SELF-CORRECTION WITH INTEGRITY: FROM TRIGGER-INDUCED GRADIENT ALIGNMENT TO DATA-PRESERVING BACKDOOR DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Poisoned samples are often learned earlier than clean samples, a phenomenon already exploited by backdoor defenses. However, how trigger-induced structure contributes to this learning advantage remains insufficiently understood. To this end, we propose Self-Correction with Integrity (SCI), which is a clean-set-free backdoor defense guided by a theoretical analysis of early learning. We decompose early training gradients to examine the roles of shared target labels and recurring trigger features. This analysis explains how these two factors can produce mutually reinforcing updates that favor early learning of the trigger-target association. Specifically, based on this analysis, SCI first uses low-capacity models to construct a suspicious subset with high poisoned-sample recall. It then uses a model trained on the remaining data to relabel suspicious samples according to their semantic content, weakening trigger-target associations. Next, SCI reconstructs the relabeled samples under the guidance of task-relevant features to suppress residual trigger information while preserving semantic content. Finally, the corrected samples are merged with the remaining data to form a purified training set that preserves data integrity for subsequent model training. Experiments on CIFAR-10 and ImageNet subsets show SCI achieves effective backdoor mitigation while maintaining clean accuracy. Comparisons with sample removal further demonstrate the accuracy benefit of retaining corrected samples. These results show that a mechanistic understanding of early learning can guide backdoor defense through sample correction, retaining the full training set without auxiliary clean data.

1 INTRODUCTION

Backdoor attacks compromise machine learning models by injecting poisoned samples with hidden triggers into their training data Li et al. (2025). A model trained on such data can behave normally on clean inputs while predicting an attacker-specified label when a trigger is present Yan et al. (2025); Zheng et al. (2025). Therefore, training-time defenses are important for limiting the influence of poisoned samples while preserving the model's ability to learn the intended task.

Recent training-time defenses identify poisoned samples through differences in prediction behavior or learned representations. MSPC examines prediction consistency under input scaling to identify suspicious data without auxiliary clean samples Pal et al. (2024). PIPD combines loss-based initialization with progressive separation in feature space to improve poisoned-sample isolation Chen et al. (2024). Other methods recover useful supervision from potentially poisoned data. AIBD corrects suspicious labels using predictions obtained through adversarial perturbations Yin et al. (2025), while VIBE infers clean pseudolabels through variational inference Sabolić et al. (2025). DataElixir extends correction to the inputs themselves by using diffusion models to suppress trigger features and recover benign content Zhou et al. (2024). These approaches have advanced both poisoned-sample identification and the recovery of useful training information.

Despite this progress, the early-learning advantage exploited by loss-based defenses remains insufficiently explained. Observing that poisoned samples reach low training losses early does not establish how recurring trigger features shape gradient updates beyond the effect of a shared target label. Understanding this distinction would help explain why poisoned samples can become distinguishable during early training. Such an explanation could also inform how to correct suspicious samples while preserving the useful semantic content that sample removal would discard. Thus, our central research question is how a mechanistic explanation of backdoor early learning can guide sample correction without auxiliary clean data while retaining the full training set.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this question, we propose Self-Correction with Integrity (SCI), which is a clean-set-free backdoor defense guided by a mechanistic analysis of early learning. We decompose early training gradients to examine the roles of shared target labels and recurring trigger features. This analysis explains how trigger features can contribute additional gradient alignment beyond the shared-label effect, producing mutually reinforcing updates that favor early learning of the trigger-target association. Specifically, SCI first uses low-capacity models to exploit this learning advantage and construct a suspicious subset with high poisoned-sample recall. This subset may also contain clean samples, thus the subsequent correction does not assume a perfect separation. To recover useful supervision, SCI uses a model trained on the remaining data to relabel suspicious samples according to their semantic content, weakening trigger-target associations. Since relabeling leaves the inputs unchanged, SCI then reconstructs these samples to suppress residual trigger information. Task-relevant features from the relabeling model serve as anchors to preserve semantic content during reconstruction. Finally, the corrected samples are merged with the remaining data to form a purified training set that retains all examples for subsequent model training.

We evaluate SCI on CIFAR-10 and ImageNet subsets under representative backdoor attacks and compare it with prior defenses. Comparisons on CIFAR-10 with matched source images and target labels support the contribution of trigger features to early gradient alignment beyond the shared-label effect. Across the evaluated attack types and poisoning rates, SCI achieves high poisoned-sample recall and effective backdoor mitigation without auxiliary clean data. Compared with removing the same suspicious samples, SCI achieves higher clean accuracy while retaining the full training set, demonstrating the practical value of sample correction.

In summary, our contributions are as follows.

- We propose Self-Correction with Integrity (SCI), which is a clean-set-free backdoor defense grounded in a mechanistic explanation of early learning. Our theoretical analysis decomposes early training gradients to explain how recurring trigger features can strengthen gradient alignment beyond the shared-label effect, favoring early learning of the trigger-target association.

- We develop a sample correction pipeline that uses low-capacity models to construct a suspicious subset with high poisoned-sample recall. Semantic relabeling weakens trigger-target associations, while input reconstruction uses task-relevant feature anchors to suppress residual trigger information and preserve semantic content. The corrected samples are merged back to retain the full training set.

- We conduct extensive experiments on CIFAR-10 and ImageNet-100 under five representative backdoor attacks and compare SCI with six prior defenses. The results demonstrate high poisoned-sample recall and effective backdoor mitigation without auxiliary clean data. Controlled gradient comparisons support the proposed mechanism, while comparisons with sample removal show the clean-accuracy benefit of retaining corrected samples.

2 Related Work

Recent backdoor research has moved beyond fixed visible triggers toward more adaptive, stealthy, and deployment-oriented threats. Recent attacks improve evasiveness through sub-partitioned triggers Cheng et al. (2024), prompt-level or embedding-level manipulation in CLIP and multimodal contrastive models Bai et al. (2024); Liang et al. (2024), physical or frequency-domain trigger designs Yin et al. (2024); Yuan et al. (2025), and poisoning of self-supervised or large-scale models Chen et al. (2025); Souly et al. (2025). Recent defenses also cover a wide range of settings, including test-time detecting and repairing Guan et al. (2024), black-box sample detoxification Yang et al. (2024); Miah & Bi (2026), vision-language knowledge alignment Ishmam & Thomas (2024), entropy-based poisoned dataset separation Xu et al. (2025), variational training on poisoned data Sabolić et al. (2025), activation-space sanitization Amula et al. (2025), and post-training optimization Wei et al. (2025). These methods have improved empirical robustness, but most of them still treat poisoned-sample separability as a useful signal rather than explaining why such separability arises during early training. Many also rely on clean reference data, test-time intervention, external foundation models, model-level access, or removal of suspicious samples. SCI differs from these works by providing a theoretical account of backdoor early learning from trigger-induced low-entropy structure and gradient alignment. This analysis directly motivates a clean-set-free prior defense that constructs a high-recall suspicious subset and repairs it through relabeling and reconstruction, rather than discarding suspicious training data. We have provided more detailed explanations in Appendix A.

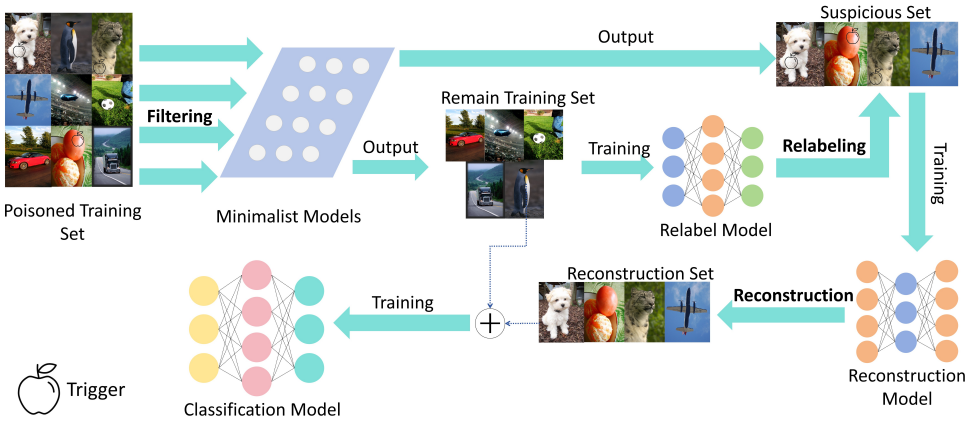


Figure 1: Overview of SCI. SCI consists of three main steps: filtering, relabeling, and reconstruction.

3 Proposed SCI

3.1 Problem Setting and Defense Goal

We consider a training set $\mathcal{D} = \{(u_i, y_i)\}_{i=1}^{n}$ that may contain poisoned samples. For analysis, we write $\mathcal{D} = \mathcal{D}_c \cup \mathcal{D}_p$, where the two subsets contain clean and poisoned samples, respectively. An attacker can modify a fraction of the training inputs and assign them an attacker-specified target label. The attacker does not control the defender's training procedure. The defender observes only $\mathcal{D}$ and does not know which samples are poisoned. Neither the trigger nor the poisoning rate is available, and no auxiliary clean dataset is provided.

SCI constructs a suspicious subset $\mathcal{S}$ and corrects its samples before training the final task model. The remaining subset is denoted by $\mathcal{R} = \mathcal{D} \setminus \mathcal{S}$. Our goal is to weaken trigger-target associations while retaining information useful for the intended task. Each training record is retained exactly once in the resulting dataset $\mathcal{D}^*$. Thus, data integrity refers to retaining the full set of training records, while semantic preservation is an objective of the correction process.

3.2 Overview

SCI connects a mechanistic analysis of early learning with sample correction, as illustrated in Figure 1. The analysis examines how recurring trigger features contribute to gradient alignment beyond the effect of a shared target label. Motivated by this mechanism, low-capacity models construct a suspicious subset with an emphasis on poisoned-sample recall. A model trained on $\mathcal{R}$ then provides semantic labels for the selected samples. Its task-relevant features also guide input reconstruction to suppress residual trigger information. The corrected samples are finally merged with $\mathcal{R}$ for subsequent model training.

3.3 A Mechanistic Account of Backdoor Early Learning

Poisoned samples can be fitted earlier than clean samples, an observation already used by backdoor defenses Li et al. (2021a). We study how shared target labels and recurring trigger features can produce this behavior. The analysis considers a group of poisoned samples assigned to a common target label y_t. This assumption is used to explain the mechanism and does not provide the defender with the target label.

