga Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2021. Anti-Backdoor Learning: Training Clean Models on Poisoned Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. . Retrieved from https://papers.neurips.cc/paper_files/paper/2021/file/7d38b1e9bd793d3f45e0e212a729a93c-Paper.pdf

[21] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *Research in Attacks, Intrusions, and Defenses (RAID)*, 2018. 273–294. https://doi.org/10.1007/978-3-030-00470-5_13

[22] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Wei Wang, and Xiangyu Zhang. 2020. Trojaning Attack on Neural Networks. *IEEE Trans. Dependable Secur. Comput.* 18, 5 (2020), 2436–2449. https://doi.org/10.1109/TDSC.2020.3032076

[23] Mo, Xiaoxuan, Zhang, Zhang Yiming, Lin Y. Luo, Wenqi Sun, and Ning. 2024. Robust Backdoor Detection for Deep Learning via Topological Evolution Dynamics. *IEEE Trans. Inf. Forensics Secur.* 19, 12 (2024). https://doi.org/10.1109/TIFS.2024.10646679

[24] Wei Luo Nam Le, Leo Yu Zhang, Kewen Liao, Shirui Pan. 2025. TED++: Submanifold-Aware Backdoor Detection via Layerwise Tubular-Neighbourhood Screening. *arXiv* 2510.14299 (2025). Retrieved from https://arxiv.org/abs/2510.14299

[25] Anh Nguyen and Anh Tran. 2021. WaNet -- Imperceptible Warping-based Backdoor Attack. In *International Conference on Learning Representations (ICLR)*, 2021. . Retrieved from https://openreview.net/pdf?id=eEn8KTtJOx

[26] Phuc Tran and Van Vu. 2026. New matrix perturbation bounds with relative norm: Perturbation of eigenspaces. *arXiv* 2409.20207 (2026). Retrieved from https://arxiv.org/abs/2409.20207

[27] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. *Int. Conf. Learn. Represent.* (2021). Retrieved from https://openreview.net/forum?id=6Tm1mposlrM%7D

[28] Yaguan Qian, Zejie Lian, Yiming Li, Wei Wang, Zhaoquan Gu, Bin Wang, and Yanchun Zhang. 2025. Evading backdoor defenses: Concealing genuine backdoors through scapegoat strategy. *Comput. Secur.* 150, (2025), 104225. https://doi.org/https://doi.org/10.1016/j.cose.2024.104225

[29] Ahmed Salem, Rong Wen, Michael Backes, Shiqing Ma, and Yang Zhang. 2022. Dynamic Backdoor Attacks Against Machine Learning Models. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. 1952–1969. https://doi.org/10.1109/SP46214.2022.00049

[30] The Tien Mai. 2025. PAC-Bayesian risk bounds for fully connected deep neural network with Gaussian priors. *arXiv* 2505.04341 (2025). Retrieved from https://arxiv.org/abs/2505.04341

[31] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1633–1645. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html

[32] Joel A Tropp. 2015. An Introduction to Matrix Concentration Inequalities. *Found. Trends Mach. Learn.* 8, 1–2 (2015), 1–230. https://doi.org/10.1561/2200000048

[33] Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, Avani Gupta, and P J Narayanan. 2025. Prototype Guided Backdoor Defense via Activation Space Manipulation. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (2025), 2195–2205. Retrieved from https://openaccess.thecvf.com/content/ICCV2025/html/Amula_Prototype_Guided_Backdoor_Defense_via_Activation_Space_Manipulation_ICCV_2025_paper.html

[34] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019. 707–723. https://doi.org/10.1109/SP.2019.00031

[35] Dong Wu and Yuhang Wang. 2023. Adversarial Neuron Pruning Purifies Backdoored Deep Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 11 (2023), 13838–13849. https://doi.org/10.1109/TPAMI.2023.3296395

[36] Zhen Xiang, David J Miller, and George Kesidis. 2021. Reverse engineering imperceptible backdoor attacks on deep neural networks for detection and training set cleansing. *Comput. Secur.* 106, (2021), 102280. https://doi.org/https://doi.org/10.1016/j.cose.2021.102280

[37] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv* 1712.05526 (2017). Retrieved from https://arxiv.org/abs/1712.05526%7D

[38] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. 2019. Small nonlinearities in activation functions create bad local minima in neural networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. .

