Margin-Guided Clean-Label Backdoor Attacks with Weak Implantation and Strong Activation

Yi Yang, Jinyang Huang*, Feng-Qi Cui, Binbin Liu, Lei Shi, Weihua Li,

Quan Bai, Xiaokang Zhou and Meng Li*

***Abstract*—Training deep neural networks on third-party data exposes learning pipelines to clean-label backdoor attacks, where poisoned samples preserve their semantic labels and can evade routine human inspection. However, clean-label backdoors remain difficult to implant under strict perturbation budgets. Since poisoned samples are drawn from the target class and retain their correct labels, victim models can fit them using natural target-class features, while a small trigger may provide insufficient decision evidence and be ignored. Existing attacks therefore face a tension between poisoning-stage stealthiness and attack effectiveness: reducing the train-time perturbation improves imperceptibility, but weakens the learned trigger–target association. This paper presents MISA, a margin-guided clean-label backdoor attack with weak implantation and strong activation. MISA implants a small bounded trigger into target-class samples during training and applies an amplified version of the same trigger at inference time. Instead of optimizing the trigger through target-label confidence, MISA enlarges the logit gap between the target class and the strongest competing class, directly targeting the decision-level condition for target-class prediction. Analytically, we show that enlarging this margin monotonically tightens an upper bound on the target-label cross-entropy, while a first-order margin expansion explains why inference-time amplification can strengthen activation when the trigger direction consistently increases the target margin. We evaluate MISA on CIFAR-10, CIFAR-100, and ImageNet-100, showing that it outperforms representative clean-label attacks and remains effective against evaluated model-, training-, and data-oriented defenses. These results demonstrate that margin-guided weak implantation provides an effective route to clean-label backdoor attacks under strict training-time perturbation constraints while preserving poisoning-stage visual stealthiness.**

***Index Terms*—Backdoor attack, Clean-label attack, AI security**

I. Introduction

Deep neural networks (DNNs) are commonly trained on large-scale datasets collected from public repositories, third-party data providers, or distributed users [1], [2], [3], [4]. While this data-driven paradigm reduces the cost of data collection and annotation, it also exposes the training pipeline to data-poisoning threats. Backdoor attacks exploit this vulnerability by injecting maliciously modified samples into the training set. A model trained on the poisoned data behaves normally on clean inputs, but predicts an adversary-specified target label whenever a predefined trigger appears at inference time [5], [6], [7].

Backdoor attacks are commonly categorized into dirty-label and clean-label attacks according to whether poisoned labels are modified. Dirty-label attacks poison non-target samples and relabel them as the target class [5], [6], [8], creating a strong trigger–label association but also causing image-label inconsistency that can be exposed by human inspection or data auditing [9]. Clean-label attacks avoid this weakness by poisoning only target-class samples while keeping their labels unchanged [9], [10], [11]. Since poisoned image-label pairs remain semantically consistent, clean-label attacks are more stealthy in outsourced or user-contributed training pipelines.

However, clean-label attacks are substantially harder to implant. In dirty-label attacks, the trigger often becomes the most reliable feature explaining why visually non-target samples are assigned the target label. In clean-label attacks, poisoned samples already contain natural target-class features. Thus, during standard supervised training, the victim model can fit these samples using their original semantic evidence without necessarily learning the injected trigger. This competition between natural class semantics and trigger-induced feature is a key reason why clean-label backdoors are difficult to implant [9], [12].

Existing clean-label attacks have addressed this problem from different perspectives. Label-Consistent attacks weaken target-class semantics before trigger insertion [9]; SIG uses sinusoidal signals as clean-label triggers [10]; hidden-trigger and feature-collision methods construct poisoned samples that mimic triggered non-target samples in feature space [13], [14]; and recent methods improve clean-label poisoning through invisible-noise triggers [11], target-oriented universal triggers [15], adaptive trigger generators [16], frequency-domain constrained triggers [17], or theoretically motivated trigger constructions [18]. These methods have substantially advanced clean-label backdoor attacks, but an important question remains: *what makes a weak clean-label trigger learnable during training and reliably activatable during inference?*

This question is important because clean-label attacks impose different requirements at training and inference time. During training, poisoned samples may be inspected by human annotators, automatic filters, or data sanitization methods;

Yi Yang, Jinyang Huang, Binbin Liu, Lei Shi and Meng Li are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, China.

Feng-Qi Cui is with School of Information Science and Technology, University of Science and Technology of China, Hefei, 230026, China.

Weihua Li is with School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, 1010, China.

Quan Bai is with School of Information and Communication Technology, University of Tasmania, Hobart, 7001, Australia.

Xiaokang Zhou is with the Faculty of Business Data Science, Kansai University, Osaka 565-8585, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan

*Corresponding author: Jinyang Huang (Email: hjy@hfut.edu.cn), Meng Li (Email: mengli@hfut.edu.cn)

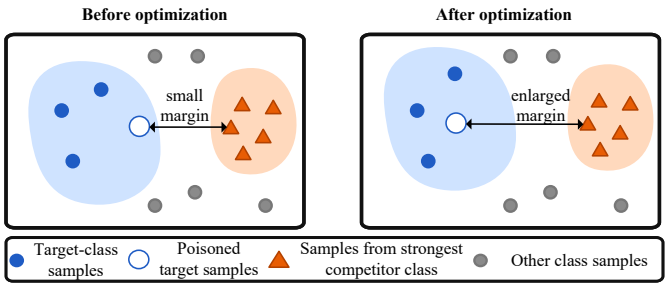


Fig. 1: Conceptual illustration of margin-guided trigger optimization. MISA optimizes the trigger to enlarge the target margin between the target class and the strongest competing class, making the poisoned target samples more strongly associated with the target label.

thus, the trigger should remain weak and inconspicuous. At inference time, however, the attacker controls the input and can strengthen the trigger to activate the implanted backdoor. Prior clean-label attacks such as Narcissus [15] exploit this asymmetry by optimizing a target-class trigger and amplifying it during inference. Nevertheless, their trigger optimization is mainly driven by target-class confidence. Under a small perturbation budget, this can be insufficient: once target-class confidence becomes saturated on target-class samples, the optimization signal weakens, while the trigger still needs to compete with natural target-class features and remain effective after inference-time amplification.