Training gradients. Let $h_\theta(u) \in \mathbb{R}^C$ denote the logits for C classes and let $q_\theta(u) = \mathrm{softmax}(h_\theta(u))$. For cross-entropy loss $\ell_\theta(u,y) = -\log q_\theta(y \mid u)$, define $\rho = |\mathcal{D}_p|/n$. When both subsets are nonempty, uniform sampling from $\mathcal{D}$ gives

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{(u,y)\in\mathcal{D}} \ell_\theta(u,y) = (1-\rho)\mathcal{L}_c(\theta) + \rho\mathcal{L}_p(\theta), \tag{1}$$

where $\mathcal{L}_c$ and $\mathcal{L}_p$ are the mean losses within their respective subsets. The decomposition is used only for analysis. SCI does not require ρ or the hidden partition. The gradient of an individual sample is

$$g_\theta(u,y) = J_\theta(u)^\top \big(q_\theta(u) - e_y\big), \qquad J_\theta(u) = \frac{\partial h_\theta(u)}{\partial \theta}, \tag{2}$$

where e_y is the one-hot label vector. This expression separates the label-dependent loss signal from its mapping into parameter space.

From shared labels to a shared update direction. Early in training, predictions may remain close to a common probability vector q_0. For samples assigned to y_t, the loss signal is then close to $a_t = q_0 - e_{y_t}$. Their parameter gradients still depend on their individual Jacobians, so a shared label alone does not establish parameter alignment. With the source inputs and target labels held fixed, adding a trigger changes this feature-dependent mapping while preserving the reference signal a_t. This identifies where an additional trigger contribution can enter the update.

Suppose the trigger induces a learnable Jacobian component J_{trig} shared across the poisoned group. Write $J_\theta(z) = J_{\text{trig}} + \Delta_z$ and $q_\theta(z) = q_0 + \delta_z$. The gradient can then be expressed as

$$g_\theta(z, y_t) = v + r_z, \qquad v = J_{\text{trig}}^\top a_t, \tag{3}$$

$$r_z = \Delta_z^\top a_t + J_\theta(z)^\top \delta_z. \tag{4}$$

The common vector v captures the interaction between the trigger-related component and the target label. The residual r_z includes variation in other features and predictions. If $v \neq 0$ and $\|r_z\|_2 \leq \varepsilon \|v\|_2$ for every sample in the group, with $0 \leq \varepsilon < \sqrt{2} - 1$, then

$$g_\theta(z, y_t)^\top g_\theta(z', y_t) \geq (1 - 2\varepsilon - \varepsilon^2)\|v\|_2^2 > 0. \tag{5}$$

The bound follows by expanding $(v + r_z)^\top (v + r_{z'})$ and bounding the three residual terms with the Cauchy–Schwarz inequality. It gives an explicit sufficient condition for mutually reinforcing gradients. The condition depends on the features expressed by the current model and is not implied by attack success on another model.

Implications for early learning. Let $\bar{g}_p = \nabla_\theta \mathcal{L}_p$ and $\bar{g}_c = \nabla_\theta \mathcal{L}_c$. Pairwise agreement contributes to the aggregate poisoned gradient through the exact identity

$$\|\bar{g}_p\|_2^2 = \frac{1}{|\mathcal{D}_p|^2} \sum_{(z,y_t),(z',y_t) \in \mathcal{D}_p} g_\theta(z, y_t)^\top g_\theta(z', y_t). \tag{6}$$

At a differentiable point, for a sufficiently small gradient step $\theta^+ = \theta - \eta\big((1-\rho)\bar{g}_c + \rho \bar{g}_p\big)$, a first-order expansion gives

$$\mathcal{L}_p(\theta^+) - \mathcal{L}_p(\theta) = -\eta\left(\rho\|\bar{g}_p\|_2^2 + (1-\rho)\bar{g}_p^\top \bar{g}_c\right) + o(\eta). \tag{7}$$

Thus, coherent poisoned gradients can reinforce learning of the trigger-target association. The net loss change also depends on the poisoning rate and interaction with clean gradients. This explains a route to early learning under stated conditions without requiring larger individual poisoned-sample gradients. It does not establish a universal ordering of clean and poisoned learning rates. The appendix expands the derivation and separates the trigger contribution through matched source inputs and target labels.

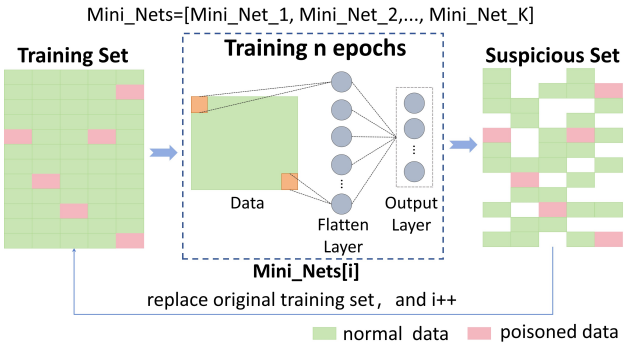


Figure 2: Iterative filtering in SCI. Each round retains correctly classified samples as candidates for subsequent correction.

3.4 Filtering

The analysis motivates using low-capacity models to expose trigger-related patterns that are readily learned. We refer to these models as minimalist models because their capacity is restricted relative to the classification task. The intended separation arises when they capture recurring trigger features before fitting much of the remaining task structure. Clean samples with readily learned features can also be selected.

Filtering starts from $\mathcal{S}_0 = \mathcal{D}$. At round k, a newly initialized minimalist model f_{θ_k} is trained on $\mathcal{S}_{k-1}$ with the supplied labels and cross-entropy loss. SCI retains the correctly classified records

$$\begin{aligned} \mathcal{S}_k &= \left\{(u, y) \in \mathcal{S}_{k-1} \mid \arg\max_c q_{\theta_k}(c \mid u) = y\right\}, \\ \mathcal{S} &= \mathcal{S}_T. \end{aligned} \tag{8}$$

Here, T is the number of filtering rounds. This rule requires neither a loss threshold nor knowledge of the poisoning rate. The CIFAR-10 experiments use two rounds. Additional architecture and training details appear in Appendix C.

The resulting subset is a candidate set for correction, as illustrated in Figure 2. Filtering quality is measured by

$$\mathrm{TPR} = \frac{|\mathcal{S} \cap \mathcal{D}_p|}{|\mathcal{D}_p|}, \qquad \mathrm{FPR} = \frac{|\mathcal{S} \cap \mathcal{D}_c|}{|\mathcal{D}_c|}, \tag{9}$$

when the corresponding ground-truth subset is nonempty. These metrics are used for evaluation and are unavailable to the defender during filtering. SCI prioritizes recall because correction operates only on $\mathcal{S}$. Missed poisoned samples remain in $\mathcal{R}$ and can affect both relabeling and final training. Additional rounds may reduce false positives while also excluding poisoned candidates, so increasing T does not guarantee better defense.

3.5 Semantic Relabeling

We train a relabeling model M_{rel} on $\mathcal{R}$ using its supplied labels. When the filter selects poisoned samples more frequently than clean samples, $\mathcal{R}$ has a lower poison concentration than $\mathcal{D}$. This provides a basis for learning task semantics from the remaining data, although $\mathcal{R}$ may still contain poisons. Class balancing accounts for unequal numbers of retained examples across represented classes without changing the label space.

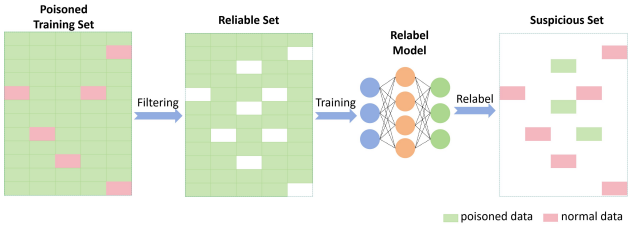


Figure 3: Semantic relabeling. A model trained on the remaining subset provides labels for the suspicious samples.

Let $q_{\mathrm{rel}}(c \mid u)$ denote the class probabilities of M_{rel}. As shown in Figure 3, the model supplies labels for all selected records

$$\widetilde{y}_u = \arg\max_c q_{\mathrm{rel}}(c \mid u), \qquad \widetilde{\mathcal{S}} = \big\{(u, \widetilde{y}_u) \mid (u, y) \in \mathcal{S}\big\}. \tag{10}$$

Relabeling aims to weaken trigger-target associations by recovering supervision consistent with semantic content. It applies to the whole suspicious subset, including any selected clean samples. Its reliability depends on the learned semantic predictions. The next stage addresses trigger information that remains in the inputs after their labels have been reassigned.

3.6 Feature-Guided Reconstruction

SCI uses the relabeling model to guide reconstruction of the selected inputs. Let $\phi_{\mathrm{rel}}(u)$ be its feature map immediately before global max pooling. For each input, $\mathcal{K}(u)$ contains the indices of the top-K activated channels. The selected features serve as anchors for retaining the evidence used by M_{rel} to predict the semantic label. Their usefulness therefore depends on the quality of the relabeling model.

A reconstruction network M_{rec} produces $\widehat{u} = M_{\mathrm{rec}}(u)$. With M_{rel} fixed, we optimize

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\widetilde{\mathcal{S}}|} \sum_{(u,\widetilde{y}_u)\in\widetilde{\mathcal{S}}} \Big(\alpha_1 \mathcal{L}_{\mathrm{label}}(u) + \alpha_2 \mathcal{L}_{\mathrm{feat}}(u) + \alpha_3 \mathcal{L}_{\mathrm{SSIM}}(u) + \alpha_4 \mathcal{L}_{\mathrm{TV}}(u)\Big), \tag{11}$$

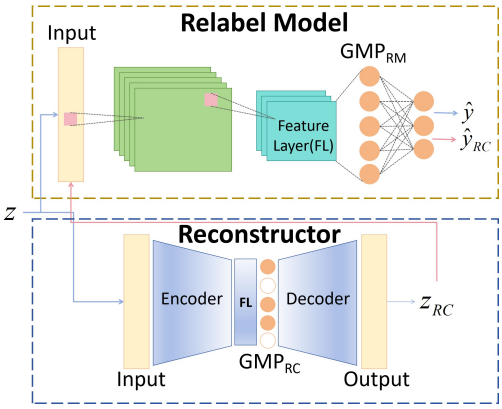


Figure 4: Feature-guided reconstruction. Semantic labels and selected feature channels constrain the corrected input, while structural regularization limits visual distortion.

where the coefficients are nonnegative. The semantic constraints are

$$\mathcal{L}_{\mathrm{label}}(u) = -\log q_{\mathrm{rel}}(\widetilde{y}_u \mid \widehat{u}), \tag{12}$$

$$\mathcal{L}_{\mathrm{feat}}(u) = \big\|\phi_{\mathrm{rel}}(u)_{\mathcal{K}(u)} - \phi_{\mathrm{rel}}(\widehat{u})_{\mathcal{K}(u)}\big\|_F^2. \tag{13}$$

The feature norm is taken over the selected channels and their spatial locations. Label consistency maintains the relabeled class, while feature consistency anchors the reconstruction to task-relevant evidence. The SSIM term preserves coarse visual structure. Total variation regularization discourages spatial

artifacts. Together, these constraints guide input correction while limiting changes to useful semantic content. The anchors specify evidence to preserve while allowing other input details to change. Appendix E gives the remaining loss definitions.