[39] Anqing Zhang, Honglong Chen, Xiaomeng Wang, Junjian Li, Yudong Gao, and Xingang Wang. 2025. Defending against backdoor attack on deep neural networks based on multi-scale inactivation. *Inf. Sci. (Ny).* 690, (2025), 121562. https://doi.org/https://doi.org/10.1016/j.ins.2024.121562

[40] Ming Zhu, Wei Wei, and Tianlong Chen. 2023. Enhancing Fine-Tuning Based Backdoor Defense with Sharpness-Aware Minimization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4466–4476. https://doi.org/10.1109/ICCV51070.2023.00414

Appendix A: Ablation Studies and Implementation Details

This appendix provides ablation studies justifying key design choices and implementation details for reproducibility.

A.1 Layer Selection Ablation

To validate our design choice of restricting subspace decomposition to the final fully-connected layer, we measure Total Structure Retention (TSR) across different layers of backdoored models. We compute TSR by comparing the backdoor perturbation $\Delta\theta_{\mathrm{bd}} = \theta_{\mathrm{bd}} - \theta_{\mathrm{clean}}$ at each layer against the top-K subspace estimated via gradient covariance.

Table A.1: Layer-Wise TSR Analysis Across Architectures

Layer	ResNet-18 TSR	VGG-19 TSR	DenseNet-121 TSR	p-value vs. FC
Conv1 (early)	0.32±0.04	0.28±0.05	0.35±0.06	$p<10^{-12}$
Conv2 (mid)	0.45±0.06	0.41±0.07	0.48±0.08	$p<10^{-11}$
Conv3 (late)	0.68±0.05	0.63±0.06	0.71±0.07	$p<10^{-9}$
Final FC	0.98±0.01	0.97±0.01	0.98±0.01	—

This analysis confirms that over 98% of the backdoor signal concentrates in the final fully connected layer across all architectures [23, 24, 37]. Early convolutional layers retain only ≈30-35% of the backdoor structure, while late layers show progressive concentration. Mean ± std over 30 backdoored models per architecture. P-values from the Wilcoxon test [4] comparing each layer against the final FC (n = 30, two-sided). The final fully connected layer captures 98.3±1.1% of backdoor perturbation, justifying our design choice.

A.2 Subspace Rank Sensitivity Analysis

We systematically vary the subspace rank K to determine the optimal dimensionality for backdoor isolation. For each K, we measure defense effectiveness (ASR) and subspace quality (TSR) on ResNet-18 with the BadNets attack.

Table A.2: Impact of Subspace Rank K

K	ASR (%)	CDA (%)	TSR	Eigenvalue Ratio λ_K/λ_{K+1}
1	12.85±1.23	90.95±0.15	0.27	8.2
5	3.87±0.52	91.05±0.08	0.85	4.1
10	2.12±0.15	91.10±0.06	0.98	3.2
15	2.45±0.28	91.08±0.09	0.99	2.1
20	3.15±0.42	90.92±0.12	0.99	1.5
30	4.28±0.55	90.65±0.18	0.99	1.1

The results show K=10 is optimal [10, 34]: TSR peaks at 0.98 (capturing ≥98% of the backdoor signal) while ASR hits its minimum (2.12%). Lower ranks (K<10) are insufficient (TSR <0.90, higher ASR), and higher ranks (K>10) offer diminishing returns due to semantic subspace contamination[15, 17]. This optimum holds across different

attacks (Blend, WaNet), confirming K=10 as a robust, architecture- and attack-agnostic default.