To address this problem, we propose MISA, ***M**argin-guided weak **I**mplantation with **S**trong **A**ctivation*, a clean-label backdoor attack that optimizes a weak train-time trigger as a decision-margin-increasing direction. MISA implants a small bounded trigger into target-class samples during training, while applying an amplified version of the same trigger at inference time. As illustrated in Fig. 1, the core of MISA is a target-margin objective that enlarges the logit gap between the target class and its strongest competing class. Unlike probability-oriented objectives that mainly increase the softmax confidence of the target label, the margin objective directly strengthens the decision-level condition for target-class prediction. In this way, MISA explains and improves weak implantation and strong activation by aligning the trigger direction with the target decision margin.

After the trigger is optimized, the adversary constructs the poisoned training set by adding the weak trigger to a small subset of target-class samples without modifying their labels. The victim then trains a model normally using its standard supervised training. At inference time, the attacker applies the same trigger to non-target inputs and amplifies its magnitude within a controlled range. Since the optimized trigger has already been associated with the target class during training, amplification strengthens its decision-level effect and more reliably activates the implanted backdoor.

We conduct extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-100 under practical clean-label poisoning settings. MISA is compared with seven representative clean-label backdoor attacks under the same trigger perturbation budget. The results show that MISA achieves strong attack success while preserving clean accuracy. We further evaluate different victim architectures, poisoning budgets, perturbation magnitudes, and inference-time amplification factors. Ablation studies validate the importance of margin-guided trigger optimization and inference-time amplification. We also evaluate MISA against representative model-oriented, training-oriented, and data-oriented backdoor defenses.

In summary, this paper makes the following contributions:

- To the best of our knowledge, we are the first to explicitly formulate clean-label backdoor attacks under strict ℓ_∞ poisoning budgets as a weak implantation and strong activation problem. This formulation separates the stealth requirement at training time from the attack-effectiveness requirement at inference time.
- MISA, a margin-guided clean-label backdoor attack, is proposed to realize this formulation. The key idea is to optimize the trigger as decision-level evidence for the target class by enlarging the logit gap between the target class and its strongest competing class.
- We further provide rigorous analytical justification for the proposed target-margin objective. The analysis shows that the margin-guided trigger optimization reduces an upper bound on the target-label cross-entropy loss, supports inference-time amplification, and mitigates the saturation issue of probability-based objectives.
- We evaluate MISA across multiple datasets, victim architectures, poisoning budgets, trigger strengths, and defense methods. The results show that MISA achieves higher attack success rates than representative clean-label attacks while preserving clean accuracy, maintaining poisoning-stage visual stealthiness, and remaining effective against the evaluated defenses.

II. Related Work

A. Backdoor Attacks

1) Dirty-label Backdoor Attacks: Backdoor attacks implant hidden behaviors into models during training. A backdoored model behaves normally on clean inputs, but predicts an adversary-specified target label when a trigger appears at inference time. Early attacks, such as BadNets, poison the training set by stamping a predefined trigger onto selected samples and changing their labels to the target class [5]. Subsequent studies improve trigger stealthiness, flexibility, and robustness through blended patterns [6], invisible perturbations [19], [20], warping-based transformations [21], frequency-domain perturbations [22], [23], and so on.

Most of these attacks follow the dirty-label setting, where poisoned samples are assigned labels inconsistent with their original semantics. Although this setting often yields high attack success rates, the resulting image-label inconsistency can be exposed by human inspection or data auditing. This limitation motivates clean-label backdoor attacks, where poisoned samples preserve their ground-truth labels.

2) Clean-label Backdoor Attacks: Clean-label backdoor attacks poison training samples without changing their labels. Compared with poisoned-label attacks, this setting is more

stealthy against human inspection, but also more difficult to realize. Since poisoned samples are usually drawn from the target class, they already contain natural semantic features that support the target label. During training, the victim model may therefore fit the target class using these intrinsic features and ignore the injected trigger.

Existing clean-label attacks address this difficulty from different perspectives. Label-Consistent attacks perturb target-class samples before trigger insertion, weakening the original class evidence and encouraging the model to rely on the trigger [9]. SIG uses sinusoidal signals as clean-label triggers [10]. Hidden Trigger Backdoor Attacks construct poisoned target-class samples whose feature representations collide with triggered source-class samples [13], while Sleeper Agent improves hidden-trigger attacks for models trained from scratch through gradient matching, data selection, and surrogate retraining [14]. Invisible Poison conceals trigger information as imperceptible noise to improve visual stealth [11]. Narcissus shows that arbitrary triggers are inefficient in the clean-label setting and optimizes a target-oriented universal trigger using only target-class data [15]. More recent methods further improve clean-label attack effectiveness or stealthiness through theoretically motivated trigger construction [18], generator-based optimization [16], and frequency-domain trigger suppression [17].

These methods also differ in the space where the trigger is defined. MISA focuses on bounded image-space clean-label attacks, where the poisoned training samples are produced by adding an image-space perturbation whose magnitude can be explicitly controlled under an ℓ_∞ budget. This setting provides a common basis for comparing train-time stealth, because all poisoned samples can be constrained to have the same maximum pixel-wise perturbation. In contrast, semantic or high-level clean-label attacks, such as attribute-level triggers [12] and style-level triggers [24], modify images through structured transformations rather than bounded pixel-wise perturbations. Although these attacks are closely related to clean-label backdoor research, their trigger strength is not naturally measured by the same ℓ_∞ budget. We therefore regard them as complementary to the bounded image-space setting studied in this work.