After reconstruction, SCI forms

$$\widehat{\mathcal{S}} = \left\{ (M_{\text{rec}}(u), \widetilde{y}_u) \mid (u, \widetilde{y}_u) \in \widetilde{\mathcal{S}} \right\}, \tag{14}$$

$$\mathcal{D}^* = \mathcal{R} \cup \widehat{\mathcal{S}}, \qquad |\mathcal{D}^*| = |\mathcal{D}|. \tag{15}$$

The union preserves training-record identities, including records whose corrected values coincide. The final task model is trained on $\mathcal{D}^*$. This completes purification through correction while retaining every training record.

4 Experiments

4.1 Experimental Setup

We evaluate SCI on CIFAR-10 Krizhevsky et al. (2009) and two subsets of ImageNet Deng et al. (2009), namely ImageNet-100 and ImageNet-30. We use ResNet-50 He et al. (2016) on ImageNet-100 and ResNet-18 for the CIFAR-10 comparisons with recent defenses. The ImageNet-30 comparisons use a common backbone across methods. We compare SCI with Spectre Hayase et al. (2021), AC Chen et al. (2018) and STRIP Gao et al. (2019). Additional baselines include CT Qi et al. (2023), ASSET Pan et al. (2023) and ABL Li et al. (2021a). We also include the recent defenses AIBD Yin et al. (2025) and VIBE Sabolić et al. (2025). Detailed experimental settings are provided in Appendix F.

4.2 Trigger-Associated Alignment Beyond a Shared Label

A shared target label can increase gradient agreement even without a trigger. To examine the additional contribution of trigger features, we compare four equally sized groups on CIFAR-10 with 5% poisoning. Clean-all contains clean examples from all classes, while Clean-target contains examples from the target class. Label-only assigns the target label to non-target images without modifying their inputs. Poisoned uses the same source images and assigned labels as Label-only, with the trigger added. We measure pairwise cosine similarity between per-sample parameter gradients during minimalist-model training. The measurements cover iterations 150–250 across all ten target classes and three random seeds.

Table 1: Gradient alignment on CIFAR-10 with 5% poisoning. Label-only and Poisoned share source images and assigned labels, allowing the trigger-associated difference to be examined.

Attack	Clean-all	Clean-target	Label-only	Poisoned
BadNet	0.018 ± 0.011	0.139 ± 0.031	0.146 ± 0.035	$\mathbf{0.274 \pm 0.047}$
Blended	0.015 ± 0.009	0.127 ± 0.034	$\mathbf{0.153 \pm 0.041}$	0.231 ± 0.052
WaNet	$\mathbf{0.022 \pm 0.013}$	0.134 ± 0.039	0.141 ± 0.037	0.191 ± 0.046
ISSBA	0.013 ± 0.010	0.116 ± 0.036	0.137 ± 0.042	0.169 ± 0.049
Color-backdoor	0.019 ± 0.012	$\mathbf{0.143 \pm 0.033}$	0.149 ± 0.039	0.218 ± 0.051

Table 1 shows higher mean gradient alignment in Clean-target and Label-only than in Clean-all, indicating that shared labels contribute to gradient agreement. Adding a trigger further increases mean alignment across all five attacks while keeping the source images and assigned labels fixed. This controlled comparison supports our claim that trigger features strengthen alignment beyond the shared-label effect, making early updates more mutually reinforcing. The increase also occurs for ISSBA, suggesting that this contribution does not require an identical trigger across samples. These findings support the gradient mechanism that motivates SCI's filtering stage.

The training curves in Figure 5 reveal a growing accuracy gap at 0.2% poisoning. The minimalist model eventually fits most poisoned samples while its clean-sample accuracy remains low. This separation supports SCI's use of correct predictions to select a suspicious subset with high poisoned-sample recall. The gap develops gradually, highlighting the importance of sufficient training to ex-

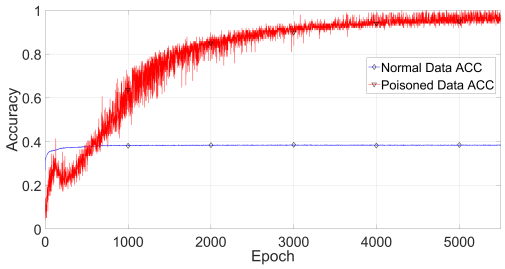


Figure 5: Training accuracy of the minimalist model on clean and poisoned samples at 0.2% poisoning. The accuracy gap develops over training and supports SCI's correctness-based filtering.

pose the filtering signal when poisoning is rare. Some clean samples are also selected, which motivates correcting the suspicious subset to preserve useful training information.

4.3 The Value of Correcting Selected Samples

The alignment signal motivates selecting suspicious samples, but selection alone does not establish the value of correction. We therefore compare correction variants on CIFAR-10 under Blended with 0.5% poisoning. All defended variants use the same $\mathcal{S}$ and remaining subset $\mathcal{R} = \mathcal{D} \setminus \mathcal{S}$. The classifier and training schedule are fixed, and the variants share random seeds. Filtering-only removes $\mathcal{S}$. Relabeling-only corrects its labels while keeping the original inputs. Reconstruction-only changes the selected inputs while retaining their supplied labels. Full SCI applies both corrections and returns the samples to the training set.

Table 2: Component ablation on CIFAR-10 under Blended with 0.5% poisoning. All defended variants share the same filtering split, with Filtering-only removing the selected samples.

Variant	ASR ↓	ACC ↑	Retained data ↑
No defense	98.61 ± 0.25	$\mathbf{94.73 \pm 0.11}$	**100**%
Filtering-only	$\mathbf{2.73 \pm 0.41}$	90.12 ± 0.36	91.94%
Relabeling-only	31.84 ± 1.27	93.88 ± 0.18	**100**%
Reconstruction-only	58.21 ± 2.06	93.26 ± 0.24	**100**%
Full SCI	3.45 ± 0.63	94.22 ± 0.15	**100**%

Table 3: Backdoor mitigation and clean accuracy on ImageNet-100 with ResNet-50.

	BadNet (1%)		Blended (0.5%)		WaNet (4%)		ISSBA (2%)		Color-backdoor (5%)	
	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑
No defense	100	92.61	98.47	93.13	100	92.34	99.71	93.31	99.16	91.37
Spectre	40.16	92.19	9.34	92.79	98.63	89.27	12.67	92.74	90.31	88.85
AC	79.28	80.47	87.23	82.41	89.87	79.34	81.58	81.81	89.65	87.36
STRIP	78.60	82.32	94.21	87.23	96.14	90.34	30.54	91.41	97.36	83.68
CT	5.17	92.06	27.85	88.58	4.64	90.89	**2.74**	92.69	15.34	90.04
ASSET	**2.68**	**92.74**	**6.14**	92.73	**3.07**	92.42	7.11	93.03	**9.34**	91.16
ABL	88.61	91.50	91.63	89.17	94.34	90.47	40.84	88.31	94.92	89.85
SCI	12.42	92.46	10.64	**93.42**	8.64	**92.63**	6.61	**93.38**	11.34	**92.14**

The ablation in Table 2 isolates the contribution of correction by keeping the filtering split fixed. Relabeling and reconstruction each leave substantial backdoor behavior when applied alone, but combining them sharply reduces ASR while preserving clean accuracy. This supports the complementary roles of correcting labels and modifying inputs. Compared with removing the same selected samples, full SCI achieves higher clean accuracy while retaining every training record, with a small increase in ASR. This trade-off demonstrates the value of recovering useful training information through sample correction.

We also examine whether this retention trade-off persists across attacks. In a separate five-attack operating-point study, SCI selects 98.6%–100% of poisoned samples with FPRs of 7.6%–10.4%. Using the same selected set for removal and repair, repair increases ACC under every attack and also leaves higher ASR. These observations extend the value of retaining corrected records beyond the single component ablation. Appendix G provides the individual filtering statistics and retention results.

Figure 6 pairs suspicious inputs with their reconstructions. The reconstructions retain recognizable object shapes while changing local patterns and image texture. Some visible artifacts remain, so visual appearance alone cannot establish that a backdoor has been removed. The examples illustrate the input changes permitted by semantic guidance, whose downstream value is tested by the component and SPD experiments.

4.4 Backdoor Mitigation Across Datasets

ImageNet-100. The ImageNet-100 evaluation in Table 3 tests SCI on a larger classification task with ResNet-50. SCI substantially reduces ASR across all five attacks while keeping clean accuracy close to or above the corresponding undefended model. The reductions under WaNet and ISSBA extend this evidence to geometric warping and sample-specific triggers, supporting the applicability of the correction pipeline across different trigger constructions.

Among the evaluated defenses, SCI achieves the highest clean accuracy in four settings, although ASSET generally attains lower ASR. This trade-off highlights SCI's ability to preserve task performance while substantially mitigating



Figure 6: Suspicious inputs and their reconstructions. The examples show local changes while retaining recognizable object structure.

backdoor behavior. Together with the controlled component ablation, these results support correcting and retaining selected samples as a way to preserve useful training information during defense.

Recent defenses. For standard attacks, the comparison in Table 4 indicates that SCI suppresses backdoor behavior while maintaining high clean accuracy on both datasets. SCI retains accuracy close to the undefended model on CIFAR-10 and exceeds it on ImageNet-30. Although VIBE achieves better average ACC and ASR, SCI also provides effective mitigation, supporting the applicability of its correction pipeline beyond the component ablation.