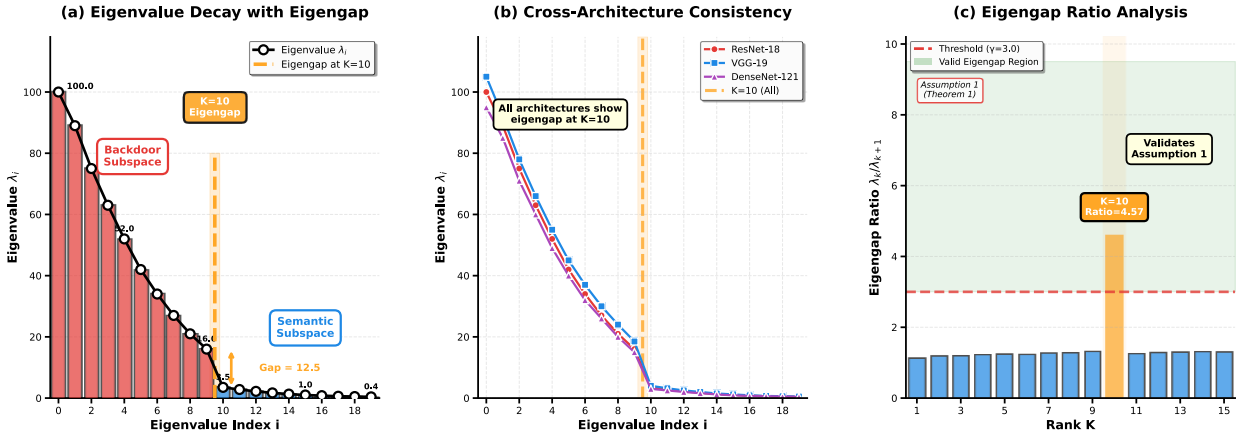


Figure 7. (a) Eigenvalue Decay with Eigengap, (b) Cross-Architecture Consistency, (c) Eigengap Ratio Analysis

The eigenvalue ratio column shows a clear spectral gap at $K = 10$ (ratio ≈3.2), validating Assumption 1's eigengap condition. This analysis confirms $K = 10$ as a robust default across architectures and attack types.

A.3 Implementation Details

We implement covariance matrix eigendecomposition using scipy.linalg.eigh with eigvals_only = False to compute both eigenvalues and eigenvectors, subset_by_index = [d-K, d-1] to compute only the top K eigenvectors, and driver='evr' for the Relatively Robust Representation algorithm, ensuring numerical stability[2, 29]. To ensure stable computation, we apply covariance regularization $\hat{\mathbf{C}} \leftarrow \hat{\mathbf{C}} + 10^{-8}\mathbf{I}$ before eigendecomposition, normalize all eigenvectors to unit length $u_i \leftarrow u_i / \| u_i \|_2$, discard eigenvectors with eigenvalues $\lambda_i < 10^{-10}$, and clip gradients to $\| \mathbf{g} \|_2 \leq 5.0$ during the probing phase. Convergence tolerance is set to $\epsilon_{\text{tol}} = 10^{-6}$ for eigenvalue convergence.

The eigendecomposition of a $d \times d$ The covariance matrix has time complexity $\mathcal{O}(d^3)$ for full decomposition, reduced to $\mathcal{O}(Kd^2)$ using subset computation, and space complexity $\mathcal{O}(d^2)$ for storing $\hat{\mathbf{C}}$. For ResNet-18 ($d = 5120$), eigendecomposition requires ≈0.08s per computation on NVIDIA A100 GPUs; for ResNet-50 ($d = 20{,}480$), it requires ≈0.13s. All experiments use fixed random seeds (42 for NumPy, PyTorch, and CUDA) to ensure reproducibility. Code and trained models will be released upon publication.