MISA follows the bounded image-space clean-label setting, but differs from prior attacks in its optimization objective. Existing optimized triggers commonly encourage the model to increase target-class confidence or to align poisoned samples with target-related representations. In contrast, MISA directly optimizes the target margin, which is more more directly captures the decision-level condition for target-class prediction. Consequently, MISA first learns a trigger direction that is aligned with the target decision margin, making the weak train-time trigger learnable during poisoning and more effective when the same trigger is amplified at inference time.

B. Backdoor Defenses

A broad line of work has investigated defenses against backdoor attacks at different stages of the learning pipeline. Model-oriented defenses inspect a trained model by reverse-engineering potential triggers, analyzing abnormal neuron behaviors, or detecting anomalous class-wise statistics [25], [26], [27]. Training-oriented defenses intervene during training or repair a suspicious model after training, for example by isolating suspicious low-loss samples or fine-tuning the model with robust optimization [28], [29]. Data-oriented defenses attempt to identify poisoned samples based on feature-space anomalies, higher-order feature statistics, provenance signals, or inference-time consistency [30], [31], [32], [33], [34].

Although these defenses are effective against several conventional backdoor attacks, clean-label attacks remain challenging because poisoned samples preserve semantic-label consistency and may not form obvious outliers in input, feature, or model-behavior space. This challenge is particularly relevant to our setting, where the training-time trigger is weak, low-amplitude, and distributed over the image. We therefore evaluate MISA against representative model-, training-, and data-oriented defenses in Section VI.

III. Threat Model and Background

A. Threat Model

We consider a poison-only clean-label backdoor attack against image classification models. The victim collects a training dataset from external or partially trusted sources and trains a classifier using a standard supervised learning pipeline. The adversary can inject a small number of poisoned samples into the training set, but cannot modify the victim's model architecture, training objective, optimizer, data augmentation strategy, or deployment procedure.

1) Adversary's goal.: The adversary specifies a target label t and aims to implant a hidden association between a trigger and this target class. A successful backdoored model should satisfy two requirements. First, it should maintain high accuracy on clean test inputs, so that the attack does not noticeably affect normal model behavior. Second, when the adversary applies the trigger to non-target test inputs, the model should predict the target label with high probability.

2) Adversary's capability.: The adversary is restricted to clean-label poisoning. Specifically, the adversary can modify only a limited subset of target-class training samples, and the labels of these samples must remain unchanged. Thus, each poisoned sample remains semantically consistent with its assigned label. The adversary does not poison non-target samples and does not perform label flipping.

The poisoned samples are generated by adding an additive trigger to selected target-class images. Since training data may be inspected before being accepted into the training pipeline, the trigger used during poisoning should remain small and inconspicuous. In our setting, this requirement is enforced by bounding the training-time trigger magnitude.

3) Adversary's knowledge.: We assume that the adversary has access to the clean training dataset for training a surrogate model. The surrogate model is used only during attack construction. The adversary does not require access to the final victim model before poisoning and does not control the victim's training process. The victim model may use a different architecture from the surrogate model.

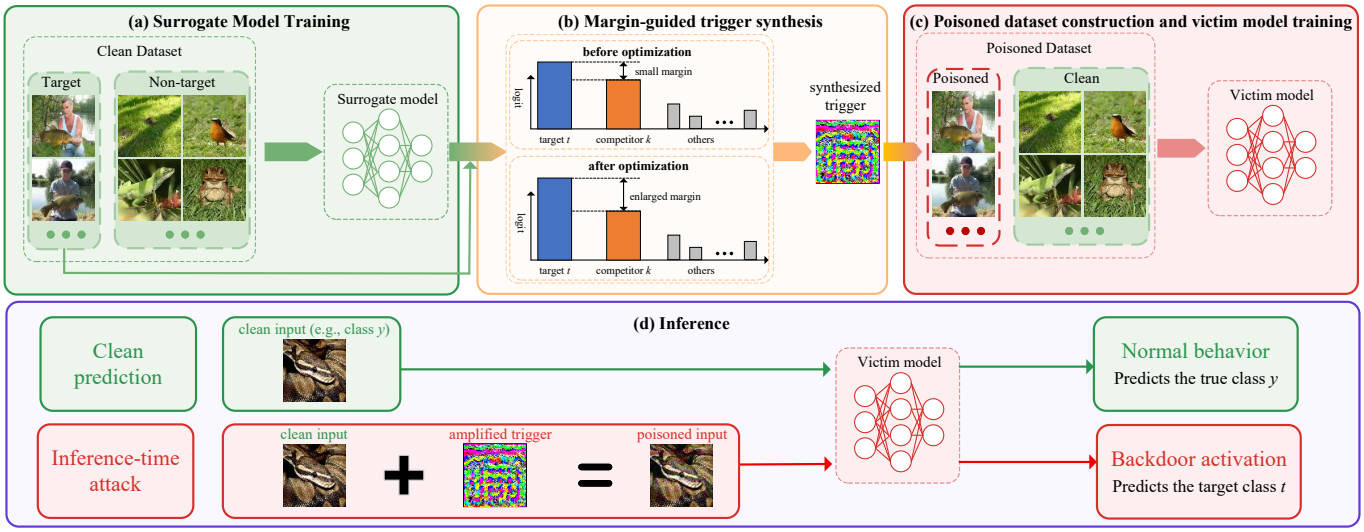


Fig. 2: Overview of the proposed MISA framework. In Stage 1, a surrogate model is trained on the clean dataset. In Stage 2, target-class samples and the surrogate model are used to synthesize a trigger through margin-guided optimization. In Stage 3, the optimized trigger is implanted into selected target-class samples to construct a clean-label poisoned dataset, which is then used by the victim to train the backdoored model. In Stage 4, during inference, the trigger is amplified and applied to test samples to activate the backdoor, causing the victim model to predict the adversary-specified target class.

4) Training–inference asymmetry.: We distinguish the role of the trigger during training and inference. During training, the trigger serves as a weak implantation signal and must remain small for data-auditing stealthiness. During inference, the adversary controls the input and can amplify the same trigger direction to more strongly activate the implanted backdoor.