SCI's comparative advantage becomes clearer under SPD, which combines pixel-domain and frequency-domain backdoors. SCI achieves the lowest ASR on both datasets while keeping clean accuracy close to VIBE. This change in the ASR ranking indicates that strong performance on standard attacks does not necessarily translate into the strongest mitigation of a compound attack. The results support SCI's value under SPD and motivate the following analysis of whether input reconstruction provides additional protection beyond label correction.

Table 4: Comparison with recent defenses at 10% poisoning.

		Standard attacks		SPD	
Dataset	Defense	ACC ↑	ASR ↓	ACC ↑	ASR ↓
CIFAR-10	No defense	94.17	99.47	**94.4**	99.9
	AIBD	85.20	0.43	84.8	8.5
	VIBE	**94.43**	**0.27**	94.1	4.0
	SCI	94.10	0.67	93.9	**1.7**
ImageNet-30	No defense	90.83	99.97	94.6	99.8
	AIBD	93.63	0.83	93.0	7.8
	VIBE	**96.87**	**0.10**	**96.4**	4.3
	SCI	96.30	0.47	96.0	**1.9**

Correction under SPD. To examine how SCI handles SPD, the stage analysis in Table 5 connects filtering quality with downstream attack success. Filtering captures most poisoned samples on both datasets, allowing subsequent correction to address most of the poisoned training data. The selected subset also contains clean samples, reinforcing the need to preserve useful content during correction without assuming perfect separation.

Despite this high recall, relabeling alone leaves residual backdoor behavior. Adding reconstruction substantially reduces ASR on both datasets, supporting the value of correcting inputs alongside their assigned labels. This consistent improvement provides evidence that input reconstruction contributes beyond label correction, helping explain SCI's effectiveness under SPD.

The profiles in Figure 7 reveal a marked difference between clean and poisoned gradients under Color-backdoor. The poisoned-sample profiles have larger absolute values across most displayed coordinates, and this difference persists even at low poisoning rates. The uneven distribution of these values indicates that the gradient response varies across model coordinates, providing a more detailed view of how poisoned and clean samples affect optimization.

Table 5: Filtering quality and the contribution of reconstruction under SPD at 10% poisoning.

Dataset	TPR ↑	FPR ↓	Relabel-only ASR ↓	Full SCI ASR ↓
CIFAR-10	**97.1**	8.2	**9.8**	**1.7**
ImageNet-30	95.4	**7.5**	11.6	1.9

These magnitude profiles complement the directional alignment analysis in Table 1. Larger absolute values alone do not establish faster learning, since the resulting updates also depend on the poisoning rate and whether gradients

reinforce or oppose one another. Figure 7 documents differences in gradient strength, while the matched comparison in Table 1 tests the additional alignment associated with trigger features beyond shared labels. Together, they provide complementary evidence for examining the proposed early-learning mechanism.

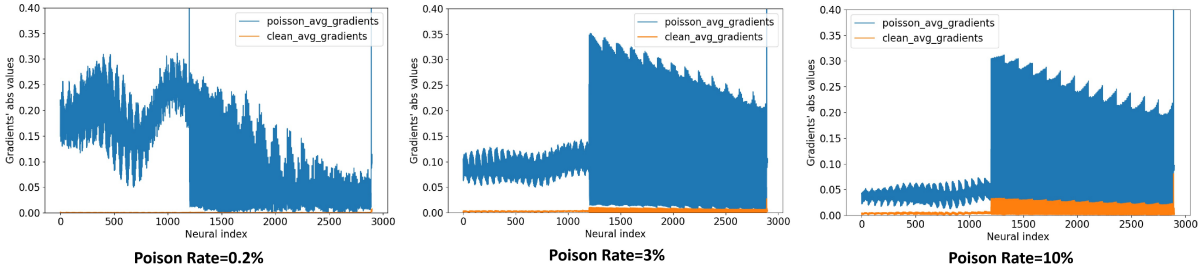


Figure 7: Absolute gradient-coordinate profiles of clean and poisoned samples under Color-backdoor. Panels compare different poisoning rates with the learning rate fixed at 0.001.

4.5 Adaptive Attack and Filtering Limitations

The adaptive evaluation in Table 6 tests SCI's dependence on its filtering signals. The adaptive attack reduces gradient alignment and target-class confidence in minimalist models while retaining high attack success before defense. The accompanying decline in recall indicates that an effective backdoor need not produce a signal that these models can readily learn. Because both signals are targeted jointly, this experiment does not isolate their individual contributions.

Table 6: Adaptive Blended evaluation on CIFAR-10 targeting gradient alignment and target-class confidence in minimalist models. Pre-ASR and Post-ASR measure attack success before and after SCI.

Attack	Align ↓	Mini Fit ↓	Pre-ASR ↓	TPR ↑	Post-ASR ↓	ACC ↑
Standard Blended	0.59	0.92	99.2	**99.1**	**6.7**	**93.1**
Adaptive Blended	**0.26**	**0.46**	**94.8**	63.5	24.1	92.8

The adaptive attack remains effective before defense while reducing both filtering signals and poisoned-sample recall. Its higher post-defense ASR shows that correction cannot compensate fully for missed poisons. These poisons remain in $\mathcal{R}$ and can influence the relabeling model before entering final training. SCI still reduces adaptive ASR from 94.8% to 24.1%, but this residual attack success is substantial. The experiment therefore identifies a concrete limitation of the filtering design. Because the attack changes alignment and fitting jointly, it does not isolate their individual effects.

Table 7: Poisoned-sample recall on ImageNet-100 under BadNet and Blended.

Attack	Poison Rate	Spectre	AC	STRIP	CT	ASSET	ABL	**SCI**
BadNet	5%	91.64 (0.8)	91.72 (0.4)	78.08 (4.6)	**100(0)**	**100(0)**	85.24 (1.1)	**100(0)**
	3%	95.13 (0.2)	87.53 (0.6)	79.33 (5.4)	**100(0)**	**100(0)**	84.67 (1.2)	**100(0)**
	1%	97.06 (1.1)	76.40 (3.6)	58.04 (4.3)	**100(0)**	**100(0)**	75.23 (1.7)	**100(0)**
	0.5%	91.26 (0.1)	62.40 (5.3)	37.80 (5.6)	**100(0)**	99.80 (0.2)	66.84 (2.3)	**100(0)**
Blended	5%	89.08 (0.3)	66.88 (1.1)	37.52 (5.8)	**100(0)**	99.60 (0.2)	93.76 (2.1)	**100(0)**
	3%	93.47 (1.4)	65.53 (3.8)	25.47 (4.7)	**100(0)**	**100(0)**	94.33 (2.5)	**100(0)**
	1%	99.20 (0.4)	45.40 (3.0)	27.60 (3.1)	98.80 (0.6)	98.60 (1.1)	85.00 (0.7)	**100(0)**
	0.5%	97.60 (0.2)	40.80 (1.3)	36.00 (2.2)	97.40 (0.7)	98.40 (0.6)	75.60 (0.4)	**100(0)**

4.6 Filtering Recall Across Poisoning Rates

Table 7 reports recall across poisoning rates for BadNet and Blended. SCI retains high recall throughout these fixed-trigger settings. Other defenses also reach full recall at several operating points, so this comparison establishes the effectiveness of SCI's candidate construction without implying a unique detection advantage. Its practical role is to place poisoned records in the subset available for correction. False-positive behavior is assessed separately below.

5 Conclusion and Future Work

We presented SCI, a backdoor defense that produces a corrected training set without auxiliary clean data while retaining all training samples. Our analysis explains how shared target labels and recurrent trigger features can jointly produce coherent gradients during early training. Class-matched experiments support additional trigger-associated alignment beyond the shared-label effect. Guided by this analysis, SCI uses low-capacity models to construct a suspicious subset and corrects it through semantic relabeling and input reconstruction. Controlled ablations demonstrate the complementary roles of these correction stages and quantify the clean-accuracy benefit of repair over removal. Experiments on CIFAR-10 and ImageNet subsets show effective backdoor mitigation with full sample retention, while adaptive attacks reveal the limits of filtering. Future work will strengthen robustness to adaptive attacks and explore correction mechanisms for large language models.

6 AI USE STATEMENT

AI tools were used for language editing and writing assistance during manuscript preparation. The authors take responsibility for the final manuscript, experimental methodology, and reported results.

REFERENCES

Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, Avani Gupta, and PJ Narayanan. Prototype guided backdoor defense via activation space manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2195–2205, 2025.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24239–24250, 2024.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Tuo Chen, Jie Gui, Minjing Dong, Ju Jia, Lanting Fang, and Jian Liu. Backdooring self-supervised contrastive learning by noisy alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3684–3693, 2025.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Yiming Chen, Haiwei Wu, and Jiantao Zhou. Progressive poisoned data isolation for training-time backdoor defense. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11425–11433, 2024.

Siyuan Cheng, Guanhong Tao, Yingqi Liu, Guangyu Shen, Shengwei An, Shiwei Feng, Xiangzhe Xu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Lotus: Evasive and resilient backdoor attacks through sub-partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24798–24809, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pp. 113–125, 2019.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Jiyang Guan, Jian Liang, and Ran He. Backdoor defense via test-time detecting and repairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24564–24573, 2024.

Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pp. 4129–4139. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Alvi Md Ishmam and Christopher Thomas. Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24820–24830, 2024.

Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8133–8142, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021a.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021b.

Zhaozheng Li, Jiahe Lan, Zheng Yan, and Erol Gelenbe. Backdoor attacks and defense mechanisms in federated learning: A survey. *Information Fusion*, pp. 103248, 2025.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24645–24654, 2024.

Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The" beatrix"resurrections: Robust backdoor detection via gram matrices. *Network and Distributed System Security Symposium*, 2023.

Abdullah Arafat Miah and Yu Bi. Lite-bd: A lightweight black-box backdoor defense via reviving multi-stage image transformations. *arXiv preprint arXiv:2602.07197*, 2026.

Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Soumyadeep Pal, Yuguang Yao, Ren Wang, Bingquan Shen, and Sijia Liu. Backdoor secrets unveiled: Identifying backdoor data with optimized scaled prediction consistency. In *International conference on learning representations*, volume 2024, pp. 14061–14083, 2024.

Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. Asset: Robust backdoor data detection across a multiplicity of deep learning paradigms. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2725–2742, 2023.

Xiangyu Qi, Tinghao Xie, Jiachen T Wang, Tong Wu, Saeed Mahloujifar, and Prateek Mittal. Towards a proactive ml approach for detecting backdoor poison samples. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1685–1702, 2023.

Ivan Sabolić, Matej Grcić, and Siniša Šegvić. Seal your backdoor with variational defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 752–764, 2025.