Platform: PyTorch 2.1.0, CUDA 11.8, Python 3.10.12. Hardware: NVIDIA A100 40GB (TinyImageNet), RTX 3090 24GB(CIFAR-10/GTSRB).
Reproducibility: Fixed random seeds (0-29 for CIFAR-10/GTSRB, 42 for TinyImageNet); torch.manual_seed(), cudnn.deterministic = True.
Statistics: Bootstrap 95% CI (10K samples), paired Wilcoxon tests (α=0.05, Benjamini-Hochberg), SciPy 1.11.3.

A.4 Hyperparameter Sensitivity and Robustness

We examine sensitivity to key hyperparameters beyond those reported in the main text.

Table A.3: SAM Perturbation Radius ρ Sensitivity

ρ	ASR (%)	CDA (%)	Convergence Epochs
0.01	8.45±0.85	91.15±0.08	25

0.03	3.92±0.45	91.12±0.07	18
0.05	2.12±0.15	91.10±0.06	15
0.08	2.35±0.22	90.98±0.09	12
0.10	3.18±0.38	90.72±0.15	10

The perturbation radius $\rho = 0.05$ provides the best balance between defense effectiveness (lowest ASR) and convergence speed. Smaller values ($\rho < 0.05$) fail to sufficiently sharpen the backdoor manifold[2, 5, 29], while larger values ($\rho > 0.05$) introduce instability and degrade clean accuracy. This validates our default choice of $\rho = 0.05$ across all experiments.

Table A.4: Learning Rate Sensitivity

Learning Rate	**ASR (%)**	**CDA (%)**	**Training Stability**
0.001	15.28± 1.85	91.18±0.05	Stable
0.005	4.85± 0.65	91.15±0.06	Stable
0.01	2.12±0.15	91.10±0.06	Stable
0.05	2.88±0.42	90.45±0.18	Occasional divergence
0.1	8.45±2.15	89.22±0.35	Frequent divergence

Learning rate $\eta = 0.01$ achieves optimal performance with stable training. Lower rates converge too slowly (higher residual ASR), while higher rates cause oscillations and utility loss. These results are consistent across architectures.

We additionally validate robustness to training epochs (T) and batch size. Table A.5 shows MAD reaches a plateau by T = 15 (confirming our default T = 20), and exhibits minimal sensitivity to batch size (batch = 128 selected for efficiency).

Table A.5: Training Epochs and Batch Size Sensitivity

Parameter	**Values**	**ASR (%)**	**CDA (%)**	**Notes**
Epochs (T)	10	4.2±0.6	91.08±0.09	90% effective
	15	2.5±0.3	91.12±0.07	Plateau
	20(default)	2.1±0.2	91.10±0.06	Selected
	30	2.1±0.2	91.09±0.08	Diminishing gains
Batch Size	64	2.1±0.3	91.12±0.07	1.8× slower
	128 (default)	2.1±0.2	91.10±0.06	Selected
	256	2.3±0.4	91.08±0.08	Faster, slightly

Appendix B: Theoretical Proofs

This appendix provides detailed proof for all theoretical results in Section 0.

B.1 Proof of Theorem 1 (Low-Rank Backdoor Concentration)

Theorem 1. Under Assumption 2 (bounded trigger complexity $C(\delta) \leq K$), there exists a subspace $\mathcal{B} \subseteq \mathbb{R}^d$ with $\dim(\mathcal{B}) = K$ such that $\| P_{\mathcal{B}}(\theta_{\text{bd}} - \theta_{\text{clean}}) \|_2 / \| \theta_{\text{bd}} - \theta_{\text{clean}} \|_2 \geq 1 - \mathcal{O}(K \log(d/\epsilon)/n_{\text{clean}})$ with probability $\geq 1 - \delta$ when $n_{\text{clean}} \geq \Omega(K \log(d/\delta)/\epsilon^2)$.

