A Theoretical Analysis of Backdoor Learning as Simplicity-Biased Optimization Dynamics

Hong Cheng[1], **Qiang Zhou**[1,2,3]*, **Sinno Jialin Pan**[4]
[1]Southeast University
[2]Purple Mountain Laboratories
[3]Engineering Research Center of Blockchain Application, Supervision and Management (Southeast University), Ministry of Education
[4]The Chinese University of Hong Kong
chenghongcs@seu.edu.cn, zhouqiang@u.nus.edu, sinnopan@cuhk.edu.hk

Abstract

Backdoor attacks implant a trigger-target association into a model, causing malicious behavior at test time while largely preserving clean performance. Despite extensive empirical study, a unified explanation for why standard training dynamics learn backdoors so effectively remains largely missing. We argue that backdoor learning can be understood as a consequence of simplicity-biased optimization dynamics: trigger features are dynamically simpler than semantic features under stochastic gradient descent (SGD) because they induce more coherent gradient alignment, more stable gating behavior, and stronger early-stage amplification. To formalize this perspective, we introduce a directional data model that separates semantic and trigger features and study a one-hidden-layer ReLU network trained by SGD. Our analysis identifies a frozen-gate signal that governs group-wise early-stage loss decrease up to controlled gate-drift and gate-flip errors, yielding a quantitative explanation for faster poisoned-sample fitting and its dependence on the poisoning ratio. Beyond optimization dynamics, we show that this early-stage bias can induce trigger-dominant neurons with selective activation on poisoned inputs, explaining why vanilla clean fine-tuning can attenuate backdoor behavior without necessarily erasing the underlying trigger-related representation. Empirically, we validate the predicted signatures of early-stage optimization bias, representation-level trigger dominance, and post-training residual behavior using ResNet-18 on standard image classification benchmarks.

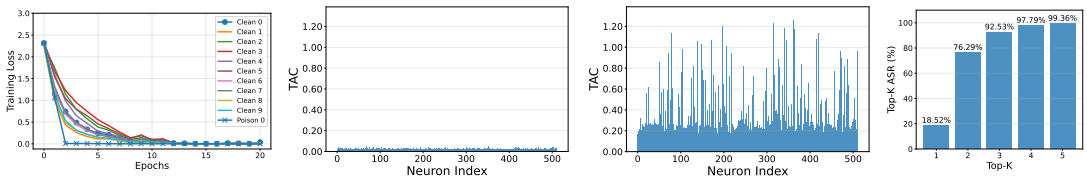
1 Introduction

Backdoor attacks implant a trigger-target association into a model through a small fraction of poisoned training data, causing the trained model to behave normally on clean inputs while predicting an attacker-specified target label whenever the trigger is present [2, 5]. This threat has motivated extensive research on backdoor attacks and defenses [28, 4], yet a fundamental question remains largely unresolved: *why do standard neural network training dynamics learn backdoors so effectively, even when poisoned data are rare?*

A growing body of empirical evidence suggests that backdoor learning is not merely a final-model phenomenon, but is already shaped by the early optimization process: (i) poisoned samples are often fitted faster than clean samples during the early stage of training [13, 32]; (ii) a small subset of neurons can become highly responsive to trigger patterns while remaining relatively inactive on clean inputs [15, 29, 33]; and (iii) post-training defenses such as clean fine-tuning can weaken backdoor

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Loss dynamics (b) TAC: clean vs. backdoored (c) Top-K ASR

Figure 1: Motivating empirical observations on CIFAR-10 under BadNets [5] with a poisoning ratio of $\alpha = 5\%$. (a) We compare class-wise clean training-loss trajectories with the poisoned-group trajectory targeting label 0. The rapid loss decrease of the poisoned group is not simply explained by clean samples from the target class. (b) The neuron-wise trigger-activated change (TAC) profile from [33] shows that the backdoored model contains a small subset of neurons with much larger TAC values than the clean model. (c) After clean fine-tuning, Top-1 attack success rate (ASR) is substantially reduced, but Top-K ASR remains high for larger K, indicating that the target label still receives a high rank under triggered inputs.

behavior but may not fully erase the underlying trigger-related representation [36, 21]. Figure 1 summarizes three representative observations on CIFAR-10 [11] under BadNets [5] that motivate our dynamics-level analysis. The corresponding experimental details are provided in Appendix C.

Together, these observations suggest that backdoor learning is not only an output-level phenomenon. It is shaped by early-stage optimization, leaves activation-level trigger sensitivity in the learned representation, and can persist as residual target-label preference after vanilla clean fine-tuning. This motivates a unified explanation that connects early-stage optimization dynamics, representation-level bias, and post-training defense behavior.

Our perspective. We argue that backdoor learning can be understood as a consequence of *simplicity-biased optimization dynamics*. The key idea is that trigger features often behave as dynamically simpler directions than semantic features under stochastic gradient descent (SGD). In this paper, semantic features refer to representation components that are predictive of the ground-truth label, whereas trigger features refer to representation components introduced by the backdoor pattern and predictive of the attacker-specified target label on poisoned samples. Because trigger-related components are cross-sample consistent and directly aligned with the target label on poisoned samples, they behave as dynamically simpler directions, inducing more coherent gradient alignment, more stable gating behavior, and stronger early-stage amplification than semantic features. As a result, SGD can amplify trigger-aligned directions rapidly, even when the poisoning ratio is small.

To formalize this perspective, we introduce a directional data model that separates semantic and trigger features, and study a one-hidden-layer ReLU network trained by SGD. Our analysis identifies a *frozen-gate signal* that captures the group-wise early-stage optimization bias induced by trigger-aligned directions. We show that the one-step loss decrease is governed by this frozen-gate signal up to controlled gate-drift and gate-flip errors. This yields a quantitative explanation for why poisoned samples can exhibit faster early-stage loss decrease than clean samples. We further validate the predicted optimization signatures on standard image classification benchmarks and show that the same mechanism leads to representation-level consequences, including trigger-dominant neurons and residual trigger-related structure after vanilla clean fine-tuning.

Related work. Our work is related to three lines of research. First, backdoor attacks and defenses have studied how to implant or mitigate trigger-target associations in neural networks [2, 5, 26, 15, 14, 31, 20]. Second, recent studies have identified training-time and neuron-level signatures of backdoor learning, including faster fitting of poisoned samples and trigger-sensitive neurons [33, 13, 8, 32]. Recent theoretical studies have further analyzed backdoor success from adaptability, statistical, and CNN-specific perspectives [30, 27, 12]. Third, theoretical and empirical studies on simplicity bias and shortcut learning suggest that neural networks trained by gradient-based methods tend to favor simpler or easier-to-fit functions and cues [25, 22, 17, 24, 7]. However, existing simplicity-bias analyses are not designed for malicious poisoning settings, and existing backdoor studies often treat faster poisoned fitting, trigger-sensitive neurons, and the persistence of trigger-related behavior after post-training mitigation as separate empirical phenomena. In contrast, our goal is to provide a unified dynamics-level explanation that connects these phenomena through early-stage optimization dynamics. Detailed related work is deferred to Appendix D.

Contributions. Our main contributions are summarized as follows:

- We propose a simplicity-biased optimization perspective on backdoor learning, showing that trigger features act as dynamically simpler directions than semantic features under SGD.
- We provide a formal early-stage characterization of SGD dynamics through a frozen-gate signal, showing that poisoned samples exhibit a strictly faster loss decrease than clean samples when the signal gap dominates the gate-stability errors.
- We empirically validate the predicted early-stage optimization-bias signatures of poisoned versus clean samples in practical deep networks on standard image classification benchmarks.
- We connect these optimization dynamics to representation-level and post-training consequences, including the emergence of trigger-dominant neurons and the residual trigger-related representation after vanilla clean fine-tuning.

2 Setup and Directional Backdoor Model

We introduce a stylized directional model that isolates the competition between semantic and trigger-related representations. The model is not intended to reproduce all low-level details of any particular data modality. Instead, it abstracts the learned representation into two predictive directions and studies how SGD amplifies them during the early stage of training. Additional discussion on the binary prototype, the target-vs-rest interpretation, and the multi-class setting is provided in Appendix C.3.

Directional backdoor model. We consider a binary classification problem with semantic label $y \in \{\pm 1\}$ and attacker-specified target label $\widetilde{y} \in \{\pm 1\}$. We assume $\Pr(y=1) = \Pr(y=-1) = 1/2$, so that $\mathbb{E}[y] = 0$. Each sample is associated with a poisoning indicator $z \sim \mathrm{Bernoulli}(\alpha)$, where $z = 0$ denotes a clean sample, $z = 1$ denotes a poisoned sample, and $\alpha \in (0,1)$ is the poisoning ratio. The actual training label for each sample is

$$\widehat{y} = (1-z)y + z\widetilde{y}. \tag{1}$$

Hence, the actual training label $\widehat{y}$ equals the ground-truth label y for clean samples ($z = 0$) and the attacker-specified target label $\widetilde{y}$ for poisoned samples ($z = 1$).

We use the term *semantic feature* to refer to a representation component whose signed correlation with the ground-truth label is positive. In the stylized model, this component is summarized by a unit semantic direction $\mathbf{u}_s$ with signal strength $\mu_s > 0$. We use the term *trigger feature* to refer to a representation component introduced by the backdoor pattern whose signed correlation with the target label is positive on poisoned samples. This component is summarized by a unit trigger direction $\mathbf{u}_t$ with signal strength $\mu_t > 0$. These are representation-level abstractions rather than specific pixels or handcrafted patterns. The representation vector $\mathbf{x} \in \mathbb{R}^d$ is generated as

$$\mathbf{x} = y\mu_s\mathbf{u}_s + z\widetilde{y}\mu_t\mathbf{u}_t + \boldsymbol{\xi}, \tag{2}$$

where $\boldsymbol{\xi}$ is isotropic Gaussian noise independent of (y, z).

Assumption 1 (Orthogonal semantic and trigger directions)**.** *The semantic direction $\mathbf{u}_s$ and the trigger direction $\mathbf{u}_t$ are orthogonal, i.e., $\langle \mathbf{u}_s, \mathbf{u}_t \rangle = 0$.*

Assumption 1 is introduced to *separate the optimization dynamics* from trivial statistical correlations between semantic and trigger features, and to simplify the directional decomposition in our analysis. In high-dimensional settings, approximate orthogonality is natural, and small nonzero correlations can be absorbed into lower-order terms or effective signal strengths μ_s and μ_t, leaving our results essentially unchanged. Under this model and Assumption 1, it holds that $\mathbb{E}[\widehat{y}\mathbf{x} \mid z = 0] = \mu_s\mathbf{u}_s$ and $\mathbb{E}[\widehat{y}\mathbf{x} \mid z = 1] = \mu_t\mathbf{u}_t$. This decomposition provides the group-wise supervised signals along different representation directions, serving as the starting point for our dynamics analysis.

Training setup. We adopt a one-hidden-layer ReLU network as a minimal nonlinear model for studying early-stage optimization dynamics under SGD, following a standard theoretical setup widely used in prior work on simplicity bias, and over-parameterized neural networks [25, 3, 19]:

$$f\big(\mathbf{x}; \{\mathbf{w}_j\}_{j=1}^k\big) = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} v_j \mathrm{ReLU}\big(\mathbf{x}^\top \mathbf{w}_j\big), \tag{3}$$

where $\mathrm{ReLU}(u) = \max\{0, u\}$, $\mathbf{w}_j \in \mathbb{R}^d$ are trainable hidden-layer weights, and $v_j \in \{\pm 1\}$ are fixed output signs drawn independently as Rademacher random variables. Fixing the output weights isolates the hidden-layer representation learning dynamics. The $1/\sqrt{k}$ normalization ensures non-degenerate network output and gradients with respect to the width k.

The network is trained only on the hidden-layer weights using SGD with the hinge loss

$$\ell(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j\}_{j=1}^k) = \max\{0, 1 - \widehat{y} f(\mathbf{x}; \{\mathbf{w}_j\}_{j=1}^k)\}. \tag{4}$$

Here $\mathbf{w}_j^t$ denotes the value of the hidden-layer weight $\mathbf{w}_j$ at iteration t.

Assumption 2 (Gaussian small initialization)**.** *The hidden-layer weights are initialized independently as $\mathbf{w}_j^0 \sim \mathcal{N}(0, \sigma_w^2 \mathbf{I})$ for $j \in [k]$. We choose σ_w sufficiently small so that the event $\mathcal{E}_{\mathrm{init}} = \{\max_{j\in[k]} \|\mathbf{w}_j^0\| \le R_0\}$ holds with high probability, where R_0 is chosen sufficiently small to ensure the early-stage active-loss regime.*

Assumption 2 serves two roles in the early-stage analysis. The small initialization scale, through the bound R_0, helps ensure that the hinge loss remains active during the early stage, while the Gaussian form provides the anti-concentration needed to control ReLU gate flips.

3 Early-Stage Simplicity-Biased Optimization Dynamics

In this section, we analyze the early-stage optimization dynamics induced by SGD under the directional backdoor model introduced in Section 2. Our goal is not to characterize the final converged solution, but to understand how different feature directions are amplified at different rates during the early stage of training. The analysis proceeds in four steps: we first establish an early-stage regime in which the hinge loss remains active for both clean and poisoned samples (Lemma 1); we then characterize the population signal structure and the initial SGD drift along semantic and trigger directions (Lemmas 2 and 3); next, we analyze ReLU gating and bound the probability of gate flips over the early-stage horizon (Lemma 4); finally, we combine these ingredients to quantitatively characterize early-stage loss dynamics (Theorem 1), deriving a strict separation between poisoned and clean samples and its dependence on the poisoning ratio (Corollaries 1 and 2). Formal proofs of all theoretical results are deferred to Appendix A.

Active-loss regime. We first formalize the early-stage horizon on which the hinge loss remains active. This ensures that the SGD updates have a uniform form, which allows us to analyze directional drift and ReLU gating around initialization.

Lemma 1 (Active-loss regime)**.** *Consider a one-hidden-layer ReLU network trained by SGD on the hinge loss, with a training set $\mathcal{D} = \{(\mathbf{x}_i, \widehat{y}_i)\}_{i=1}^N$ generated by* (2) *and* (1)*. Fix $\delta \in (0, 1)$ and define the bounded-input event $\mathcal{E}_B \triangleq \{\max_{i\in[N]} \|\mathbf{x}_i\| \le B\}$, where $B \triangleq (\mu_s + \mu_t) + \sigma(\sqrt{d} + \sqrt{2\log(4N/\delta)})$. Then $\Pr(\mathcal{E}_B) \ge 1 - \delta$. Under Assumption 2 and conditioned on $\mathcal{E}_B$, suppose further that $\sqrt{k} B R_0 \le c_0$ for some sufficiently small universal constant $c_0 > 0$. For SGD with step size $\eta > 0$, define $T \triangleq \lfloor c_1/(\eta B^2) \rfloor$ for some universal constant $c_1 > 0$. Then, for all iterations $t \in \{0, 1, \ldots, T\}$ and all training samples $i \in [N]$, we have $\widehat{y}_i f(\mathbf{x}_i; \{\mathbf{w}_j^t\}_{j=1}^k) < 1$. Hence the hinge loss remains active throughout the first T iterations.*

Lemma 1 is technical but important: it identifies an early-stage horizon on which loss gradients do not vanish. All subsequent dynamics statements are made on this horizon.

Population signals and initial drift. For each group $z \in \{0, 1\}$, define the population signal vector $\mathbf{m}_z \triangleq \mathbb{E}[\widehat{y}\mathbf{x} \mid z]$. The following lemma identifies the clean and poisoned supervised signals induced by the directional backdoor model.

Lemma 2 (Population signal decomposition)**.** *Under the directional backdoor model in* (2) *and the training-label rule in* (1)*, it holds that $\mathbf{m}_0 = \mu_s \mathbf{u}_s$ and $\mathbf{m}_1 = \mu_t \mathbf{u}_t$. In particular, $\mathbb{E}[\widehat{y}\langle \mathbf{x}, \mathbf{u}_s\rangle \mid z = 0] = \mu_s$ and $\mathbb{E}[\widehat{y}\langle \mathbf{x}, \mathbf{u}_t\rangle \mid z = 1] = \mu_t$. Moreover, under Assumption 1, the cross-directional correlations vanish,* i.e.*, $\mathbb{E}[\widehat{y}\langle \mathbf{x}, \mathbf{u}_t\rangle \mid z = 0] = 0$ and $\mathbb{E}[\widehat{y}\langle \mathbf{x}, \mathbf{u}_s\rangle \mid z = 1] = 0$.*

Lemma 2 shows that clean samples induce a supervised signal along the semantic direction, whereas poisoned samples induce a supervised signal along the trigger direction. The next lemma connects these population signals to the initial SGD drift.

Lemma 3 (Initial SGD drift alignment)**.** *Consider a one-hidden-layer ReLU network trained by SGD on the hinge loss with step size η. Work under the conditions of Lemma 1 so that the hinge loss is active at iteration $t = 0$. Assume the hidden-layer weights satisfy Assumption 2, and the output signs $\{v_j\}_{j=1}^k \subseteq \{\pm 1\}$ are fixed. Let $\Delta \mathbf{w}_j^0(z) \triangleq \mathbf{w}_j^1 - \mathbf{w}_j^0$ denote the SGD update of neuron j at iteration $t = 0$ when the sampled example satisfies $z \in \{0, 1\}$. Define $\Delta \overline{\mathbf{w}}^0(z) \triangleq \frac{1}{k} \sum_{j=1}^k \Delta \mathbf{w}_j^0(z)$ and $\bar{v} \triangleq \frac{1}{k} \sum_{j=1}^k v_j$. Then, taking expectation over the initialization and the conditional data distribution given z, it holds that $\mathbb{E}[\Delta \overline{\mathbf{w}}^0(z)] = \frac{\eta}{2\sqrt{k}} \bar{v} \mathbf{m}_z$. In particular, by Lemma 2, it holds that $\mathbb{E}[\Delta \overline{\mathbf{w}}^0(0)] = \frac{\eta \mu_s}{2\sqrt{k}} \bar{v} \mathbf{u}_s$ and $\mathbb{E}[\Delta \overline{\mathbf{w}}^0(1)] = \frac{\eta \mu_t}{2\sqrt{k}} \bar{v} \mathbf{u}_t$.*

Lemma 3 shows that, at initialization, clean and poisoned examples drive the averaged hidden-layer update along different representation directions. In a ReLU network, this ungated supervised signal is further modulated by activation patterns, which motivates the frozen-gate signal below.