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, et al. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of dnns for robust backdoor contamination detection. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1541–1558, 2021.

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 *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723. IEEE, 2019.

Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Mitigating backdoor attack by injecting proactive defensive backdoor. *Advances in Neural Information Processing Systems*, 37:80674–80705, 2024.

Shaokui Wei, Jiayin Liu, and Hongyuan Zha. Backdoor mitigation by distance-driven detoxification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4465–4474, 2025.

Binyan Xu, Fan Yang, Xilin Dai, Di Tang, and Kehuan Zhang. Clip-guided backdoor defense through entropy-based poisoned dataset separation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7415–7423, 2025.

Nan Yan, Yuqing Li, Xiong Wang, Jing Chen, Kun He, and Bo Li. Embedx:embedding-based cross-trigger backdoor attack against large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 241–257, 2025.

Yanxin Yang, Chentao Jia, DengKe Yan, Ming Hu, Tianlin Li, Xiaofei Xie, Xian Wei, and Mingsong Chen. Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification. *Advances in Neural Information Processing Systems*, 37:121236–121264, 2024.

Jia-Li Yin, Weijian Wang, Wei Lin, Ximeng Liu, et al. Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9508–9516, 2025.

Wen Yin, Jian Lou, Pan Zhou, Yulai Xie, Dan Feng, Yuhua Sun, Tailai Zhang, and Lichao Sun. Physical backdoor: Towards temperature-based backdoor attacks in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12733–12743, 2024.

Shunjie Yuan, Xinghua Li, Xuelin Cao, Haiyan Zhang, Mengyao Zhu, and Robert H Deng. Spd: Shallow backdoor protecting deep backdoor against backdoor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4029–4038, 2025.

Jingyi Zheng, Tianyi Hu, Tianshuo Cong, and Xinlei He. Cl-attack: Textual backdoor attacks via cross-lingual triggers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26427–26435, 2025.

Jiachen Zhou, Peizhuo Lv, Yibing Lan, Guozhu Meng, Kai Chen, and Hualong Ma. Dataelixir: Purifying poisoned dataset to mitigate backdoor attacks via diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21850–21858, 2024.

A Related Work

We compare SCI with representative backdoor defenses across several dimensions.

Table 8: Comparison of representative backdoor defenses. "W/o Clean" indicates that no auxiliary clean dataset is required. "No Ext." indicates that the method does not rely on an external foundation model or reference model. "Repair/Retain" indicates whether suspicious training samples are corrected and retained rather than removed. "Prior" indicates whether the method operates before training the final task model. "Theory" indicates whether the method provides a theoretical account of backdoor early learning.

Method	W/o Clean	No Ext.	Repair/Retain	Prior	Theory	Main Defense Type
ABL Li et al. (2021a)	●	●	○	○	○	Training dynamics
ASSET Pan et al. (2023)	○	●	○	●	○	Poison detection
CT Qi et al. (2023)	○	●	○	●	○	Confusion training
Beatrix Ma et al. (2023)	●	●	○	●	○	Activation statistics
Spectre Hayase et al. (2021)	○	●	○	○	○	Spectral detection
STRIP Gao et al. (2019)	●	●	○	○	○	Inference detection
Neural Cleanse Wang et al. (2019)	●	●	○	○	○	Trigger inversion
SCAn Tang et al. (2021)	○	●	○	●	○	Statistical detection
PDB Wei et al. (2024)	○	●	○	●	○	Proactive defense
TTBD Guan et al. (2024)	●	●	○	○	○	Test-time repair
SampDetox Yang et al. (2024)	●	●	○	○	○	Sample detoxification
CGD Xu et al. (2025)	●	○	△	●	○	CLIP-guided defense
VIBE Sabolić et al. (2025)	●	●	△	●	○	Variational relabeling
PGBD Amula et al. (2025)	○	●	○	○	○	Post-hoc sanitization
D3 Wei et al. (2025)	○	●	○	○	○	Post-training detoxification
SCI (ours)	●	●	●	●	●	Theory-guided self-correction

●: supported; ○: not supported; △: partially or conditionally supported. CGD is marked as partial for "Repair/Retain" because it reuses poisoned data with CLIP-guided correction, but does not explicitly reconstruct suspicious inputs. VIBE is marked as partial because it corrects latent labels and trains on poisoned data, but does not remove trigger features from suspicious samples. For test-time and post-hoc defenses, "Repair/Retain" refers to training-set integrity rather than input-level test-time repair.

B Notation Description

Table 9 summarizes the main symbols used in this paper.

C Detailed Analysis of Backdoor Early Learning

This appendix expands the mechanistic analysis in the main text. The argument concerns early optimization in a specified model. It distinguishes shared-label effects from trigger-related feature concentration and makes the conditions for coherent parameter updates explicit. The resulting explanation motivates SCI's filtering design, while filtering performance remains an empirical property of the chosen model and attack.

Table 9: Notation summary.

Symbol	Description
$\mathcal{D}$	Original training set, which may contain both clean and poisoned samples.
$\mathcal{D}_c$	Clean subset of the training data.
$\mathcal{D}_p$	Poisoned subset of the training data.
(x, y)	A clean sample x with its ground-truth label y.
(z, y_t)	A poisoned sample z assigned to the attacker-specified target label y_t.
f_θ	Classification model parameterized by θ.
$q_\theta(y\|u)$	Predicted probability of label y for input u under model f_θ.
$\ell(u, y; \theta)$	Training loss of sample (u, y), typically cross-entropy loss.
$\nabla_\theta \ell(u, y; \theta)$	Parameter gradient contributed by sample (u, y).
$J_\theta(u)$	Jacobian of the model output with respect to parameters θ for input u.
$H(\cdot)$	Shannon entropy of a distribution.
$\mathcal{S}$	Suspicious subset constructed by the filtering stage. It aims to enrich poisoned samples but may also contain clean samples.
$\mathcal{R}$	Reliable subset after filtering, defined as $\mathcal{R} = \mathcal{D} \setminus \mathcal{S}$. It has a lower poison concentration when TPR exceeds FPR than the original training set but is not assumed to be perfectly clean.
M_{rel}	Relabeling model trained on the reliable subset $\mathcal{R}$.
$\widetilde{y}_u$	Semantic label assigned to suspicious sample u by M_{rel}.
$\widetilde{\mathcal{S}}$	Relabeled suspicious subset, defined as $\{(u, \widetilde{y}_u) \mid u \in \mathcal{S}\}$.
M_{rec}	Reconstruction network used to generate corrected suspicious samples.
$\widehat{u}$	Reconstructed sample, defined as $\widehat{u} = M_{\text{rec}}(u)$.
$\phi_{\text{rel}}(u)$	Feature map of input u extracted from M_{rel}, typically from the layer before global max pooling.
$\mathcal{K}(u)$	Indices of the top-K activated feature channels in $\phi_{\text{rel}}(u)$.
$\phi_{\text{rel}}(u)_{\mathcal{K}(u)}$	Top-K semantic feature anchors selected from M_{rel}.
$\widehat{\mathcal{S}}$	Reconstructed suspicious subset, defined as $\{(M_{\text{rec}}(u), \widetilde{y}_u) \mid (u, \widetilde{y}_u) \in \widetilde{\mathcal{S}}\}$.
$\mathcal{D}^*$	Final purified training set, defined as $\mathcal{D}^* = \mathcal{R} \cup \widehat{\mathcal{S}}$.
$\mathcal{L}_{\text{rec}}$	Reconstruction objective used to train M_{rec}.
$\mathcal{L}_{\text{label}}$	Label consistency loss for preserving the relabeled semantic class after reconstruction.
$\mathcal{L}_{\text{feat}}$	Semantic feature consistency loss based on top-K feature anchors from M_{rel}.
$\mathcal{L}_{\text{SSIM}}$	Structural similarity loss for preserving coarse visual structure.
$\mathcal{L}_{\text{TV}}$	Total variation loss for reducing high-frequency reconstruction artifacts.
α_i	Weight coefficient for the i-th term in the reconstruction objective.

C.1 Gradient Factorization and Label Concentration

Let $h_\theta(u) \in \mathbb{R}^C$ be the logits and let $q_\theta(u) = \text{softmax}(h_\theta(u))$. For cross-entropy,

$$\ell_\theta(u, y) = -h_\theta(u)_y + \log \sum_{c=1}^{C} \exp\big(h_\theta(u)_c\big). \tag{16}$$

Differentiating with respect to the logits gives $\nabla_h \ell_\theta(u, y) = q_\theta(u) - e_y$. The chain rule therefore yields

$$g_\theta(u, y) = \nabla_\theta \ell_\theta(u, y) = J_\theta(u)^\top \big(q_\theta(u) - e_y\big), \qquad J_\theta(u) = \frac{\partial h_\theta(u)}{\partial \theta}. \tag{17}$$

The Jacobian in this expression is the logit Jacobian throughout the analysis.

Write the early prediction as $q_\theta(u) = q_0 + \delta_u$, where q_0 is a common reference probability vector. For the poisoned group assigned to y_t, the average logit gradient is

$$\frac{1}{|\mathcal{D}_p|} \sum_{(z, y_t) \in \mathcal{D}_p} \nabla_h \ell_\theta(z, y_t) = q_0 - e_{y_t} + \bar{\delta}_p. \tag{18}$$

For the clean group, let π_c be the empirical proportion of class c and define $\pi = \sum_{c=1}^{C} \pi_c e_c$. Its average logit gradient is

$$\frac{1}{|\mathcal{D}_c|} \sum_{(x, y) \in \mathcal{D}_c} \nabla_h \ell_\theta(x, y) = q_0 - \pi + \bar{\delta}_c. \tag{19}$$

Here, $\bar{\delta}_p$ and $\bar{\delta}_c$ are the corresponding mean prediction deviations. If q_0 is close to π and the deviations are small, averaging across clean labels can reduce the logit gradient. Samples assigned to a common target keep a common label-dependent direction.

This concentration does not require larger individual logit gradients. At uniform predictions, $q_0 = \mathbf{1}/C$, every label satisfies

$$\|q_0 - e_y\|_2^2 = 1 - \frac{1}{C}. \tag{20}$$

Nor does cancellation of average logit gradients imply cancellation of parameter gradients. To see this, let $\bar{J}_c = \mathbb{E}_{\mathcal{D}_c}[J_\theta(x)]$, where the expectation is an empirical average. An exact expansion gives

$$\begin{aligned} \bar{g}_c = \bar{J}_c^\top (q_0 - \pi) - \mathbb{E}_{\mathcal{D}_c}\big[(J_\theta(x) - \bar{J}_c)^\top (e_y - \pi)\big] \\ + \mathbb{E}_{\mathcal{D}_c}\big[J_\theta(x)^\top \delta_x\big]. \end{aligned} \tag{21}$$

Even if $q_0 = \pi$, dependence between the label and the Jacobian can leave a nonzero parameter update. This is why the feature-dependent mapping must be analyzed alongside the target-label effect.

