SCAR: CLASS-WISE ASSESSMENT AND SHORTCUT-GUIDED MODEL REVISION FOR BACKDOOR DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks pose a severe security threat to deep neural networks (DNNs) by implanting hidden behaviors activated by specific triggers. For downstream users acquiring models from untrusted sources, reliable deployment entails assessing model integrity and, when necessary, repairing compromised models while preserving clean-task performance. The underlying backdoor mechanism embeds an additional trigger-to-target association into the model. Despite remaining largely dormant during normal inference, this hidden association can manifest as class-specific irregularities in classifier-head weights and prediction responses, while trigger activation induces directional changes in the feature space. Together, these complementary signatures reveal backdoor shortcuts at the class level without explicit trigger reconstruction. Building on this insight, we propose Shortcut-aware Class-wise Assessment and Revision (SCAR), a unified framework for post-training backdoor assessment and targeted model revision. SCAR combines classifier-weight anomalies with transformation-induced prediction preferences to assess model integrity and identify suspicious target classes. For each identified class, SCAR estimates a shortcut-related feature direction from class-specific weights and feature variability, and selectively revises the corresponding classifier-head parameters to suppress backdoor behavior. Extensive experiments across four datasets, ten attacks, and four model architectures demonstrate the effectiveness and generality of SCAR for both backdoor assessment and model recovery. Code will be released upon acceptance.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable performance across a broad range of perception and decision-making tasks (LeCun et al., 2015; He et al., 2015; Mnih et al., 2015). Yet increasing data and computational demands have encouraged reliance on weakly curated web data, outsourced training, and third-party pretrained models, making acquired models difficult to trust (Hong et al., 2022; Carlini et al., 2024). Backdoor attacks pose a particularly stealthy threat in this setting: by poisoning a small fraction of training data, an adversary can establish a trigger-to-target association that redirects triggered inputs to an attacker-specified class while preserving benign performance (Manoj & Blum, 2021). Consequently, the presence of hidden trigger-activated behaviors calls for post-acquisition inspection and, when necessary, model repair (Gu et al., 2019).

Existing post-training defenses primarily address two complementary tasks: *model inspection* and *model repair* (Gao et al., 2024). Model inspection determines whether a model is compromised and, when possible, identifies the associated target class through abnormal trigger responses (Wang et al., 2019) or model-internal signatures (Fu et al., 2023; Zhang et al., 2025). Model repair seeks to suppress backdoor behavior while preserving benign performance, using techniques such as adversarial unlearning (Zeng et al., 2022), neuron/channel pruning (Li et al., 2023; Yang et al., 2026), and knowledge distillation (Li et al., 2021a; Pang et al., 2023). Several defenses further integrate inspection and repair into a single workflow, connecting model assessment with subsequent mitigation before deployment or at test time (Guan et al., 2024; Yuan et al., 2026; Li et al., 2026a).

Despite this progress, applying existing defenses remains nontrivial for downstream users of untrusted models. First, available data and computational resources may be limited. In many model-acquisition scenarios, the defender has neither the original training data nor trusted labels, but only a

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.

small, potentially contaminated unlabeled downstream set. This makes approaches requiring trusted auxiliary data, costly trigger reconstruction, iterative optimization, or architecture-specific interventions less practical and portable across deployment settings. Second, model inspection and repair are often loosely coupled. Identifying a compromised model or suspicious target class does not directly reveal how the model should be revised. This disconnect becomes more consequential when trusted data and extensive optimization are unavailable, requiring inspection evidence to directly guide lightweight and targeted repair. A practical defense should therefore connect class-wise assessment with architecture-agnostic model revision.

Key Insight. These challenges motivate a shared, class-wise basis for both assessment and repair. For a targeted backdoor, the affected target class provides such a basis: its decision rule carries backdoor-related evidence, while its classifier parameters offer a direct handle for revision. Backdoor learning embeds a trigger-to-target shortcut alongside the intended task mapping. Although largely dormant on benign inputs, this shortcut leaves complementary signatures in the target class, including irregularities in classifier weights and preferential responses under mild input transformations (Fig. 1). These signals enable identification of suspicious target classes without explicitly reconstructing the trigger.

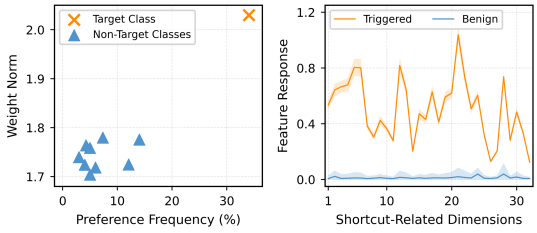

Figure 1: Backdoor-induced traces in a ResNet-18 on CIFAR-10 under the Blended attack. Left: the target class shows anomalous classifier-weight norm and prediction preference under mild Gaussian noise. Right: triggered inputs exhibit stronger responses than their original counterparts along shortcut-related feature dimensions.

For an identified class, its classifier weight together with feature-wise variability from downstream representations provides a proxy for the shortcut-related feature direction. Revising the classifier response along this direction suppresses the shortcut while confining the intervention to the affected class. This shared class-wise basis therefore connects assessment to targeted revision, allowing the evidence used for identification to directly guide repair.

Building on this insight, we propose **S**hortcut-aware **C**lass-wise **A**ssessment and **R**evision (SCAR), a unified post-training framework that couples backdoor assessment with targeted model revision. SCAR combines static classifier-weight irregularities and dynamic class preferences under mild input transformations to identify suspicious target classes, then derives a shortcut-related feature direction from class-specific weights and downstream representation statistics to guide lightweight classifier-head revision. SCAR operates solely on unlabeled downstream data, requiring neither trigger access, original training data, trusted labels, nor additional model training.

Our contributions are summarized as follows:

- We identify the suspicious target class as a natural bridge between backdoor assessment and model repair, providing a unified way to connect where a backdoor is manifested with where the model should be revised.
- We develop SCAR, a trigger-free post-training defense that uses complementary class-wise evidence to localize the target and translates this evidence into shortcut-guided classifier-head revision. The resulting repair requires no original training data, trusted labels, trigger access, or additional model training.
- We conduct a comprehensive evaluation across four datasets, ten attack methods, and four model architectures, showing that SCAR consistently supports effective and general backdoor assessment and model repair.

2 Related Work

Backdoor Attacks. Backdoor attacks implant hidden malicious behaviors into neural networks, causing compromised models to behave normally on benign inputs while producing attacker-specified predictions when triggered (Li et al., 2022; Liu et al., 2018b). Existing attacks broadly fall into data-poisoning and weight-poisoning attacks (Hong et al., 2022; Li et al., 2026b). Data-poisoning attacks span diverse designs, such as visible triggers (Gu et al., 2019), imperceptible or

sample-specific triggers (Doan et al., 2021; Nguyen & Tran, 2021), clean-label poisoning (Souri et al., 2022), and defense-aware attacks (Cheng et al., 2024), whereas weight-poisoning attacks directly modify model parameters (Rakin et al., 2020; Hong et al., 2022). In this work, we focus primarily on data-poisoning attacks, a widely used mechanism for backdoor implantation, while a preliminary investigation of weight-poisoning attacks is provided in Appendix A.2.

Post-training Backdoor Inspection and Repair. Post-training backdoor defenses broadly encompass model inspection and repair. Model inspection methods identify compromised models using trigger reconstruction (Wang et al., 2019), or intrinsic and structural signatures such as learned responses (Kolouri et al., 2020; Xu et al., 2021), network topology (Zheng et al., 2021), energy discrepancies (Gao et al., 2024), and abnormal logit margins (Wang et al., 2024). Repair methods suppress embedded backdoors through pruning or fine-tuning (Wu & Wang, 2021; Li et al., 2023), knowledge distillation (Li et al., 2021a), adversarial unlearning (Zeng et al., 2022), or feature-space adjustment (Min et al., 2023), with recent methods further reshaping activation geometry or logit margins (Amula et al., 2025; Yang et al., 2026). However, existing approaches often require iterative optimization or internal model intervention, and detection results do not directly determine how repair should be performed. In contrast, SCAR uses class-wise assessment evidence to guide targeted classifier-head revision without requiring the original training data or additional model training.

3 Problem Formulation

Preliminaries. This work considers a K-category classification problem over an input space $\mathcal{X}$ and a label space $\mathcal{Y} = \{1, \ldots, K\}$, with $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ denoting the training set. A DNN classifier f consists of a feature extractor h followed by a final classifier head. Given an input image x, its penultimate representation $h(x) \in \mathbb{R}^d$ is mapped to the output logits and predicted label as

$$f(x) = Wh(x) + b \in \mathbb{R}^K, \qquad \hat{y}(x) = \arg\max_{c \in \mathcal{Y}} f_c(x), \tag{1}$$

where $W = [w_1^\top, \ldots, w_K^\top] \in \mathbb{R}^{K \times d}$ and $b \in \mathbb{R}^K$ denote the classifier weight matrix and bias vector, respectively. When the training is untrusted, however, the resulting classifier may be compromised by a backdoor attack, as described next.

Threat Model. We consider data-poisoning backdoor attacks, where an attacker applies a trigger transformation $g_t(\cdot)$ to a subset of training samples. Given an original sample (x, y), its poisoned counterpart is $(g_t(x), y_p)$, where y_p may be altered or preserved depending on the attack strategy. Training on such samples implants backdoor behavior, after which the compromised model may be distributed through untrusted third-party sources or model-sharing platforms.

Defender's Goal and Capability. We consider a downstream user acting as the defender and obtains a pretrained classifier f from an untrusted third party or model-sharing platform, without access to its training process or knowledge of whether it has been compromised. The defender has white-box access to f, but no trusted clean samples or labels. Instead, only a small unlabeled set drawn from the downstream input distribution is available, which may contain contaminated samples. No individual sample is assumed to be clean; we only assume that benign samples constitute the majority of the set. The defender further has no trusted reference model or identified poisoned samples, and no knowledge of the trigger, target class, or attack strategy. Under these constraints, the defender aims to assess whether the acquired model is compromised and, when necessary, recover it by suppressing backdoor behavior while preserving benign predictive utility.