Frozen-gate signal. For $j \in [k]$, define the ReLU gate at iteration t as $A_j^t(\mathbf{x}) = \mathbb{I}\{(\mathbf{w}_j^t)^\top \mathbf{x} > 0\}$. To capture joint ReLU activation, define the empirical joint gating factor $G_t(\mathbf{x}_1, \mathbf{x}_2) \triangleq \frac{1}{k} \sum_{j=1}^k A_j^t(\mathbf{x}_1) A_j^t(\mathbf{x}_2)$, which quantifies the empirical co-activation rate of two inputs at iteration t. Let $(\mathbf{x}_1, \widehat{y}_1)$ and $(\mathbf{x}_2, \widehat{y}_2)$ be two IID samples drawn from the conditional distribution given $z_1 = z_2 = z \in \{0, 1\}$. Define the group-wise frozen-gate signal as

$$h_z = \mathbb{E}\big[\widehat{y}_1 \widehat{y}_2 \mathbf{x}_1^\top \mathbf{x}_2 G_0(\mathbf{x}_1, \mathbf{x}_2) \mid z_1 = z_2 = z\big], \tag{5}$$

where G_0 is the joint gating factor at initialization. The expectation defining h_z is over the two data samples with the realized initialization fixed. Thus, h_z couples signed-feature agreement with initial gate overlap. This signal has an exact neural tangent kernel (NTK) interpretation [10]. Let $\boldsymbol{\theta}$ collect the trainable hidden-layer weights and write $f_{\boldsymbol{\theta}}$ for the network. Since the output weights are fixed and satisfy $v_j^2 = 1$, the finite-width NTK, using the gate convention above, satisfies

$$K_t(\mathbf{x}_1, \mathbf{x}_2) \triangleq \langle \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}_1), \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{x}_2) \rangle|_{\boldsymbol{\theta} = \boldsymbol{\theta}^t} = \mathbf{x}_1^\top \mathbf{x}_2 \, G_t(\mathbf{x}_1, \mathbf{x}_2).$$

Consequently, h_z is an unnormalized within-group signed alignment under K_0.

This identity does not assume that $K_t = K_0$ during training. The backdoor-specific question concerns how an injected, target-aligned trigger changes this group-wise supervised alignment. Appendix B analyzes the associated feature–gate geometry and gives an exact gated-signal representation. To relate the initial signal to subsequent early-stage loss decreases, the analysis must control both gate drift and the residual from ReLU gate crossings. The next lemma provides the gate-stability estimate used in this approximation.

Lemma 4 (Early-stage gate stability)**.** *Work under the conditions of Lemma 1 and Assumption 2, and condition on $\mathcal{E}_B$. Fix any neuron $j \in [k]$ and any $t \le T$. Then for any fixed vector $\mathbf{x} \in \mathbb{R}^d$ independent of $\mathbf{w}_j^0$, it holds that*

$$\Pr\big(A_j^t(\mathbf{x}) \neq A_j^0(\mathbf{x}) \mid \mathcal{E}_B\big) \le \frac{C}{\sigma_w} \mathbb{E}\big[\|\mathbf{w}_j^t - \mathbf{w}_j^0\| \mid \mathcal{E}_B\big]. \tag{6}$$

This bound holds provided the weight drift is independent of the initialization margin, or alternatively when the drift is replaced by a deterministic upper bound. In particular, if $\|\mathbf{w}_j^t - \mathbf{w}_j^0\| \le t\eta B/\sqrt{k}$, then $\Pr\big(A_j^t(\mathbf{x}) \neq A_j^0(\mathbf{x}) \mid \mathcal{E}_B\big) \le Ct\eta B/(\sigma_w \sqrt{k})$. Moreover, for any fixed pair $(\mathbf{x}_1, \mathbf{x}_2)$, the joint-gate deviation between time t and initialization is bounded by $2Ct\eta B/(\sigma_w \sqrt{k})$.

Lemma 4 controls the probability that ReLU gates differ from their initial states over the early-stage horizon, rather than assuming that the gates remain fixed. Since this bound grows with the accumulated weight drift, the frozen-gate approximation is informative when the resulting deviation remains small. This control of joint-gate deviations is used in Theorem 1 to relate the initial frozen-gate signal to subsequent early-stage loss decreases, up to a controlled approximation error.

Main dynamics result. Define the population hinge loss of group $z \in \{0, 1\}$ at iteration t as $L_z^t \triangleq \mathbb{E}[\ell(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j^t\}_{j=1}^k) \mid z]$, where $(\mathbf{x}, \widehat{y})$ is an independent draw from the conditional distribution given z. The following result characterizes the early-stage one-step loss decrease for each group.

Theorem 1 (Early-stage loss dynamics)**.** *Work under the setup of Section 2, the active-loss regime in Lemma 1, and the gate-stability conditions in Lemma 4. Then there exists a universal constant $C > 0$ such that, for every $z \in \{0, 1\}$ and every $t \leq T - 1$,*

$$\left|\mathbb{E}\left[L_z^t - L_z^{t+1} \mid z_t = z, \mathcal{E}_B\right] - \eta h_z\right| \leq 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}. \tag{7}$$

Theorem 1 provides a group-symmetric characterization of early-stage optimization: for either $z \in \{0, 1\}$, the conditional expected one-step loss decrease is governed by the corresponding frozen-gate signal h_z, up to controlled errors from gate drift and gate flips. The theorem neither assumes nor guarantees a poisoned-group advantage. Rather, which group enjoys a larger conditional expected loss decrease depends on whether the corresponding signal gap dominates these errors, as made explicit in Corollary 1.

Poisoned-clean separation. For $z \in \{0, 1\}$, define $D_z^t \triangleq \mathbb{E}[L_z^t - L_z^{t+1} \mid z_t = z, \mathcal{E}_B]$. The main theorem immediately yields a sufficient condition under which poisoned samples exhibit a strictly larger early-stage loss decrease than clean samples.

Corollary 1 (Group-wise loss separation)**.** *Work under the conditions of Theorem 1. Then, for any $t \leq T - 1$, it holds that $|D_z^t - \eta h_z| \leq 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}$. In particular, if $h_1 - h_0 > 8C\frac{(t+1)\eta B^3}{\sigma_w\sqrt{k}}$, then $D_1^t > D_0^t$.*

Corollary 1 thus gives a sufficient condition for a poisoned-group advantage, not an unconditional ordering between groups. The backdoor-specific mechanism lies in how the injected, target-aligned trigger changes the poisoned group's data geometry: its shared component can strengthen signed-feature agreement and joint ReLU gating overlap, leading to a larger frozen-gate signal h_1. This advantage need not follow merely from a larger nominal trigger signal strength; differences in group geometry can matter even when semantic and trigger strengths are comparable. Conversely, a sufficiently large signal gap in the opposite direction would favor the clean group. Appendix B discusses the relevant geometric ingredients, while Figure 4 reports corresponding empirical diagnostics.

Effect of poisoning ratio. Corollary 1 establishes a conditional group-wise comparison. In actual SGD, samples are drawn from a mixture of clean and poisoned data according to the poisoning ratio. The next corollary lifts this comparison to the mixture-level expectation conditioned on $\mathcal{E}_B$, and characterizes how α modulates the overall early-stage loss dynamics.

Corollary 2 (Poisoning-ratio-weighted loss dynamics)**.** *Work under the conditions of Theorem 1. Let $\Pr(z_t = 1) = \alpha$ and $\Pr(z_t = 0) = 1 - \alpha$. Then, for any $t \leq T - 1$,*

$$\left|(1-\alpha)D_0^t + \alpha D_1^t - \eta((1-\alpha)h_0 + \alpha h_1)\right| \leq 4C(t+1)\eta^2 B^3/(\sigma_w\sqrt{k}). \tag{8}$$

Moreover, if the gap condition $\alpha h_1 - (1-\alpha)h_0 > 4C(t+1)\eta B^3/(\sigma_w\sqrt{k})$ holds for some $t \leq T-1$, then $\alpha D_1^t > (1-\alpha)D_0^t$.

Corollary 2 explains how the global early-stage loss decay depends on the poisoning ratio. For fixed h_0 and h_1, the left-hand side of the gap condition increases linearly with α, so a larger poisoning ratio makes the condition easier to satisfy and leads to a stronger poisoned-dominant loss decrease. Section 4 empirically verifies this trend by varying α.

4 Empirical Validation of Early-Stage Optimization Bias

We next investigate whether the early-stage optimization bias characterized in Section 3 exhibits corresponding signatures in practical deep networks. The theory predicts that poisoned samples can exhibit a larger frozen-gate signal and, consequently, a larger early-stage loss decrease. We evaluate this prediction on practical deep networks using ResNet-18 [6] on CIFAR-10 and CIFAR-100 [11] under two representative backdoor attacks, BadNets [5] and TrojanNN [16].The empirical quantity $h_z^{\bar{t}}$ is a diagnostic analogue of the theoretical frozen-gate signal h_z: it is computed using target-versus-rest signed penultimate-layer representations and Gaussian probe gates fixed at a reference epoch, rather than the trained network's actual ReLU gates. The poisoned-minus-clean early-stage loss-decrease gap $\Delta D^{\bar{t}} \triangleq D_1^{\bar{t}} - D_0^{\bar{t}}$ is measured using multi-class cross-entropy between epochs $\bar{t}$ and $\bar{t} + 1$. Full experimental settings and diagnostic details are provided in Appendices C.1 and C.3.

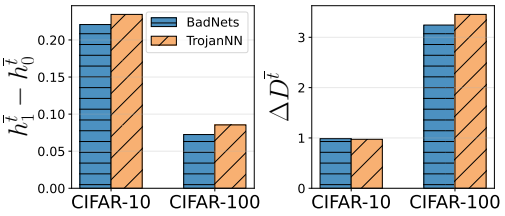


Figure 2: Empirical verification of early-stage optimization bias with poisoning ratio $\alpha = 5\%$. The two panels show the frozen-gate signal gap $h_1^{\bar{t}} - h_0^{\bar{t}}$ and the early-stage loss decrease gap $\Delta D^{\bar{t}}$ at checkpoint $\bar{t}$.

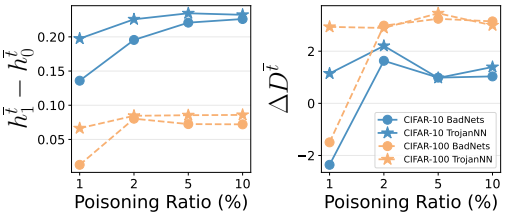


Figure 3: Effect of poisoning ratio on early-stage optimization bias. We vary $\alpha \in \{1\%, 2\%, 5\%, 10\%\}$ and report the frozen-gate signal gap $h_1^{\bar{t}} - h_0^{\bar{t}}$ and the early-stage loss decrease gap $\Delta D^{\bar{t}}$.

Fixed poisoning ratio. Figure 2 reports the results under the default poisoning ratio $\alpha = 5\%$. Across all dataset–attack pairs, the empirical frozen-gate signal gap $h_1^{\bar{t}} - h_0^{\bar{t}}$ is positive, indicating stronger within-group signed alignment under the fixed probe gates for poisoned samples. The early-stage loss-decrease gap $\Delta D^{\bar{t}}$ is likewise positive, indicating faster fitting of poisoned samples over the evaluated interval. Together, these observations provide qualitative empirical evidence consistent with the signal–loss relationship characterized by Theorem 1 and Corollary 1, without directly testing their finite-width error bounds.

Effect of poisoning ratio. To examine the robustness of the early-stage bias, we vary the poisoning ratio $\alpha \in \{1\%, 2\%, 5\%, 10\%\}$ and track both the frozen-gate signal gap $h_1^{\bar{t}} - h_0^{\bar{t}}$ and the loss decrease gap $\Delta D^{\bar{t}}$ across dataset–attack settings, as shown in Figure 3. As α increases, the frozen-gate signal gap generally becomes larger, while the loss decrease gap becomes more consistently positive. The loss decrease gap is not expected to be strictly monotone, as it is measured under the multi-class loss and therefore also reflects class competition and optimization landscape effects. These observations qualitatively complement the poisoning-ratio-weighted perspective of Corollary 2. That corollary considers the effect of α for fixed group-wise signals, whereas varying α in these experiments also changes the learned representations and the empirical signal gaps.

Diagnostic decomposition of each group. To further diagnose the origin of the observed frozen-gate signal gap, we examine two diagnostic factors associated with the empirical frozen-gate signal

$$h_z^{\bar{t}} = \mathbb{E}\big[\widehat{y}_1 \widehat{y}_2 \, \mathbf{x}_1^\top \mathbf{x}_2 G_0(\mathbf{x}_1, \mathbf{x}_2) \,\big|\, z\big].$$

Specifically, $\|\mathbf{m}_z^{\bar{t}}\|_2^2 = \mathbb{E}\big[\widehat{y}_1 \widehat{y}_2 \, \mathbf{x}_1^\top \mathbf{x}_2 \,\big|\, z\big]$ captures the magnitude of signed-feature alignment within group z at checkpoint $\bar{t}$, while $g_z = \mathbb{E}\big[G_0(\mathbf{x}_1, \mathbf{x}_2) \,\big|\, z\big]$ measures the consistency of joint ReLU activations. Figure 4 reports these two factors for BadNets on CIFAR-10 and CIFAR-100 across poisoning ratios. The poisoned group ($z = 1$) consistently exhibits larger values for both quantities, alongside a larger measured frozen-gate signal. These factors are informative diagnostics of the observed asymmetry, but their separate averages neither provide an exact factorization of $h_z^{\bar{t}}$ nor establish the ordering of $h_1^{\bar{t}}$ and $h_0^{\bar{t}}$ by themselves; see Appendix B. The corresponding TrojanNN plots are provided in Appendix C.3.

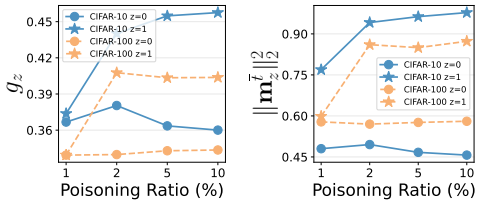


Figure 4: Diagnostic factors underlying the frozen-gate signal under BadNets. The two panels show the group-wise joint gating factor g_z and the squared norm of the empirical signed-feature mean $\|\mathbf{m}_z^{\bar{t}}\|_2^2$ across poisoning ratios. The poisoned group consistently exhibits larger values, indicating both higher joint gating consistency and stronger signed-feature alignment.

5 From Representation-Level Bias to Defense Implications

The early-stage analysis presented in Section 3 suggests that poisoned samples can induce a stronger trigger-aligned optimization signal than clean samples. In this section, we translate this optimization-level asymmetry into representation-level consequences and examine its implications for post-training defenses. While Theorem 1 characterizes early-stage group-wise loss decrease, Propositions 1 and 2

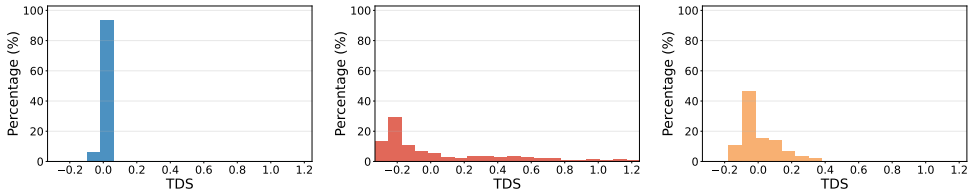


Figure 5: TDS distribution on CIFAR-10 under BadNets. Compared with the clean model, the backdoored model exhibits a heavier right tail, indicating that trigger dominance is concentrated in a small subset of neurons. Clean fine-tuning attenuates this right tail but may leave residual trigger-related structure.

concern representation-level consequences within this regime. The subsequent behavior under vanilla clean fine-tuning is examined empirically rather than established by these theoretical results.

Trigger-dominant representation. The theory predicts that trigger-aligned directions can be amplified early because they provide coherent label-aligned signals and stable gate overlap. In deep networks, this mechanism should appear as a small subset of neurons whose responses are disproportionately dominated by the trigger. The following proposition formalizes how the group-wise frozen-gate asymmetry can be inherited by a nontrivial subset of hidden neurons.

Proposition 1 (Trigger-biased reinforcement of neurons)**.** *Work under the conditions of Theorem 1 and condition on the early-stage event $\mathcal{E}_B$. For each neuron $j \in [k]$, define the trigger-dominance gap $\Delta_j^t \triangleq \mu_t \langle \mathbf{w}_j^t, \mathbf{u}_t \rangle - \mu_s \langle \mathbf{w}_j^t, \mathbf{u}_s \rangle$. Then there exists a nontrivial subset of neurons $\mathcal{J} \subseteq [k]$ such that, for each $j \in \mathcal{J}$, the one-step increment $\Delta_j^{t+1} - \Delta_j^t$ has positive conditional expectation in the early stage, up to the same controlled early-stage error as in Theorem 1. In particular, whenever the corresponding neuron-wise frozen-gate contribution dominates this error term, the neurons in $\mathcal{J}$ experience a systematic reinforcement bias toward the trigger direction in expectation.*

Proposition 1 shows that the optimization bias favoring the poisoned group is not only reflected in faster loss decrease. It can also be imprinted into the hidden representation through neuron-level reinforcement biased toward the trigger direction. We next explain how neurons exhibiting trigger-biased reinforcement give rise to selective activation on poisoned inputs.

Proposition 2 (Thresholded selective activation of trigger-biased neurons)**.** *Work under the conditions of Proposition 1 and fix a threshold $\tau \geq 0$. Then for any neuron $j \in \mathcal{J}$ and any iteration $t \leq T$ in the early-stage regime, poisoned inputs induce an additional trigger-aligned shift in the pre-activation relative to clean inputs. In particular, whenever the resulting trigger-dominant shift is sufficiently positive compared with the corresponding thresholded fluctuation level, it holds that*

$$\Pr\big((\mathbf{w}_j^t)^\top \mathbf{x} > \tau \mid z = 1, \mathcal{E}_B\big) > \Pr\big((\mathbf{w}_j^t)^\top \mathbf{x} > \tau \mid z = 0, \mathcal{E}_B\big). \tag{9}$$

Together, Propositions 1 and 2 provide a representation-level interpretation of the early-stage optimization bias in Section 3: the same optimization bias that favors the poisoned group can first induce trigger-biased neurons and then make these neurons more likely to cross a moderate activation threshold on poisoned inputs.

Implications for vanilla clean fine-tuning. Post-training defenses such as vanilla clean fine-tuning update the model using only clean samples [15, 14, 35]. Our early-stage analysis does not characterize the subsequent clean fine-tuning trajectory or establish whether trigger-related representations are fully removed. Since clean-only fine-tuning uses no triggered examples, its objective does not explicitly penalize the trigger–target association. Nevertheless, updates to shared parameters may indirectly weaken or even remove trigger-related features. Thus, whether such representations persist after fine-tuning is an empirical question rather than a consequence of our early-stage theory. Even when Top-1 ASR is substantially reduced, the target label may remain highly ranked on triggered inputs, as reflected by non-negligible Top-K ASR [36, 21]. These observations motivate examining representation-level changes alongside output-level attack success, rather than treating reduced Top-1 ASR as evidence of complete representation erasure.