B. Notation and Poisoning Setup

We next introduce the notation and poisoning setup used throughout the paper. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the clean training set, where $x_i \in [0,1]^{C\times H\times W}$ is an image and $y_i \in \mathcal{Y} = \{1, \dots, K\}$ is its ground-truth label. Let $t \in \mathcal{Y}$ be the adversary-specified target label. The target-class subset is denoted as $\mathcal{D}_t = \{(x_i, y_i) \in \mathcal{D} : y_i = t\}$.

1) Clean-label poisoning.: The adversary selects a subset of target-class samples

$$\mathcal{S}_p \subseteq \mathcal{D}_t, \qquad |\mathcal{S}_p| = m,$$

where m is the number of poisoned samples. For each selected sample $(x_i, t) \in \mathcal{S}_p$, the adversary constructs a poisoned image by adding a universal additive trigger δ:

$$x_i^{\delta} = \mathrm{clip}(x_i + \delta, 0, 1). \tag{1}$$

Here, the same trigger δ is shared by all poisoned samples. The label is preserved as t, satisfying the clean-label constraint. Thus, the poisoned training set is

$$\mathcal{D}_p = (\mathcal{D} \setminus \mathcal{S}_p) \cup \{(x_i^{\delta}, t) : (x_i, t) \in \mathcal{S}_p\}.$$

After the poisoned training set is constructed, it is supplied to the victim, who then trains a model using its standard training algorithm: $f_v = \mathcal{A}(\mathcal{D}_p)$, where $\mathcal{A}$ denotes the victim training algorithm, f_v is the resulting victim model. In the poison-only setting, the adversary can only modify the training data supplied to $\mathcal{A}$.

2) Training-time trigger constraint.: To preserve training-time stealthiness, the trigger used for poisoning is constrained by a small perturbation budget: $\|\delta\|_\infty \le \epsilon$, where ϵ controls the maximum training-time trigger magnitude. This constraint reflects the requirement that poisoned training samples remain inconspicuous under human inspection or automated data auditing. The specific perturbation budget used in our experiments is provided in Section V.

3) Inference-time trigger amplification.: At inference time, the adversary applies the same trigger direction to a non-target test input x, with an amplification factor $\alpha \ge 1$:

$$x^{\alpha\delta} = \mathrm{clip}(x + \alpha\delta, 0, 1). \tag{2}$$

When $\alpha = 1$, the inference-time trigger has the same magnitude as the training-time trigger. When $\alpha > 1$, the trigger is amplified to more strongly activate the learned backdoor.

Directly optimizing the post-training attack behavior is difficult because it depends on the victim's training algorithm and the final trained model. Therefore, practical clean-label attacks usually rely on surrogate objectives during attack construction. In the next section, we introduce a margin-guided surrogate objective that explicitly enlarges the target-class logit gap, enabling weak training-time implantation and stronger inference-time activation.

IV. Method

This section presents MISA, a margin loss-guided clean-label backdoor attack. The attack is designed to address the tension between stealthy training-time poisoning and reliable inference-time activation. During training, the trigger is constrained to be small so that poisoned samples remain inconspicuous. To ensure that this weak perturbation can still

be learned by the victim model, we optimize it by a margin-guided loss. At inference time, the same trigger direction can be amplified to more strongly activate the implanted backdoor. In the rest of this section, we will first describe the detailed attack workflow, then we will present the analysis of why the proposed method is effective.

A. Attack Workflow

MISA follows a weak-implantation and strong-activation design. In the training phase, the adversary injects a small perturbation into a small subset of target-class samples, preserving both label consistency and visual stealthiness. To ensure that this weak perturbation remains learnable, we optimize the trigger direction on a surrogate model to enlarge the target-class logit margin. During inference, the same trigger can be amplified to more reliably activate the implanted backdoor. The overview of MISA is illustrated in Fig. 2.

1) Surrogate model training: Given a clean training set $\mathcal{D}$, the adversary first trains a surrogate classifier f_s. The surrogate is used only for trigger construction and is not involved in victim training. We denote the surrogate as a feature extractor $h_s(\cdot)$ followed by a classification head $g_s(\cdot)$. For an input image x, the surrogate produces

$$h_s(x) \in \mathbb{R}^d, \qquad z_s(x) = g_s(h_s(x)) \in \mathbb{R}^K,$$

where $z_s(x)$ is the logit vector and $z_{s,k}(x)$ denotes the logit of class k.

2) Margin-guided trigger synthesis: For a trigger δ, the triggered sample is denoted as x^δ in Eq. (1). We optimize the trigger to enlarge the logit gap between the target class and the strongest non-target class. Specifically, the trigger is obtained by minimizing the negative target margin:

$$\mathcal{L}_{\text{margin}}(\delta) = -\mathbb{E}_{x\sim\mathcal{D}_t}\left[z_{s,t}(x^\delta) - \max_{k\neq t} z_{s,k}(x^\delta)\right]. \tag{3}$$

Equivalently, we perform projected mini-batch gradient ascent on the target margin $z_{s,t}(x^\delta) - \max_{k\neq t} z_{s,k}(x^\delta)$ under an ℓ_∞ constraint $\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$. The projection $\text{Proj}_\Delta(\cdot)$ can be implemented by clipping each dimension of δ into $[-\epsilon, +\epsilon]$. The optimization procedure is summarized in Algorithm 1.

3) Poisoned dataset construction and victim model training: After obtaining the optimized trigger δ^*, the adversary selects a small subset $\mathcal{S}_p \subseteq \mathcal{D}_t$ of target-class samples and constructs poisoned images by:

$$x_{\text{train}}^{\delta^*} = \text{clip}(x_{\text{train}} + \delta^*, 0, 1), \qquad (x_{\text{train}}, t) \in \mathcal{S}_p.$$

The labels are kept unchanged as t, so the poisoned samples remain clean-label. The poisoned target-class samples are then supplied to the victim together with the remaining training data. The victim trains its model using its standard training pipeline, without any modification to the architecture, loss function, optimizer, or augmentation strategy.