4 Method

In this section, we present **S**hortcut-aware **C**lass-wise **A**ssessment and **R**evision (SCAR), a post-training framework for backdoor assessment and model recovery. SCAR first assesses the model using static weight anomalies and dynamic prediction preferences exposed by mild input transformations and localizes suspicious target classes. For these classes, SCAR filters downstream samples exhibiting unusually strong preferences toward them and robustly estimates feature statistics from the retained data to reduce bias in shortcut estimation. Using these robust statistics together with class-specific weights, SCAR estimates the corresponding shortcut-related feature directions, refines them using class geometry, and selectively revises the associated classifier-head parameters. The overall framework is illustrated in Fig. 2.

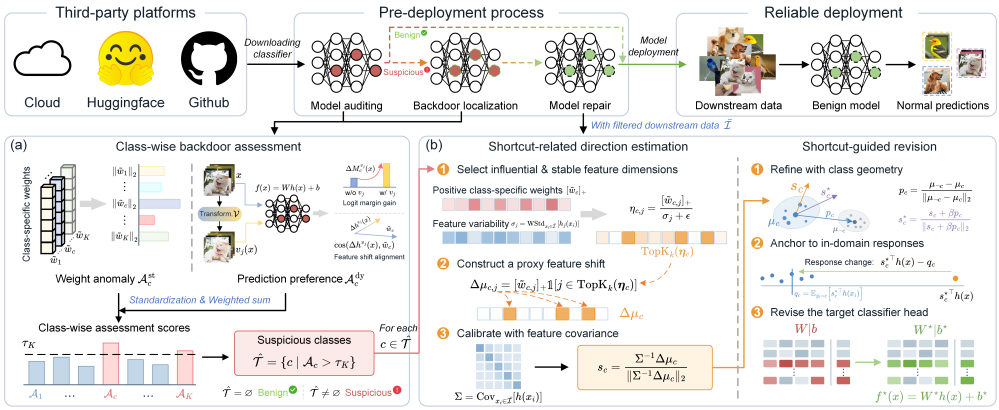

Figure 2: Illustration of SCAR within the pre-deployment workflow, from untrusted model acquisition to auditing, repair, and reliable deployment. (a) Class-wise backdoor assessment combines classifier-weight anomalies and transformation-induced prediction preferences to localize suspicious target classes. (b) Shortcut-guided revision estimates shortcut-related feature directions using filtered downstream data and performs targeted classifier-head revision.

4.1 Class-wise Backdoor Assessment

Backdoor training introduces a trigger-to-target shortcut beyond ordinary semantic learning, which can leave distinctive traces in target-class weights and induce abnormal prediction preferences toward the target class under mild input transformations. Accordingly, we assess each class using *static weight anomaly* and *dynamic prediction preference*.

Static Weight Anomaly. Let w_c denote the classifier weight for class c. Removing the component shared across classes yields its class-specific weight:

$$\tilde{w}_c = w_c - \frac{1}{K}\sum_{i=1}^{K} w_i. \tag{2}$$

We quantify the weight anomaly by comparing its L_2 magnitude with those of the remaining classes:

$$\mathcal{A}_c^{\mathrm{st}} = \left| \frac{\|\tilde{w}_c\|_2 - \mathrm{Med}_{i\neq c}(\|\tilde{w}_i\|_2)}{\kappa\ \mathrm{MAD}_{i\neq c}\left(\|\tilde{w}_i\|_2\right)} \right|. \tag{3}$$

Here, $\mathrm{Med}(\cdot)$ and $\mathrm{MAD}(\cdot)$ denote the median and median absolute deviation, respectively, and $\kappa = 1.4826$ is the standard Gaussian-consistency factor for MAD (Rousseeuw & Hubert, 2018); a small constant is added to the denominator for numerical stability.

Dynamic Prediction Preference. Given the available unlabeled downstream samples, let $\mathcal{I} = \{x_i\}_{i=1}^{N}$, with pseudo-labels $\hat{y}_i = \hat{y}(x_i)$ assigned by the acquired model. A fixed set of mild, attack-agnostic transformations $\mathcal{V} = \{v_j\}_{j=1}^{J}$ is then applied to reveal latent prediction preferences toward each class. For an input x, the logit margin of class c is $M_c(x) = f_c(x) - \max_{i\neq c} f_i(x)$, with its margin and representation changes under $v \in \mathcal{V}$ given by

$$\Delta M_c^v(x) = M_c(v(x)) - M_c(x), \qquad \Delta h^v(x) = h(v(x)) - h(x). \tag{4}$$

A preference toward class c is reflected by both an increased logit margin and a representation shift aligned with its class-specific weight $\tilde{w}_c$:

$$a_c^v(x) = [\Delta M_c^v(x)]_+ \,[\cos(\Delta h^v(x), \tilde{w}_c)]_+ \,\mathbb{1}[\hat{y}(x) \neq c], \tag{5}$$

where $[\cdot]_+ = \max(\cdot, 0)$, $\cos(\cdot,\cdot)$ denotes cosine similarity, and $\mathbb{1}[\cdot]$ the indicator function. We aggregate the responses across samples and transformations to obtain the prediction preference score:

$$\mathcal{A}_c^{\mathrm{dy}} = \frac{1}{|\mathcal{I}_{\neg c}|J} \sum_{x_i \in \mathcal{I}_{\neg c}} \sum_{j=1}^{J} a_c^{v_j}(x_i), \tag{6}$$

where $\mathcal{I}_{\neg c} = \{x_i \in \mathcal{I} \mid \hat{y}_i \neq c\}$ denotes the samples not predicted as class c.

Joint Backdoor Assessment. The static weight anomaly and dynamic prediction preference provide complementary evidence of class-wise abnormality. We normalize and combine them into a joint evidence, then standardize it across classes to obtain the final class-wise assessment score:

$$\mathcal{E}_c = \log\left(1 + \frac{(\mathcal{A}_c^{\mathrm{st}})^2}{\sum_{i=1}^{K}(\mathcal{A}_i^{\mathrm{st}})^2} + \frac{(\mathcal{A}_c^{\mathrm{dy}})^2}{\sum_{i=1}^{K}(\mathcal{A}_i^{\mathrm{dy}})^2}\right), \qquad \mathcal{A}_c = \frac{\mathcal{E}_c - \mathrm{Med}_{i\neq c}(\mathcal{E}_i)}{\kappa\ \mathrm{MAD}_{i\neq c}(\mathcal{E}_i)}. \tag{7}$$

We set the detection threshold to $\tau_K = \Phi^{-1}(1-\delta/K)$, where $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function of the standard normal distribution and δ is a preset significance level. The suspicious target set is $\hat{\mathcal{T}} = \{c \in \mathcal{Y} \mid \mathcal{A}_c > \tau_K\}$. A non-empty $\hat{\mathcal{T}}$ indicates a suspicious model and specifies the classes for subsequent revision; otherwise, SCAR leaves the model unchanged.

4.2 Shortcut-Guided Model Revision

Backdoor training embeds a trigger-to-target shortcut that largely preserves benign behavior. We seek feature directions where trigger-induced changes are distinguishable from normal task variation. Since downstream unlabeled data may be contaminated, we first filter samples with unusually strong preferences toward suspicious classes and then use robust estimators to derive shortcut-related statistics.

For each downstream sample x, we assess its preference for each suspicious class $c \in \hat{\mathcal{T}}$ using two cues: the representation projection onto the class-specific weight, which reflects alignment with the class decision direction, and the logit margin $M_c(x)$, which may become abnormally large for triggered inputs as the backdoor raises the target-class logit while suppressing competing logits (Wang et al., 2024; Mu et al., 2023). The projection is computed as $r_c(x) = h(x)^\top \frac{\tilde{w}_c}{\|\tilde{w}_c\|_2}$. Samples whose $[M_c(x)]_+[r_c(x)]_+$ exceeds a preset filtering threshold τ_{f} for any suspicious class are regarded as potentially contaminated and filtered out, yielding the retained set $\bar{\mathcal{I}}$; its effectiveness is further analyzed in Appendix A.1.

Shortcut Feature Direction Estimation. We first consider an oracle setting with access to the trigger transformation g_t. For a suspicious class c, the trigger-induced feature shift and reference feature covariance are

$$\Delta\mu_c^{\mathrm{ora}} = \mathbb{E}_{\substack{x_i\in\bar{\mathcal{I}}\\ \hat{y}_i\neq c}}\left[h(g_t(x_i)) - h(x_i)\right], \qquad \Sigma = \mathrm{Cov}_{x_i\in\bar{\mathcal{I}}}[h(x_i)]\,. \tag{8}$$

For robust empirical estimation, $\Delta\mu_c^{\mathrm{ora}}$ is computed using a 1% two-sided trimmed mean of the feature differences (Mair & Wilcox, 2020), which further reduces sensitivity to residual contamination after filtering. A backdoor shortcut is expected to induce a directional feature shift beyond ordinary representation variation. We therefore seek a direction s that maximizes this contrast:

$$\mathcal{J}(s) = \frac{\left(s^\top \Delta\mu_c^{\mathrm{ora}}\right)^2}{s^\top \Sigma s}. \tag{9}$$

Since $\mathcal{J}(s)$ is invariant to the scale of s, we impose $s^\top\Sigma s = 1$ without loss of generality. Choosing the orientation such that $s^\top\Delta\mu_c^{\mathrm{ora}} \geq 0$, maximizing $\mathcal{J}(s)$ is equivalent to

$$\max_s\ s^\top \Delta\mu_c^{\mathrm{ora}} \qquad \text{s.t.} \qquad s^\top\Sigma s = 1. \tag{10}$$

To solve this constrained problem, we introduce the Lagrangian $\mathcal{L}(s,\lambda) = s^\top\Delta\mu_c^{\mathrm{ora}} - \frac{\lambda}{2}\left(s^\top\Sigma s - 1\right)$, where λ is the Lagrange multiplier. The stationary condition with respect to s gives $\nabla_s\mathcal{L} = \Delta\mu_c^{\mathrm{ora}} - \lambda\Sigma s = 0$, implying $s \propto \Sigma^{-1}\Delta\mu_c^{\mathrm{ora}}$. Normalization then gives the oracle direction

$$s_c^{\mathrm{ora}} = \frac{\Sigma^{-1}\Delta\mu_c^{\mathrm{ora}}}{\|\Sigma^{-1}\Delta\mu_c^{\mathrm{ora}}\|_2}. \tag{11}$$

A small diagonal loading term is added to Σ before inversion for positive definiteness and numerical stability.