Empirical evidence. To quantify these effects in practical networks, we use the Trigger Dominance Score (TDS), which measures how strongly a neuron responds to triggered inputs relative to clean semantic variation. TDS provides an empirical diagnostic of trigger-dominant activation associated

Table 1: Output-level performance on CIFAR-10 for clean training, backdoor training under BadNets, and vanilla clean fine-tuning.

Model	ACC (%)	ASR (%)
Clean	85.40	–
Backdoored	85.14	100.00
Fine-tuned	79.00	18.52

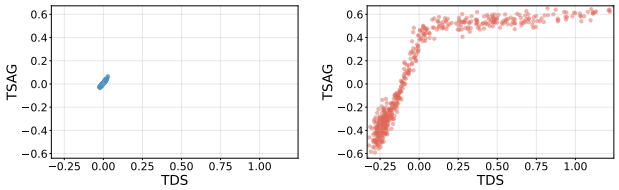


Figure 6: TDS–TSAG relationship on CIFAR-10 under BadNets. Left: clean model; right: backdoored model. High-TDS neurons concentrate in the high-TSAG region, indicating selective activation on poisoned inputs.

with the reinforcement mechanism in Proposition 1, rather than directly measuring its parameter-level condition. We also use the Thresholded Selective Activation Gap (TSAG) to quantify the selective-activation behavior considered in Proposition 2. Detailed definitions of TDS and TSAG, together with the TrojanNN results omitted from the main text due to space constraints, are provided in Appendix C.4.

Before examining the representation-level diagnostics in detail, we first summarize the output-level behavior of the models used in the following analysis. Table 1 reports the main task accuracy (ACC) and ASR on CIFAR-10 for the clean model, the backdoored model under BadNets, and the corresponding model after vanilla clean fine-tuning.

Figure 5 reports the TDS distributions at the final checkpoint on CIFAR-10 under BadNets. The clean model remains concentrated near zero, whereas the backdoored model exhibits a substantially heavier right tail, indicating that backdoor training gives rise to a nontrivial subset of trigger-dominant neurons. After vanilla clean fine-tuning, this tail is clearly compressed, showing that fine-tuning attenuates the prominence of these neurons. However, the fine-tuned distribution does not fully collapse to the clean distribution, indicating that the trigger-related representational footprint is weakened but not fully removed. This representation-level pattern is consistent with the output-level results in Table 1: vanilla clean fine-tuning reduces Top-1 ASR from 100% to 18.52%. However, as shown in Figure 1(c), the Top-2 ASR remains as high as 76.29% after vanilla clean fine-tuning, indicating that the target label is still highly ranked under triggered inputs. Together, these output-level and representation-level diagnostics indicate that, under the evaluated fine-tuning protocol, backdoor behavior is attenuated while measurable trigger-related signatures remain.

Figure 6 examines the relationship between TDS and TSAG at the final checkpoint on CIFAR-10 under BadNets. The clean model remains concentrated near the origin, whereas the backdoored model exhibits a pronounced right-upper spread, indicating that neurons with higher TDS also tend to have larger TSAG. Together, these results show that the coupling between trigger dominance and selective activation is substantially strengthened after backdoor training, providing empirical support for the representation-level link from Proposition 1 to Proposition 2.

Next, we examine whether the post-training defense behavior discussed above is reflected in practice. Specifically, we compare clean, backdoored, and clean fine-tuned models to examine how trigger-related activation signatures change after vanilla clean fine-tuning. Figure 7 provides a condensed group-wise view under BadNets. Neurons are ranked by their TDS in the backdoored model, and we report the average TDS and TSAG over the top 1%, top 5%, random 5%, and bottom 5% subsets.

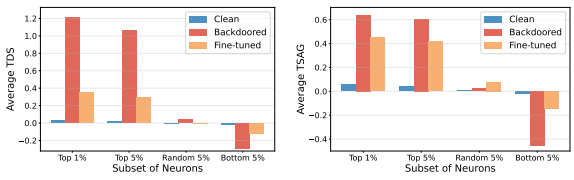


Figure 7: Group-wise TDS and TSAG summaries on CIFAR-10 under BadNets. Neurons are grouped by their TDS ranking in the backdoored model, and clean, backdoored, and clean fine-tuned models are compared at the final checkpoint.

The backdoored model exhibits substantially larger TDS and TSAG on the top-ranked subsets than the clean baseline, showing that backdoor training concentrates both trigger dominance and selective activation into a small subset of neurons. After vanilla clean fine-tuning, both quantities are markedly reduced on these top-ranked subsets,

indicating that fine-tuning suppresses the most trigger-related part of the representation. However, the fine-tuned model still remains above the clean baseline, suggesting that the trigger-dominant and selectively activated structure is weakened but not fully removed. By contrast, the differences across models are much smaller on the random and bottom-ranked subsets. This group-wise comparison is consistent with the full-distribution evidence in Figure 5 and further indicates that, under the evaluated protocol, clean fine-tuning attenuates but does not eliminate the measured trigger-related activation signatures. The corresponding TrojanNN group-wise TDS and TSAG results are provided in Appendix C.4.

6 Discussion and Limitations

Scope of the theoretical model. Our theoretical analysis uses a stylized binary directional model to isolate semantic–trigger competition rather than reproduce the complexity of natural data. Its clean-label signal lies along a single semantic direction, making the underlying classification structure deliberately simple. However, the learner is a finite-width ReLU network with trainable hidden-layer weights and input-dependent gates that evolve during training; its optimization dynamics are therefore not those of a fixed linear classifier. The orthogonality of $\mathbf{u}_s$ and $\mathbf{u}_t$ serves to separate the two signals analytically. The mechanism of interest is the conditional early-stage advantage of a coherent, target-aligned trigger signal, not exact orthogonality. More generally, within-group gradient coherence suggests an analogous local perspective for nonlinear models, but this observation does not extend our quantitative bounds beyond the stated directional-model and ReLU-network assumptions.

Relation to multi-class backdoor learning. Although the theory is developed in a binary prototype, our empirical analysis is conducted on original multi-class networks. The target-versus-rest view is used only as a diagnostic interpretation of a trained multi-class model, not as a binary training procedure. This distinction is important: the observed early poisoned loss advantage, frozen-gate signal gap, trigger-dominant neurons, and post-training residual structure all arise in the original multi-class models. Thus, the binary prototype should be viewed as a mechanism-isolating theory, while the multi-class experiments show that the same signatures persist in standard deep-network training settings. Extending the theory to a genuine multi-class setting would require modeling coupled logits, class-dependent semantic prototypes, and their interaction with ReLU gating, which we leave as future work.

Limitations and implications. Our results characterize early-stage training dynamics, not the full training trajectory, the final converged model, or the dynamics of subsequent clean fine-tuning. This focus isolates the regime in which trigger-aligned directions can acquire an optimization advantage and trigger-related representations begin to emerge. We do not propose a new attack or defense. Rather, our theoretical results characterize this mechanism under the stated assumptions, while experiments on practical networks show corresponding signatures and, under the evaluated vanilla clean fine-tuning protocol, residual trigger-related activations after attenuation. These post-training observations do not establish that clean fine-tuning must leave residual representations in general. Future directions include multi-class theory, longer-horizon dynamics, and defenses targeting trigger-related representations.

Acknowledgments and Disclosure of Funding

Qiang Zhou was supported by the National Natural Science Foundation of China under Grant No. 62106045, the Jiangsu Provincial Double-Innovation Doctor Program under Grant No. JSS-CBS20220185, and the Fundamental Research Funds for the Central Universities under Grant No. 2242026K30047. This work was supported by the Big Data Computing Center of Southeast University. Sinno J. Pan is supported by the JC STEM Lab of Machine Learning and Symbolic Reasoning funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, abs/1712.05526, 2017. URL `http://arxiv.org/abs/1712.05526`.

[3] Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[4] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 45(2):1563–1580, 2023.

[5] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017. URL `http://arxiv.org/abs/1708.06733`.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael Curtis Mozer. On the foundations of shortcut learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

[8] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[9] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 125–136, 2019.

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8580–8589, 2018.

[11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

[12] Boqi Li and Weiwei Liu. A theoretical analysis of backdoor poisoning attacks in convolutional neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2024.

[13] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 14900–14912, 2021.

[14] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[15] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of International Symposium on Research in Attacks, Intrusions and Defenses*, pages 273–294, 2018.

[16] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *Proceedings of Annual Network and Distributed System Security Symposium (NDSS)*, 2018.

[17] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 12978–12991, 2021.

[18] Naren Manoj and Avrim Blum. Excess capacity and backdoor poisoning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[19] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *CoRR*, abs/1804.06561, 2018. URL `http://arxiv.org/abs/1804.06561`.

[20] Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[21] Rui Min, Zeyu Qin, Nevin L. Zhang, Li Shen, and Minhao Cheng. Uncovering, explaining, and mitigating the superficial safety of backdoor defense. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[22] Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, and Boaz Barak. SGD on neural networks learns functions of increasing complexity. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3491–3501, 2019.

[23] Guillermo Valle Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[24] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? A study from the parameter-space perspective. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[25] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[26] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *Proceedings of IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019.

[27] Ganghua Wang, Xun Xian, Ashish Kundu, Jayanth Srinivasa, Xuan Bi, Mingyi Hong, and Jie Ding. Demystifying poisoning backdoor attacks from a statistical perspective. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

[28] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[29] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 16913–16925, 2021.

[30] Xun Xian, Ganghua Wang, Jayanth Srinivasa, Ashish Kundu, Xuan Bi, Mingyi Hong, and Jie Ding. Understanding backdoor attacks through the adaptability hypothesis. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 37952–37976, 2023.

[31] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.

[32] Qi Zhao and Christian Wressnegger. Two sides of the same coin: Learning the backdoor to remove the backdoor. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 22804–22812, 2025.

[33] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 175–191, 2022.

[34] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Pre-activation distributions expose backdoor neurons. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[35] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4443–4454, 2023.

[36] Mingli Zhu, Siyuan Liang, and Baoyuan Wu. Breaking the false sense of security in backdoor defense through re-activation attack. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

List of Appendices

A Proofs of Main Theoretical Results **15**
A.1 Proof of Lemma 1 . . . 15
A.2 Proof of Lemma 2 . . . 16
A.3 Proof of Lemma 3 . . . 16
A.4 Proof of Lemma 4 . . . 17
A.5 Proof of Theorem 1 . . . 18
A.6 Proof of Corollary 1 . . . 19
A.7 Proof of Corollary 2 . . . 19
A.8 Proof of Proposition 1 . . . 20
A.8.1 Proof of Lemma A1 . . . 21
A.9 Proof of Proposition 2 . . . 22

B Geometry and Structure of the Frozen-Gate Signal **22**
B.1 Signed-Feature Agreement . . . 23
B.2 Gaussian Halfspace Geometry and Joint Gating . . . 23
B.3 Gated-Signal Representation and Group-Wise Comparison . . . 24

C Experimental Protocols and Additional Results **25**
C.1 Common Experimental Setup . . . 26
C.2 Additional Motivating Observations . . . 26
C.3 Empirical Verification of Early-Stage Optimization Bias . . . 29
C.3.1 Empirical Quantities and Checkpoint Protocol . . . 29
C.3.2 Verification Protocol under Fixed and Varying Poisoning Ratios . . . 30
C.3.3 Group-Wise Diagnostic Decomposition . . . 30
C.4 Representation-Level Diagnostics and Post-Training Defense . . . 31
C.4.1 Output-Level Performance of Evaluated Models . . . 31
C.4.2 Diagnostic Metrics and Evaluation Protocol . . . 32
C.4.3 Trigger-Dominance and Selective Activation under TrojanNN . . . 33
C.4.4 Early Checkpoint Evolution of Representation-Level Bias . . . 33
C.4.5 Group-Wise Representation Diagnostics under TrojanNN . . . 34
C.4.6 Output-Level and Representation-Level Evaluation after Vanilla Clean Fine-Tuning . . . 35

D Detailed Related Work **36**

A Proofs of Main Theoretical Results

A.1 Proof of Lemma 1

Proof. We prove the lemma in two parts: first, we establish a high-probability bound on $\max_{i\in[N]} \|\mathbf{x}_i\|$; second, we show that the network output remains below the hinge margin throughout the early-stage horizon.

Fix $\delta' \in (0,1)$. By (2), each sample satisfies

$$\mathbf{x}_i = y_i\mu_s\mathbf{u}_s + z_i\widetilde{y}\mu_t\mathbf{u}_t + \boldsymbol{\xi}_i, \qquad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Using $\|\mathbf{u}_s\| = \|\mathbf{u}_t\| = 1$, $|y_i| = |\widetilde{y}| = 1$, and $z_i \in \{0,1\}$, the triangle inequality gives

$$\|\mathbf{x}_i\| \le \mu_s + \mu_t + \|\boldsymbol{\xi}_i\|. \tag{A1}$$

Write $\boldsymbol{\xi}_i = \sigma\mathbf{g}_i$, where $\mathbf{g}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the function $f(\mathbf{g}) = \|\mathbf{g}\|$ is 1-Lipschitz w.r.t. the Euclidean norm. By Gaussian concentration for Lipschitz functions [1, Theorem 5.6], for any $t \ge 0$,

$$\Pr(\|\mathbf{g}\| \ge \mathbb{E}[\|\mathbf{g}\|] + t) \le \exp(-t^2/2).$$

Applying Jensen's inequality and $\mathbb{E}\big[\|\boldsymbol{g}\|^2\big] = d$, we have $\mathbb{E}[\|\boldsymbol{g}\|] \le \sqrt{d}$. Therefore, setting $t = \sqrt{2\log(2N/\delta')}$ gives

$$\Pr\Big(\|\mathbf{g}_i\| \ge \sqrt{d} + \sqrt{2\log(2N/\delta')}\Big) \le \frac{\delta'}{2N}. \tag{A2}$$

A union bound over $i \in [N]$ implies that, with probability at least $1 - \delta'/2$,

$$\max_{i\in[N]} \|\mathbf{g}_i\| \le \sqrt{d} + \sqrt{2\log(2N/\delta')}.$$

Multiplying both sides by σ and combining it with (A1), we obtain

$$\max_{i\in[N]} \|\mathbf{x}_i\| \le (\mu_s + \mu_t) + \sigma\Big(\sqrt{d} + \sqrt{2\log(2N/\delta')}\Big) \tag{A3}$$

with probability at least $1 - \delta'/2$. Taking $\delta' = 2\delta$, this is exactly the bounded-input event $\mathcal{E}_B$ with

$$B \triangleq (\mu_s + \mu_t) + \sigma\Big(\sqrt{d} + \sqrt{2\log(4N/\delta)}\Big),$$

and hence $\Pr(\mathcal{E}_B) \ge 1 - \delta$.

We now condition on $\mathcal{E}_B$. Recall that

$$f\big(\mathbf{x}; \{\mathbf{w}_j\}_{j=1}^k\big) = \frac{1}{\sqrt{k}} \sum_{j=1}^k v_j \mathrm{ReLU}\big(\mathbf{x}^\top \mathbf{w}_j\big), \qquad v_j \in \{\pm 1\}.$$

When the hinge loss is active, a valid subgradient is

$$\nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j\}_{j=1}^k\big) = -\frac{\widehat{y}}{\sqrt{k}} v_j \mathbf{x} \mathbb{I}\big\{\mathbf{x}^\top \mathbf{w}_j > 0\big\},$$

and when the loss is inactive we take the subgradient to be $\mathbf{0}$. Thus,

$$\big\|\nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j\}_{j=1}^k\big)\big\| \le \frac{1}{\sqrt{k}} \|\mathbf{x}\|.$$

Conditional on $\mathcal{E}_B$, every SGD step satisfies

$$\big\|\nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}_i, \widehat{y}_i; \{\mathbf{w}_j\}_{j=1}^k\big)\big\| \le \frac{B}{\sqrt{k}}. \tag{A4}$$

The update

$$\mathbf{w}_j^{t+1} = \mathbf{w}_j^t - \eta \nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}_{i_t}, \widehat{y}_{i_t}; \{\mathbf{w}_j^t\}_{j=1}^k\big)$$

then gives

$$\|\mathbf{w}_j^{t+1}\| \le \|\mathbf{w}_j^t\| + \eta \frac{B}{\sqrt{k}}.$$

Iterating and using the initialization event in Assumption 2, we obtain

$$\|\mathbf{w}_j^t\| \le \|\mathbf{w}_j^0\| + t\eta\frac{B}{\sqrt{k}} \le R_0 + t\eta\frac{B}{\sqrt{k}}. \tag{A5}$$

For any training sample $\mathbf{x}_i$ and any $t \le T$, using $\mathrm{ReLU}(u) \le |u|$ and Cauchy–Schwarz gives

$$\left|f\big(\mathbf{x}_i; \{\mathbf{w}_j^t\}_{j=1}^k\big)\right| \le \frac{1}{\sqrt{k}}\sum_{j=1}^k |\mathbf{x}_i^\top \mathbf{w}_j^t| \le \frac{B}{\sqrt{k}}\sum_{j=1}^k \|\mathbf{w}_j^t\|.$$

Applying (A5) to all neurons yields

$$\left|f(\mathbf{x}_i; \{\mathbf{w}_j^t\}_{j=1}^k)\right| \le \sqrt{k}BR_0 + t\eta B^2.$$

By the lemma assumption, $\sqrt{k}BR_0 \le c_0$. Moreover, for $T = \lfloor c_1/(\eta B^2)\rfloor$, all $t \le T$ satisfy $t\eta B^2 \le c_1$. Choosing universal constants $c_0, c_1 > 0$ such that $c_0 + c_1 < 1$, we obtain

$$\left|f\big(\mathbf{x}_i; \{\mathbf{w}_j^t\}_{j=1}^k\big)\right| < 1 \qquad \text{for all } i \in [N],\ t \le T.$$

Thus $\widehat{y}_i f(\mathbf{x}_i; \{\mathbf{w}_j^t\}_{j=1}^k) < 1$ for all training samples and all $t \le T$. Hence the hinge loss remains active throughout the first T iterations. □

A.2 Proof of Lemma 2

Proof. We analyze the two cases $z = 0$ and $z = 1$.

Case $z = 0$. For a clean sample, $\widehat{y} = y$ and

$$\mathbf{x} = y\mu_s\mathbf{u}_s + \boldsymbol{\xi}, \qquad \mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}.$$

Therefore,

$$\mathbb{E}[\widehat{y}\mathbf{x} \mid z = 0] = \mathbb{E}[y(y\mu_s\mathbf{u}_s + \boldsymbol{\xi})] = \mu_s\mathbf{u}_s.$$

This gives $\mathbb{E}[\widehat{y}\langle\mathbf{x}, \mathbf{u}_s\rangle \mid z = 0] = \mu_s$.