4) Inference-time activation: At inference time, the adversary applies the same trigger direction to a test sample. To strengthen the activation of the learned backdoor, the trigger can be amplified by a factor $\alpha \geq 1$:

$$x_{\text{test}}^{\alpha\delta^*} = \text{clip}(x_{\text{test}} + \alpha\delta^*, 0, 1).$$

Algorithm 1: Trigger Generation Algorithm

Input: $z_s(\cdot)$ (surrogate logit function);
$\mathcal{D}_t$ (target-class training samples);
t (target label);
$\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$ (trigger constraint set)
Output: δ^* (the optimized trigger)
Parameters: $\mathcal{I}$ (total iteration number); η (step size)

$\delta_0 \leftarrow 0^{C\times H\times W}$;
for $i = 0$ **to** $\mathcal{I} - 1$ **do**
  Sample a mini-batch $\mathcal{B}_i$ from $\mathcal{D}_t$;
  $J_i(\delta) = \frac{1}{|\mathcal{B}_i|}\sum_{x\in\mathcal{B}_i}\left[z_{s,t}(x^\delta) - \max_{k\neq t} z_{s,k}(x^\delta)\right]$
  $g_i = \nabla_\delta J_i(\delta)|_{\delta=\delta_i}$
  $\delta_{i+1} \leftarrow \text{Proj}_\Delta(\delta_i + \eta g_i)$
$\delta^* = \delta_{\mathcal{I}}$
return δ^*

When $\alpha = 1$, the inference-time trigger has the same magnitude as the training-time trigger. When $\alpha > 1$, the trigger is amplified to increase the target-class evidence induced by the trigger. This training–inference asymmetry allows the attack to use a weak and stealthy trigger during poisoning while using a stronger activation signal during inference.

B. Analysis

We analyze why margin-guided trigger optimization provides a decision-directed signal for clean-label backdoor learning. The analysis has three goals. First, we relate the target margin to the target-label cross-entropy loss and show that increasing the margin tightens a standard margin-based upper bound on the loss. Second, we give a local explanation of inference-time trigger amplification in terms of the effective perturbation $\alpha\delta$. Third, we compare the margin objective with the cross-entropy-based trigger optimization used in related attacks such as Narcissus [15]. The analysis is intended as an explanation rather than a global guarantee; the transfer of the optimized trigger direction to the victim model and to attack-time samples is validated empirically in Sec. V.

Let $z_k^g(u)$ denote the logit of class k produced by model g for input u. For target class t, we denote the target margin as

$$M_t^g(u) = z_t^g(u) - \max_{k\neq t} z_k^g(u). \tag{4}$$

When the model is clear from context, we omit the superscript g. For analytical clarity, we write the triggered input as $x+\delta$. In the actual implementation, pixel clipping may make the perturbation input-dependent; therefore, the following analysis describes local behavior of the optimized trigger direction before empirical validation under clipping.

1) Target margin and weak-trigger implantation: In a clean-label backdoor attack, poisoned samples are drawn from the target class and retain their correct labels. This improves label consistency and visual stealthiness, but also creates a learning challenge: because the poisoned samples already contain genuine target-class semantics, a victim model may fit them using ordinary class features and ignore a weak trigger.

MISA addresses this issue by optimizing the trigger to enlarge the target margin, making the trigger more directly relevant to target-class fitting.

Proposition IV.1 (Target-margin bound on target-label loss)**.** *For a K-class classifier g and an input u labeled as the target class t, the cross-entropy loss satisfies*

$$\mathcal{L}_{\mathrm{CE}}(g(u), t) \leq \log\left(1 + (K-1)\exp(-M_t(u))\right). \quad (5)$$

Moreover, the right-hand side is strictly decreasing in $M_t(u)$.

Proof. The target-label cross-entropy can be written as

$$\mathcal{L}_{\mathrm{CE}}(g(u), t) = \log\left(1 + \sum_{k \neq t} \exp(z_k(u) - z_t(u))\right). \quad (6)$$

By the definition of $M_t(u)$, we have $z_k(u) - z_t(u) \leq -M_t(u)$ for all $k \neq t$. Hence,

$$\sum_{k \neq t} \exp(z_k(u) - z_t(u)) \leq (K-1)\exp(-M_t(u)), \quad (7)$$

which gives Eq. (5). The derivative of the bound with respect to $M_t(u)$ is negative, so the bound decreases monotonically as the target margin increases. □

Proposition IV.1 shows that maximizing the target margin is consistent with reducing an upper bound on the target-label loss. Therefore, for target-class poisoned samples, increasing $M_t(x+\delta)$ makes the trigger compatible with target-class fitting. If the trigger induces a positive margin gain, i.e.,

$$M_t(x+\delta) \geq M_t(x) + \gamma, \qquad \gamma > 0, \quad (8)$$

then Proposition IV.1 gives a tighter target-label loss bound for the triggered sample.

This result does not prove that the victim model must learn the trigger as a shortcut. Rather, it explains why a margin-optimized trigger is more informative than an arbitrary weak perturbation: it moves target-class samples in a direction that increases the target logit relative to the strongest competing non-target logit. Under a small perturbation budget, this decision-directed signal is important because the trigger has limited capacity and can otherwise be ignored during training.

2) Local view of inference-time amplification: We next analyze why applying a stronger version of the same trigger at inference time can increase attack success. Let f_v denote the trained victim model and consider an attack-time input x. In a local neighborhood where the strongest competing class remains unchanged, the target margin $M_t^{f_v}(x)$ is differentiable with respect to x. For an amplification factor α such that the effective perturbation $\alpha\delta$ remains locally small, the first-order Taylor expansion gives

$$M_t^{f_v}(x + \alpha\delta) = M_t^{f_v}(x) + \alpha \nabla_x M_t^{f_v}(x)^\top \delta + o(\|\alpha\delta\|). \quad (9)$$

Thus, if $\nabla_x M_t^{f_v}(x)^\top \delta > 0$, then increasing the trigger strength initially increases the target margin to first order.