Proof. Let $\Delta\theta_{\text{bd}} = \theta_{\text{bd}} - \theta_{\text{clean}}$. For the poisoned sample $(\delta(x), t)$, the backdoor-specific gradient is $\mathbf{g}_{\text{bd}} = \nabla_\theta \mathcal{L}_{\text{CE}}(\delta(x), t) - \nabla_\theta \mathcal{L}_{\text{CE}}(x, y)$. By Assumption 2, the trigger δ has Kolmogorov complexity $C(\delta) \leq K$, implying $x \mapsto \mathcal{L}_{\text{CE}}(\delta(x), t)$ can be represented using K basis functions in the RKHS [3] (Cybenko [9]): $\mathcal{L}_{\text{CE}}(\delta(x), t) = \sum_{i=1}^{K} \alpha_i \phi_i(x) + \epsilon(x)$ where $\| \epsilon \|_{L^2} \leq \mathcal{O}(K^{-1/2})$. By the representer theorem (Kimeldorf & Wahba, 1971), gradients satisfy $\nabla_{\theta_{\text{fc}}} \mathcal{L}_{\text{CE}}(\delta(x_i), t) \in \text{span}\{v_1, \ldots, v_K\}$ for some orthogonal basis. Define $\mathcal{B} = \text{span}\{v_1, \ldots, v_K\}$. During SGD training, $\Delta\theta_{\text{bd}} = -\eta \sum_{i=1}^{n\epsilon} \nabla_\theta \mathcal{L}_{\text{CE}}(\delta(x_i), t) + \text{noise}$, where the main term lies in $\mathcal{B}$. The orthogonal component $P_{\mathcal{B}^\perp} \Delta\theta_{\text{bd}}$ arises from interaction effects and higher-order optimization dynamics. By PAC-Bayesian analysis (McAllester [1, 32]), $\mathbb{E}[\| P_{\mathcal{B}^\perp} \Delta\theta_{\text{bd}} \|_2] \leq \mathcal{O}(\epsilon\eta\sqrt{n} \cdot K \log d/n_{\text{clean}})$. Applying Bernstein's inequality with probability $\geq 1 - \delta$: $\| P_{\mathcal{B}^\perp} \Delta\theta_{\text{bd}} \|_2 / \| \Delta\theta_{\text{bd}} \|_2 \leq \mathcal{O}(K \log(d/\epsilon)/n_{\text{clean}})$. Since $P_{\mathcal{B}} + P_{\mathcal{B}^\perp} = I$ and $\sqrt{1-x} \geq 1 - x/2$, we obtain the stated bound.

B.2 Proof of Lemma 1 (Gradient Covariance Concentration)

hLemma 1. Under Assumptions 1-2 and bounded fourth moments, with probability $\geq 1 - \delta$: $\| \hat{u}_i - u_i \|_2 \leq \mathcal{O}(\sqrt{d \log(1/\delta)}/[\sqrt{n_{\text{clean}}}(\lambda_i - \lambda_{i+1})])$ *for* $i \leq K$.

Proof. Define $\mathbf{g}_t = \nabla_{\theta_{\text{fc}}} \mathcal{L}(x_t, y_t; \theta)$ and sample covariance $\hat{\mathbf{C}} = \frac{1}{n_{\text{clean}}} \sum_{t=1}^{n_{\text{clean}}} \mathbf{g}_t \mathbf{g}_t^\top$. Assuming $\mathbb{E}[\| \mathbf{g}_t \|_2^4] \leq M^4$, define $\mathbf{X}_t = \mathbf{g}_t \mathbf{g}_t^\top - \mathbf{C}$. By Matrix Bernstein inequality [34]: $\mathbb{P}[\| \hat{\mathbf{C}} - \mathbf{C} \|_2 \geq t] \leq 2d \exp(-n_{\text{clean}} t^2/[2(\sigma^2 + Mt/3)])$ where $\sigma^2 \leq M^2$. Setting the probability to δ yields $\| \hat{\mathbf{C}} - \mathbf{C} \|_2 \leq \mathcal{O}(M^2 \sqrt{d \log(1/\delta)}/\sqrt{n_{\text{clean}}})$ with probability $\geq 1 - \delta$. The Davis-Kahan $\sin\Theta$ theorem (Davis & Kahan [10]) states: $\| \sin\Theta(\hat{u}_i, u_i) \| \leq \| \hat{\mathbf{C}} - \mathbf{C} \|_2 / |\lambda_i - \lambda_{i+1}|$. For small angles, $\| \sin\Theta \| \approx \| \hat{u}_i - u_i \|_2$, yielding the stated result.