Case $z = 1$. For a poisoned sample, $\widehat{y} = \widetilde{y}$ and

$$\mathbf{x} = y\mu_s\mathbf{u}_s + \widetilde{y}\mu_t\mathbf{u}_t + \boldsymbol{\xi}.$$

Using $\mathbb{E}[y] = 0$, $\widetilde{y}^2 = 1$, and $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$, we obtain

$$\mathbb{E}[\widehat{y}\mathbf{x} \mid z = 1] = \mathbb{E}[\widetilde{y}(y\mu_s\mathbf{u}_s + \widetilde{y}\mu_t\mathbf{u}_t + \boldsymbol{\xi})] = \mu_t\mathbf{u}_t.$$

This gives $\mathbb{E}[\widehat{y}\langle\mathbf{x}, \mathbf{u}_t\rangle \mid z = 1] = \mu_t$.

Finally, under Assumption 1, $\langle\mathbf{u}_s, \mathbf{u}_t\rangle = 0$. Taking inner products of the above identities with the opposite direction gives

$$\mathbb{E}[\widehat{y}\langle\mathbf{x}, \mathbf{u}_t\rangle \mid z = 0] = 0, \qquad \mathbb{E}[\widehat{y}\langle\mathbf{x}, \mathbf{u}_s\rangle \mid z = 1] = 0.$$

This proves the lemma. □

A.3 Proof of Lemma 3

Proof. Under Lemma 1, the hinge loss is active at $t = 0$. Therefore, for a sampled example $(\mathbf{x}, \widehat{y})$,

$$\nabla_{\mathbf{w}_j}\ell\big(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j^0\}_{j=1}^k\big) = -\frac{\widehat{y}}{\sqrt{k}}v_j\mathbf{x}\mathbb{I}\big\{\mathbf{x}^\top\mathbf{w}_j^0 > 0\big\},$$

and hence

$$\Delta\mathbf{w}_j^0(z) = \mathbf{w}_j^1 - \mathbf{w}_j^0 = \frac{\eta}{\sqrt{k}}\widehat{y}v_j\mathbf{x}\mathbb{I}\big\{\mathbf{x}^\top\mathbf{w}_j^0 > 0\big\}. \tag{A6}$$

Conditioned on any fixed nonzero $\mathbf{x}$, the scalar $\mathbf{x}^\top\mathbf{w}_j^0$ is centered Gaussian by Assumption 2. Therefore,

$$\mathbb{E}_{\mathbf{w}_j^0}\big[\mathbb{I}\big\{\mathbf{x}^\top\mathbf{w}_j^0 > 0\big\} \mid \mathbf{x}\big] = \frac{1}{2}.$$

Taking expectation over $\mathbf{w}_j^0$ in (A6) yields

$$\mathbb{E}_{\mathbf{w}_j^0}\big[\Delta \mathbf{w}_j^0(z) \mid \mathbf{x}, \widehat{y}\big] = \frac{\eta}{2\sqrt{k}} \widehat{y} v_j \mathbf{x}.$$

Now taking conditional expectation over the data given z, we obtain

$$\mathbb{E}[\Delta \mathbf{w}_j^0(z) \mid z] = \frac{\eta}{2\sqrt{k}} v_j \mathbb{E}[\widehat{y}\mathbf{x} \mid z] = \frac{\eta}{2\sqrt{k}} v_j \mathbf{m}_z.$$

Averaging over j gives

$$\mathbb{E}[\Delta \overline{\mathbf{w}}^0(z)] = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}[\Delta \mathbf{w}_j^0(z) \mid z] = \frac{\eta}{2\sqrt{k}} \left(\frac{1}{k}\sum_{j=1}^{k} v_j\right) \mathbf{m}_z = \frac{\eta}{2\sqrt{k}} \bar{v} \mathbf{m}_z.$$

This proves the claimed drift identity. The final specialization follows from Lemma 2. □

A.4 Proof of Lemma 4

Proof. Fix any $\mathbf{x}$ independent of $\mathbf{w}_j^0$, and define

$$\mathbf{q}_j^t \triangleq \mathbf{w}_j^t - \mathbf{w}_j^0.$$

If $A_j^t(\mathbf{x}) \neq A_j^0(\mathbf{x})$, then $\mathbf{x}^\top \mathbf{w}_j^0$ and $\mathbf{x}^\top \mathbf{w}_j^t$ have opposite signs. Hence

$$\big|\mathbf{x}^\top \mathbf{w}_j^0\big| \leq \big|\mathbf{x}^\top \mathbf{q}_j^t\big| \leq \|\mathbf{q}_j^t\| \, \|\mathbf{x}\|. \tag{A7}$$

The first bound in the lemma follows by conditioning on a deterministic upper bound for $\|\mathbf{q}_j^t\|$. Indeed, under Assumption 2, conditioned on $\mathbf{x}$, the scalar $\mathbf{x}^\top \mathbf{w}_j^0$ is distributed as $\sigma_w \|\mathbf{x}\| G$, where $G \sim \mathcal{N}(0,1)$. Therefore, for any $a \geq 0$,

$$\Pr\Big(\big|\mathbf{x}^\top \mathbf{w}_j^0\big| \leq a\|\mathbf{x}\| \,\Big|\, \mathbf{x}\Big) = \Pr\left(|G| \leq \frac{a}{\sigma_w}\right) \leq C\frac{a}{\sigma_w}, \tag{A8}$$

for a universal constant $C > 0$. If the drift is independent of the initialization margin, the same argument gives

$$\Pr\big(A_j^t(\mathbf{x}) \neq A_j^0(\mathbf{x}) \mid \mathcal{E}_B\big) \leq \frac{C}{\sigma_w} \mathbb{E}\big[\|\mathbf{w}_j^t - \mathbf{w}_j^0\| \mid \mathcal{E}_B\big].$$

Alternatively, this bound holds directly whenever the drift is replaced by a deterministic upper bound.

Under $\mathcal{E}_B$, Lemma 1 gives

$$\big\|\nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}_{i_s}, \widehat{y}_{i_s}; \{\mathbf{w}_j^s\}_{j=1}^k\big)\big\| \leq \frac{B}{\sqrt{k}}$$

for every step s. Therefore, for any $t \leq T$,

$$\|\mathbf{q}_j^t\| \leq \sum_{s=0}^{t-1} \eta \big\|\nabla_{\mathbf{w}_j} \ell\big(\mathbf{x}_{i_s}, \widehat{y}_{i_s}; \{\mathbf{w}_j^s\}_{j=1}^k\big)\big\| \leq \frac{t\eta B}{\sqrt{k}}. \tag{A9}$$

Combining (A7), (A8), and (A9) gives

$$\Pr\big(A_j^t(\mathbf{x}) \neq A_j^0(\mathbf{x}) \mid \mathcal{E}_B\big) \leq C\frac{t\eta B}{\sigma_w \sqrt{k}}.$$

For any fixed pair $(\mathbf{x}_1, \mathbf{x}_2)$, we have

$$\big|A_j^t(\mathbf{x}_1)A_j^t(\mathbf{x}_2) - A_j^0(\mathbf{x}_1)A_j^0(\mathbf{x}_2)\big| \leq \mathbb{I}\{A_j^t(\mathbf{x}_1) \neq A_j^0(\mathbf{x}_1)\} + \mathbb{I}\{A_j^t(\mathbf{x}_2) \neq A_j^0(\mathbf{x}_2)\}. \tag{A10}$$

Taking expectation conditioned on $\mathcal{E}_B$ and applying the gate-flip bound to both $\mathbf{x}_1$ and $\mathbf{x}_2$ yields

$$\big|\mathbb{E}[A_j^t(\mathbf{x}_1)A_j^t(\mathbf{x}_2) \mid \mathcal{E}_B] - \mathbb{E}[A_j^0(\mathbf{x}_1)A_j^0(\mathbf{x}_2) \mid \mathcal{E}_B]\big| \leq 2C\frac{t\eta B}{\sigma_w \sqrt{k}}.$$

This proves the lemma. □

A.5 Proof of Theorem 1

Proof. Condition throughout on $\mathcal{E}_B$. By Lemma 1, for all $t \le T$, the hinge loss is active for all training samples. Hence

$$\ell\big(\mathbf{x}, \widehat{y}; \{\mathbf{w}_j^t\}_{j=1}^k\big) = 1 - \widehat{y} f\big(\mathbf{x}; \{\mathbf{w}_j^t\}_{j=1}^k\big).$$

Therefore, for each $z \in \{0, 1\}$,

$$L_z^t - L_z^{t+1} = \mathbb{E}\big[\widehat{y}\big(f\big(\mathbf{x}; \{\mathbf{w}_j^{t+1}\}_{j=1}^k\big) - f\big(\mathbf{x}; \{\mathbf{w}_j^t\}_{j=1}^k\big)\big) \mid z\big], \tag{A11}$$

where $(\mathbf{x}, \widehat{y})$ is an independent conditional draw given z.

At iteration t, SGD samples $(\mathbf{x}_t, \widehat{y}_t)$. Under the active-loss regime, the update is

$$\mathbf{w}_j^{t+1} = \mathbf{w}_j^t + \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j \mathbf{x}_t A_j^t(\mathbf{x}_t).$$

Fix a test input $\mathbf{x}$, and define

$$a_j \triangleq (\mathbf{w}_j^t)^\top \mathbf{x}, \qquad b_j \triangleq (\mathbf{w}_j^{t+1} - \mathbf{w}_j^t)^\top \mathbf{x} = \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j \mathbf{x}_t^\top \mathbf{x} A_j^t(\mathbf{x}_t).$$

Then $A_j^t(\mathbf{x}) = \mathbb{I}\{a_j > 0\}$. Define the ReLU residual

$$r_j \triangleq \mathrm{ReLU}(a_j + b_j) - \mathrm{ReLU}(a_j) - A_j^t(\mathbf{x}) b_j. \tag{A12}$$

If the gate does not change between a_j and $a_j + b_j$, then $r_j = 0$. If it changes, a direct case analysis gives

$$|r_j| \le |b_j| \, \mathbb{I}\big\{A_j^{t+1}(\mathbf{x}) \ne A_j^t(\mathbf{x})\big\}.$$

Thus

$$\mathrm{ReLU}(a_j + b_j) - \mathrm{ReLU}(a_j) = A_j^t(\mathbf{x}) b_j + r_j, \qquad |r_j| \le |b_j| \, \mathbb{I}\big\{A_j^{t+1}(\mathbf{x}) \ne A_j^t(\mathbf{x})\big\}. \tag{A13}$$

Summing over neurons and using $v_j^2 = 1$, we obtain

$$f\big(\mathbf{x}; \{\mathbf{w}_j^{t+1}\}_{j=1}^k\big) - f\big(\mathbf{x}; \{\mathbf{w}_j^t\}_{j=1}^k\big) = \eta \widehat{y}_t \mathbf{x}_t^\top \mathbf{x} G_t(\mathbf{x}, \mathbf{x}_t) + R_t(\mathbf{x}, \mathbf{x}_t), \tag{A14}$$

where

$$R_t(\mathbf{x}, \mathbf{x}_t) \triangleq \frac{1}{\sqrt{k}} \sum_{j=1}^k v_j r_j.$$

Using (A13) and $|\mathbf{x}_t^\top \mathbf{x}| \le B^2$ on $\mathcal{E}_B$, we have

$$|R_t(\mathbf{x}, \mathbf{x}_t)| \le \eta B^2 \frac{1}{k} \sum_{j=1}^k A_j^t(\mathbf{x}_t) \mathbb{I}\big\{A_j^{t+1}(\mathbf{x}) \ne A_j^t(\mathbf{x})\big\}. \tag{A15}$$

Moreover,

$$\mathbb{I}\{A_j^{t+1}(\mathbf{x}) \ne A_j^t(\mathbf{x})\} \le \mathbb{I}\big\{A_j^{t+1}(\mathbf{x}) \ne A_j^0(\mathbf{x})\big\} + \mathbb{I}\big\{A_j^t(\mathbf{x}) \ne A_j^0(\mathbf{x})\big\}.$$

Applying Lemma 4 at times t and $t+1$ gives

$$\mathbb{E}\big[\mathbb{I}\big\{A_j^{t+1}(\mathbf{x}) \ne A_j^t(\mathbf{x})\big\} \mid \mathcal{E}_B\big] \le 2C \frac{(t+1)\eta B}{\sigma_w \sqrt{k}}.$$

Substituting into (A15) yields

$$\mathbb{E}[|R_t(\mathbf{x}, \mathbf{x}_t)| \mid \mathcal{E}_B] \le 2C \frac{(t+1)\eta^2 B^3}{\sigma_w \sqrt{k}}. \tag{A16}$$

Next, we replace G_t by G_0. By Lemma 4 and (A10),

$$\mathbb{E}[|G_t(\mathbf{x}, \mathbf{x}_t) - G_0(\mathbf{x}, \mathbf{x}_t)| \mid \mathcal{E}_B] \le 2C \frac{t\eta B}{\sigma_w \sqrt{k}}.$$

Therefore,

$$\mathbb{E}\Big[\big|\eta \widehat{y} \widehat{y}_t \mathbf{x}_t^\top \mathbf{x} (G_t(\mathbf{x}, \mathbf{x}_t) - G_0(\mathbf{x}, \mathbf{x}_t))\big| \,\Big|\, \mathcal{E}_B\Big] \le 2C \frac{t\eta^2 B^3}{\sigma_w \sqrt{k}}. \tag{A17}$$

Combining (A11), (A14), (A16), and (A17), and then conditioning on $z_t = z$, we get

$$\big|\mathbb{E}\big[L_z^t - L_z^{t+1} \mid z_t = z, \mathcal{E}_B\big] - \eta \mathbb{E}\big[\widehat{y} \widehat{y}_t \mathbf{x}_t^\top \mathbf{x} G_0(\mathbf{x}, \mathbf{x}_t) \mid z_t = z\big]\big| \le 4C \frac{(t+1)\eta^2 B^3}{\sigma_w \sqrt{k}}.$$

Since $(\mathbf{x}, \widehat{y})$ is an independent conditional draw from group z, and $(\mathbf{x}_t, \widehat{y}_t)$ is the SGD sample satisfying $z_t = z$, the leading term equals ηh_z. This proves (7). □

A.6 Proof of Corollary 1

Proof. Fix any $t \le T-1$ and $z \in \{0,1\}$. By Theorem 1 and the definition of D_z^t,

$$|D_z^t - \eta h_z| \le 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}. \tag{A18}$$

Applying (A18) to $z=1$ and $z=0$, we obtain

$$D_1^t \ge \eta h_1 - 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}, \qquad D_0^t \le \eta h_0 + 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

Subtracting gives

$$D_1^t - D_0^t \ge \eta(h_1 - h_0) - 8C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

Under the separation condition

$$h_1 - h_0 > 8C\frac{(t+1)\eta B^3}{\sigma_w\sqrt{k}},$$

the right-hand side is strictly positive. Hence $D_1^t > D_0^t$. □

A.7 Proof of Corollary 2

Proof. Fix any $t \le T-1$. By the law of total expectation over $z_t \in \{0,1\}$ and conditioning on $\mathcal{E}_B$,

$$(1-\alpha)D_0^t + \alpha D_1^t = (1-\alpha)\mathbb{E}\big[L_0^t - L_0^{t+1} \mid z_t = 0, \mathcal{E}_B\big] + \alpha\mathbb{E}\big[L_1^t - L_1^{t+1} \mid z_t = 1, \mathcal{E}_B\big].$$

By Theorem 1, for each $z \in \{0,1\}$ there exists an error e_z such that

$$D_z^t = \eta h_z + e_z, \qquad |e_z| \le 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

Therefore,

$$(1-\alpha)D_0^t + \alpha D_1^t = \eta((1-\alpha)h_0 + \alpha h_1) + (1-\alpha)e_0 + \alpha e_1.$$

Using the triangle inequality and $\alpha \in [0,1]$,

$$|(1-\alpha)e_0 + \alpha e_1| \le (1-\alpha)|e_0| + \alpha|e_1| \le 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

This proves

$$\big|(1-\alpha)D_0^t + \alpha D_1^t - \eta((1-\alpha)h_0 + \alpha h_1)\big| \le 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

It remains to prove the weighted dominance statement. Again by Theorem 1,

$$D_1^t \ge \eta h_1 - 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}, \qquad D_0^t \le \eta h_0 + 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}}.$$

Multiplying the first inequality by α and the second by $1-\alpha$, and subtracting, gives

$$\alpha D_1^t - (1-\alpha)D_0^t \ge \eta(\alpha h_1 - (1-\alpha)h_0) - 4C\frac{(t+1)\eta^2 B^3}{\sigma_w\sqrt{k}},$$

where the single error term appears because $\alpha + (1-\alpha) = 1$. Under the gap condition

$$\alpha h_1 - (1-\alpha)h_0 > 4C\frac{(t+1)\eta B^3}{\sigma_w\sqrt{k}},$$

the right-hand side is strictly positive. Hence $\alpha D_1^t > (1-\alpha)D_0^t$, as claimed. □

A.8 Proof of Proposition 1

We first state an auxiliary lemma showing that, under a mild non-degeneracy condition, Gaussian initialization produces a nontrivial subset of neurons whose frozen gates are responsive to the trigger-aligned direction. Throughout this proof, without loss of generality, we absorb the target sign $\widetilde{y}$ into the trigger direction $\mathbf{u}_t$, so that poisoned inputs have a positive trigger-aligned shift. The following condition should be read as a neuron-wise non-degeneracy condition ensuring that the trigger-aligned offset is not cancelled by semantic fluctuation and noise on a positive-probability initialization event.

Lemma A1 (Existence of a trigger-responsive subset)**.** *Work under the conditions of Theorem 1 and condition on the early-stage event $\mathcal{E}_B$. For each neuron $j \in [k]$, define its frozen-gate trigger-response coefficient by*

$$\gamma_j \triangleq v_j \mathbb{E}\big[\widehat{y}(\mu_t \langle \mathbf{x}, \mathbf{u}_t \rangle - \mu_s \langle \mathbf{x}, \mathbf{u}_s \rangle) A_j^0(\mathbf{x}) \mid \mathcal{E}_B\big]. \tag{A19}$$

Assume that the trigger-aligned offset dominates the semantic fluctuation and noise on a positive-probability initialization event, so that there exist constants $c_{\mathrm{resp}}, c_{\mathrm{sub}} > 0$, independent of k, such that

$$\Pr(\gamma_j \geq c_{\mathrm{resp}} \mid \mathcal{E}_B) \geq c_{\mathrm{sub}}. \tag{A20}$$

Define

$$\mathcal{J} \triangleq \{j \in [k] : \gamma_j \geq c_{\mathrm{resp}}\}. \tag{A21}$$

Then, with probability at least $1 - \exp(-ck)$ for some universal constant $c > 0$,

$$|\mathcal{J}| \geq \frac{c_{\mathrm{sub}}}{2} k. \tag{A22}$$

In particular, $\mathcal{J}$ is nontrivial with high probability.