Estimating the Shift without Triggers. The oracle shift $\Delta\mu_c^{\mathrm{ora}}$ depends on the true trigger and is unavailable in practice. We therefore approximate it using class-specific weights and the per-dimension feature standard deviations estimated over $\bar{\mathcal{I}}$, prioritizing dimensions with strong influence on c but limited variation. Accordingly,

$$\sigma_j = \mathrm{WStd}_{x_i\in\bar{\mathcal{I}}}\left[h_j(x_i)\right], \qquad \eta_{c,j} = \frac{[\tilde{w}_{c,j}]_+}{\sigma_j + \epsilon}, \qquad \Delta\mu_{c,j} = [\tilde{w}_{c,j}]_+ \mathbb{1}\left[j \in \mathrm{TopK}_k(\boldsymbol{\eta}_c)\right], \tag{12}$$

where $\mathrm{WStd}[\cdot]$ denotes the standard deviation after 1% two-sided winsorization (Mair & Wilcox, 2020), $\boldsymbol{\eta}_c = (\eta_{c,1}, \ldots, \eta_{c,d})$, and $\epsilon > 0$ ensures numerical stability; TopK_k selects the k largest entries. Replacing $\Delta\mu_c^{\mathrm{ora}}$ in Eq. (11) with the proxy shift $\Delta\mu_c$ yields the practical direction s_c.

Shortcut-Guided Classifier Revision. The estimated direction s_c captures shortcut-related responses and serves as the primary revision direction. To better preserve class geometry underlying normal predictions, we further refine it using the direction from the class-c center toward the remaining samples, yeilding the final revision direction:

$$p_c = \frac{\mu_{\neg c} - \mu_c}{\|\mu_{\neg c} - \mu_c\|_2}, \qquad s_c^\star = \frac{s_c + \beta p_c}{\|s_c + \beta p_c\|_2}, \tag{13}$$

where μ_c and $\mu_{\neg c}$ denote the representation centers of samples in $\bar{\mathcal{I}}$ predicted as c and the remaining classes, respectively, and β is a correction strength hyperparameter. To anchor the revision to normal class-c responses, we use the reference projection $q_c = \mathbb{E}_{\substack{x_i \in \bar{\mathcal{I}} \\ \hat{y}_i = c}} \left[s_c^{\star\top} h(x_i) \right]$ as a reference. The centered response $s_c^{\star\top} h(x) - q_c$ then serves as the correction signal for the final classifier revision.

4.3 Revised Model Construction

For each suspicious class $c \in \hat{\mathcal{T}}$, we scale the correction strength according to its relative anomaly score and revise the corresponding logit as

$$\gamma_c^\star = \gamma \frac{\mathcal{A}_c}{\sqrt{\sum_{i \in \hat{\mathcal{T}}} (A_i)^2}}, \qquad f_c^\star(x) = f_c(x) - \gamma_c^\star \left(s_c^{\star\top} h(x) - q_c \right), \tag{14}$$

where γ is a hyperparameter controlling the overall revision strength. The logit revision is equivalent to updating the corresponding classifier-head parameters as

$$w_c^\star = w_c - \gamma_c^\star s_c^\star, \qquad b_c^\star = b_c + \gamma_c^\star q_c. \tag{15}$$

The final revised model is $f^\star(x) = W^\star h(x) + b^\star$, where only the parameters of classes in $\hat{\mathcal{T}}$ are updated and all others remain unchanged.

5 Experimental Results

5.1 Experimental Setup

Implementation. We set the significance level to $\delta = 0.05$, which determines the threshold τ_K for class-wise backdoor assessment. The filtering threshold τ_{f} is set to 30, 45, 80, and 15 for CIFAR-10, CIFAR-100, TinyImageNet, and ImageNet-1k, respectively. For shortcut direction estimation, we retain the top 25% feature dimensions ranked by shortcut relevance, with the geometry calibration strength β set to 0.25, 0.25, 8, and 2 on the four datasets, respectively. The revision strength γ is set to 128 on CIFAR-10 and 512 for the remaining datasets. All experiments are implemented in PyTorch (Paszke et al., 2019) and conducted on a workstation with dual 2.90 GHz CPUs, 192 GB RAM, and four NVIDIA RTX 3090 GPUs.

Datasets and Model Architectures. We conduct experiments on four benchmark datasets spanning three input resolutions: CIFAR-10 and CIFAR-100 (32×32) (Krizhevsky et al., 2009), TinyImageNet (64×64) (Le et al., 2015), and ImageNet-1k (224×224) (Deng et al., 2009). We pair these datasets with ResNet-18 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2015), ViT-B/16 (Dosovitskiy et al., 2021), and ResNet-50 (He et al., 2016), respectively.

Backdoor Attacks. We evaluate our approach against ten representative attacks, including BadNets (Gu et al., 2019), Blended (Chen et al., 2017), TrojanNN (Liu et al., 2018b), Label-consistent (LC) (Turner et al., 2019), SIG (Barni et al., 2019), SSBA (Li et al., 2021b), WaNet (Nguyen & Tran, 2021), BPP (Wang et al., 2022), LOTUS (Cheng et al., 2024), and Grond (Xu et al., 2025).

Detection and Defense Baselines. For backdoor detection, we select six baselines: ABS (Liu et al., 2019), BAN (Xu et al., 2024), MM-BD (Wang et al., 2024), TRODO (Mirzaei et al., 2024), BARBIE (Zhang et al., 2025), and SPD (Zhao et al., 2026). For backdoor defense, we consider

Table 1: Comparison of backdoor model detection performance across different methods on the CIFAR-10 and CIFAR-100 benchmarks.

Attack	ABS		BAN		MM-BD		TRODO		BARBIE		SPD		Ours	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10/ResNet-18														
BadNets	33.33	92.52	14.29	97.05	38.10	94.78	85.71	50.11	38.10	88.44	**0.00**	**100.00**	**0.00**	99.77
Blended	**0.00**	**100.00**	9.52	99.09	38.10	91.61	66.67	65.53	85.71	66.67	**0.00**	**100.00**	**0.00**	99.77
TrojanNN	**0.00**	95.24	**0.00**	**100.00**	**0.00**	**100.00**	100.00	31.52	42.86	90.70	85.71	62.13	**0.00**	**100.00**
LC	100.00	21.32	19.05	85.71	**0.00**	**100.00**	95.24	40.36	57.14	84.13	**0.00**	**100.00**	57.14	90.70
SIG	100.00	65.65	14.29	89.57	**0.00**	**100.00**	100.00	24.49	71.43	79.82	**0.00**	**100.00**	**0.00**	99.77
SSBA	23.81	91.50	**0.00**	**100.00**	90.48	58.73	85.71	60.09	85.71	61.22	61.90	83.67	14.29	97.51
WaNet	100.00	45.35	9.52	98.87	90.48	49.21	100.00	32.88	85.71	64.40	**0.00**	**100.00**	**0.00**	**100.00**
BPP	100.00	76.19	14.29	97.51	95.24	35.15	100.00	13.83	38.10	86.62	61.90	86.17	**0.00**	**100.00**
LOTUS	100.00	1.81	14.29	91.61	95.24	68.03	85.71	65.76	85.71	66.21	95.24	59.18	**5.24**	**96.69**
Grond	100.00	63.72	19.05	91.61	90.48	47.39	90.48	61.68	52.38	81.63	90.48	53.51	**10.24**	**92.60**
Average	65.71	65.33	11.43	95.10	53.81	74.49	90.95	44.63	64.29	76.98	39.52	84.47	**8.69**	**97.68**
CIFAR-100/VGG-16														
BadNets	23.81	92.97	**0.00**	**100.00**	**0.00**	**100.00**	61.90	78.00	80.95	68.48	4.76	99.09	**0.00**	**100.00**
Blended	38.10	84.69	**0.00**	**100.00**	95.24	75.17	80.95	68.93	95.24	55.78	85.71	58.28	**0.00**	**100.00**
TrojanNN	76.19	82.20	**0.00**	**100.00**	**0.00**	96.60	80.95	72.11	80.95	69.61	90.48	46.49	**0.00**	**100.00**
LC	85.71	50.34	42.86	94.10	71.43	87.07	85.71	43.99	100.00	47.62	**0.00**	**100.00**	**0.00**	**100.00**
SIG	100.00	36.96	**0.00**	**100.00**	95.24	41.95	90.48	53.06	100.00	32.65	**0.00**	96.37	**0.00**	**100.00**
SSBA	85.71	39.12	**0.00**	**100.00**	95.24	32.88	90.48	70.75	71.43	64.63	80.95	63.95	**0.00**	**100.00**
WaNet	85.71	66.10	**0.00**	**100.00**	95.24	52.38	90.48	42.63	100.00	45.80	**0.00**	99.55	**0.00**	**100.00**
BPP	4.76	98.41	**0.00**	**100.00**	80.95	65.08	85.71	50.57	**0.00**	98.41	**0.00**	**100.00**	**0.00**	**100.00**
LOTUS	85.71	51.02	**0.00**	**99.55**	80.95	59.07	80.95	79.82	90.48	55.33	90.48	70.75	4.24	96.01
Grond	85.71	54.65	61.90	69.39	95.24	59.52	47.62	81.41	95.24	50.34	90.48	49.89	**22.65**	**89.43**
Average	67.14	65.65	10.48	96.30	70.95	66.97	79.52	64.13	81.43	58.87	44.29	78.44	**2.69**	**98.54**

ten post-training methods: Fine-Pruning (FP) (Liu et al., 2018a), NC (Wang et al., 2019), NAD (Li et al., 2021a), FT-SAM (Zhu et al., 2023), RNP (Li et al., 2023), MM-BM (Wang et al., 2024), TSBD (Lin et al., 2024), D3 (Wei et al., 2025), PGBD (Amula et al., 2025), and LMR (Yang et al., 2026). Further details of these baselines are provided in Appendix A.3.