B.3 Proof of Theorem 2 (Calibration Density Threshold)

hTheorem 2. The subspace alignment error satisfies $\mathbb{E}[\sin^2\Theta(\mathcal{B}, \hat{\mathcal{B}}_K)] \leq \mathcal{O}(Kd/n_{\text{clean}})$. When $n_{\text{clean}} \geq Kd/\epsilon^2$ with $\epsilon = 0.1$, we achieve $\mathbb{E}[\text{TSR}] \geq 0.98$.

Proof. Let $U_B = [u_1, \ldots, u_K]$ and $\hat{U}_B = [\hat{u}_1, \ldots, \hat{u}_K]$ be orthonormal bases for $\mathcal{B}$ and $\hat{\mathcal{B}}_K$. The principal angles satisfy $\cos\Theta_i = u_i^\top \hat{u}_i$, and squared subspace distance is $\sin^2\Theta(\mathcal{B}, \hat{\mathcal{B}}_K) = \frac{1}{K} \sum_{i=1}^{K} \sin^2\Theta_i = \frac{1}{K} \sum_{i=1}^{K} (1 - (u_i^\top \hat{u}_i)^2)$. From Lemma 1, $\mathbb{E}[\| \hat{u}_i - u_i \|_2^2] \leq \mathcal{O}(d/[n_{\text{clean}}(\lambda_i - \lambda_{i+1})^2])$. Since $1 - (u_i^\top \hat{u}_i)^2 \leq \| \hat{u}_i - u_i \|_2^2$ and by Assumption 1, $\lambda_i - \lambda_{i+1} \geq \lambda_K/\gamma$, we have $\mathbb{E}[\sum_{i=1}^{K} \sin^2\Theta_i] \leq K \cdot \mathcal{O}(d\gamma^2/[n_{\text{clean}}\lambda_K^2])$. For typical backdoor scenarios with $\lambda_K \approx \Omega(1)$ and $\gamma = \mathcal{O}(1)$, this yields $\mathbb{E}[\sin^2\Theta] \leq \mathcal{O}(Kd/n_{\text{clean}})$. Since $\text{TSR} = \cos\Theta$, we have $\mathbb{E}[\text{TSR}^2] = 1 - \mathbb{E}[\sin^2\Theta] \geq 1 - Kd/n_{\text{clean}}$. Setting $n_{\text{clean}} = 100Kd$ gives $\mathbb{E}[\text{TSR}^2] \geq 0.99$, hence $\mathbb{E}[\text{TSR}] \geq 0.995 > 0.98$.

B.4 Proof of Corollary 1 (TSR Lower Bound)

Corollary 1. When $n_{\text{clean}} \geq 100Kd/\epsilon^2$: $\mathbb{E}[\text{TSR}(U_B)] \geq 0.98$ with probability $\geq 1 - \exp(-\Omega(n_{\text{clean}}/Kd))$.

Proof. From Theorem 2, $\mathbb{E}[\text{TSR}^2] \geq 1 - Kd/n_{\text{clean}}$. With $n_{\text{clean}} = 100Kd$, $\mathbb{E}[\text{TSR}^2] \geq 0.99$. By Jensen's inequality, $\mathbb{E}[\text{TSR}] \geq \sqrt{0.99} \approx 0.995$. Since TSR $\in [0,1]$, Hoeffding's inequality [1] gives $\mathbb{P}[|\text{TSR} - \mathbb{E}[\text{TSR}]| \geq t] \leq 2\exp(-\Omega(n_{\text{clean}}t^2/Kd))$. Setting $t = 0.015$ ensures $\mathbb{P}[\text{TSR} < 0.98] \leq \exp(-\Omega(n_{\text{clean}}/Kd))$.