We now prove Proposition 1 using Lemma A1.

Proof of Proposition 1. Fix a neuron $j \in [k]$ and decompose its hidden weight along the semantic and trigger directions:

$$\mathbf{w}_j^t = a_{j,s}^t \mathbf{u}_s + a_{j,t}^t \mathbf{u}_t + \mathbf{w}_j^{t,\perp},$$

where $\mathbf{w}_j^{t,\perp}$ is orthogonal to both $\mathbf{u}_s$ and $\mathbf{u}_t$. By definition,

$$\Delta_j^t = \mu_t a_{j,t}^t - \mu_s a_{j,s}^t.$$

Under the active-loss regime in Lemma 1, the one-step SGD update is

$$\mathbf{w}_j^{t+1} = \mathbf{w}_j^t + \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j \mathbf{x}_t A_j^t(\mathbf{x}_t).$$

Taking inner products with $\mathbf{u}_s$ and $\mathbf{u}_t$, we obtain

$$\langle \mathbf{w}_j^{t+1} - \mathbf{w}_j^t, \mathbf{u}_s \rangle = \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j \langle \mathbf{x}_t, \mathbf{u}_s \rangle A_j^t(\mathbf{x}_t),$$

and

$$\langle \mathbf{w}_j^{t+1} - \mathbf{w}_j^t, \mathbf{u}_t \rangle = \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j \langle \mathbf{x}_t, \mathbf{u}_t \rangle A_j^t(\mathbf{x}_t).$$

Therefore,

$$\Delta_j^{t+1} - \Delta_j^t = \frac{\eta}{\sqrt{k}} \widehat{y}_t v_j (\mu_t \langle \mathbf{x}_t, \mathbf{u}_t \rangle - \mu_s \langle \mathbf{x}_t, \mathbf{u}_s \rangle) A_j^t(\mathbf{x}_t). \tag{A23}$$

The right-hand side is a signed directional signal modulated by the current ReLU gate. By Lemma 4, $A_j^t(\mathbf{x}_t)$ can be replaced by $A_j^0(\mathbf{x}_t)$ up to the same gate-stability order as in Theorem 1. Taking conditional expectation given $\mathcal{E}_B$, we get

$$\mathbb{E}[\Delta_j^{t+1} - \Delta_j^t \mid \mathcal{E}_B] = \frac{\eta}{\sqrt{k}} \gamma_j + \mathrm{Err}_{j,t}, \tag{A24}$$

where γ_j is defined in (A19), and

$$|\mathrm{Err}_{j,t}| \leq C_{\mathrm{err}} \frac{(t+1)\eta^2 B^3}{\sigma_w \sqrt{k}} \tag{A25}$$

for a universal constant $C_{\mathrm{err}} > 0$.

By Lemma A1, with probability at least $1 - \exp(-ck)$, there exists a nontrivial subset $\mathcal{J} \subseteq [k]$ with

$$|\mathcal{J}| \geq \frac{c_{\mathrm{sub}}}{2} k$$

such that $\gamma_j \geq c_{\mathrm{resp}}$ for every $j \in \mathcal{J}$. For such neurons,

$$\mathbb{E}[\Delta_j^{t+1} - \Delta_j^t \mid \mathcal{E}_B] \geq \frac{\eta}{\sqrt{k}} c_{\mathrm{resp}} - C_{\mathrm{err}} \frac{(t+1)\eta^2 B^3}{\sigma_w \sqrt{k}}.$$

Thus, whenever the neuron-wise frozen-gate contribution dominates the gate-stability error, namely

$$c_{\mathrm{resp}} > C_{\mathrm{err}} \frac{(t+1)\eta B^3}{\sigma_w},$$

the right-hand side is strictly positive. Therefore,

$$\mathbb{E}[\Delta_j^{t+1} - \Delta_j^t \mid \mathcal{E}_B] > 0 \qquad \text{for all } j \in \mathcal{J}.$$

This proves the trigger-biased reinforcement claim. If the main text uses the notation $\mathcal{J}_t$, we may take $\mathcal{J}_t = \mathcal{J}$ for each early-stage iteration t. □

A.8.1 Proof of Lemma A1

Proof. The key point is that, under Gaussian initialization, a positive fraction of neurons are initially more responsive to the trigger-aligned direction than to the semantic direction. To see this, consider the initialization event

$$\mathcal{R}_j \triangleq \big\{ v_j = +1,\ \langle \mathbf{w}_j^0, \mathbf{u}_t \rangle \geq 2a_0,\ |\langle \mathbf{w}_j^0, \mathbf{u}_s \rangle| \leq a_0 \big\}, \tag{A26}$$

where $a_0 = c_0 \sigma_w$ for a fixed constant $c_0 > 0$. Here $v_j = +1$ corresponds to being aligned with the target sign after absorbing $\widetilde{y}$ into $\mathbf{u}_t$.

By Assumption 2, $\mathbf{w}_j^0 \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$. Under Assumption 1, the two coordinates $\langle \mathbf{w}_j^0, \mathbf{u}_t \rangle$ and $\langle \mathbf{w}_j^0, \mathbf{u}_s \rangle$ are independent centered Gaussians with variance σ_w^2. Since v_j is an independent Rademacher sign, we have

$$\Pr(\mathcal{R}_j) = \frac{1}{2} \Pr(Z \geq 2c_0) \Pr(|Z| \leq c_0),$$

where $Z \sim \mathcal{N}(0,1)$. This probability is a positive constant independent of k.

On the event $\mathcal{R}_j$, the initialization has a positive projection on the trigger direction and a controlled projection on the semantic direction. For an input

$$\mathbf{x} = y\mu_s \mathbf{u}_s + z\widetilde{y}\mu_t \mathbf{u}_t + \boldsymbol{\xi},$$

the initial pre-activation decomposes as

$$\langle \mathbf{w}_j^0, \mathbf{x} \rangle = y\mu_s \langle \mathbf{w}_j^0, \mathbf{u}_s \rangle + z\widetilde{y}\mu_t \langle \mathbf{w}_j^0, \mathbf{u}_t \rangle + \langle \mathbf{w}_j^0, \boldsymbol{\xi} \rangle.$$

After absorbing $\widetilde{y}$ into $\mathbf{u}_t$, poisoned inputs receive a positive trigger-aligned deterministic offset of size at least $2\mu_t a_0$, while the semantic contribution is controlled by $\mu_s a_0$. Under the stated non-degeneracy condition, this trigger-aligned offset is not overwhelmed by the semantic fluctuation or noise. Hence, on a positive-probability subset of $\mathcal{R}_j$, the random variable inside the expectation defining γ_j has positive conditional expectation bounded below by a constant $c_{\mathrm{resp}} > 0$. Equivalently,

$$\mathcal{R}_j \subseteq \{\gamma_j \geq c_{\mathrm{resp}}\}$$

up to an adjustment of constants under the conditioning on $\mathcal{E}_B$. Since $\mathcal{E}_B$ only restricts the data norm on a high-probability event and does not alter the Gaussian symmetry of the initialization, the lower bound in (A20) follows after adjusting constants.

Finally, the indicators $\mathbb{I}\{\gamma_j \geq c_{\mathrm{resp}}\}$ are IID across neurons with success probability at least c_{sub}. A standard multiplicative Chernoff bound gives

$$\Pr\Big(|\mathcal{J}| < \frac{c_{\mathrm{sub}}}{2} k \Big) \leq \exp(-ck)$$

for some universal constant $c > 0$. This proves the lemma. □

A.9 Proof of Proposition 2

Proof. As in the proof above, absorb the target sign $\widetilde{y}$ into $\mathbf{u}_t$, so that poisoned inputs have a positive trigger-aligned shift. Fix a neuron $j \in \mathcal{J}$ and an iteration $t \leq T$. Decompose

$$\mathbf{w}_j^t = a_{j,t}^t \mathbf{u}_t + a_{j,s}^t \mathbf{u}_s + \mathbf{w}_j^{t,\perp},$$

where $\mathbf{w}_j^{t,\perp}$ is orthogonal to both $\mathbf{u}_t$ and $\mathbf{u}_s$. The trigger-dominance gap is

$$\Delta_j^t = \mu_t a_{j,t}^t - \mu_s a_{j,s}^t.$$

By Proposition 1, neurons in $\mathcal{J}$ experience early-stage reinforcement toward the trigger direction in expectation. Hence, when this accumulated reinforcement makes Δ_j^t sufficiently positive relative to the fluctuation level, poisoned inputs induce a larger threshold-crossing probability than clean inputs.

Let

$$\zeta_j^t \triangleq \langle \mathbf{w}_j^t, \boldsymbol{\xi} \rangle.$$

Conditioned on $\mathbf{w}_j^t$, we have

$$\zeta_j^t \sim \mathcal{N}(0, \sigma^2 \|\mathbf{w}_j^t\|^2).$$

For poisoned samples $(z = 1)$,

$$\mathbf{x} = y\mu_s \mathbf{u}_s + \mu_t \mathbf{u}_t + \boldsymbol{\xi},$$

and hence

$$(\mathbf{w}_j^t)^\top \mathbf{x} = \mu_t a_{j,t}^t + y\mu_s a_{j,s}^t + \zeta_j^t.$$

For clean samples $(z = 0)$,

$$\mathbf{x} = y\mu_s \mathbf{u}_s + \boldsymbol{\xi},$$

and therefore

$$(\mathbf{w}_j^t)^\top \mathbf{x} = y\mu_s a_{j,s}^t + \zeta_j^t.$$

Thus, conditional on the semantic label y, the poisoned pre-activation is obtained from the clean pre-activation by adding the trigger-aligned shift $\mu_t a_{j,t}^t$. The semantic term $y\mu_s a_{j,s}^t$ and the Gaussian fluctuation ζ_j^t are shared in the comparison. The quantity Δ_j^t measures the trigger-aligned advantage relative to the semantic component. When this advantage is sufficiently positive compared with the thresholded fluctuation level, the monotonicity of Gaussian tail probabilities implies

$$\Pr\Big((\mathbf{w}_j^t)^\top \mathbf{x} > \tau \mid z = 1, \mathcal{E}_B\Big) > \Pr\Big((\mathbf{w}_j^t)^\top \mathbf{x} > \tau \mid z = 0, \mathcal{E}_B\Big).$$

This proves the claimed thresholded selective activation ordering. □

B Geometry and Structure of the Frozen-Gate Signal

Let $\mathbf{W}^0 = (\mathbf{w}_1^0, \ldots, \mathbf{w}_k^0)$ denote the initialization, independent of the data. This section examines how supervised feature structure and initial ReLU gating enter the frozen-gate signal. We make the conditioning on the realized initialization explicit:

$$h_z = \mathbb{E}\big[\widehat{y}_1 \widehat{y}_2 \mathbf{x}_1^\top \mathbf{x}_2 G_0(\mathbf{x}_1, \mathbf{x}_2) \mid z_1 = z_2 = z, \mathbf{W}^0\big]. \tag{A27}$$

Here, the expectation is over two conditionally IID data samples with the gates fixed; $\mathbb{E}_{\mathbf{W}^0}$ denotes a separate expectation over initialization. We first examine signed-feature agreement and Gaussian halfspace geometry, then derive an exact gated-signal representation and an initialization-average decomposition of the group-wise signal gap. These results clarify the structure of the quantities compared in Corollary 1, without asserting a universal ordering between the clean and poisoned groups.

B.1 Signed-Feature Agreement

We first show that the expected signed-feature agreement within each group equals the squared norm of the corresponding population signal.

Lemma A2 (Signed-feature agreement equals squared population signal)**.** *Work under the directional backdoor model in* (2) *and the training-label rule in* (1)*. For each group $z \in \{0,1\}$, let $\mathbf{m}_z \triangleq \mathbb{E}[\widehat{y}\mathbf{x} \mid z]$. Draw two IID samples $(\mathbf{x}_1, \widehat{y}_1)$ and $(\mathbf{x}_2, \widehat{y}_2)$ from the conditional distribution given $z_1 = z_2 = z$. Then*

$$\mathbb{E}\big[\widehat{y}_1\widehat{y}_2\mathbf{x}_1^\top\mathbf{x}_2 \mid z_1 = z_2 = z\big] = \|\mathbf{m}_z\|^2.$$

In particular, under Lemma 2, $\|\mathbf{m}_0\|^2 = \mu_s^2$ and $\|\mathbf{m}_1\|^2 = \mu_t^2$.

Proof. Let $\mathbf{s} \triangleq \widehat{y}\mathbf{x}$. Since the two samples are conditionally IID given $z_1 = z_2 = z$, we have

$$\begin{aligned}\mathbb{E}\big[\widehat{y}_1\widehat{y}_2\mathbf{x}_1^\top\mathbf{x}_2 \mid z_1 = z_2 = z\big] &= \mathbb{E}\big[\mathbf{s}_1^\top\mathbf{s}_2 \mid z_1 = z_2 = z\big] \\ &= \mathbb{E}[\mathbf{s}_1 \mid z]^\top\mathbb{E}[\mathbf{s}_2 \mid z] = \|\mathbf{m}_z\|^2.\end{aligned}$$

The final statement follows from Lemma 2. □

Lemma A2 shows that the expected signed-feature agreement equals $\|\mathbf{m}_z\|^2$. For the poisoned group, $\widehat{y}\mathbf{x}$ contains a consistent trigger-aligned contribution along $\mathbf{u}_t$, while the semantic term averages out under the balanced-label assumption. Thus, poisoned samples can remain diverse in semantic content while retaining a coherent supervised signal. This does not imply a universal advantage in signed-feature agreement: its magnitude is μ_t^2 for the poisoned group and μ_s^2 for the clean group. In particular, these quantities coincide when $\mu_t = \mu_s$, even though the groups may have different gating geometry.

B.2 Gaussian Halfspace Geometry and Joint Gating

We next explain the second mechanism-level factor: joint gating consistency. Under Gaussian initialization, each ReLU gate corresponds to a random halfspace. For two fixed inputs $\mathbf{x}_1$ and $\mathbf{x}_2$, the probability that they are simultaneously activated by a random neuron depends only on their normalized correlation.

Lemma A3 (Gaussian halfspace joint gating)**.** *Let $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_w^2\mathbf{I})$ with $\sigma_w > 0$, and let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ be two fixed nonzero vectors. Define*

$$\rho(\mathbf{x}_1, \mathbf{x}_2) \triangleq \frac{\mathbf{x}_1^\top\mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|} \in [-1, 1].$$

Then $\Pr(\mathbf{w}^\top\mathbf{x}_1 > 0) = 1/2$, and

$$\Pr\big(\mathbf{w}^\top\mathbf{x}_1 > 0,\, \mathbf{w}^\top\mathbf{x}_2 > 0\big) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho(\mathbf{x}_1, \mathbf{x}_2)).$$

In particular, if $\rho(\mathbf{x}_1, \mathbf{x}_2) \geq \rho_{\min}$ for some $\rho_{\min} \in [-1, 1]$, then

$$\Pr\big(\mathbf{w}^\top\mathbf{x}_1 > 0,\, \mathbf{w}^\top\mathbf{x}_2 > 0\big) \geq \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho_{\min}) \triangleq c_g(\rho_{\min}).$$

Proof. Let $u_1 \triangleq \mathbf{x}_1^\top\mathbf{w}$ and $u_2 \triangleq \mathbf{x}_2^\top\mathbf{w}$. Since $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_w^2\mathbf{I})$, the pair (u_1, u_2) is a centered bivariate Gaussian with

$$\mathrm{Var}(u_1 = \sigma_w^2\|\mathbf{x}_1\|^2, \quad \mathrm{Var}(u_2) = \sigma_w^2\|\mathbf{x}_2\|^2, \quad \mathrm{Cov}(u_1, u_2) = \sigma_w^2\mathbf{x}_1^\top\mathbf{x}_2.$$

After normalization, $s_1 \triangleq u_1/(\sigma_w\|\mathbf{x}_1\|)$ and $s_2 \triangleq u_2/(\sigma_w\|\mathbf{x}_2\|)$ form a centered bivariate Gaussian with unit variances and correlation

$$\mathrm{Corr}(s_1, s_2) = \rho(\mathbf{x}_1, \mathbf{x}_2).$$

The positive scale σ_w cancels after normalization and does not affect the signs of u_1 and u_2. For convenience, write $\rho \triangleq \rho(\mathbf{x}_1, \mathbf{x}_2)$. A standard Gaussian representation gives $s_1 = \beta$ and $s_2 = \rho\beta + \sqrt{1-\rho^2}\gamma$, where $\beta, \gamma \sim \mathcal{N}(0, 1)$ are independent. By symmetry, $\Pr(s_1 > 0) = 1/2$.

To compute the joint probability, write $(\beta, \gamma) = R(\cos\Theta, \sin\Theta)$, where $R \geq 0$ and Θ is uniform on $[0, 2\pi)$. The event $\{s_1 > 0\}$ corresponds to $\cos\Theta > 0$. Let $\Omega \triangleq \arccos(\rho) \in [0, \pi]$. Then

$$s_2 = \rho\beta + \sqrt{1-\rho^2}\gamma = R\cos(\Theta - \Omega).$$

Thus $\{s_2 > 0\}$ corresponds to $\cos(\Theta - \Omega) > 0$. The intersection of the two half-circle events has angular length $\pi - \Omega$, and hence

$$\Pr(s_1 > 0, s_2 > 0) = \frac{\pi - \Omega}{2\pi} = \frac{1}{2} - \frac{\Omega}{2\pi}.$$

Since $\Omega = \arccos(\rho) = \frac{\pi}{2} - \arcsin(\rho)$, we obtain

$$\Pr(s_1 > 0, s_2 > 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho).$$

The lower bound under $\rho \geq \rho_{\min}$ follows from the monotonicity of $\arcsin(\cdot)$. □

Lemma A3 shows that joint activation probability is monotone in the normalized input correlation. A shared trigger can change the poisoned group's angular geometry and can increase its joint gating overlap. However, the lemma describes a single random gate, or equivalently the initialization expectation of G_0 for a fixed input pair; it is not a deterministic lower bound for every finite-width realization. Moreover, signed-feature agreement and gating overlap depend on the same input pair. Larger separate averages of these quantities do not, by themselves, establish a larger expectation of their product. The next subsection treats this coupling directly.

B.3 Gated-Signal Representation and Group-Wise Comparison

We express the frozen-gate signal through gated supervised means. Its initialization average then separates into an ungated signal term and variation of the supervised mean across initial gates.