Setup. For attack construction, clean-label attacks (LC, SIG, and Grond) poison 50% of the target-class training samples, whereas all other attacks use a 10% poisoning rate over the full training set. Unless otherwise specified, we set the contamination rate of the unlabeled downstream set to 10% in subsequent experiments. We follow the configurations recommended by BackdoorBench (Wu et al., 2025) when available and the original recommended settings otherwise. Model-level detection uses 30 clean models and 30 backdoored models for each attack. To compute the dynamic prediction preference score $\mathcal{A}_c^{\mathrm{dy}}$, SCAR applies five fixed, attack-agnostic transformation: Gaussian noise, Cutout, Mixup, low-pass filtering, and sinusoidal warping. Throughout the defense process, SCAR uses unlabeled downstream data amounting to 1% of the corresponding training set.

Evaluation Metrics. Following established evaluation practice (Mazeika et al., 2023; Li et al., 2023), we evaluate backdoor detection using the area under the receiver operating characteristic curve (AUROC) and the false positive rate at 95% true positive rate (FPR95), and backdoor defense using attack success rate (ASR) and clean accuracy (ACC). All metrics are reported as percentages.

5.2 Experimental Results

Performance on Backdoor Detection. Tab. 1 presents the backdoor model detection performance on CIFAR-10 and CIFAR-100 benchmarks. When SCAR identifies multiple suspicious target classes, a detection is considered successful only if the top-ranked class matches the ground-truth target. Under this criterion, SCAR outperforms competing approaches in most cases, reducing the average FPR95 by $2.74\%/7.79\%$ and improving the average AUROC by $2.58\%/2.24\%$ over the next-best method on CIFAR-10/CIFAR-100, respectively. These results demonstrate the effectiveness of SCAR in reliably detecting backdoored models and localizing their target classes.

Performance on Backdoor Defense. Tab. 2 presents the backdoor defense performance across four benchmark settings and ten attacks. SCAR consistently achieves the lowest average ASR on all

Table 2: Comparison of backdoor defense performance across different methods on the CIFAR-10, CIFAR-100, TinyImageNet, and ImageNet-1k benchmarks.

Defense	BadNets		Blended		TrojanNN		LC		SIG		SSBA		WaNet		BPP		LOTUS		Grond		Average	
	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑
CIFAR-10/ResNet-18																						
No Defense	97.69	94.65	99.90	95.02	100.00	94.81	98.31	94.27	99.84	93.52	100.00	94.93	98.77	93.80	100.00	94.45	96.80	93.90	97.17	94.38	98.85	94.37
FP	6.79	94.28	78.00	94.02	45.67	93.91	4.86	93.18	1.13	93.14	1.00	93.98	1.86	92.99	16.70	93.68	3.80	92.98	96.31	93.78	25.61	93.59
NC	0.61	93.70	1.73	93.17	1.19	93.33	1.71	93.38	99.84	92.12	100.00	94.29	98.79	93.79	100.00	**94.44**	96.80	93.90	97.16	93.37	59.78	93.55
NAD	1.28	92.47	8.00	93.18	5.52	92.26	2.84	92.44	0.53	92.66	2.91	91.75	1.17	92.87	1.00	93.93	5.70	91.29	98.44	93.83	12.74	92.67
FT-SAM	0.94	94.36	10.76	94.36	3.81	**94.21**	8.34	93.44	3.39	93.35	1.47	93.29	0.72	93.70	1.64	94.13	5.70	93.11	95.26	93.82	13.20	**93.78**
RNP	1.89	90.88	26.92	91.83	1.51	92.50	0.82	93.58	56.62	88.18	0.13	93.56	**0.41**	90.44	0.36	92.44	15.00	93.35	76.72	91.14	18.04	91.79
MM-BM	0.86	94.46	77.36	**95.02**	1.19	91.68	0.96	92.26	84.48	88.50	74.62	89.77	1.08	90.68	2.06	90.10	1.70	89.27	72.33	89.41	31.66	91.11
TSBD	0.86	93.77	4.06	94.11	3.17	94.04	4.60	92.93	0.56	92.66	1.63	93.51	0.60	93.73	1.07	93.99	4.40	91.93	87.84	93.39	10.88	93.41
D3	0.17	93.86	35.87	93.98	25.89	93.83	0.93	92.71	37.77	92.16	2.22	93.55	10.18	92.72	29.33	93.65	11.50	92.51	95.27	93.09	24.91	93.21
PGBD	0.72	90.87	5.40	91.75	4.71	90.83	98.30	**94.26**	0.89	88.60	1.88	92.40	0.86	92.36	1.80	90.98	**0.30**	88.62	26.38	74.59	14.12	89.53
LMR	0.84	94.03	6.47	93.05	4.20	93.01	9.14	93.01	2.00	**93.49**	1.50	92.81	1.21	91.67	1.36	93.26	62.30	92.89	99.92	93.75	18.89	93.10
Ours	**0.13**	**94.50**	**0.01**	94.15	**0.95**	93.90	**0.02**	92.32	**0.21**	92.14	**0.10**	**94.58**	2.08	**93.88**	**0.01**	93.01	5.00	**93.96**	**8.39**	**93.84**	**1.69**	93.63
CIFAR-100/VGG-16																						
No Defense	93.37	73.57	99.88	73.25	99.88	72.49	99.05	73.26	85.54	72.88	100.00	73.15	96.63	72.31	99.99	71.42	94.00	73.35	94.67	71.84	96.30	72.75
FP	42.26	72.20	95.99	71.48	98.49	71.55	99.43	72.63	9.21	72.50	0.13	71.63	13.48	71.82	0.13	68.99	4.00	72.26	92.58	**73.22**	45.57	71.83
NC	0.16	69.66	0.01	68.05	0.18	68.96	0.41	69.65	0.15	64.52	0.23	65.39	0.21	68.59	0.08	65.07	1.00	65.91	7.86	65.02	1.03	67.08
NAD	0.08	70.18	10.95	69.59	44.82	68.87	5.95	72.34	1.26	70.23	0.16	69.52	0.26	70.16	0.07	64.98	29.00	72.14	93.19	70.71	18.57	69.87
FT-SAM	0.23	66.59	0.06	65.74	0.37	66.50	0.64	66.64	0.23	66.08	0.22	67.06	0.28	67.45	22.21	**71.42**	2.00	65.86	94.67	71.84	12.09	67.52
RNP	**0.00**	69.71	**0.00**	68.92	2.91	67.48	**0.00**	72.86	**0.00**	72.23	**0.00**	70.41	0.17	67.18	0.12	67.07	**0.00**	68.08	63.03	67.71	6.62	69.17
MM-BM	20.42	72.67	99.91	71.69	98.74	**72.56**	**0.00**	72.38	85.62	72.50	62.51	71.04	96.63	71.34	49.63	71.41	93.00	**73.40**	94.64	71.79	70.11	**72.08**
TSBD	0.04	69.99	0.04	69.90	0.03	69.46	0.45	70.18	0.31	69.33	0.33	69.03	0.29	69.79	0.07	64.90	1.00	70.83	54.22	64.62	5.68	68.80
D3	0.01	71.82	**0.00**	71.78	**0.00**	71.33	0.01	72.59	0.65	72.25	0.03	70.99	**0.11**	72.04	0.08	68.86	2.00	72.76	91.71	72.72	9.46	71.71
PGBD	0.01	61.17	0.27	60.23	0.89	60.54	99.05	72.28	1.66	64.43	0.16	68.14	0.68	63.42	0.20	65.47	2.00	63.66	89.60	68.03	19.45	64.74
LMR	11.99	70.27	60.39	69.85	97.10	69.92	22.01	71.02	0.88	72.31	0.31	70.78	0.23	**72.07**	0.16	69.51	4.00	72.32	88.21	70.77	28.53	70.88
Ours	0.32	**73.01**	**0.00**	**72.47**	**0.00**	72.28	**0.00**	**72.90**	2.80	**72.63**	**0.00**	**72.14**	0.25	71.91	**0.00**	69.64	0.56	72.82	**5.35**	70.54	**0.93**	72.03
TinyImageNet/ViT-B/16																						
No Defense	100.00	88.90	99.98	88.51	99.93	87.62	100.00	88.32	96.98	88.70	100.00	88.52	98.06	88.04	99.85	87.16	100.00	88.02	90.04	80.01	98.48	87.38
FP	79.79	73.60	50.49	74.60	98.16	73.89	0.03	73.37	1.02	73.78	0.10	73.56	0.22	73.68	49.89	72.93	**0.00**	72.83	82.49	67.61	36.22	72.99
NC	0.04	78.09	2.11	78.83	1.16	78.22	0.12	79.16	1.91	79.46	0.14	79.11	0.21	79.86	0.18	78.14	4.00	79.43	17.01	75.54	2.69	78.58
NAD	99.84	85.38	99.49	85.99	99.78	83.74	99.93	86.13	89.47	86.00	99.25	85.45	44.56	85.36	99.08	84.30	90.00	85.61	82.12	80.05	90.35	84.80
FT-SAM	57.72	80.19	86.70	79.16	97.52	80.30	54.99	80.41	63.97	81.05	84.55	80.59	1.91	79.63	52.18	78.32	46.00	80.22	68.59	75.82	61.41	79.57
RNP	100.00	88.18	99.98	87.13	99.93	87.61	100.00	**88.34**	96.93	88.60	100.00	88.49	98.02	88.12	0.10	86.22	100.00	88.02	90.19	**80.13**	88.52	**87.08**
MM-BM	99.98	88.82	99.98	87.22	99.93	87.61	98.90	85.50	97.01	88.72	100.00	88.53	98.16	88.00	99.53	**87.15**	86.00	88.01	90.17	80.01	96.97	86.96
TSBD	74.41	78.54	75.28	77.28	96.20	79.05	48.78	80.54	72.64	79.82	34.60	77.91	1.23	78.01	76.73	81.36	2.00	78.62	58.84	75.31	54.07	78.64
D3	**0.00**	76.91	0.05	75.70	0.10	75.56	0.05	76.10	0.23	76.03	0.54	75.42	**0.12**	76.64	0.32	75.60	0.05	75.20	73.35	72.05	7.48	75.52
PGBD	66.06	82.99	99.98	87.22	99.93	87.63	100.00	88.32	93.08	88.70	100.00	88.52	13.43	84.41	95.61	83.95	54.00	85.03	83.59	79.93	80.57	85.67
LMR	88.14	83.09	90.98	83.00	99.51	82.15	78.77	83.17	66.34	83.22	92.88	82.79	1.16	82.87	88.86	81.71	30.00	83.28	77.31	77.63	71.40	82.29
Ours	9.40	**88.84**	**0.02**	**87.23**	**0.00**	**87.81**	**0.01**	86.72	**0.16**	**88.75**	**0.00**	**88.65**	5.07	**88.36**	**0.05**	85.35	**0.00**	**88.03**	**1.81**	78.56	**1.87**	86.83
ImageNet-1k/ResNet-50																						
No Defense	100.00	76.79	99.98	76.69	99.94	76.84	99.57	76.87	95.65	76.89	99.98	76.80	91.56	76.82	99.76	75.01	100.00	75.79	91.19	76.82	97.76	76.53
FP	0.01	59.80	0.02	60.07	0.03	59.71	0.01	59.76	0.02	59.96	0.06	59.66	0.01	59.84	0.01	59.48	**0.00**	59.98	19.89	59.78	2.01	59.80
NC	0.12	60.16	0.26	60.52	1.20	60.48	0.01	60.52	**0.00**	60.10	**0.00**	60.42	0.22	60.54	**0.00**	59.84	1.95	60.64	10.73	60.48	1.45	60.37
NAD	**0.00**	69.87	1.11	70.10	0.02	70.00	**0.00**	69.93	0.44	69.88	4.17	69.92	0.01	70.12	**0.00**	69.21	**0.00**	70.14	40.72	69.96	4.65	69.91
FT-SAM	0.01	76.39	0.69	76.68	73.94	**76.84**	**0.00**	76.87	93.29	76.60	99.99	76.31	91.63	76.71	99.76	74.81	14.00	**75.79**	91.19	76.63	56.45	76.36
RNP	0.01	76.39	99.84	76.59	73.94	76.83	0.01	76.87	93.29	76.55	99.99	76.43	91.63	76.72	99.77	74.81	14.00	75.63	91.19	76.83	66.37	76.36
MM-BM	0.01	76.35	99.86	76.62	61.04	76.73	**0.00**	76.56	92.82	76.57	99.99	76.39	89.92	76.52	99.88	74.65	16.00	75.34	91.17	76.85	65.07	76.26
TSBD	0.01	76.39	0.10	63.42	0.52	64.78	**0.00**	76.87	0.04	63.57	0.34	64.83	**0.00**	64.79	**0.00**	62.80	14.00	**75.79**	17.85	64.29	3.29	67.75
D3	0.01	**76.79**	91.84	76.68	73.94	76.83	0.01	76.87	93.29	76.50	0.02	58.01	91.63	76.78	0.01	57.42	14.00	75.78	15.17	58.07	37.99	70.97
PGBD	0.01	76.39	22.96	68.59	2.03	68.70	**0.00**	76.87	5.98	68.62	41.39	67.01	91.63	76.78	1.32	68.95	**0.00**	72.11	25.88	66.66	19.12	71.07
LMR	0.01	71.52	27.31	71.60	0.66	71.02	0.01	71.70	6.58	69.43	47.58	71.45	0.01	71.44	0.02	70.59	**0.00**	71.52	51.50	69.54	13.37	70.98
Ours	**0.00**	76.44	**0.00**	**76.70**	**0.00**	76.17	4.30	**76.88**	**0.00**	**76.65**	**0.00**	**76.46**	**0.00**	**76.79**	9.87	**74.91**	**0.00**	75.63	**0.00**	**76.86**	**1.42**	**76.35**