This condition is related to MISA's objective. MISA learns a universal trigger on a surrogate model f_s by maximizing the average target margin over target-class samples:

$$\delta^*_{\mathrm{MISA}} = \arg\max_{\delta \in \Delta} \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} M_t^{f_s}(x+\delta). \quad (10)$$

Under a local linearization, this objective encourages positive average alignment between the trigger direction and the surrogate margin gradient:

$$\mathbb{E}_{x \sim \mathcal{D}_t}\left[\nabla_x M_t^{f_s}(x)^\top \delta\right] > 0. \quad (11)$$

If this margin-increasing direction transfers from the surrogate model to the victim model, and from target-class optimization samples to attack-time samples, then amplifying the same trigger can further increase the victim model's target margin.

This analysis is local with respect to the effective perturbation $\alpha\delta$, rather than the amplification factor α alone. In our setting $\|\delta\|$ is set very small, for example later in the experiment the train-time trigger satisfies $\|\delta\|_\infty \leq 4/255$, so inference-time amplification gives $\|\alpha\delta\|_\infty \leq 4\alpha/255$. Nevertheless, the first-order approximation may become less accurate when the amplified input leaves the local linear region, changes the strongest competing class, crosses activation boundaries, or is affected by pixel clipping. Therefore, we use Eq. (9) as a local explanation of the amplification mechanism and validate the effect empirically under practical amplification factors.

3) Comparison with cross-entropy-based trigger optimization: MISA is related to Narcissus [15], which also synthesizes a universal additive trigger using stochastic gradient descent and uses inference-time trigger amplification to increase attack performance. The main difference is the trigger-optimization objective. Narcissus minimizes the target-label cross-entropy (CE objective) loss on target-class samples:

$$\delta^*_{\mathrm{Nar}} = \arg\min_{\delta \in \Delta} \frac{1}{|\mathcal{D}_t|} \sum_{x \in \mathcal{D}_t} \mathcal{L}_{\mathrm{CE}}\big(f_s(x+\delta), t\big). \quad (12)$$

This objective encourages the trigger to increase the target-class confidence of target-class samples. In contrast, MISA directly optimizes the target margin in Eq. (10), focusing on the logit gap that determines whether the target class dominates the strongest competing class.

The distinction is not that cross-entropy is unrelated to the margin. Proposition IV.1 shows that they are connected. Rather, cross-entropy objective in Eq.(12) optimizes the margin indirectly through softmax-normalized probabilities, which may lead to weak update signals when target-class samples are already classified with high confidence. To see this, let

$$p_k(x+\delta) = \frac{\exp(z_k(x+\delta))}{\sum_{\ell=1}^{K} \exp(z_\ell(x+\delta))} \quad (13)$$

be the softmax probability of class k. The gradient of the target-label cross-entropy loss with respect to the trigger is

$$\nabla_\delta \mathcal{L}_{\mathrm{CE}}(x+\delta, t) = \big(p_t(x+\delta) - 1\big)\nabla_\delta z_t(x+\delta) + \sum_{k \neq t} p_k(x+\delta)\nabla_\delta z_k(x+\delta). \quad (14)$$

For many training samples, a well-trained surrogate model may already produce $p_t(x+\delta) \approx 1$ and $p_k(x+\delta) \approx 0$ for $k \neq t$. In this case, the coefficients in Eq. (14) become small, and such samples may contribute only weakly to the trigger update. By contrast, the target-margin objective directly

TABLE I: Dataset statistics and poisoning configurations.

Dataset	#Classes	Input Size	#Poison. Images	Poisoning Rate	#Train. Images
CIFAR-10 [35]	10	32×32	200	0.40%	50,000
CIFAR-100 [35]	100	32×32	300	0.60%	50,000
IN-100 [36]	100	224×224	400	0.31%	129,395

enlarges the gap between the target logit and the strongest competing non-target logit. Let

$$j^*(x+\delta) = \arg\max_{k \neq t} z_k(x+\delta). \tag{15}$$

When the maximizer is unique, we have

$$\nabla_\delta M_t(x+\delta) = \nabla_\delta z_t(x+\delta) - \nabla_\delta z_{j^*}(x+\delta). \tag{16}$$

When the maximizer is not unique, Eq. (16) can be interpreted as a subgradient. Unlike the cross-entropy gradient, the margin gradient is not directly scaled by saturated softmax probabilities. It continues to focus the update on the currently strongest competing class, which is the most relevant class for changing the target decision.

This property is useful under small perturbation budgets. When the trigger is constrained to be visually weak, its limited capacity should be directed toward the most decision-relevant logit gap. Cross-entropy minimization can provide diminishing update signals once target-class confidence is high, whereas margin maximization continues to enlarge the separation between the target class and its nearest competitor. Thus, under the same perturbation constraint, MISA is expected to produce a stronger target-margin-increasing trigger direction when the optimized direction transfers from the surrogate model to the victim model. We empirically examine both aspects in Sec. V-C, where we measure victim-model margin gain under inference-time amplification and compare MISA with CE objective under the same attack pipeline.

V. Experiments

We evaluate MISA from four perspectives. First, we compare its attack performance with representative clean-label backdoor attacks on multiple benchmark datasets. Second, we validate the margin-amplification mechanism by measuring target-margin changes under inference-time amplification and comparing with the CE objective variant. Third, we assess poisoning-stage stealthiness qualitatively and quantitatively. Finally, we conduct ablation studies on victim architecture, poisoning budget, and train-time trigger perturbation.