Lemma A4 (Gated-signal representation and initialization-average decomposition)**.** *Fix a group $z \in \{0, 1\}$. Let $\mathbf{s} \triangleq \widehat{y}\mathbf{x}$ and $\mathbf{m}_z \triangleq \mathbb{E}[\mathbf{s} \mid z]$, with $|\widehat{y}| = 1$ and $M_z \triangleq \mathbb{E}[\|\mathbf{x}\| \mid z] < \infty$. For a fixed vector $\mathbf{w} \in \mathbb{R}^d$, define*

$$\mathbf{b}_z(\mathbf{w}) \triangleq \mathbb{E}\big[\widehat{y}\mathbf{x}\,\mathbb{I}\{\mathbf{w}^\top\mathbf{x} > 0\} \mid z\big].$$

For an initialization $\mathbf{W}^0$ independent of the data, write $\mathbf{b}_{j,z} \triangleq \mathbf{b}_z(\mathbf{w}_j^0)$ and $A_j^0(\mathbf{x}) = \mathbb{I}\{(\mathbf{w}_j^0)^\top\mathbf{x} > 0\}$. For two conditionally IID data samples, recall that

$$h_z = \mathbb{E}\big[\mathbf{s}_1^\top\mathbf{s}_2 G_0(\mathbf{x}_1, \mathbf{x}_2) \mid z_1 = z_2 = z, \mathbf{W}^0\big], \tag{A28}$$

where $G_0(\mathbf{x}_1, \mathbf{x}_2) = k^{-1}\sum_{j=1}^k A_j^0(\mathbf{x}_1)A_j^0(\mathbf{x}_2)$. Then, for every fixed initialization,

$$h_z = \frac{1}{k}\sum_{j=1}^k \|\mathbf{b}_{j,z}\|^2 \geq 0. \tag{A29}$$

Moreover, suppose $\mathbf{w}_1^0, \ldots, \mathbf{w}_k^0$ are independent $\mathcal{N}(\mathbf{0}, \sigma_w^2\mathbf{I})$ vectors with $\sigma_w > 0$. For $\mathbf{w}$ drawn from the same Gaussian law, define

$$V_z \triangleq \mathbb{E}_{\mathbf{w}}\left[\left\|\mathbf{b}_z(\mathbf{w}) - \frac{1}{2}\mathbf{m}_z\right\|^2\right]. \tag{A30}$$

Then

$$\mathbb{E}_{\mathbf{W}^0}[h_z] = \frac{1}{4}\|\mathbf{m}_z\|^2 + V_z. \tag{A31}$$

In particular,

$$\mathbb{E}_{\mathbf{W}^0}[h_z] \geq \frac{1}{4}\|\mathbf{m}_z\|^2. \tag{A32}$$

Proof. First fix $\mathbf{W}^0$. The two data samples remain conditionally independent given their group and the initialization. Since $|\widehat{y}| = 1$ and $M_z < \infty$,

$$\mathbb{E}[\|\mathbf{s}_1\|\|\mathbf{s}_2\| \mid z_1 = z_2 = z, \mathbf{W}^0] = M_z^2 < \infty.$$

Expanding G_0 and using conditional independence therefore gives

$$h_z = \frac{1}{k}\sum_{j=1}^{k}\mathbb{E}\Big[\big(\mathbf{s}_1 A_j^0(\mathbf{x}_1)\big)^\top\big(\mathbf{s}_2 A_j^0(\mathbf{x}_2)\big) \mid z_1 = z_2 = z, \mathbf{W}^0\Big] = \frac{1}{k}\sum_{j=1}^{k}\mathbf{b}_{j,z}^\top\mathbf{b}_{j,z},$$

proving (A29).

For every fixed $\mathbf{w}$,

$$\|\mathbf{b}_z(\mathbf{w})\| \le \mathbb{E}\big[\|\mathbf{x}\|\mathbb{I}\{\mathbf{w}^\top\mathbf{x} > 0\} \mid z\big] \le M_z.$$

Hence $\mathbf{b}_z(\mathbf{w})$ has a finite second moment over initialization. For every fixed nonzero $\mathbf{x}$, Gaussian symmetry gives

$$\mathbb{E}_{\mathbf{w}}[\mathbb{I}\{\mathbf{w}^\top\mathbf{x} > 0\}] = \frac{1}{2}.$$

At $\mathbf{x} = \mathbf{0}$, the signed feature is zero. Independence and $M_z < \infty$ permit interchange of expectations, yielding

$$\mathbb{E}_{\mathbf{w}}[\mathbf{b}_z(\mathbf{w})] = \mathbb{E}\big[\widehat{y}\mathbf{x}\,\mathbb{E}_{\mathbf{w}}[\mathbb{I}\{\mathbf{w}^\top\mathbf{x} > 0\}] \mid z\big] = \frac{1}{2}\mathbf{m}_z.$$

Consequently, expanding the squared norm around this mean gives

$$\mathbb{E}_{\mathbf{w}}[\|\mathbf{b}_z(\mathbf{w})\|^2] = \left\|\frac{1}{2}\mathbf{m}_z\right\|^2 + \mathbb{E}_{\mathbf{w}}\left[\left\|\mathbf{b}_z(\mathbf{w}) - \frac{1}{2}\mathbf{m}_z\right\|^2\right] = \frac{1}{4}\|\mathbf{m}_z\|^2 + V_z,$$

where the cross term vanishes because $\mathbb{E}_{\mathbf{w}}[\mathbf{b}_z(\mathbf{w}) - \mathbf{m}_z/2] = \mathbf{0}$. Taking the initialization expectation in (A29) and using the identical distribution of the initial weights proves (A31). The lower bound (A32) follows from $V_z \ge 0$. □

Group-wise signal comparison. Under Lemma 2, $\|\mathbf{m}_0\|^2 = \mu_s^2$ and $\|\mathbf{m}_1\|^2 = \mu_t^2$. Subtracting the two instances of (A31) gives

$$\mathbb{E}_{\mathbf{W}^0}[h_1 - h_0] = \frac{\mu_t^2 - \mu_s^2}{4} + (V_1 - V_0). \tag{A33}$$

The gap thus separates into differences in ungated signal strength and gated-mean variation. When $\mu_t = \mu_s$,

$$\mathbb{E}_{\mathbf{W}^0}[h_1 - h_0] = V_1 - V_0.$$

Equal nominal strengths therefore do not, by themselves, force equal initialization-average signals. The quantity V_z depends on the joint input–label distribution and its selection by initial gates; it is neither the mean overlap g_z nor a measure of gate changes during training.

Remark 1 (Scope of the comparison)**.** *The identity* (A29) *holds for each fixed initialization, whereas* (A31) *and* (A33) *average over initialization. The decomposition alone does not determine the sign of $V_1 - V_0$; establishing that sign requires further properties of the group distributions. An initialization-average ordering does not by itself verify the realized signal-gap condition in Corollary 1. These statements use the population data distribution and the unconditioned Gaussian initialization law. The finite first-moment assumption holds for the directional Gaussian model.*

A shared target-aligned trigger can modify both supervised feature structure and gating, but Corollary 1 still requires a positive realized signal gap large enough to dominate the approximation errors. The quantities $\|\mathbf{m}_z^{\bar{t}}\|_2^2$ and g_z in Section 4 diagnose feature alignment and gating overlap. They should be interpreted alongside the measured frozen-gate signal, not as an exact factorization of h_z or direct measurements of V_z.

C Experimental Protocols and Additional Results

The experimental appendix is organized as follows. Appendix C.1 describes the common setup used across all experiments. Appendix C.2 provides additional empirical observations that motivated the theoretical formulation but are omitted from the main text due to space constraints. Appendix C.3 gives the detailed protocol and additional results for validating early-stage optimization bias, including a group-wise diagnostic decomposition. Appendix C.4 reports representation-level diagnostics and post-training defense results, which are discussed jointly because post-training behavior is evaluated through the same trigger-sensitive representation measures.

C.1 Common Experimental Setup

Unless otherwise specified, all empirical results in Appendix C follow a shared experimental protocol. We use the same datasets, backbone architecture, training pipeline, and targeted backdoor setting across experiments, and vary only the components needed for each diagnostic. Section-specific quantities, such as the frozen-gate signal, early-stage loss decrease, TDS, and TSAG, are defined in the corresponding subsections. All experiments were run on an NVIDIA Tesla V100 GPU.

Datasets, model, and training. We conduct experiments on CIFAR-10 and CIFAR-100 [11]. All models use ResNet-18 [6] as the backbone and are trained from scratch using SGD for 100 epochs with learning rate 0.01 and batch size 128. We do not use data augmentation unless explicitly stated. For representation-level diagnostics, we extract features from the last hidden layer before the classifier and treat each hidden dimension as a neuron or channel.

Targeted backdoor poisoning. We consider two representative targeted backdoor attacks: BadNets [5] and TrojanNN [16]. For BadNets, we use a fixed manually specified trigger following the common BackdoorBench-style configuration [28], with trigger size 3×3. For TrojanNN, we follow the standard optimized-trigger setup, where the trigger is generated to activate selected neurons of a clean model before being used for backdoor implantation [16]; the trigger size is fixed as 8×8. The attack target label is fixed as $\widetilde{y} = 0$ in all experiments.

Training samples are divided into a clean group and a poisoned group. Clean samples are trained with their original labels, while poisoned samples are stamped with the trigger and assigned the target label $\widetilde{y}$. The poisoning ratio α denotes the fraction of poisoned samples in the training set. Unless otherwise specified, we use the default poisoning ratio $\alpha = 5\%$. In poisoning-ratio sensitivity experiments, we use $\alpha \in \{1\%, 2\%, 5\%, 10\%\}$.

Output-level evaluation. We report main-task accuracy (ACC) and attack success rate (ASR). ACC is measured on the clean test set. ASR is measured by stamping the trigger on test samples and evaluating whether the model predicts the target label $\widetilde{y}$. When computing ASR, we exclude test samples whose original label is already the target label. For post-training analysis, we also report Top-K ASR, where an attack is counted as successful if the target label appears among the Top-K predicted classes under triggered inputs. Standard ASR is therefore equivalent to Top-1 ASR.

Representation-level evaluation. Unless otherwise specified, representation-level diagnostics are computed at the penultimate layer. Given a clean input and its trigger-stamped counterpart, we use neuron-wise activation changes to quantify trigger sensitivity. In particular, the Trigger-Activated Change (TAC) [33] metric measures the change in a neuron's activation induced by the trigger. Larger TAC values indicate stronger trigger sensitivity and therefore a greater potential contribution to backdoor behavior.

Vanilla clean fine-tuning protocol. For post-training defense analysis, we deliberately focus on vanilla clean fine-tuning as the simplest post-training intervention, rather than enhanced variants with additional regularization [14] or sharpness-aware objectives [35]. We sample a clean subset whose size is 5% of the original training set and fine-tune the backdoored model using only this clean subset. Unless otherwise specified, clean fine-tuning uses SGD with learning rate 0.005 for 10 epochs. The fine-tuned model is then evaluated using both output-level metrics and representation-level diagnostics.

C.2 Additional Motivating Observations

This subsection provides additional empirical observations that motivate our simplicity-biased optimization view. These observations are not used as assumptions in the theoretical analysis. Instead, they illustrate recurring phenomena in backdoor learning that our framework aims to explain: poisoned samples often enjoy a faster early-stage loss decrease, trigger-sensitive neurons emerge during backdoor training, and vanilla clean fine-tuning can weaken but does not necessarily remove all trigger-related signatures. As output-level context, Table A2 shows that the BadNets backdoored model achieves 100.00% ASR on CIFAR-10, while vanilla clean fine-tuning reduces ASR to 18.52%. This substantial but incomplete reduction motivates the representation-level analyses below.

Faster loss decrease on poisoned samples. We first examine the training dynamics of clean and poisoned samples during backdoor training. Figure A1 plots the average training loss of the clean and poisoned groups on CIFAR-10 under the BadNets attack with different poisoning ratios. Although

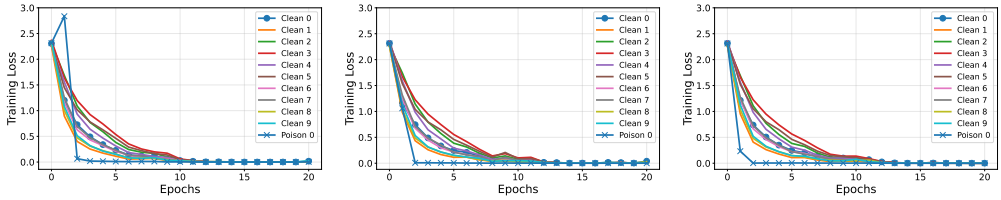


Figure A1: Average training loss on poisoned and clean samples under different poisoning ratios on CIFAR-10 with BadNets. From left to right, $\alpha = 1\%$, 5%, and 10%. The poisoned group shows a faster early-stage loss decrease across poisoning ratios, and the separation becomes more pronounced as α increases.

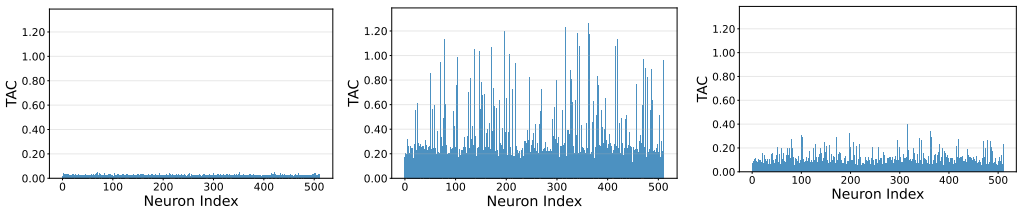


Figure A2: Neuron-wise TAC profiles at the penultimate layer on CIFAR-10 for a clean model, a BadNets-poisoned model, and the same poisoned model after vanilla clean fine-tuning, with poisoning ratio $\alpha = 5\%$. Compared with the clean model, the backdoored model contains a small subset of neurons with much larger TAC values, indicating the emergence of trigger-sensitive neurons. Vanilla clean fine-tuning reduces the overall TAC magnitude, but the fine-tuned model still retains a residual group of high-TAC neurons compared with the clean model.

poisoned samples account for only a small fraction of the training set, their average loss decreases noticeably faster in the early stage. This observation suggests an early-stage optimization advantage of poisoned samples during backdoor training.

The same qualitative behavior already appears at $\alpha = 1\%$, and the early-stage separation between the poisoned and clean groups becomes more pronounced as the poisoning ratio increases. This trend is consistent with Corollary 2, which shows that the mixture-level loss dynamics become increasingly dominated by the poisoned-group signal as α grows.

Trigger-sensitive neurons and residual signatures after vanilla clean fine-tuning. We next examine whether backdoor training leaves localized signatures at the representation level, and whether these signatures are removed by vanilla clean fine-tuning. We use the Trigger-Activated Change (TAC) [33] metric to quantify how strongly each hidden unit responds to the trigger. For a fixed hidden unit, TAC measures the average activation change between a clean input and its trigger-stamped counterpart. Larger TAC values indicate stronger trigger sensitivity.

Figure A2 compares the neuron-wise TAC profiles at the penultimate layer of three models on CIFAR-10: a clean model, a BadNets-poisoned model, and the same poisoned model after clean fine-tuning. We discuss the comparison from three perspectives.

Backdoored model versus clean model. The clean model has uniformly small TAC values across neurons, suggesting that most hidden units are insensitive to the trigger pattern. In contrast, the backdoored model contains a small subset of neurons with much larger TAC values than the rest. This indicates that backdoor training does not spread trigger sensitivity uniformly across the representation, but instead concentrates it into a localized subset of trigger-sensitive neurons.

Fine-tuned model versus backdoored model. Vanilla clean fine-tuning substantially reduces the overall TAC magnitude compared with the backdoored model, indicating that many trigger-sensitive responses are weakened by clean-only post-training updates. This is consistent with the output-level results in Table A2, where vanilla clean fine-tuning reduces the ASR under BadNets from 100.00% to 18.52%.

Table A1: Effect of pruning high-TAC neurons on CIFAR-10 under the BadNets attack with poisoning ratio $\alpha = 5\%$. Pruning neurons with TAC above the mean TAC of the poisoned model sharply reduces ASR, indicating that high-TAC neurons are functionally related to the learned backdoor behavior.

Model	ACC	ASR
Backdoored model	84.76%	100.00%
After pruning high-TAC neurons	77.46%	2.83%

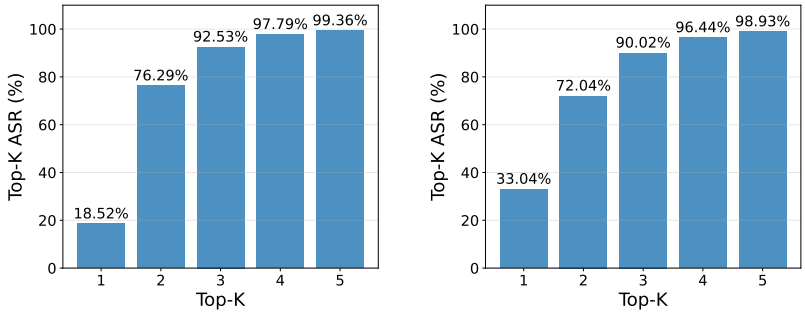


Figure A3: Top-K ASR of clean fine-tuned backdoored models on CIFAR-10 with poisoning ratio $\alpha = 5\%$ under BadNets (left) and TrojanNN (right). In both cases, clean fine-tuning substantially reduces the Top-1 ASR, but the target label remains highly ranked under triggered inputs, as reflected by the rapid increase of Top-K ASR as K grows. This suggests that vanilla clean fine-tuning weakens the backdoor at the decision level but may still leave a residual output-level preference toward the target label.

Fine-tuned model versus clean model. Despite this reduction, the neuron-wise TAC profile of the fine-tuned model still contains several elevated responses compared with the clean model. This suggests that vanilla clean fine-tuning may attenuate the learned backdoor behavior without fully restoring the trigger-insensitive representation of a clean model. In other words, some trigger-sensitive neurons can remain as residual representation-level signatures after fine-tuning.

To test whether these high-TAC neurons are functionally related to the backdoor behavior, we perform a simple pruning-based diagnostic [15]. Specifically, we prune neurons whose TAC values exceed the mean TAC of the poisoned model and evaluate the resulting model without additional retraining. This procedure is not intended as a full defense method, but as an intervention to test whether the localized subset of neurons with elevated TAC values contributes to the attack behavior.

As shown in Table A1, pruning high-TAC neurons reduces ASR from 100.00% to 2.83%, while the clean accuracy decreases from 84.76% to 77.46%. The sharp reduction in ASR indicates that the high-TAC neurons are strongly associated with the learned backdoor behavior. At the same time, the nontrivial accuracy drop shows that these neurons may also participate in benign prediction to some extent, so TAC pruning should be interpreted as a diagnostic intervention rather than a practical post-training defense.