C.2 A Shared Trigger Component and Its Residual

Suppose the current model expresses a common trigger-related component in the poisoned-sample Jacobians. For a fixed parameter vector θ, write

$$J_\theta(z) = J_{\text{trig}} + \Delta_z, \qquad a_t = q_0 - e_{y_t}. \tag{22}$$

The existence of a useful common component is a modeling condition. Its relevance depends on the trigger and on the features represented by the model at this stage of training. The residual Δ_z contains the remaining input-dependent variation. Substituting this decomposition into the gradient gives the exact expression

$$g_\theta(z, y_t) = (J_{\text{trig}} + \Delta_z)^\top (a_t + \delta_z) = v + r_z, \tag{23}$$

$$v = J_{\text{trig}}^\top a_t, \qquad r_z = \Delta_z^\top a_t + J_\theta(z)^\top \delta_z. \tag{24}$$

In particular, the residual satisfies

$$\|r_z\|_2 \le \|\Delta_z\|_{\text{op}} \|a_t\|_2 + \|J_\theta(z)\|_{\text{op}} \|\delta_z\|_2, \tag{25}$$

where $\|\cdot\|_{\text{op}}$ denotes the spectral operator norm. The bound identifies two sources of variation that can weaken a common update direction. A shared feature contribution with sufficiently small residual feature and prediction terms can therefore support a common update direction. A shared Jacobian component is ineffective for this argument if $J_{\text{trig}}^\top a_t = 0$.

The term low-entropy structure describes concentration in the assigned labels and in trigger-related features. The quantitative analysis uses the shared gradient component and its residual. It does not infer a gradient bound from a measured Shannon entropy or assume that all clean images have more complex visual structure than triggered images.

A sufficient condition for positive alignment. Assume $v \neq 0$ and $\|r_z\|_2 \le \varepsilon \|v\|_2$ for all samples in the group. For two such samples,

$$g_\theta(z, y_t)^\top g_\theta(z', y_t) = \|v\|_2^2 + v^\top r_z + v^\top r_{z'} + r_z^\top r_{z'} \tag{26}$$

$$\ge \|v\|_2^2 - \|v\|_2 \|r_z\|_2 - \|v\|_2 \|r_{z'}\|_2 - \|r_z\|_2 \|r_{z'}\|_2 \tag{27}$$

$$\ge (1 - 2\varepsilon - \varepsilon^2) \|v\|_2^2. \tag{28}$$

When $0 \le \varepsilon < \sqrt{2} - 1$, the lower bound is positive. The triangle inequality also gives

$$(1 - \varepsilon)\|v\|_2 \le \|g_\theta(z, y_t)\|_2 \le (1 + \varepsilon)\|v\|_2. \tag{29}$$

Dividing the positive inner-product bound by an upper bound on the product of the norms yields

$$\cos\big(g_\theta(z, y_t), g_\theta(z', y_t)\big) \ge \frac{1 - 2\varepsilon - \varepsilon^2}{(1 + \varepsilon)^2}. \tag{30}$$

For the mean gradient, the reverse triangle inequality gives

$$\|\bar{g}_p\|_2 = \|v + \bar{r}_p\|_2 \ge (1 - \varepsilon)\|v\|_2. \tag{31}$$

These inequalities are sufficient conditions for coherence in the chosen poisoned group. They do not establish that the residual bound holds for every attack or that the group is more aligned than every clean group. The threshold on ε is an analytical sufficient condition, not an estimated operational boundary for SCI.

C.3 Separating Trigger Effects with Matched Labels

To distinguish the trigger contribution from the shared-label effect, consider a source input x_i and its triggered version z_i at the same model parameters. Assign both inputs the target label y_t. Define the gradients

$$b_i = g_\theta(x_i, y_t), \qquad p_i = g_\theta(z_i, y_t). \tag{32}$$

Using the same reference vector q_0, their exact difference is

$$\begin{aligned} p_i - b_i = &\big(J_\theta(z_i) - J_\theta(x_i)\big)^\top a_t \\ &+ J_\theta(z_i)^\top \delta_{z_i} - J_\theta(x_i)^\top \delta_{x_i}. \end{aligned} \tag{33}$$

Both examples have the same label-dependent reference signal. Their difference records changes induced by the trigger in the Jacobian and in the prediction. When prediction deviations are small and the Jacobian differences share a component, the triggered gradients can acquire a common additional direction.

An additional common direction does not automatically increase every pairwise cosine. If $p_i = b_i + d_i$, then

$$p_i^\top p_j - b_i^\top b_j = b_i^\top d_j + d_i^\top b_j + d_i^\top d_j. \tag{34}$$

The first two terms can have either sign, and normalization also affects the cosine. Matched controls therefore test the additional alignment empirically. Our matched comparison uses the same source images and target labels for the Label-only and Poisoned groups. The triggers are the input difference between the two groups.

C.4 Aggregate Updates and Loss Changes

Let $m = |\mathcal{D}_p|$ and write g_i for a poisoned-sample gradient. The squared norm of their mean is

$$\begin{aligned} \|\bar{g}_p\|_2^2 = &\frac{1}{m^2} \sum_{i=1}^{m} \|g_i\|_2^2 \\ &+ \frac{2}{m^2} \sum_{i<j} \|g_i\|_2 \|g_j\|_2 \cos(g_i, g_j). \end{aligned} \tag{35}$$

For fixed individual norms, larger pairwise agreement increases the aggregate norm. In actual training, both the directions and norms can change. The expression therefore separates the contribution of alignment from the size of individual gradients.

For uniform sampling over the training records, the full gradient is

$$g = (1-\rho)\bar{g}_c + \rho \bar{g}_p. \tag{36}$$

At a differentiable point, the update $\theta^+ = \theta - \eta g$ gives

$$\mathcal{L}_p(\theta^+) - \mathcal{L}_p(\theta) = -\eta A_p + o(\eta), \tag{37}$$

$$A_p = \rho \|\bar{g}_p\|_2^2 + (1-\rho)\bar{g}_p^\top \bar{g}_c. \tag{38}$$

If $\mathcal{L}_p$ has a β_p-Lipschitz gradient along the update segment, the descent lemma makes the remainder explicit

$$\mathcal{L}_p(\theta^+) \le \mathcal{L}_p(\theta) - \eta A_p + \frac{\beta_p \eta^2}{2} \|g\|_2^2. \tag{39}$$

For $\beta_p > 0$ and $A_p > 0$, a step satisfying $0 < \eta < 2A_p/(\beta_p \|g\|_2^2)$ decreases the poisoned-group loss. Thus, coherent poisoned gradients can support early loss reduction, but their contribution is weighted by the poisoning rate. Opposing clean gradients can weaken or reverse the effect.

The corresponding coefficient for the clean-group loss is

$$A_c = (1-\rho)\|\bar{g}_c\|_2^2 + \rho \bar{g}_p^\top \bar{g}_c. \tag{40}$$

A larger first-order decrease in the poisoned-group loss requires $A_p > A_c$, or equivalently

$$\rho \|\bar{g}_p\|_2^2 - (1-\rho)\|\bar{g}_c\|_2^2 + (1-2\rho)\bar{g}_p^\top \bar{g}_c > 0. \tag{41}$$

Gradient coherence alone does not establish this inequality. Neither the local loss expansion nor this comparison guarantees a particular ordering of classification accuracy over the full training trajectory. They explain how alignment can contribute to early learning when the other optimization terms permit it.

With an unbiased mini-batch gradient $\widehat{g}$, the first-order expected loss change is still $-\eta A_p$. If the smoothness bound holds along the sampled update segments and the second moment is finite, the second-order bound instead involves $\mathbb{E}\|\widehat{g}\|_2^2$. This dependence captures the additional effect of mini-batch variation.

D Filtering Dynamics and the Remaining Subset

D.1 Conditional Enrichment Across Rounds

The candidate sets satisfy $\mathcal{S}_k \subseteq \mathcal{S}_{k-1}$. For nonempty clean and poisoned portions of $\mathcal{S}_{k-1}$, define the empirical retention fractions

$$r_{p,k} = \frac{|\mathcal{S}_k \cap \mathcal{D}_p|}{|\mathcal{S}_{k-1} \cap \mathcal{D}_p|}, \qquad r_{c,k} = \frac{|\mathcal{S}_k \cap \mathcal{D}_c|}{|\mathcal{S}_{k-1} \cap \mathcal{D}_c|}. \tag{42}$$

Let ρ_{k-1} be the poisoned fraction in $\mathcal{S}_{k-1}$. Whenever $\mathcal{S}_k$ is nonempty, counting the retained records gives

$$\rho_k = \frac{\rho_{k-1} r_{p,k}}{\rho_{k-1} r_{p,k} + (1-\rho_{k-1}) r_{c,k}}. \tag{43}$$

For $0 < \rho_{k-1} < 1$, enrichment occurs exactly when $r_{p,k} > r_{c,k}$. The identity does not assume independence across rounds. It also exposes the cost of repeated filtering, since the cumulative rates satisfy

$$\mathrm{TPR}_T = \prod_{k=1}^{T} r_{p,k}, \qquad \mathrm{FPR}_T = \prod_{k=1}^{T} r_{c,k}, \tag{44}$$

when the corresponding conditional fractions are defined. If one portion becomes empty, its cumulative retention is zero in later rounds. An additional round can improve the poisoned fraction of the selected set while reducing its overall poisoned-sample recall.

D.2 Contamination of the Remaining Data

Let $t = \mathrm{TPR}$ and $f = \mathrm{FPR}$ for the final selected set. The selection fraction is

$$\frac{|\mathcal{S}|}{|\mathcal{D}|} = \rho t + (1-\rho) f. \tag{45}$$

If $\mathcal{R}$ is nonempty, its poisoned fraction is

$$\rho_{\mathcal{R}} = \frac{\rho(1-t)}{\rho(1-t) + (1-\rho)(1-f)}. \tag{46}$$

For $0 < \rho < 1$, this fraction is below ρ exactly when $t > f$. A lower concentration is therefore conditional on the filtering operating point. It does not make the relabeling model immune to the poisons that remain. Since reconstruction is applied only to $\mathcal{S}$, such missed samples also remain in the final dataset.