A. Experimental Setup

We evaluate MISA on CIFAR-10, CIFAR-100, and ImageNet-100 (IN-100). CIFAR-10 and CIFAR-100 are standard benchmarks widely used in backdoor attack evaluation. IN-100 is constructed from the first 100 alphabetically sorted classes of ImageNet-1K. Table I summarizes the dataset statistics and the poisoning configurations.

The target class is set to class 0. To preserve train-time stealthiness, **we restrict the train-time trigger perturbation budget to $\epsilon = 4/255$**. All experiments are conducted on a server equipped with a single NVIDIA RTX 4090 GPU. The trigger is optimized on the clean ResNet-18 surrogate model for 1000 steps using Adam with a batch size of 256 and a learning rate of 0.03. During trigger optimization, the surrogate model is frozen and only the trigger perturbation is updated. The default victim model is ResNet-18 [37]. Surrogate and victim models are trained using SGD with a batch size of 256, an initial learning rate of 0.1, and cosine annealing. CIFAR-10 and CIFAR-100 models are trained for 200 epochs, while IN-100 models are trained for 100 epochs. We use dataset-specific preprocessing pipelines. For CIFAR-10 and CIFAR-100, images are processed at 32×32 with standard training-time spatial augmentation and normalization. For IN-100, images are processed at 224×224 with standard spatial augmentation and normalization.

TABLE II: Performance comparison of clean-label backdoor attacks. ACC and ASR are reported in percentages.

Method	CIFAR-10			CIFAR-100			IN-100		
	ACC	ASR		ACC	ASR		ACC	ASR	
		$\alpha=1$	$\alpha=5$		$\alpha=1$	$\alpha=5$		$\alpha=1$	$\alpha=5$
Clean (unpoisoned)	95.14	–	–	77.96	–	–	73.78	–	–
SIG	94.81	0.32	0.67	77.25	0.19	0.21	72.83	2.00	1.56
LC	94.71	0.63	0.67	77.64	0.10	0.11	72.00	1.50	2.00
Invisible Poison	95.20	0.56	1.07	74.80	1.35	16.18	72.64	0.02	0.08
GenBound	94.60	0.11	0.18	77.72	1.22	1.96	72.94	0.96	11.00
Hybrid	95.48	0.57	0.29	77.86	0.19	1.80	71.45	0.10	0.48
COMBAT	95.39	2.08	76.21	76.99	0.98	13.52	71.40	2.67	37.03
Narcissus	95.14	2.32	78.24	77.81	5.97	83.31	72.04	2.55	73.17
MISA	95.30	5.08	97.41	77.91	15.09	94.55	73.66	5.25	95.27

We report two standard metrics: **clean accuracy (ACC)**, defined as the classification accuracy on clean test samples, and **attack success rate (ASR)**, defined as the proportion of triggered test samples classified into the target class.

B. Attack Performance

We compare MISA with seven representative clean-label backdoor attacks whose train-time image-space perturbations can be explicitly controlled under an ℓ_∞ budget: LC [9], SIG [10], Narcissus [15], Invisible Poison [11], GenBound [18], Hybrid [17], and COMBAT [16]. All methods are evaluated under the same train-time perturbation budget of $4/255$, which enables a direct comparison of train-time stealth. We exclude style or atrribute trigger attacks because their triggers are not naturally constrained by the same image-space ℓ_∞ budget.

All baselines are implemented using their official codebases or published specifications. At inference time, we report ASR under the original trigger strength ($\alpha = 1$) and a stronger activation setting ($\alpha = 5$). This amplified setting evaluates weak-implantation and strong-activation behavior, rather than inference-time imperceptibility. For surrogate-based methods, we use a ResNet-18 surrogate matching the default victim architecture. For Narcissus, we evaluate its complete-information variant optimized on the full training dataset as a stronger reference baseline.

Table II reports the attack performance on CIFAR-10, CIFAR-100, and IN-100. MISA achieves the highest ASR on all three datasets while maintaining clean accuracy close to

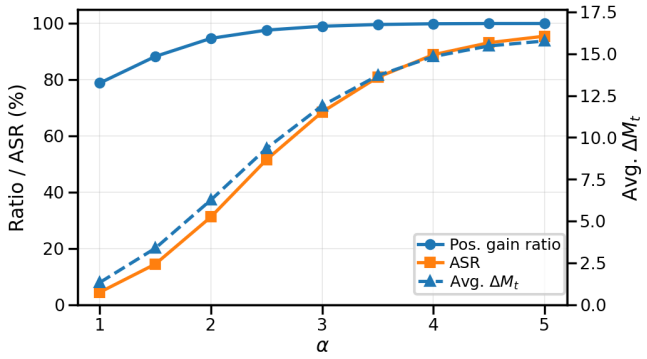


Fig. 3: Empirical validation of the margin-amplification mechanism on CIFAR-10.

that of the clean model. Under the same train-time perturbation budget of $4/255$, most existing clean-label attacks' triggers are not reliably learned by the victim model. Even with $\times 5$ inference-time amplification, SIG, LC, Invisible Poison, GenBound, and Hybrid obtain low ASRs on most datasets.

COMBAT and Narcissus are the strongest baselines in this comparison. COMBAT reaches 76.21% ASR on CIFAR-10 and 37.03% ASR on IN-100, while Narcissus achieves 83.31% ASR on CIFAR-100 and 73.17% ASR on IN-100. In contrast, MISA obtains 97.41%, 94.55%, and 95.27% ASR on CIFAR-10, CIFAR-100, and IN-100, respectively. Its clean accuracy remains comparable to the clean model across all datasets, indicating that the improved attack effectiveness does not come at the cost of benign classification performance.

The advantage of MISA is more pronounced on the challenging datasets. CIFAR-100 contains fewer samples per class, making clean-label poisoning harder under a fixed budget, while IN-100 contains higher-resolution images with richer visual variations. Existing methods often struggle in these settings because a weak trigger must compete with strong natural class features. By contrast, MISA remains effective, suggesting that the proposed margin-guided trigger optimization builds a more reliable association between the trigger and the target class. These results support the weak implantation and strong activation design of MISA, where the trigger is kept small during training for stealthiness, and is then strengthened at inference time through amplification.