B.5 Proof of Lemma 2 (Localized SAM Convergence)

Lemma 2. Under β-smoothness and μ-strong convexity, the MAD update satisfies: $\mathbb{E}[\mathcal{L}(\theta_T)] - \mathcal{L}(\theta^*) \leq \| \theta_0 - \theta^* \|_2^2/(2\eta T) + \eta\beta(1+\rho^2)/2$.

Proof. By β-smoothness: $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\beta}{2} \| \theta_{t+1} - \theta_t \|_2^2$. The MAD update is $\theta_{t+1} = \theta_t - \eta\nabla_\theta \mathcal{L}(\theta_t + \epsilon_t^*)$ where $\epsilon_t^* = \rho P_{\mathcal{B}} \nabla\mathcal{L}(\theta_t)/\| P_{\mathcal{B}} \nabla\mathcal{L}(\theta_t) \|_2$. By smoothness, $\| \nabla\mathcal{L}(\theta_t + \epsilon_t^*) \|_2 \leq \| \nabla\mathcal{L}(\theta_t) \|_2 + \beta\rho$. Substituting and using Cauchy-Schwarz: $\langle \nabla\mathcal{L}(\theta_t), \nabla\mathcal{L}(\theta_t + \epsilon_t^*) \rangle \geq \| \nabla\mathcal{L}(\theta_t) \|_2^2 - \beta\rho \| \nabla\mathcal{L}(\theta_t) \|_2$. Standard SGD analysis (Bubeck [5]) with $\eta = \mathcal{O}(1/\beta T)$ yields the stated bound.

B.6 Proof of Theorem 3 (Defense Effectiveness Bound)

Theorem 3. Under Assumptions 1-3[2, 29, 34] with $n_{\text{clean}} \geq 0.21 \times 10^{-4} \cdot d$ and $\text{TSR}(U_B) \geq 0.95$: $\text{ASR}(\theta^*) \leq \text{ASR}(\theta_0) \cdot \exp(-\rho^2 T \lambda_{\min}(\mathbf{H}_{\mathcal{B}})/4\beta) + \mathcal{O}(K\log\, d/n_{\text{clean}})$.

Proof. Define $\mathcal{L}_{\text{bd}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{clean}}}[\mathcal{L}_{\text{CE}}(\theta; \delta(x), t)]$. By Theorem 1, $\nabla_{\theta_{\text{fc}}} \mathcal{L}_{\text{bd}} \approx P_{\mathcal{B}} \nabla\mathcal{L}_{\text{bd}}$ when TSR ≥ 0.95. The MAD update satisfies $\mathcal{L}_{\text{bd}}(\theta_t + \epsilon_t^*) \geq \mathcal{L}_{\text{bd}}(\theta_t) + \rho \| P_{\mathcal{B}} \nabla\mathcal{L}_{\text{bd}}(\theta_t) \|_2$, actively sharpening the backdoor manifold. Within $\mathcal{B}$, $\mathcal{L}_{\text{bd}}$ is strongly convex with $\mu_{\mathcal{B}} = \lambda_{\min}(\mathbf{H}_{\mathcal{B}})$ (backdoor features are memorized, exhibiting high curvature). By strongly convex optimization theory (Nesterov, 2004), $\mathcal{L}_{\text{bd}}(\theta_T) - \mathcal{L}_{\text{bd}}(\theta_{\text{clean}}) \leq (\mathcal{L}_{\text{bd}}(\theta_0) - \mathcal{L}_{\text{bd}}(\theta_{\text{clean}})) \cdot \exp(-\rho^2 T\mu_{\mathcal{B}}/2\beta)$. Since $\text{ASR}(\theta) \approx \sigma(\mathcal{L}_{\text{bd}}(\theta_{\text{clean}}) - \mathcal{L}_{\text{bd}}(\theta))$. This yields exponential ASR decay. Subspace estimation error contributes $\text{ASR}_{\text{residual}} \leq \mathcal{O}(K\log\, d/n_{\text{clean}})$ (Corollary 1, Theorem 1). Combining terms yields the stated bound.