Residual output-level preference after vanilla clean fine-tuning. The residual representation-level sensitivity also appears at the output level. Figure A3 reports Top-K ASR after vanilla clean fine-tuning under both BadNets and TrojanNN on CIFAR-10. Although vanilla clean fine-tuning substantially reduces the Top-1 ASR in both cases, the target label remains highly ranked under triggered inputs, as reflected by the rapid increase of Top-K ASR as K grows. In particular, the Top-2 ASR reaches 76.29% under BadNets and 72.04% under TrojanNN. This indicates that even when the trigger no longer forces the target label to be the top prediction, it can still push the target label close to the decision boundary, leaving a residual output-level backdoor signature.

Taken together, these observations motivate the central mechanism studied in this work. Backdoor training gives poisoned samples an early-stage optimization advantage, concentrates trigger sensitivity into a localized subset of neurons, and leaves residual trigger-related signatures that may persist

after vanilla clean fine-tuning. The theoretical and empirical results in the main text explain these phenomena through the lens of early-stage simplicity-biased optimization dynamics.

C.3 Empirical Verification of Early-Stage Optimization Bias

This subsection provides the experimental details for validating the early-stage optimization bias studied in Section 4. The goal is to clarify how the empirical frozen-gate signal, the early-stage loss decrease gap, and the group-wise diagnostic factors are computed in practical multi-class neural networks. The main text reports the core validation results due to space constraints. Here, we provide the full measurement protocol and additional diagnostic decomposition results.

C.3.1 Empirical Quantities and Checkpoint Protocol

Target-versus-rest signed representation. Although the theoretical analysis is developed in a binary directional model, the empirical evaluation is performed on the original multi-class classifiers. To connect the multi-class model to the binary mechanism studied in the theory, we use a target-versus-rest signed representation. For a training sample with actual training label $\widehat{y}$, define

$$\delta(\widehat{y}, \widetilde{y}) = \begin{cases} +1, & \widehat{y} = \widetilde{y}, \\ -1, & \widehat{y} \neq \widetilde{y}. \end{cases}$$

Let $\phi_t(\mathbf{x})$ denote the penultimate-layer representation of input $\mathbf{x}$ at epoch t. We define the signed representation

$$\mathbf{s}^t(\mathbf{x}) = \delta(\widehat{y}, \widetilde{y})\phi_t(\mathbf{x}).$$

This construction is the empirical target-versus-rest counterpart of the signed feature $\widehat{y}\mathbf{x}$ in the binary directional model.

Group-wise signed-feature mean. For each group $z \in \{0, 1\}$, where $z = 0$ denotes clean samples and $z = 1$ denotes poisoned samples, we compute the empirical signed-feature mean

$$\mathbf{m}_z^t = \frac{1}{n_z} \sum_{i: z_i = z} \mathbf{s}^t(\mathbf{x}_i).$$

The squared norm $\|\mathbf{m}_z^t\|_2^2$ measures the strength of within-group signed-feature alignment. A larger value indicates that samples in the group induce more coherent signed representations.

Frozen-gate reference and probe gates. We fix the frozen-gating reference epoch as $t_0 = 1$. For each trained model, we extract penultimate-layer representations at epoch t_0 and instantiate random probe gates in the representation space. Specifically, we draw probe weights

$$\mathbf{a}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad j \in [k],$$

and define the frozen probe gate

$$A_j^0(\mathbf{x}) = \mathbb{I}\big(\mathbf{a}_j^\top \phi_{t_0}(\mathbf{x}) > 0\big).$$

The corresponding empirical frozen-gate overlap is

$$G_0(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{k} \sum_{j=1}^{k} A_j^0(\mathbf{x}_1) A_j^0(\mathbf{x}_2).$$

The probe gates are used to instantiate the frozen-gate kernel in representation space, rather than to reproduce the exact gates of the trained network. Unless otherwise specified, we use probe width $k = 512$.

Empirical frozen-gate signal. For each group z, the empirical frozen-gate signal at checkpoint $\bar{t}$ is estimated as

$$h_z^{\bar{t}} = \mathbb{E}\big[\delta(\widehat{y}_1, \widetilde{y})\delta(\widehat{y}_2, \widetilde{y})\phi_{\bar{t}}(\mathbf{x}_1)^\top \phi_{\bar{t}}(\mathbf{x}_2) G_0(\mathbf{x}_1, \mathbf{x}_2) \mid z_1 = z_2 = z\big].$$

In implementation, the expectation is approximated by Monte Carlo sampling within each group. Unless otherwise specified, we use $M = 10^6$ sampled within-group pairs. The frozen-gate signal gap is computed as

$$h_1^{\bar{t}} - h_0^{\bar{t}}.$$

A positive gap indicates that the poisoned group induces a stronger frozen-gate supervised signal than the clean group.

Early-stage loss decrease gap. We set the early-stage checkpoint as $\bar{t} = 1$. Let L_z^t denote the average multi-class cross-entropy loss of group z at epoch t. The group-wise loss decrease from epoch $\bar{t}$ to epoch $\bar{t}+1$ is

$$D_z^{\bar{t}} = L_z^{\bar{t}} - L_z^{\bar{t}+1}.$$

The poisoned-minus-clean loss decrease gap is

$$\Delta D^{\bar{t}} = D_1^{\bar{t}} - D_0^{\bar{t}} = \left(L_1^{\bar{t}} - L_1^{\bar{t}+1}\right) - \left(L_0^{\bar{t}} - L_0^{\bar{t}+1}\right).$$

A positive value of $\Delta D^{\bar{t}}$ means that the poisoned group obtains a larger average loss decrease than the clean group in the early stage. Since the empirical loss is the standard multi-class cross-entropy loss, this quantity is used as a qualitative counterpart of the theoretical hinge-loss decrease, rather than as an exact numerical instantiation of Theorem 1.

C.3.2 Verification Protocol under Fixed and Varying Poisoning Ratios

Fixed poisoning ratio. For the fixed-ratio verification in Figure 2, we use the default poisoning ratio $\alpha = 5\%$. For each dataset-attack pair, we train the corresponding backdoored model, extract the penultimate-layer representations at epochs 1 and 2, and compute both the frozen-gate signal gap $h_1^{\bar{t}} - h_0^{\bar{t}}$ and the early-stage loss decrease gap $\Delta D^{\bar{t}}$ using the protocol above. The two quantities are evaluated jointly: the frozen-gate signal gap tests the predicted signal-level asymmetry, while the loss decrease gap tests whether this asymmetry is reflected in the early-stage optimization dynamics.

Varying poisoning ratios. For the poisoning-ratio sensitivity study in Figure 3, we vary $\alpha \in \{1\%, 2\%, 5\%, 10\%\}$ and keep the same datasets, attacks, backbone, and checkpoint protocol. For each dataset-attack-poisoning-ratio combination, we train an independent backdoored model and compute $h_1^{\bar{t}} - h_0^{\bar{t}}$ and $\Delta D^{\bar{t}}$. The loss decrease gap is not expected to be strictly monotone in α, because it is measured under a multi-class loss and can also reflect class competition and finite-sample optimization effects. Nevertheless, the overall trend is used to assess whether larger poisoning ratios strengthen the poisoned-group contribution, as predicted by Corollary 2.

C.3.3 Group-Wise Diagnostic Decomposition

Diagnostic factors. To further diagnose why the poisoned group can have a larger frozen-gate signal, we examine two group-wise factors associated with $h_z^{\bar{t}}$. The first is the signed-feature alignment magnitude $\|\mathbf{m}_z^{\bar{t}}\|_2^2$, which measures how coherently the group aligns with the target-versus-rest signed direction. The second is the average frozen-gate overlap

$$g_z = \mathbb{E}[G_0(\mathbf{x}_1, \mathbf{x}_2) \mid z_1 = z_2 = z],$$

which measures the consistency of joint ReLU activation within group z. Although we do not claim an exact factorization of $h_z^{\bar{t}}$, larger values of $\|\mathbf{m}_z^{\bar{t}}\|_2^2$ and g_z provide a natural diagnostic explanation for why a group may exhibit a larger frozen-gate signal. Appendix B gives a complementary theoretical explanation under a Gaussian directional model, showing how stronger trigger-aligned mean separation and more concentrated trigger geometry can jointly increase the poisoned-group frozen-gate signal.

The main text reports the BadNets diagnostic factors g_z and $\|\mathbf{m}_z^{\bar{t}}\|_2^2$ due to space constraints. Here, we provide the complete group-wise diagnostic decomposition for both BadNets and TrojanNN, including the frozen-gate signal $h_z^{\bar{t}}$. Figure A4 shows the results on CIFAR-10 and CIFAR-100 across poisoning ratios. For both attacks, the poisoned group consistently exhibits larger values than the clean group in the signed-feature alignment magnitude, the joint gating factor, and the frozen-gate signal. This pattern indicates that the poisoned-group advantage in $h_z^{\bar{t}}$ is supported by coordinated differences in signed-feature alignment and gating consistency.

Among the diagnostic factors, the separation in $\|\mathbf{m}_z^{\bar{t}}\|_2^2$ is typically the most pronounced, suggesting that stronger signed-feature alignment is a major driver of the observed frozen-gate signal gap. At the same time, $g_1 > g_0$ indicates that poisoned samples tend to share more consistent frozen-gate activation patterns. Together, these two effects provide a coherent diagnostic explanation for why

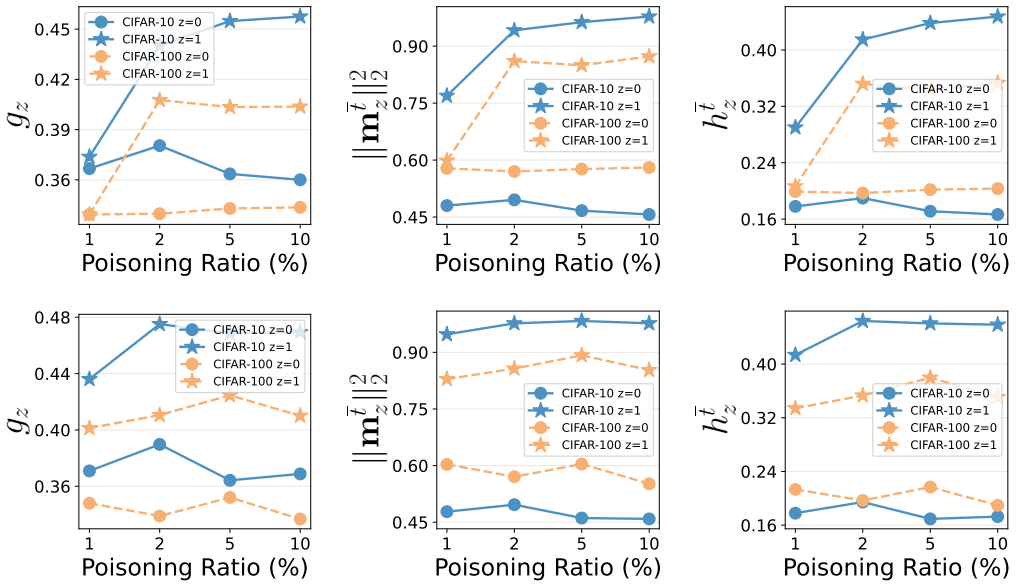


Figure A4: Complete group-wise diagnostic decomposition under BadNets (first row) and TrojanNN (second row). The three columns report the group-wise joint gating factor g_z, the squared norm of the empirical signed-feature mean $\|\mathbf{m}_z^{\bar{t}}\|_2^2$, and the frozen-gate signal $h_z^{\bar{t}}$, respectively. Across poisoning ratios on CIFAR-10 and CIFAR-100, the poisoned group (*i.e.*, $z = 1$) consistently shows larger values than the clean group (*i.e.*, $z = 0$), indicating that its larger frozen-gate signal is accompanied by stronger signed-feature alignment and higher joint gating consistency under both attacks.

the poisoned group attains a larger frozen-gate signal, namely $h_1^{\bar{t}} > h_0^{\bar{t}}$, under both BadNets and TrojanNN. This empirical pattern mirrors the Gaussian-geometry analysis in Appendix B, where a trigger-aligned mean shift and higher within-trigger geometric coherence increase the poisoned-group signal in the frozen-gate regime.

This decomposition should be interpreted as a diagnostic analysis rather than an exact algebraic factorization of $h_z^{\bar{t}}$. Its role is to show that the poisoned-group advantage in the frozen-gate signal is supported by two mechanism-level ingredients: stronger target-aligned signed-feature agreement and more consistent frozen-gate co-activation.

C.4 Representation-Level Diagnostics and Post-Training Defense

This subsection provides the experimental details for the representation-level diagnostics and post-training defense analysis in the main text. Section 5 reports representative BadNets results due to space constraints. Here, we first summarize the output-level behavior of the evaluated models, and then provide the full definitions of the diagnostic metrics, the evaluation protocol, early-checkpoint evidence for the formation of representation-level bias, and the corresponding TrojanNN results.

C.4.1 Output-Level Performance of Evaluated Models

Before examining the representation-level diagnostics, we first summarize the output-level behavior of the models used in the following analysis. Table A2 reports ACC and ASR on CIFAR-10 for the clean model, the backdoored models under BadNets and TrojanNN, and their corresponding models after clean fine-tuning. Under both attacks, backdoor training preserves clean accuracy while achieving 100.00% ASR, confirming that the implanted backdoors are effective. Vanilla clean fine-tuning substantially reduces ASR under both attacks, but does not fully eliminate the attack behavior. Meanwhile, ACC also decreases after fine-tuning. These results provide the output-level context for the representation-level analyses below.

Table A2: Output-level performance on CIFAR-10 under clean training, backdoor training, and vanilla clean fine-tuning. ACC is measured on clean test inputs, and ASR is measured on triggered non-target test inputs.

Setting	Model	ACC (%)	ASR (%)
Clean	Clean	85.40	–
BadNets	Backdoored	85.14	100.00
	Fine-tuned	79.00	18.52
TrojanNN	Backdoored	85.25	100.00
	Fine-tuned	76.13	33.04

C.4.2 Diagnostic Metrics and Evaluation Protocol

Paired clean and triggered inputs. For each clean input $\mathbf{x}_i$, we construct its trigger-stamped counterpart

$$\mathbf{x}_i^{\mathrm{trig}} = \mathrm{Trigger}(\mathbf{x}_i).$$

Unless otherwise specified, paired clean-triggered inputs are constructed from non-target test samples, so that successful attack behavior can be attributed to the trigger-induced shift toward the target label rather than to samples whose original label is already $\widetilde{y}$.

Trigger-Dominance Score (TDS). Let $\phi_t(\mathbf{x}) \in \mathbb{R}^d$ denote the penultimate-layer representation at epoch t, and let $\phi_{t,j}(\mathbf{x})$ denote the post-ReLU activation of neuron j. To quantify how strongly a neuron responds to the trigger, we define the TDS of neuron j at epoch t as

$$\mathrm{TDS}_j^t = \frac{1}{N}\sum_{i=1}^{N}\Big(\phi_{t,j}(\mathbf{x}_i^{\mathrm{trig}}) - \phi_{t,j}(\mathbf{x}_i)\Big).$$

A large positive value of TDS_j^t indicates that neuron j is activated more strongly by triggered inputs than by their clean counterparts. Thus, TDS serves as an empirical proxy for the trigger-biased reinforcement described in Proposition 1.

Thresholded Selective Activation Gap (TSAG). To measure whether a neuron is selectively active on poisoned samples, we define the empirical activation probability of neuron j on group $z \in \{0, 1\}$ at threshold τ as

$$\widehat{p}_{j,z}^t(\tau) = \frac{1}{n_z}\sum_{i:z_i=z}\mathbb{I}(\phi_{t,j}(\mathbf{x}_i) > \tau),$$

where $z = 0$ denotes clean samples and $z = 1$ denotes poisoned samples. The TSAG is then defined as

$$\mathrm{TSAG}_j^t(\tau) = \widehat{p}_{j,1}^t(\tau) - \widehat{p}_{j,0}^t(\tau).$$

A positive value of $\mathrm{TSAG}_j^t(\tau)$ means that neuron j exceeds the activation threshold more frequently on poisoned samples than on clean samples.

Fixed global threshold. The threshold τ is computed once from the clean baseline model on clean samples and then fixed for all model comparisons. Specifically, we pool all penultimate-layer post-ReLU activations of the clean baseline model over the evaluated clean samples and set τ to their global median. This fixed-threshold protocol prevents the threshold from being re-calibrated separately for clean, backdoored, and fine-tuned models, which could otherwise hide backdoor-induced activation shifts.

Neuron ranking and group-wise summaries. For group-wise diagnostics, neurons are ranked according to their TDS values in the backdoored model. We then report average TDS and average TSAG over selected neuron subsets, including the top 1%, top 5%, random 5%, and bottom 5% neurons under this ranking. This protocol tests whether trigger-dominant behavior is concentrated in a small subset of neurons and whether such neurons also exhibit selective activation on poisoned samples.

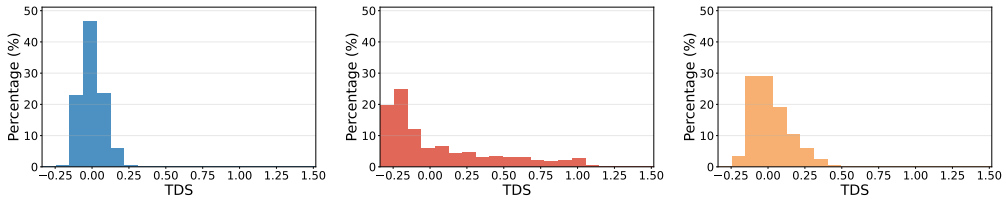


Figure A5: TDS distributions at the final checkpoint on CIFAR-10 under TrojanNN. We compare the clean model, the backdoored model, and the corresponding model after vanilla clean fine-tuning. The backdoored model exhibits a heavier positive-side tail than the clean model, indicating the emergence of trigger-dominant neurons. Vanilla clean fine-tuning attenuates this tail but does not fully restore the clean distribution.

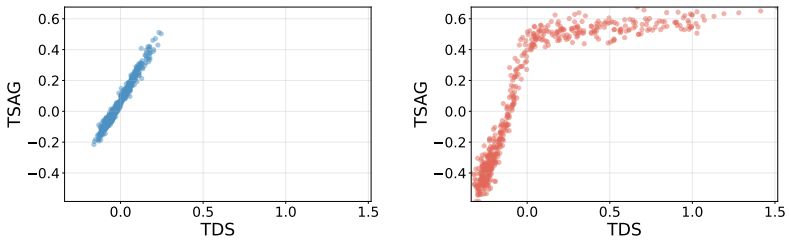


Figure A6: Relation between TDS and TSAG at the final checkpoint on CIFAR-10 under TrojanNN. The clean model already shows a mild positive spread because the TrojanNN trigger is optimized to activate selected neurons. Nevertheless, the backdoored model exhibits a substantially stronger right-upper spread, indicating that trigger-dominant neurons are also more selectively activated on poisoned samples.