D.3 Filtering Configuration

Every round initializes its minimalist model from scratch and uses the labels supplied in the candidate set. The first round uses all available training records. Later rounds train only on the preceding candidate set. The CIFAR-10 experiments use $T = 2$, with approximately ten training epochs per round. The first model has one convolutional layer followed by a linear output layer, with 9,038 parameters. Its layer sizes are

Layer	Output shape	Parameters
Input	$32 \times 32 \times 3$	0
Convolution	$30 \times 30 \times 1$	28
Flatten	900	0
Linear output	10	9,010

The number of rounds is denoted by T to distinguish it from the number of reconstruction anchors K. The mechanism is most useful when the chosen filters express the recurring trigger feature. Clean shortcuts can increase false positives, while a trigger that is difficult for these filters to learn can reduce recall even if it succeeds on the final model.

E Details of the Reconstruction Objective

For a relabeled input $(u, \widetilde{y}_u)$, the reconstruction network produces $\widehat{u} = M_{\mathrm{rec}}(u)$. The relabeling model remains fixed during reconstruction. The assigned label and feature targets are computed from the original input. The same selected channel indices $\mathcal{K}(u)$ are applied to the original and reconstructed feature maps.

Semantic constraints. The label and feature terms are

$$\mathcal{L}_{\text{label}}(u) = -\log q_{\text{rel}}(\widetilde{y}_u \mid \widehat{u}), \tag{47}$$

$$\mathcal{L}_{\text{feat}}(u) = \left\| \phi_{\text{rel}}(u)_{\mathcal{K}(u)} - \phi_{\text{rel}}(\widehat{u})_{\mathcal{K}(u)} \right\|_F^2. \tag{48}$$

The Frobenius norm is equivalent to the Euclidean norm after vectorizing the selected feature tensor. Only M_{rec} is optimized. Gradients are propagated through the fixed relabeling model to the reconstructed input. The targets from u and its top-K selection remain fixed. These constraints preserve the class evidence supplied by the relabeling model. They can also preserve unwanted features if that model relies on a residual trigger.

Structural regularization. The structural similarity term is

$$\mathcal{L}_{\text{SSIM}}(u) = 1 - \text{SSIM}(u, \widehat{u}). \tag{49}$$

It encourages the reconstruction to retain coarse image structure. For an image with d_{in} channels and spatial dimensions $H \times W$, the isotropic total variation term can be written as

$$\mathcal{L}_{\text{TV}}(u) = \sum_{c=1}^{d_{\text{in}}} \sum_{i=1}^{H} \sum_{j=1}^{W} \sqrt{(\Delta_h \widehat{u}_{c,i,j})^2 + (\Delta_w \widehat{u}_{c,i,j})^2}. \tag{50}$$

Here, $\Delta_h \widehat{u}_{c,i,j} = \widehat{u}_{c,i+1,j} - \widehat{u}_{c,i,j}$ for $i < H$, and $\Delta_w \widehat{u}_{c,i,j} = \widehat{u}_{c,i,j+1} - \widehat{u}_{c,i,j}$ for $j < W$. An unavailable forward difference is zero at the image boundary. This term discourages abrupt spatial changes in the reconstruction. The training objective averages the weighted sum of the four losses over $\widetilde{\mathcal{S}}$.

Role of feature anchors. The top-K channels identify features to preserve during correction. SCI does not infer that every unselected channel represents a trigger. The label and feature constraints protect selected semantic evidence while the reconstruction and structural penalties permit changes to other input details. The effectiveness of this process is assessed through downstream backdoor mitigation and clean accuracy. Retaining all records alone does not establish that every reconstructed sample preserves its original semantics.

The reconstruction study uses $K = 60$ for a feature tensor with 64 channels.

F Experimental Setup

Data and models. We use CIFAR-10 and two ImageNet subsets. CIFAR-10 contains 60,000 RGB images from ten classes at a resolution of 32×32. ImageNet-100 contains 100 classes, with images resized to 224×224. We use ResNet-50 as the final classifier on ImageNet-100. Comparisons with recent defenses additionally use ImageNet-30, following the benchmark choice of VIBE. For these comparisons, we use ResNet-18 on CIFAR-10 and a common backbone across methods on ImageNet-30.

Attacks and defenses. We evaluate BadNet Gu et al. (2019), Blended Chen et al. (2017) and WaNet Nguyen & Tran (2021). ISSBA Li et al. (2021b) and Color-backdoor Jiang et al. (2023) extend the evaluation to sample-specific and color-space triggers. SPD (Yuan et al., 2025) uses a shallow pixel-domain backdoor to conceal a deep frequency-domain backdoor. We also construct an adaptive Blended attack that targets SCI's filtering signal. The defense baselines include Spectre Hayase et al. (2021), AC Chen et al. (2018) and STRIP Gao et al. (2019). We further compare with CT Qi et al. (2023), ASSET Pan et al. (2023) and ABL Li et al. (2021a). Recent comparisons include AIBD (Yin et al., 2025) and VIBE (Sabolić et al., 2025).

Evaluation protocols. The recent-defense comparisons use a poisoning rate of 10% on CIFAR-10 and ImageNet-30. The mechanism study and component ablation use separate poisoning rates, which are specified in their respective experiments. SCI uses no auxiliary clean data. The default CIFAR-10 configuration uses two filtering rounds.

Evaluation metrics. We report ACC and ASR after training the final classifier. For filtering, TPR measures the fraction of poisoned training samples selected into $\mathcal{S}$. FPR measures the fraction of clean training samples selected into $\mathcal{S}$. Data retention is the fraction of original training samples used for final-model training, counting both unchanged and corrected samples. This metric quantifies sample retention, while ACC measures downstream classification utility.

G SUPPLEMENTARY EXPERIMENTS

This appendix provides the complete detection results and additional analyses of filtering and correction. The main text contains the matched gradient comparison and the component ablation, together with the end-to-end and adaptive evaluations needed to assess the central claims.

G.1 PROTOCOL FOR THE MATCHED GRADIENT COMPARISON

The gradient study uses CIFAR-10 with 5% poisoning. For each target class, Label-only and Poisoned contain the same non-target source images and receive the same assigned label. Only Poisoned receives the attack trigger. Clean-target contains genuine target-class examples, and Clean-all contains clean examples across classes. The groups have equal size. Per-sample parameter gradients are compared through pairwise cosine similarity over iterations 150–250. The measurements aggregate ten target classes and three random seeds. Table 1 reports every evaluated attack. The comparison tests whether the triggered group exhibits additional alignment after controlling the source images and assigned labels.

G.2 COMPLETE DETECTION RESULTS ON IMAGENET-100

Table 10 extends the comparison to spatial warping and sample-specific triggers, together with Color-backdoor. SCI maintains high recall across these constructions, showing that the selected set is not restricted to examples with a fixed visible patch. ASSET has higher recall for Color-backdoor, which also shows that SCI's advantage depends on the attack. These measurements support the filtering stage empirically. The matched gradient experiment in the main text provides the separate evidence for the proposed mechanism.

Table 10: Poisoned-sample recall on ImageNet-100 under WaNet, ISSBA and Color-backdoor.

Attack	Poison Rate	Spectre	AC	STRIP	CT	ASSET	ABL	**SCI**
WaNet	5%	65.16 (2.1)	30.56 (6.7)	29.20 (5.3)	98.76 (0.4)	90.16 (0.2)	64.08 (3.3)	100 (0)
	3%	89.53 (2.6)	45.33 (5.3)	31.33 (8.7)	97.60 (1.9)	87.07 (3.2)	56.33 (4.1)	100 (0)
	1%	90.00 (1.2)	41.43 (0.5)	33.60 (1.2)	94.00 (1.6)	86.80 (0.5)	60.40 (0.7)	99.40 (0.2)
	0.5%	85.40 (0.4)	28.87 (10.2)	34.00 (8.3)	91.00 (0.9)	84.40 (1.3)	54.00 (2.4)	99.20 (0.1)
ISSBA	5%	95.52 (0.4)	59.44 (1.2)	75.28 (0.8)	96.86 (1.6)	99.35 (0.2)	88.56 (2.1)	100 (0)
	3%	90.07 (1.3)	34.07 (2.1)	70.53 (1.4)	87.79 (2.3)	97.07 (0.8)	86.27 (3.7)	99.06 (0.2)
	1%	78.20 (2.9)	24.00 (10.8)	67.20 (1.2)	83.56 (3.7)	95.80 (1.8)	65.60 (8.9)	99.20 (0.2)
	0.5%	60.40 (0.3)	12.80 (5.6)	58.20 (3.8)	82.54 (5.9)	89.20 (2.1)	54.00 (10.2)	98.80 (0.1)
Color-backdoor	5%	91.28 (1.4)	77.56 (2.1)	35.12 (6.7)	98.04 (1.0)	100 (0)	81.12 (3.8)	98.60 (0.2)
	3%	89.07 (2.2)	74.27 (0.6)	32.93 (3.5)	97.33 (0.9)	99.00 (0.2)	79.47 (6.3)	97.80 (1.8)
	1%	87.60 (1.9)	71.80 (4.7)	30.80 (13.8)	97.00 (2.1)	98.80 (0.1)	77.00 (3.8)	96.40 (1.1)
	0.5%	85.60 (3.5)	69.20 (5.7)	28.40 (9.7)	96.40 (2.1)	98.40 (0.1)	75.60 (4.1)	97.80 (0.1)

G.3 FILTERING OPERATING POINTS AND DATA RETENTION

Recall must be interpreted together with the amount of clean data selected. Table 11 reports both rates and the size of $\mathcal{S}$ for five attack settings. For rates expressed as fractions, the selected proportion satisfies

$$\frac{|\mathcal{S}|}{|\mathcal{D}|} = \rho\,\mathrm{TPR} + (1-\rho)\,\mathrm{FPR}, \tag{51}$$

where ρ is the poisoning rate. The selected sets include a material number of clean records despite their high poisoned-sample recall. This explains why the effect of selection on final training must be evaluated through correction and retention, in addition to detection rates.

Table 12 compares removing and correcting the same selected records within each attack. Repair improves clean accuracy in every case, while removal leaves lower ASR. The mean differences computed from the individual entries are 1.50 percentage points in ACC and 1.80 points in ASR. The repeated direction across attacks supports the usefulness of retaining corrected records and quantifies the associated mitigation trade-off. The comparison includes both selected clean examples and poisoned examples, so it does not isolate the utility of either group.