Table 3: Comparison of backdoor defense performance under multi-target attacks on CIFAR-10 with ResNet-18.

Defense	BadNets		Blended		WaNet	
	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑
No defense	97.20	94.08	99.99	94.06	98.49	94.14
FP	4.23	93.36	71.34	93.25	11.13	93.80
NAD	0.80	92.84	13.76	92.86	1.54	93.22
FT-SAM	0.82	93.34	41.30	93.12	1.16	93.86
RNP	35.11	86.17	89.03	90.24	88.33	**94.19**
MM-BM	1.61	85.60	39.61	92.77	21.44	93.38
TSBD	1.10	93.48	5.51	**93.32**	0.80	93.79
D3	0.41	92.45	**0.00**	92.06	0.92	92.49
PGBD	1.28	89.60	0.83	89.04	1.15	92.35
Ours	**0.20**	**93.58**	**0.00**	93.02	**0.24**	93.14

Table 4: Comparison of backdoor defense performance under coexisting attacks on CIFAR-10 with ResNet-18.

Defense	BadNets +Blended		SIG +WaNet		BadNets +WaNet+SSBA	
	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑
No defense	99.94	94.07	99.99	94.52	100.00	93.05
NC	1.00	93.53	63.14	92.91	0.99	93.22
FP	3.14	93.72	21.54	94.07	46.53	93.74
NAD	6.39	93.47	2.82	92.59	10.64	93.28
FT-SAM	1.23	93.54	25.89	94.19	0.94	93.21
RNP	8.68	86.63	42.34	93.32	96.77	92.95
MM-BM	55.26	91.07	95.58	**94.52**	70.63	93.05
TSBD	4.20	**93.90**	20.08	94.27	4.98	93.34
D3	0.03	93.31	0.08	93.81	0.06	93.02
PGBD	3.04	89.98	0.60	93.05	98.77	93.05
LMR	1.04	91.97	77.79	92.78	56.84	92.61
Ours	**0.00**	93.52	**0.00**	93.58	**0.04**	**93.86**

benchmarks, with average ASRs below 2% while maintaining competitive benign accuracy. Its advantage remains consistent across diverse attacks, datasets, and model architectures, demonstrating robust backdoor mitigation with limited degradation in benign performance.

Performance under Multi-Target Attacks. To further evaluate SCAR under multi-target attacks, we conduct experiments on CIFAR-10 with ResNet-18 using BadNets, Blended, and WaNet, comparing against baseline defenses that support multi-target settings, where ASR is averaged over all

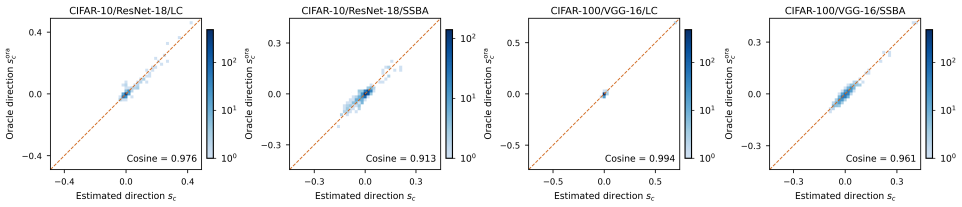

Figure 3: Visualization of the consistency between the estimated and oracle shortcut-related directions on CIFAR-10/ResNet-18 and CIFAR-100/VGG-16 under LC and SSBA attacks.

target classes. As shown in Tab. 3, SCAR achieves strong backdoor suppression while maintaining competitive benign accuracy, benefiting from its class-wise design that naturally accommodates multiple suspicious targets without modifying non-suspicious classifier parameters.

Performance under Coexisting Attacks. To evaluate SCAR under coexisting backdoor attacks, which require simultaneously suppressing multiple heterogeneous backdoors within a single model, we consider both dual-attack and triple-attack settings on CIFAR-10 with ResNet-18. Tab. 4 shows that SCAR maintains strong performance across all settings. This is because SCAR does not need to explicitly disentangle individual backdoors; instead, it captures their aggregated target-class abnormality through a shortcut-related direction for targeted revision. These results demonstrate its robustness to coexisting backdoor attacks.

Table 5: Ablation study on the two components of the class-wise assessment score with VGG-16 on CIFAR-100.

$\mathcal{A}_c^{\text{st}}$	$\mathcal{A}_c^{\text{dy}}$	FPR95↓	AUROC↑
✓		7.58	96.37
	✓	10.59	95.56
✓	✓	**2.69**	**98.54**

5.3 Ablation Studies and More Analysis

Effect of Two Components of the Assessment Score. To investigate the effect of the two components in the class-wise assessment score, we evaluate VGG-16 on CIFAR-100 using 30 backdoored models per attack across ten attacks and 30 clean models. Tab. 5 reports the average results across all attacks. Either component alone is effective, while their combination further improves detection, reducing FPR95 and increasing AUROC by at least 4.89% and 2.17%, respectively, demonstrating their complementarity for reliable backdoor detection.

Effect of Two Components of the Revision Direction. To investigate the roles of the two components in the revision direction, we conduct an ablation study on CIFAR-100 with VGG-16 and report the average ASR and ACC over ten attacks. As shown in Tab. 6, the shortcut-related direction s_c provides the dominant backdoor suppression effect, whereas the class geometry direction p_c offers weaker suppression while better preserving benign accuracy. Their combination further reduces ASR and improves ACC, indicating that incorporating the class-geometry direction helps refine the revision while better preserving benign predictions.

Table 6: Ablation study on the two components of the revision direction with VGG-16 on CIFAR-100.

s_c	p_c	ASR↓	ACC↑
✓		1.68	70.33
	✓	30.03	71.22
✓	✓	**0.93**	**72.03**

Visualization of Shortcut-Direction Consistency. To qualitatively evaluate the estimated shortcut-related direction, we compare it with an oracle direction obtained with trigger information on CIFAR-10/ResNet-18 and CIFAR-100/VGG-16 under LC and SSBA attacks. As shown in Fig. 3, the estimated direction closely matches the oracle direction, supporting the effectiveness of estimating shortcut-related directions from class-specific weights and feature-wise variability.

6 Conclusion

We presented SCAR, a unified post-training framework that connects class-wise backdoor assessment with targeted model revision for models acquired from untrusted sources. SCAR enables lightweight recovery using only unlabeled downstream data, without requiring the original training data, trusted labels, or trigger information. However, SCAR primarily targets data-poisoning backdoors, and its applicability to direct weight-poisoning attacks remains to be further investigated; see Appendix A.2 for further discussion.