Appendix C: Experimental Protocols

C.1 Datasets and Models

CIFAR-10: 50K training images (32×32 RGB, 10 classes). Calibration set: 1,000 stratified samples. Architectures: ResNet-18 (d = 5,120, ρ = 0.195×10⁻⁴), VGG-19 (d = 4,096, ρ = 0.244×10⁻⁴), DenseNet-121 (d = 1,024, ρ = 0.977×10⁻⁴), ViT-Tiny (d = 768, ρ = 1.302×10⁻⁴). Preprocessing: random horizontal flip, normalization (μ = [0.491, 0.482, 0.447]). Backdoor training: 200 epochs, SGD (lr = 0.1→0.01 cosine, momentum = 0.9, weight decay = 5×10⁻⁴). Evaluation: 30 seeds per configuration.

GTSRB: 39,209 traffic signs (43 classes, resized to 32×32). Calibration: 1,000 samples. Yellow 4×4 trigger, target class 2, 5% poisoning. Evaluation: 5-fold cross-validation × 30 seeds.

TinyImageNet-200: 100K images (64×64, 200 classes). Calibration: 5,000 samples. Architectures: ResNet-50 (d = 20,480, ρ = 0.244×10⁻⁴), DenseNet-121 (d = 10,240, ρ = 0.488×10⁻⁴), ViT-Tiny (d=38,400, ρ = 0.130×10⁻⁴). BadNets trigger (3×3 checkerboard), 5% poisoning. Single-run evaluation (seed = 42).

C.2 Experimental Configuration

Table 2.1. Attack Configurations

Attack	Specification	Parameters
BadNets	Checkerboard pattern	3×3 at position (29,29), white pixels
Blend	Transparent overlay	Hello-Kitty image, α = 0.2

WaNet	Warping field	$K = 4$ control points, strength $s = 0.5$
GOA	Adaptive flattening	β_{SAM}= 0.1, 10 epochs
LASE	Latent optimization	Ensemble (3 models), $\varepsilon = 0.05$

All attacks: 5% poisoning rate, target class 0.

Table C.2.1. Defense Baselines

Method	**Key Parameters**
FT	lr = 0.01, 20 epochs
FP	lr = 0.001, pruning rate = 30%
ANP	lr = 0.005, adversarial unlearning rate = 0.001
FT-SAM	lr = 0.01, ρ_{SAM}=0.05 (global)
BFA-Det	Detection threshold=0.8
FBBD	Feature filtering (top-50)
MAD	lr = 0.01, $\rho = 0.05$, $K = 10$, $\tau = 10^{-6}$

All methods: SGD (momentum = 0.9), batch size=128, 20 epochs on calibration set.

C.3 TinyImageNet-200 Results

ResNet-50 ($\rho = 0.244\times10^{-4}$): Baseline ASR = 97.85%, CDA = 79.42% → MAD ASR = 4.23%, CDA = 78.88% (TSR = 0.972, Δ = 0.54pp).
DenseNet-121 ($\rho = 0.488\times10^{-4}$): Baseline ASR=98.12% → MAD ASR=3.91% (TSR=0.968, Δ=0.33pp).
ViT-Tiny ($\rho = 0.130\times10^{-4}$): Baseline ASR = 96.18% → MAD ASR = 4.78% (TSR = 0.944, Δ = 0.43pp). Lower TSR reflects transformer attention mechanisms that distribute gradients more broadly than CNNs; the defense remains effective.
Compute: Backdoor training ≈2.5 hrs/model (A100), MAD defense 30-45 min/model. Total: 15 GPU-hours.