Table 11: Filtering operating points for the five-attack retention study.

Attack	Poisoning rate	TPR ↑	FPR ↓	Selected data
BadNet	1.0	100.0	8.2	9.12
Blended	0.5	100.0	7.6	8.06
WaNet	4.0	99.8	10.4	13.98
ISSBA	2.0	99.3	9.1	10.90
Color-backdoor	5.0	98.6	8.6	13.10
Mean	–	99.54	8.78	11.03

Table 12: Removal and repair using the same suspicious subset within each attack.

	Remove $\mathcal{S}$		Full SCI	
Attack	ASR ↓	ACC ↑	ASR ↓	ACC ↑
BadNet	9.84	91.28	12.42	92.46
Blended	8.77	91.74	10.64	93.42
WaNet	6.92	90.86	8.64	92.63
ISSBA	5.78	91.93	6.61	93.38
Color-backdoor	9.36	90.71	11.34	92.14
Mean	8.13	91.30	9.93	92.81

G.4 Complete Comparisons with Recent Defenses

The recent-defense comparison uses 10% poisoning on both datasets. CIFAR-10 uses ResNet-18, while the ImageNet-30 comparison uses a common backbone across methods. Table 13 gives the individual results summarized in Table 4. VIBE has lower ASR than SCI on each of the three standard attacks on both datasets. SCI has lower ASR under SPD while retaining similar clean accuracy. The complete results show that the SPD benefit is specific to the evaluated attack and is not a uniform ranking across all conditions.

Table 13: Individual results for the recent-defense comparison at 10% poisoning.

Dataset	Defense	BadNet	Blended	WaNet	SPD
CIFAR-10	No defense	94.3 / 99.6	94.0 / 99.3	94.2 / 99.5	94.4 / 99.9
	AIBD	85.3 / 0.3	84.9 / 0.7	85.4 / 0.3	84.8 / 8.5
	VIBE	94.4 / 0.1	94.6 / 0.0	94.3 / 0.7	94.1 / 4.0
	SCI	94.2 / 0.5	94.1 / 0.7	94.0 / 0.8	93.9 / 1.7
ImageNet-30	No defense	95.3 / 100.0	83.7 / 99.9	93.5 / 100.0	94.6 / 99.8
	AIBD	93.9 / 0.5	93.4 / 0.9	93.6 / 1.1	93.0 / 7.8
	VIBE	96.7 / 0.1	96.8 / 0.1	97.1 / 0.1	96.4 / 4.3
	SCI	96.4 / 0.4	96.2 / 0.5	96.3 / 0.5	96.0 / 1.9

SPD labeling diagnostic. The reported relabeling accuracies under SPD are 92.4% on CIFAR-10 and 93.6% on ImageNet-30. These measurements provide an additional check on the semantic labeling stage. Table 5 then measures whether adding input correction improves the final defended model. Attributing that improvement specifically to the deep or shallow trigger would require evaluating those components separately.

G.5 Minimalist Models and the Number of Filtering Rounds

Each filtering round initializes a new model and trains it on the current candidate set using the supplied labels. Correctly classified records become candidates for the next round. Table 14 gives the first CIFAR-10 filter, which has one convolutional layer followed by a linear classifier. Its 9,038 parameters document the restricted capacity used to construct the selected set. This architecture makes the capacity choice reproducible, while the filtering measurements determine whether it captures the attack in a given setting.

Table 14: Architecture of the first CIFAR-10 filtering model.

Layer	Output Shape	Parameters
Input	$(32, 32, 3)$	0
Conv	$(30, 30, 1)$	28
Flatten	(900)	0
Output	(10)	9010

The default CIFAR-10 setting uses two rounds and approximately ten training epochs per round. Table 15 varies the number of rounds with the same choice applied to all five attacks. Moving from one round to two reduces FPR while largely preserving recall. Further rounds produce smaller reductions in FPR and exclude more poisoned candidates, with higher downstream ASR. The experiment supports two rounds as a useful choice in these settings and shows why repeatedly narrowing the candidate set can become counterproductive.

Table 15: Filtering-round sensitivity on CIFAR-10, averaged over five attacks and five independent runs.

Rounds	TPR ↑	FPR ↓	ASR ↓
1	99.92	13.84	6.74
2	99.64	8.37	5.56
3	99.08	7.72	5.83
4	96.12	7.20	8.91

G.6 Poisoning Rate and Filtering Dynamics

Table 16 varies the poisoning rate under Color-backdoor and reports attack strength before defense alongside filtering quality. Recall remains high over most of the displayed range, while both pre-defense ASR and recall fall at the lowest rate. This joint change shows why low recall should be interpreted in the context of the attack being filtered. It does not establish that every effective attack is easy to select, as the adaptive evaluation gives an effective attack with reduced recall. The nonzero FPR also reinforces the need to evaluate the selected clean records through their downstream utility.

Table 16: Color-backdoor stress test across poisoning rates. The last column counts poisoned records left outside the selected set.

Poison Rate	ASR w/o Defense	FPR↓	TPR↑	Missed Poison Samples↓
30%	100	9.09	100	0
20%	100	15.30	100	0
5%	100	8.34	100	0
0.5%	98.74	7.50	100	0
0.2%	90.23	9.40	96.50	7
0.1%	8.89	8.79	17.00	83

G.7 Semantic Relabeling and Reconstruction

Table 17 compares semantic relabeling with random reassignment on CIFAR-10. Each reported relabeling model recovers substantially more correct labels than the random baseline. This indicates that the remaining data provide useful semantic supervision for selected records in these experiments. The comparison tests label recovery, while the fixed-split ablation in Table 2 measures how that supervision affects final accuracy and backdoor mitigation.

Figure 8 varies the number of selected feature channels in a 64-channel representation. The smaller displayed values produce stronger distortion, while larger values retain more recognizable content. The reported setting $K = 60$ lies in the latter range. This visualization supports the role of feature anchors in preserving image content, but it does not determine which K minimizes ASR. That choice requires joint measurement of mitigation and clean accuracy.

Table 17: Relabeling accuracy on CIFAR-10 for semantic prediction and random reassignment.

Relabeling model	SCI	Random
ResNet-50	92.34%	9.87%
VGG19	89.91%	13.41%
BN-Inception	91.42%	15.60%
ResNet-v2	93.54%	7.64%
ResNet-v1	92.81%	9.01%



Figure 8: Reconstructions using different numbers of feature anchors in a 64-channel representation. The displayed settings illustrate the effect of K on retained visual content.

G.8 Supplementary Gradient Profiles

The coordinate profiles below provide an additional view of the gradient signal. They display absolute values along parameter index, so their horizontal axes do not represent training time. Figure 9 shows that BadNet produces different profiles for poisoned and clean records under the displayed settings. The concentration of large values in some coordinates is consistent with a feature-dependent update signal. These magnitude profiles do not measure pairwise directional alignment, which is tested directly in Table 1.

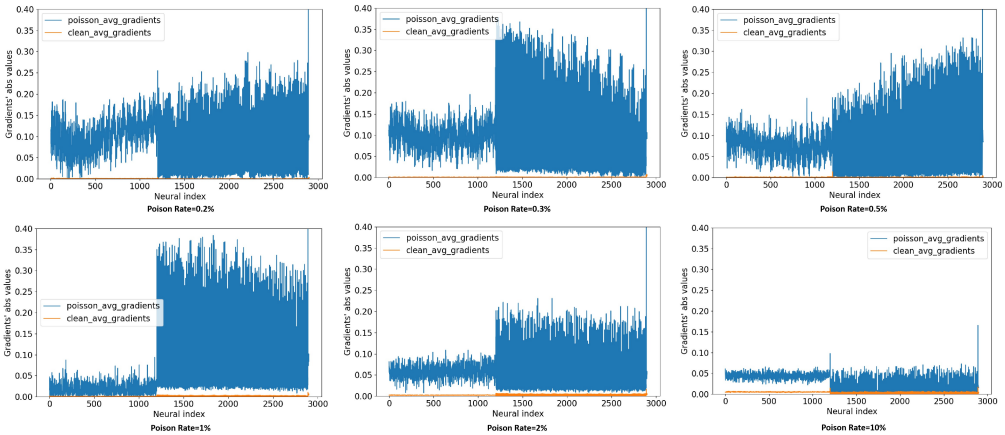


Figure 9: Absolute gradient-coordinate profiles under BadNet at different poisoning rates, using a learning rate of 0.001.

Figure 7 presents the corresponding profiles for Color-backdoor. Differences between clean and poisoned records remain visible despite the attack's different trigger construction. The shape and scale of the profiles vary across

poisoning rates, so they do not establish a monotonic relation between poisoning rate and gradient strength. Together with the BadNet profiles, they provide a descriptive check on how the optimization signal changes across attacks. The matched comparison remains the evidence for additional alignment beyond a shared label.

G.9 Adaptive-Attack Protocol

The adaptive Blended attack optimizes its trigger τ using

$$\min_{\tau} \quad \mathcal{L}_{\text{attack}}(\tau) + \lambda_1 \,\text{Align}_{\text{mini}}(\tau) + \lambda_2 \,\text{Fit}_{\text{mini}}(\tau), \qquad \lambda_1 = \lambda_2 = 0.5. \tag{52}$$

The attack loss promotes target prediction on a full-capacity surrogate. The alignment term penalizes pairwise cosine similarity between poisoned-sample gradients. The fitting term penalizes target-class confidence in independently initialized minimalist models.

G.10 Runtime Measurement

Runtime is measured from the poisoned ImageNet-100 dataset to the final defended ResNet-50 checkpoint on one RTX 5070 Laptop GPU. Data loading and mixed-precision settings are shared across methods. Batch size and final-classifier training settings are also held fixed. The no-defense baseline trains ResNet-50 for 100 epochs. SCI's measurement includes filtering and semantic relabeling, followed by reconstruction and final-model training. Table 18 shows that the complete correction pipeline costs more than standard training and ASSET, but less than CT in this setup. The comparison quantifies the training cost of producing corrected data without introducing additional modules at inference.

Table 18: Complete training time on ImageNet-100 with ResNet-50 on one RTX 5070 Laptop GPU. Overhead is relative to the 8.5-hour no-defense baseline.

Method	Runtime in hours	Overhead
No defense	8.5	–
ASSET	19.0 ± 0.4	123.5%
CT	24.4 ± 0.5	187.1%
SCI	20.5 ± 0.4	141.2%