AI USE STATEMENT

LLMs were used to assist with language refinement, manuscript organization, and selected research tasks such as brainstorming experimental designs and debugging. AI-generated outputs were treated as auxiliary suggestions and did not substitute for scientific judgment or validation. All cited literature was independently verified by the authors, and all methodological decisions, implementations, experimental results, analyses, and conclusions were independently reviewed and validated by the authors.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. It does not involve human subjects, private or sensitive user data, or potentially harmful applications. Our method is designed to improve the security and trustworthiness of deep neural networks by assessing and repairing backdoored models acquired from untrusted sources, and does not introduce or facilitate new attack capabilities. All datasets used in our experiments are publicly available and widely adopted in the backdoor literature. The study focuses on defensive model auditing and recovery, with the goal of reducing the risks associated with deploying potentially compromised models.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we report the datasets, model architectures, attack settings, evaluation metrics, implementation details, and key hyperparameters in Sec. 5 and the supplementary material. All experiments follow fixed evaluation protocols, and the main configurations used for backdoor assessment and model recovery are explicitly specified. The complete source code, configuration files, and reproduction instructions will be released publicly upon acceptance.

REFERENCES

Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, Avani Gupta, and PJ Narayanan. Prototype guided backdoor defense via activation space manipulation. In *ICCV*, pp. 2195–2205. IEEE, 2025.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *ICIP*, pp. 101–105. IEEE, 2019.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.

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.

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 *CVPR*, pp. 24798–24809. IEEE, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *ICCV*, pp. 11946–11956. IEEE, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Chong Fu, Xuhong Zhang, Shouling Ji, Ting Wang, Peng Lin, Yanghe Feng, and Jianwei Yin. {FreeEagle}: Detecting complex neural trojans in {Data-Free} cases. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 6399–6416, 2023.

Yudong Gao, Honglong Chen, Peng Sun, Zhe Li, Junjian Li, and Huajie Shao. Energy-based backdoor defense without task-specific samples and model retraining. In *ICML*, volume 235, pp. 14611–14637, 2024.

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 *CVPR*, pp. 24564–24573. IEEE, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. Handcrafted backdoors in deep neural networks. In *NeurIPS*, volume 35, pp. 8068–8080, 2022.

Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *CVPR*, pp. 298–307. IEEE, 2020.

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

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jingtai Li, Xiujiu Yuan, Jiwei Tian, Shiwei Lu, and Dengxiu Yu. Upgp:backdoor defense via unlearning perturbation and orthogonality-constraint gradient projection. *Pattern Recognition*, 176: 113211, 2026a.

Weijun Li, Ansh Arora, Xuanli He, Mark Dras, and Qiongkai Xu. Defending against backdoor attacks via module switching. In *ICLR*, volume 2026, pp. 125805–125845, 2026b.

Xiang Li, Ying Meng, Junming Chen, Lannan Luo, and Qiang Zeng. {Rowhammer-Based} trojan injection: One bit flip is sufficient for backdooring {DNNs}. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6319–6337, 2025.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021a.

Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *ICML*, pp. 19837–19854. PMLR, 2023.

Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35(1):5–22, 2022.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, pp. 16443–16452. IEEE, 2021b.

Weilin Lin, Li Liu, Shaokui Wei, Jianze Li, and Hui Xiong. Unveiling and mitigating backdoor vulnerabilities based on unlearning weight changes and backdoor activeness. In *NeurIPS*, volume 37, pp. 42097–42122, 2024.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018a.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018b.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 1265–1282, 2019.

Patrick Mair and Rand Wilcox. Robust statistical methods in r using the wrs2 package. *Behavior research methods*, 52(2):464–488, 2020.

Naren Manoj and Avrim Blum. Excess capacity and backdoor poisoning. In *NeurIPS*, volume 34, pp. 20373–20384, 2021.

Mantas Mazeika, Dan Hendrycks, Huichen Li, Xiaojun Xu, Sidney Hough, Andy Zou, Arezoo Rajabi, Qi Yao, Zihao Wang, Jian Tian, et al. The trojan detection challenge. In *NeurIPS*, pp. 279–291. PMLR, 2023.

Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *NeurIPS*, volume 36, pp. 75286–75306, 2023.

Hossein Mirzaei, Ali Ansari, Bahar D Nia, Mojtaba Nafez, Moein Madadi, Sepehr Rezaee, Zeinab S Taghavi, Arad Maleki, Kian Shamsaie, Jafar Habibi, et al. Scanning trojaned models using out-of-distribution samples. In *NeurIPS*, volume 37, pp. 132545–132582, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Miao, Rong Jin, and Gang Hua. Progressive backdoor erasing via connecting backdoor and adversarial attacks. In *CVPR*, pp. 20495–20503, 2023.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *ICLR*, 2021.

Lu Pang, Tao Sun, Haibin Ling, and Chao Chen. Backdoor cleansing with unlabeled data. In *CVPR*, pp. 12218–12227. IEEE, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8026–8037, 2019.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *CVPR*, pp. 13198–13207, 2020.

Peter J Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 8(2):e1236, 2018.

K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In *NeurIPS*, pp. 19165–19178, 2022.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

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.

Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1994–2012. IEEE, 2024.

Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *CVPR*, pp. 15054–15063. IEEE, 2022.

Shaokui Wei, Jiayin Liu, and Hongyuan Zha. Backdoor mitigation by distance-driven detoxification. In *ICCV*, pp. 4465–4474. IEEE, 2025.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Shenm Chao. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision (IJCV)*, 2025.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *NeurIPS*, volume 34, pp. 16913–16925, 2021.

Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 103–120. IEEE, 2021.

Xiaoyun Xu, Zhuoran Liu, Stefanos Koffas, Shujian Yu, and Stjepan Picek. Ban: Detecting backdoors activated by adversarial neuron noise. In *NeurIPS*, volume 37, pp. 114348–114373, 2024.

Xiaoyun Xu, Zhuoran Liu, Stefanos Koffas, and Stjepan Picek. Towards backdoor stealthiness in model parameter space. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2863–2876, 2025.

Zhiguo Yang, Dongsheng Xu, Ruizhi Zhong, Jiacheng Pi, Xingxing Huang, and Wenjie Ruan. Logit-margin repulsion for backdoor defense. In *CVPR*, pp. 34918–34928, 2026.

Haojie Yuan, Bin Benjamin Zhu, Qi Chu, Tao Gong, Nenghai Yu, and Dongmei Zhang. Abdp: Adversarial backdoor detection and purification. *IEEE Transactions on Information Forensics and Security*, 2026.

Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, 2022.

Hanlei Zhang, Yijie Bai, Yanjiao Chen, Zhongming Ma, and Wenyuan Xu 0001. Barbie: Robust backdoor detection based on latent separability. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025*, 2025.

Quan Zhao, Linkang Du, Yuntao Wang, Zhou Su, Zheng Li, Xiangshan Gao, and Yang Zhang. Unveiling the pitfalls of {Data-Free} backdoor detection against {Pre-Trained} models. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 261–280, 2026.

Songzhu Zheng, Yikai Zhang, Hubert Wagner, Mayank Goswami, and Chao Chen. Topological detection of trojaned neural networks. In *NeurIPS*, volume 34, pp. 17258–17272, 2021.

Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, pp. 4443–4454. IEEE, 2023.

Table 7: Ablation study on the margin and alignment terms in prediction preference score on CIFAR-100 with VGG-16.

Margin	Alignment	FPR95↓	AUROC↑
✓		18.33	92.92
	✓	17.33	91.86
✓	✓	**2.69**	**98.54**

Table 9: Effect of the filtering strategy under different downstream poisoning rates on CIFAR-100 with VGG-16, averaged over ten attacks.

Contam. Rate	Contam. Removed	Clean Removed	ASR↓	ACC↑
1%	71.62%	0.76%	0.74	71.96
5%	71.17%	0.76%	0.99	72.04
10%	71.18%	0.76%	0.93	72.03
20%	71.16%	0.77%	0.88	71.76
50%	71.14%	0.78%	13.32	70.95

A Appendix

A.1 More Ablation Studies and Analysis

Effect of Components in Prediction Preference Score. To investigate the roles of the two terms in the prediction preference score, we separately ablate the logit margin increase and representation-shift alignment on CIFAR-100 with VGG-16. As shown in Tab. 7, either term alone provides useful detection cues, while combining them substantially improves both FPR95 and AUROC. This indicates that jointly considering margin changes and representation alignment helps suppress incidental responses and yields more reliable evidence of abnormal prediction preference.

Effect of Input Transformation Number. To investigate the effect of the number of input transformations, we exhaustively evaluate all 31 non-empty subsets of five candidate transformations on CIFAR-100 with VGG-16. Tab. 8 reports the average performance over all subsets of size J, together with their variation. SCAR maintains strong detection performance across different values of J, indicating limited sensitivity to the number of transformations. Meanwhile, the overall performance improves as more transformations are incorporated, with all five achieving the best FPR95 and AUROC. This suggests that the transformations provide complementary views of prediction responses. We therefore use all five throughout the experiments to obtain diverse evidence without dataset-specific subset selection.

Table 8: Ablation study on the number of input transformations on CIFAR-100 with VGG-16.

J	Subsets	FPR95↓	AUROC↑
1	5	$6.07_{(\pm 0.65)}$	$94.52_{(\pm 1.53)}$
2	10	$5.33_{(\pm 0.79)}$	$95.28_{(\pm 1.59)}$
3	10	$4.30_{(\pm 0.69)}$	$95.50_{(\pm 1.27)}$
4	5	$4.40_{(\pm 0.44)}$	$96.09_{(\pm 0.42)}$
5	1	**2.69**	**98.54**

Effect of the Filtering Strategy under Different Downstream Contamination Rates. To evaluate the filtering strategy under varying downstream contamination rates, we vary the proportion of contaminated samples from 1% to 50% on CIFAR-100 with VGG-16 and report results averaged over ten attacks. We use the same filtering threshold as in the main experiments across all contamination rates. As shown in Tab. 9, the filter consistently removes about 71% of contaminated samples while discarding less than 0.8% of clean samples across all settings. SCAR maintains ASR below 1% with stable ACC for poisoning rates up to 20%. At the extreme 50% contamination level, although the removal rate remains comparable, the much larger initial number of contaminated samples leaves substantially more residual poisoned samples after filtering, which can interfere with the subsequent revision and increase ASR. In such cases, collecting additional, predominantly clean downstream data may reduce the relative influence of residual poisoned samples and provide more reliable statistics for model revision. Overall, these results demonstrate that the filtering strategy effectively suppresses contamination with minimal clean-data loss and supports robust model recovery over a broad range of downstream poisoning rates.

Table 10: Comparison of backdoor defense performance across different methods using ResNet-34 on TinyImageNet.