C.4.3 Trigger-Dominance and Selective Activation under TrojanNN

The main text reports final-checkpoint TDS and TSAG diagnostics under BadNets. We now provide the corresponding final-checkpoint results under TrojanNN, complementing the representative BadNets results reported in Section 5. Figure A5 compares the TDS distributions under TrojanNN on CIFAR-10 for the clean model, the backdoored model, and the same backdoored model after vanilla clean fine-tuning. The clean model already shows a mild positive spread, which is consistent with the optimized nature of the TrojanNN trigger. Nevertheless, the backdoored model exhibits a substantially heavier positive-side tail, indicating that backdoor training amplifies a subset of trigger-dominant neurons. Vanilla clean fine-tuning compresses this tail, but the fine-tuned distribution does not fully return to the clean baseline, suggesting a residual trigger-related representation footprint.

Figure A6 further reports the relation between TDS and TSAG under TrojanNN. Compared with the clean model, the backdoored model shows a stronger right-upper spread, meaning that neurons with larger trigger-induced activation changes also tend to be more selectively active on poisoned samples. This supports the representation-level bridge from trigger-biased reinforcement to thresholded selective activation, as described by Propositions 1 and 2.

C.4.4 Early Checkpoint Evolution of Representation-Level Bias

The final-checkpoint diagnostics in the main text and above show that backdoor training produces trigger-dominant and selectively activated neurons. We further examine whether these representation-level signatures emerge only after long training or already appear during the early stage. This distinction is important because our theory predicts that the poisoned group can obtain an early-stage optimization advantage, which should rapidly induce trigger-related representation bias.

Figure A7 reports the checkpoint-wise evolution of the TDS distribution under BadNets and TrojanNN on CIFAR-10. Trigger-dominant neurons emerge quickly: already at early checkpoints such as $t = 1$, the backdoored models contain neurons with large positive TDS values, indicating that some neurons respond more strongly to triggered inputs than to their clean counterparts. As training proceeds, the

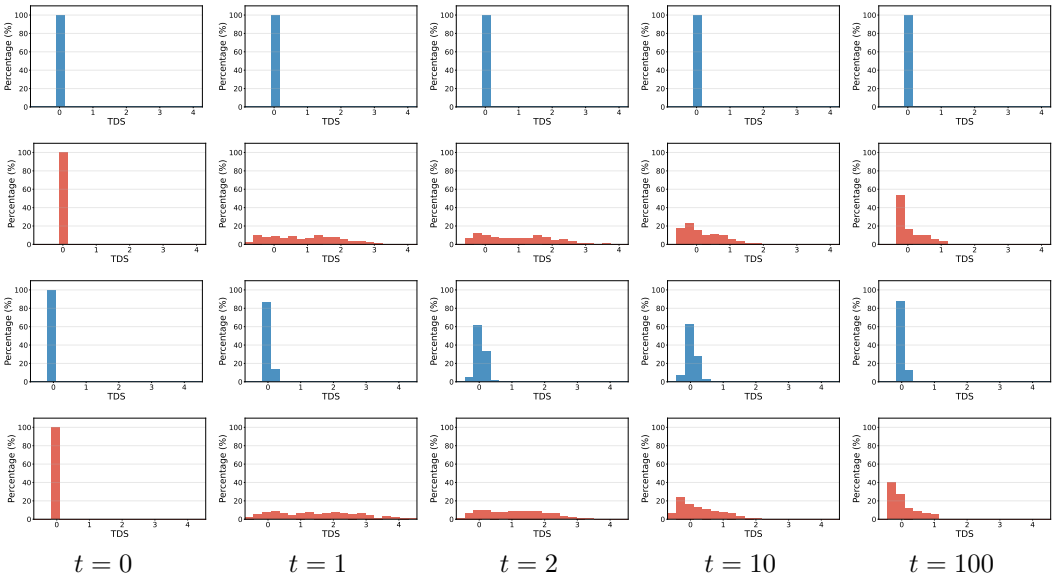


Figure A7: Checkpoint-wise evolution of TDS distributions on CIFAR-10 under BadNets (top two rows) and TrojanNN (bottom two rows). Within each attack, the upper row shows the clean model and the lower row shows the backdoored model. We report selected checkpoints $t = 0, 1, 2, 10, 100$. In both attacks, the clean and backdoored distributions are close at initialization, but the backdoored model quickly develops a positive-side separation that remains visible at later checkpoints. These results show that trigger-dominant neurons emerge early during training, consistent with the early-stage optimization bias studied in the main text.

extreme early TDS values may decrease, but the backdoored models maintain a stable positive-side separation from the clean models. In contrast, the clean models remain concentrated around zero. These results show that trigger-related representation bias is formed rapidly during early training and persists throughout later training.

Figure A8 reports the checkpoint-wise evolution of the relation between TDS and TSAG. The association between high TDS and high TSAG already appears at early checkpoints such as $t = 1$, indicating that neurons responding strongly to the trigger also tend to become selectively active on poisoned samples. Although the largest TDS values may decrease at later checkpoints, this right-upper pattern remains clearly separated from the clean model. These checkpoint-wise results directly connect the early-stage optimization bias studied in Section 3 with the representation-level consequences discussed in Section 5: backdoor-related neurons and selective activation patterns are formed rapidly during training, rather than appearing only as final-checkpoint artifacts.

These checkpoint-wise results show that representation-level backdoor signatures are formed rapidly during early training, providing a temporal link between the loss-dynamics evidence and the final representation diagnostics.

C.4.5 Group-Wise Representation Diagnostics under TrojanNN

The main text reports the group-wise TDS and group-wise TSAG summaries under BadNets. Figure A9 provides the corresponding TrojanNN results. The left panel reports the average TDS over selected neuron subsets, and the right panel reports the average TSAG over the same subsets. The subsets are formed according to the TDS ranking of the backdoored model.

Under TrojanNN, the backdoored model exhibits substantially larger average TDS on the top-ranked neuron subsets than the clean baseline, showing that trigger-dominant behavior is concentrated in a relatively small subset of neurons. After vanilla clean fine-tuning, the average TDS of these top-ranked subsets is reduced, indicating that fine-tuning weakens the trigger-dominant structure. However, the fine-tuned model remains above the clean baseline, suggesting that the structure is attenuated but not fully erased.

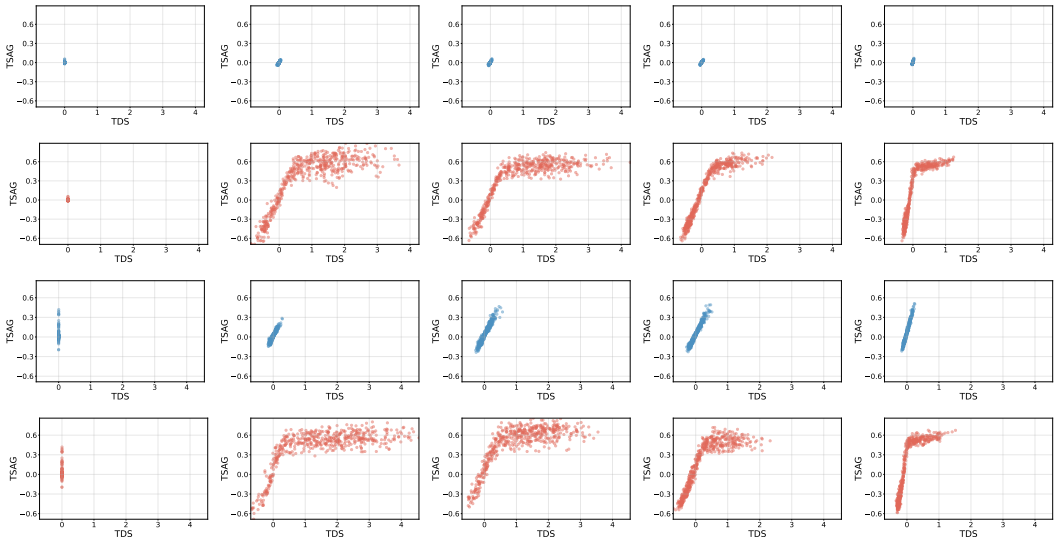


Figure A8: Checkpoint-wise evolution of the relation between TDS and TSAG on CIFAR-10 under BadNets (top two rows) and TrojanNN (bottom two rows). Within each attack, the upper row shows the clean model and the lower row shows the backdoored model. We report selected checkpoints $t = 0, 1, 2, 10, 100$. In both attacks, the backdoored model quickly develops a clearer right-upper structure than the clean model from the early stage, indicating that trigger-dominant neurons also become selectively activated on poisoned samples. This early coupling between trigger dominance and selective activation is consistent with the representation-level consequences of early-stage optimization bias.

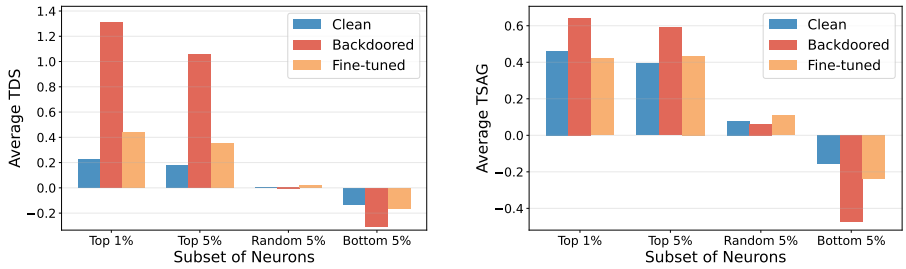


Figure A9: Group-wise representation diagnostics under TrojanNN on CIFAR-10. The left panel reports average TDS over selected neuron subsets, and the right panel reports average TSAG over the same subsets. Neurons are grouped according to the TDS ranking of the backdoored model, including the top 1%, top 5%, random 5%, and bottom 5% subsets. The backdoored model exhibits larger TDS and TSAG on the top-ranked subsets than the clean baseline. Vanilla clean fine-tuning reduces both quantities but does not fully restore them to the clean regime.

The same pattern appears in the TSAG summaries. The backdoored model has larger average TSAG on the top-ranked neuron subsets, indicating that trigger-dominant neurons are also more selectively active on poisoned samples. Clean fine-tuning reduces this selective activation, but the fine-tuned model still remains above the clean baseline. Together, these results extend the BadNets findings in the main text to TrojanNN and support the view that vanilla clean fine-tuning weakens, rather than completely removes, the trigger-related representation structure.

C.4.6 Output-Level and Representation-Level Evaluation after Vanilla Clean Fine-Tuning

Evaluation protocol. After vanilla clean fine-tuning, we evaluate both output-level and representation-level behavior. At the output level, we compute ACC, ASR, and Top-K ASR on triggered test samples

excluding samples whose original label is already $\widetilde{y}$. At the representation level, we compute TAC, TDS, and TSAG at the penultimate layer. This combined evaluation is important because vanilla clean fine-tuning may substantially reduce Top-1 ASR while still leaving residual trigger-related preferences or activation patterns.

Summary of evidence. The output-level results in Table A2 show that vanilla clean fine-tuning substantially reduces ASR under both BadNets and TrojanNN, but does not fully eliminate attack behavior. The TAC and Top-K ASR results in Appendix C.2 provide complementary evidence: neuron-wise TAC profiles show that fine-tuning reduces overall trigger sensitivity but may still leave a small subset of neurons with elevated TAC values, while Top-K ASR shows that the target label can remain highly ranked under triggered inputs even when it is no longer the top prediction. The TDS and TSAG results above provide a more fine-grained representation-level view of the same phenomenon: vanilla clean fine-tuning attenuates trigger-dominant and selectively activated neurons, but does not necessarily restore the representation to the clean regime.

D Detailed Related Work

Backdoor attacks and defenses. Backdoor attacks have been extensively studied in modern machine learning systems. Early works showed that an adversary can implant a trigger-target association into a model through poisoned training data, causing the model to behave normally on clean inputs while predicting an attacker-specified label whenever the trigger is present [2, 5]. Since then, a large body of work has developed diverse attack and defense methods, while benchmark efforts have systematized empirical evaluation protocols [28, 4]. Representative defenses include trigger inversion and detection [26], neuron- or channel-level pruning and purification [15, 29, 33], and training-based or feature-level mitigation methods [14, 31, 20].

Post-training mitigation methods are particularly relevant to our analysis. These methods attempt to weaken backdoors after training through fine-tuning, pruning, adversarial unlearning, or feature-level adjustment [15, 14, 31, 20, 35]. Recent studies further suggest that the apparent effectiveness of such defenses can be limited or unstable, and that weakened backdoor behavior may remain latent or become reactivated [36, 21]. These findings indicate that the effectiveness of post-training mitigation depends on how backdoor-related features have already been encoded during training. Our work complements this line of research by explaining why standard training can efficiently encode trigger-aligned representations and why clean fine-tuning can weaken backdoor behavior without necessarily eliminating them.

Backdoor learning dynamics and neuron-level signatures. A growing line of work has revealed recurring empirical behaviors in backdoored models during training and post-training analysis. Poisoned samples have been observed to exhibit faster loss decrease during training [13, 32], and backdoor functionality is often associated with a small subset of neurons or channels that respond strongly to trigger patterns [15, 34, 29, 33]. These observations provide important evidence about the internal evolution of backdoor learning and motivate the search for a more principled dynamical explanation.

However, existing evidence is largely observation-driven. Faster fitting of poisoned samples, trigger-sensitive neurons, and residual backdoor behavior after post-training mitigation are often studied as separate phenomena. Our work connects these phenomena through a common early-stage dynamics perspective: poisoned samples can induce a stronger frozen-gate signal, obtain faster early loss decrease, and rapidly produce trigger-dominant representation signatures.

Theoretical analyses of backdoor learning. Recent works have begun to provide theoretical explanations for why backdoor attacks can succeed. Manoj and Blum [18] study backdoor poisoning from the perspective of excess capacity, showing how overparameterized models can accommodate both clean and backdoor behavior. Xian et al. [30] propose the adaptability hypothesis to explain when models can learn backdoor mappings. Wang et al. [27] analyze poisoning backdoor attacks from a statistical perspective and characterize factors such as trigger direction and poisoning rate. Li and Liu [12] study dirty-label backdoor poisoning in two-layer CNNs under a multi-view data model, analyze the learning dynamics of feature and trigger vectors, and identify conditions for achieving both high clean accuracy and high attack success rate.

These works provide important theoretical understanding of backdoor success conditions and trigger-feature learning. Our work is complementary. Rather than focusing on capacity, statistical risk, or time-to-capture conditions for trigger vectors, we study the group-wise early-stage optimization dynamics through which poisoned samples obtain a larger frozen-gate signal, faster loss decrease, and trigger-dominant representation signatures.

Simplicity bias and shortcut learning. Our work is also related to the literature on simplicity bias and shortcut learning in neural network training. Prior studies suggest that deep networks and gradient-based training are biased toward simple functions or lower-complexity rules. Pérez et al. [23] argue that the parameter-function map of deep networks is biased toward simple functions. Nakkiran et al. [22] show that SGD on neural networks learns functions of increasing complexity over training, while Lyu et al. [17] connect gradient descent on two-layer networks to margin maximization and simplicity bias. Shah et al. [25] further show that simplicity bias can cause models to rely on simple but potentially non-robust features. Related work on shortcut learning studies how neural networks choose among shortcut cues when multiple predictive signals are available [24, 7]. These works provide important background for understanding why some features may be learned earlier or more strongly than others.

Relatedly, Ilyas et al. [9] show that standard classifiers can exploit highly predictive yet non-robust features in the data distribution, helping explain adversarial vulnerability. Our work shares the perspective that supervised learning may rely on predictive cues that do not align with human-interpretable semantics. However, our setting differs in both the origin of the cue and the research question: the trigger is deliberately injected by an attacker and associated with a target label, and need not be non-robust to small adversarial perturbations. We investigate how such a malicious cue can gain a conditional early-stage optimization advantage and examine its connection to trigger-dominant representations and empirically observed residual signatures after clean fine-tuning.

Positioning of our work. Our work differs from the above lines of research in both objective and scope. We do not propose a new attack, a new benchmark, or a new post-training defense. Instead, we seek a unified theoretical explanation for several central empirical phenomena in backdoor learning. To this end, we introduce a directional data model that separates semantic and trigger-related directions, and use it to connect simplicity-biased optimization dynamics to backdoor learning in a security-critical training scenario.

Compared with prior work, our focus is distinct in four aspects:

- Compared with empirical attack and defense studies, we do not aim to improve attack strength or defense performance, but to explain why backdoor-related features are learned efficiently by standard training.
- Compared with neuron-level diagnostic studies, we connect faster poisoned fitting, trigger-sensitive neurons, and residual post-training behavior through a single early-stage optimization mechanism.
- Compared with theoretical backdoor analyses, we focus on early-stage optimization dynamics through which backdoor-related features are formed.
- Compared with general simplicity-bias theory, we study a maliciously injected, target-aligned trigger direction rather than a benign shortcut or spurious feature.

Focusing on the early stage of training, our analysis characterizes a group-wise asymmetry whereby poisoned samples induce more coherent and more strongly aligned updates along the trigger direction. This asymmetry leads to faster loss decrease on poisoned samples, the emergence of trigger-dominant neurons, and representation-level effects that are only partially mitigated by clean fine-tuning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are consistent with the theoretical results, empirical validation, and stated scope of the submission.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses limitations including the stylized directional backdoor model, the early-stage nature of the analysis, and the gap between binary theory and multi-class experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the main text, and complete proofs of the main theoretical results are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main text and Appendix C provide the computation resources, datasets, attacks, model architecture, training setup, poisoning ratios, evaluation metrics, and diagnostic protocols needed to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The experiments use standard public datasets and the submission provides detailed protocols, but the current anonymous version does not include a public code release.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and evaluation details, including optimizer, learning rates, batch size, epochs, poisoning ratios, target label, attacks, and diagnostic metrics, are provided in the Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The submission reports consistent trends across datasets, attacks, and poisoning ratios, but does not include error bars over repeated random seeds for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The submission reports the computation resources used for the experiments in Appendix C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses standard public datasets and focuses on understanding and diagnosing backdoor learning mechanisms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This submission focuses on the theoretical analysis and empirical diagnosis of existing backdoor learning phenomena, and does not include a separate discussion of broader societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The submission does not release high-risk pretrained models, scraped datasets, or other assets with elevated misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The submission uses standard public datasets and existing attack settings, and cites the corresponding original sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The submission does not introduce a new dataset, benchmark, pretrained model, or standalone research asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The submission does not involve crowdsourcing, user studies, or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The submission does not involve human subjects or personal data, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the method, theory, or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.