Defense	BadNets		Blended		TrojanNN		LC		SIG		SSBA		WaNet		BPP		LOTUS		Grond		Average	
	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑	ASR↓	ACC↑
No Defense	100.00	61.94	99.91	62.07	99.89	61.24	99.78	63.20	96.31	62.98	100.00	61.52	95.33	62.43	99.68	58.17	96.00	62.90	55.03	63.60	94.19	62.01
FP	82.93	56.86	5.71	57.25	83.52	56.15	3.01	58.67	7.84	58.28	41.11	55.94	0.56	58.87	0.85	54.08	**0.00**	59.02	31.84	59.61	25.74	57.47
NC	35.99	58.28	0.13	59.52	0.28	59.81	0.49	61.67	2.11	58.04	0.69	59.71	0.63	61.14	0.16	53.25	**0.00**	60.25	1.32	59.21	4.18	59.09
NAD	99.86	60.85	31.51	61.07	99.71	60.68	99.89	62.36	91.78	**63.23**	27.62	60.59	1.05	**62.90**	0.18	54.16	68.00	61.90	59.27	**63.86**	57.89	61.16
FT-SAM	100.00	61.02	0.69	54.21	3.17	53.04	7.21	55.04	43.71	55.35	3.91	49.74	2.18	55.67	99.68	58.17	**0.00**	55.62	29.97	56.20	29.05	55.41
RNP	**0.00**	59.65	**0.00**	56.74	0.05	57.41	**0.00**	60.09	**0.00**	51.05	**0.00**	54.86	95.27	62.42	99.68	58.17	50.00	58.73	**1.20**	54.65	24.62	57.38
MM-BM	99.26	61.06	99.92	**62.07**	99.89	60.36	94.53	62.21	96.42	63.05	100.00	**61.45**	95.21	62.54	99.52	57.06	96.00	62.38	55.29	62.44	93.60	61.46
TSBD	0.13	58.91	5.37	58.48	17.99	57.86	16.11	60.03	9.27	59.65	23.91	58.13	0.42	59.76	**0.07**	48.20	2.00	59.74	2.98	50.14	7.83	57.09
D3	0.01	51.35	0.09	50.32	0.02	50.80	**0.00**	54.08	**0.00**	53.39	4.47	50.75	0.21	51.58	99.68	58.17	**0.00**	51.88	32.18	57.57	13.67	52.99
PGBD	11.61	53.47	1.15	55.40	24.66	44.76	0.01	49.76	11.11	44.05	2.18	48.75	0.93	55.57	1.74	44.06	2.00	51.32	18.00	57.23	7.34	50.44
LMR	30.51	59.59	2.15	60.92	67.56	59.46	99.48	62.43	75.67	61.84	10.31	60.04	**0.18**	62.69	0.23	52.58	4.00	61.72	37.56	61.02	32.77	60.23
Ours	**0.00**	**61.23**	**0.00**	60.28	**0.01**	**61.01**	**0.00**	**62.91**	**0.00**	62.57	**0.00**	60.75	0.44	**62.90**	6.48	**58.20**	4.30	**62.62**	5.85	62.60	**1.71**	**61.51**

A.2 Additional Quantitative Results

Additional Backdoor Defense Results. To further evaluate the general applicability of SCAR, we conduct additional experiments on TinyImageNet with ResNet-34. As shown in Tab. 10, although some baselines achieve the best ACC or ASR under individual attacks, their performance varies considerably across different attack settings. In contrast, SCAR maintains consistently effective backdoor mitigation while preserving competitive benign accuracy across diverse attacks, further demonstrating its broad applicability across datasets and model architectures.

Generalization to Weight-poisoning Backdoor Attacks. To investigate whether SCAR generalizes to weight-poisoning attacks, we evaluate it against TBT (Rakin et al., 2020) and OneFlip (Li et al., 2025) on CIFAR-10 with ResNet-18. Tab. 11 reports the auditing results in terms of the localized suspicious class set $\hat{\mathcal{T}}$, together with ASR and ACC after model revision. Although SCAR correctly identifies the ground-truth target class as the most suspicious class for both attacks, its subsequent revision provides only limited mitigation. This limitation likely stems from the different mechanisms used to implant the backdoor. SCAR is motivated by the observation that backdoor training introduces an additional trigger-to-target shortcut, which can leave target-class signatures in classifier weights, class preference, and feature responses. By contrast, backdoors introduced through direct weight manipulation may not produce the same learned shortcut structure and are therefore less amenable to our revision mechanism. Extending SCAR to effectively address weight-poisoning attacks therefore remains an important direction for future work, and we will release our code to facilitate further research in this direction.

Table 11: Backdoor defense performance of SCAR against the weight-poisoning attacks TBT and OneFlip, using ResNet-18 trained on CIFAR-10.

Method	TBT (Target: 2)			OneFlip (Target: 6)		
	$\hat{\mathcal{T}}$	ASR↓	ACC↑	$\hat{\mathcal{T}}$	ASR↓	ACC↑
No Defense	—	96.72	90.64	—	100.00	86.39
Ours	2, 9	64.44	91.57	6, 2	100.00	86.39

A.3 Baseline Methods and Implementation Details

Attack Baselines. We evaluate our approach against ten representative backdoor attacks spanning diverse trigger designs and poisoning mechanisms. BadNets (Gu et al., 2019) inserts a fixed patch trigger into poisoned samples and relabels them to an attacker-specified target class, whereas Blended (Chen et al., 2017) improves trigger stealthiness by blending a predefined pattern with benign images. TrojanNN (Liu et al., 2018b) generates Trojan triggers by inverting selected internal neurons and retrains the model on reverse-engineered inputs to implant backdoor behavior. To reduce the perceptibility of poisoned samples, LC (Turner et al., 2019) constructs label-consistent poisons using adversarial perturbations or generative models, SIG (Barni et al., 2019) embeds a sinusoidal signal into target-class training samples without changing their labels, and SSBA (Li et al., 2021b) employs sample-specific invisible triggers generated through image steganography. Beyond additive triggers, WaNet (Nguyen & Tran, 2021) uses subtle image warping, while BPP (Wang et al., 2022) exploits image quantization and dithering together with contrastive adversarial learning to establish an imperceptible backdoor. LOTUS (Cheng et al., 2024) partitions victim-class samples

and assigns partition-specific triggers through a trigger-focusing mechanism, making the backdoor dependent on the correspondence between samples and triggers. Grond (Xu et al., 2025) further targets parameter-space stealthiness through adversarial backdoor injection that adaptively constrains backdoor-induced parameter changes during poisoning.

Detection Baselines. We compare SCAR with six representative backdoor detection methods spanning neuron-level inspection, parameter perturbation, optimization-based analysis, and representation-level probing. ABS (Liu et al., 2019) identifies suspicious neurons by systematically stimulating internal activations and validates them through trigger reverse engineering, while BAN (Xu et al., 2024) perturbs model parameters to expose latent backdoor behavior and detects compromised models from the resulting abnormal predictions. MM-BD (Wang et al., 2024) instead optimizes inputs to estimate class-wise maximum logit margins and identifies anomalous target classes, whereas TRODO (Mirzaei et al., 2024) probes Trojan-induced blind spots using out-of-distribution samples and exploits abnormal prediction responses under adversarial perturbations. BARBIE (Zhang et al., 2025) identifies anomalous target classes through relative competition scores derived from inverted latent representations, while SPD (Zhao et al., 2026) uses abnormal convergence speeds during class-wise latent representation construction as the detection signal.

Defense Baselines. We compare SCAR with ten post-training defenses spanning neuron pruning, trigger-guided mitigation, knowledge distillation, fine-tuning, activation suppression, and parameter- or representation-level purification. Among pruning-based methods, Fine-Pruning (FP) (Liu et al., 2018a) removes neurons with low activation on clean inputs and subsequently fine-tunes the pruned model, while RNP (Li et al., 2023) exposes backdoor-related neurons through asymmetric unlearning and recovery before pruning them. Trigger- or representation-guided approaches include NC (Wang et al., 2019), which reverse-engineers class-specific triggers for backdoor mitigation; NAD (Li et al., 2021a), which transfers intermediate-layer attention from a clean-data-fine-tuned teacher; and PGBD (Amula et al., 2025), which regularizes post-hoc fine-tuning using class prototypes and target-directed prototype activation vectors. Several methods instead intervene through optimization or activation control: FT-SAM (Zhu et al., 2023) incorporates sharpness-aware minimization into clean-data fine-tuning, MM-BM (Wang et al., 2024) learns activation bounds from clean data and clips excessive neuron activations at inference time, and TSBD (Lin et al., 2024) combines clean unlearning, suspicious-subweight reinitialization, and activeness-aware fine-tuning. Finally, D3 (Wei et al., 2025) drives the model away from its compromised parameter state while constraining clean-performance degradation, whereas LMR (Yang et al., 2026) enlarges logit-margin separation and removes backdoor-related channels through selective pruning.

Implementation Details. For the above methods, we adopt the recommended configurations from BackdoorBench (Wu et al., 2025) whenever available, and otherwise strictly follow the settings reported in the original papers.

A.4 More Visualization Results

Visualization of Shortcut-Related Feature Responses. To examine whether the estimated shortcut-related directions capture backdoor-specific responses, we visualize the feature responses of triggered samples and their corresponding benign counterparts along the selected shortcut-related dimensions. We conduct this analysis on CIFAR-10 with VGG-16 and ImageNet-1k with ResNet-50 under BadNets, Blended, and SSBA attacks. As shown in Figs. 4 and 5, triggered samples generally exhibit stronger responses than their benign counterparts across these dimensions, indicating that the estimated directions effectively capture feature changes associated with the trigger-to-target shortcut. The consistent pattern across different datasets, architectures, and attacks further supports the generality of the estimated shortcut-related directions.

Visualization of Input Transformations. To illustrate the input transformations used for dynamic prediction preference, we visualize the five transformations on CIFAR-10 and ImageNet-1k. As shown in Figs. 6 and 7, they perturb the inputs from complementary perspectives, including random perturbation, local occlusion, sample mixing, frequency variation, and geometric distortion, while largely preserving the original semantic content. This diversity enables SCAR to probe prediction responses under different forms of mild input variation.

Visualization of Shortcut-Direction Consistency. To further examine the consistency between the estimated and oracle shortcut-related directions, we provide additional visualizations on CIFAR-

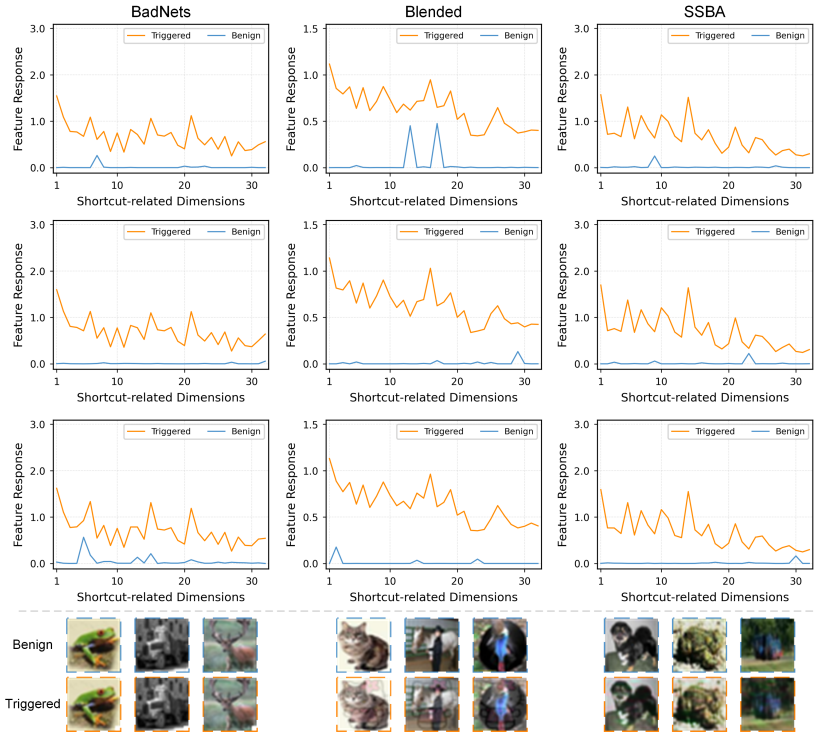

Figure 4: Visualization of triggered and corresponding benign sample responses along the estimated shortcut-related feature dimensions on CIFAR-10 with VGG-16 under BadNets, Blended, and SSBA attacks.

100 with VGG-16 and ImageNet-1k with ResNet-50 across various attacks. As shown in Fig. 8, the estimated directions exhibit clear positive correspondence with their oracle counterparts, with most feature dimensions concentrated around the diagonal. This consistency is observed across different attacks and model scales, and remains evident even on the more challenging ImageNet-1k benchmark. These results further support that SCAR can reliably capture shortcut-related feature directions without access to trigger information.

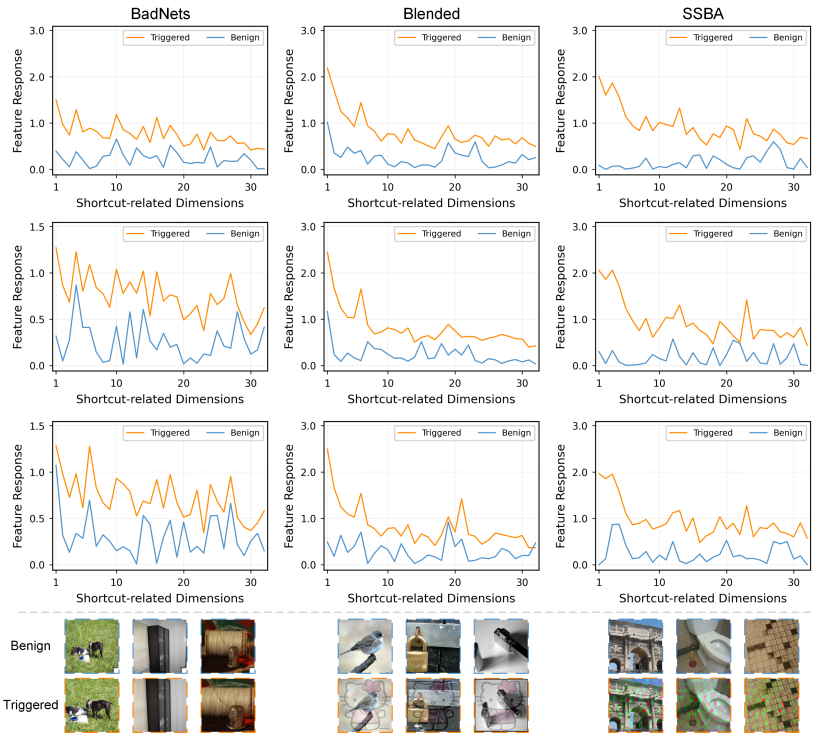

Figure 5: Visualization of triggered and corresponding benign sample responses along the estimated shortcut-related feature dimensions on ImageNet-1k with ResNet-50 under BadNets, Blended, and SSBA attacks.

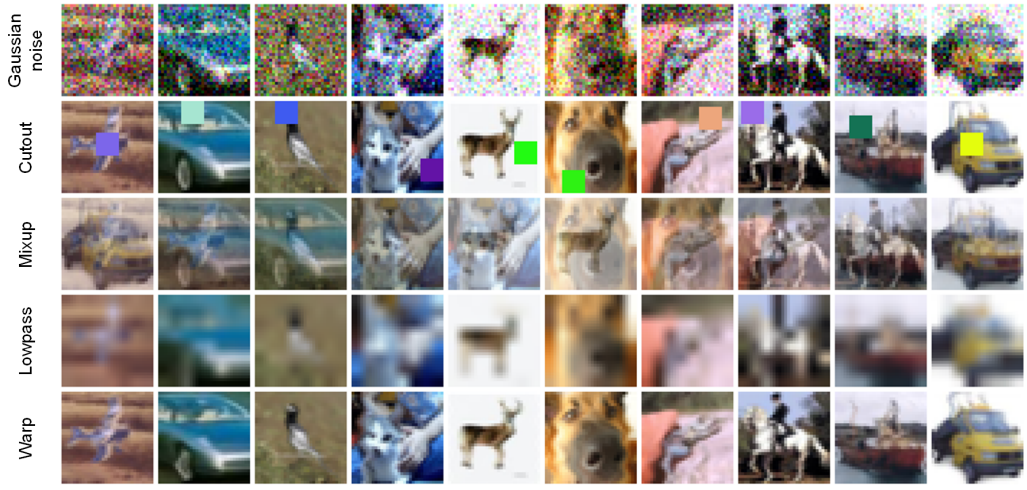

Figure 6: Visualization of the five input transformations used in SCAR on CIFAR-10.



Figure 7: Visualization of the five input transformations used in SCAR on ImageNet-1k.

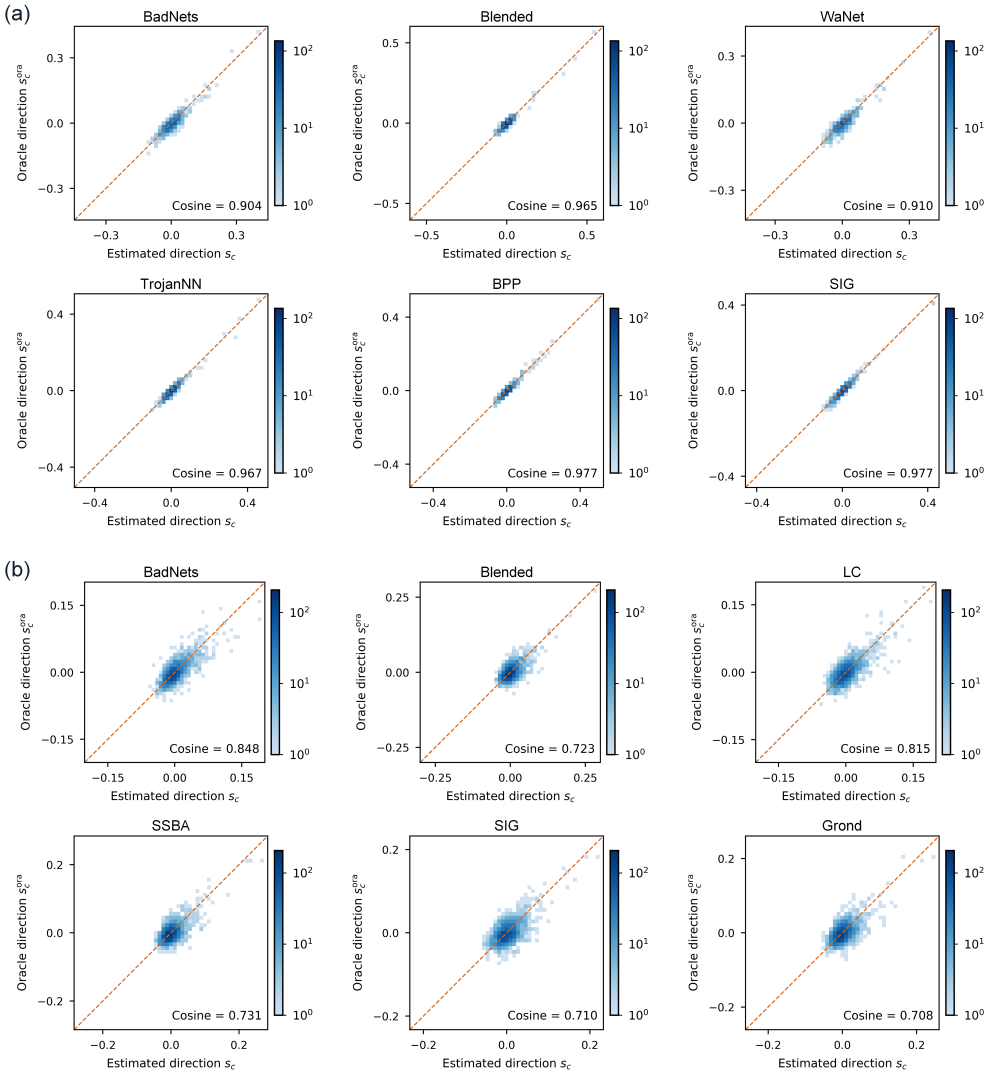

Figure 8: Visualization of the consistency between the estimated and oracle shortcut-related directions on (a) CIFAR-100 with VGG-16 and (b) ImageNet-1k with ResNet-50 under various attacks.