C. Empirical Mechanism Validation

This subsection empirically connects the analytical discussion in Sec. IV-B with the observed attack behavior using CIFAR-10. We first examine whether inference-time amplification indeed enlarges the victim-model target margin, explaining how a weak train-time trigger becomes effective after amplification. We then compare MISA with a CE-objective variant to isolate the contribution of directly optimizing the target logit margin over optimizing target-label confidence.

1) Margin-Amplification Behavior: The analysis in Sec. IV-B explains MISA as a margin-amplification mechanism rather than a global guarantee of attack success. To examine whether this mechanism holds in the trained victim model, we measure the target-margin gain on non-target CIFAR-10 test samples:

$$\Delta M_t(x, \alpha) = M_t(x + \alpha\delta) - M_t(x), \tag{17}$$

where M_t is computed using the victim model. We further define the positive margin-gain ratio as

$$R_+(\alpha) = \frac{1}{|\mathcal{D}_{\text{nt}}|} \sum_{x \in \mathcal{D}_{\text{nt}}} \mathbf{1}[\Delta M_t(x, \alpha) > 0], \tag{18}$$

where $\mathcal{D}_{\text{nt}}$ denotes the non-target test set. This ratio measures the fraction of non-target samples whose target margin increases after applying the amplified trigger.

As shown in Fig. 3, the optimized trigger direction consistently increases the victim-model target margin. The positive margin-gain ratio rapidly approaches 100%, indicating that the trigger moves almost all non-target samples in the target-margin-increasing direction. Meanwhile, the average margin gain continues to increase as α grows, showing that amplification strengthens the magnitude of this margin shift. Although $R_+(\alpha)$ saturates after moderate amplification, ASR continues to increase. This is expected because a positive margin gain only indicates directional alignment, i.e., $\Delta M_t(x, \alpha) > 0$, but does not imply that the triggered sample has crossed the target decision boundary.

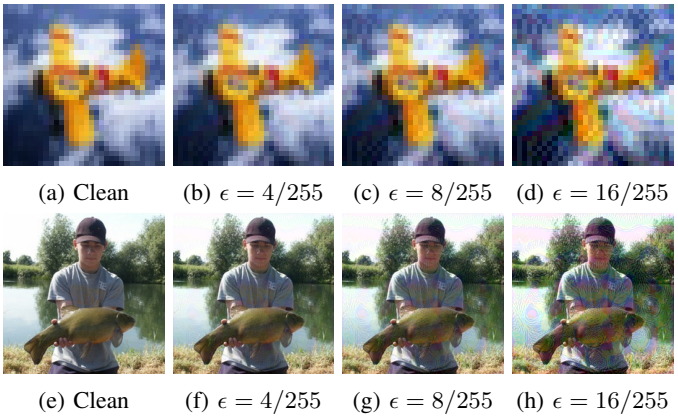
2) Comparison with Cross-Entropy Optimization: To further evaluate the effect of the margin-guided objective, we compare MISA's margin-guided objective with CE objective (Narcissus with complete information). This comparison evaluates whether directly optimizing the target margin yields a stronger amplifiable trigger direction than optimizing CE objective in Eq.(12).

TABLE III: Objective-level comparison between margin-objective and the CE objective variant on CIFAR-10. Value after "@" denotes the inference-time amplification factor α.

Objective	ASR@1	ASR@2.5	ASR@5	Avg. ΔM_t@5	Triggered M_t@5
CE (Narcissus)	2.32	22.47	78.24	12.73	3.72
Margin (MISA)	5.08	57.91	97.41	15.76	6.69

As shown in Table III, the CE objective used by Narcissus is a strong baseline under inference-time amplification, reaching 78.24% ASR at $\alpha = 5$. However, the margin-guided objective used by MISA achieves 97.41% ASR under the same setting and exhibits a stronger activation trend. In particular, MISA reaches 57.91% ASR at $\alpha = 2.5$, whereas Narcissus obtains only 22.47% ASR. This indicates that the margin-optimized trigger becomes effective under weaker amplification.

The margin statistics provide a direct explanation. At $\alpha = 5$, MISA produces a larger average target-margin gain than Narcissus, increasing from 12.73 to 15.76. The triggered target margin is also substantially larger, increasing from 3.72 to 6.69. These results show that directly optimizing the target logit gap yields a stronger decision-level trigger direction than optimizing target-label cross-entropy.



(a) Clean (b) $\epsilon = 4/255$ (c) $\epsilon = 8/255$ (d) $\epsilon = 16/255$

(e) Clean (f) $\epsilon = 4/255$ (g) $\epsilon = 8/255$ (h) $\epsilon = 16/255$

Fig. 4: Visual comparison between clean and poisoned samples. The top row shows CIFAR-10 examples, and the bottom row shows IN-100 examples. The second column corresponds to the default MISA train-time perturbation, $\epsilon = 4/255$. Larger perturbations are shown only for visual reference.

D. Poisoning-stage Visual Stealthiness

We evaluate the poisoning-stage visual stealthiness of MISA from both qualitative and quantitative perspectives. Since MISA is a clean-label attack, all poisoned training samples retain their original labels. Therefore, during data inspection, the main source of possible suspicion is the visual discrepancy between a clean training image and its poisoned counterpart. In the default setting, MISA uses a train-time perturbation budget of only $\epsilon = 4/255$, which is smaller than the commonly used $8/255$ perturbation budget in adversarial-perturbation studies [38], and is also lower than the default designed perturbation budgets $\epsilon = 16/255$ adopted by many clean-label backdoor attacks, such as Narcissus and COMBAT.

Fig. 4 visualizes clean and poisoned samples on CIFAR-10 and IN-100 under different perturbation magnitudes. The second column corresponds to the default setting used by MISA, i.e., $\epsilon = 4/255$. Under this budget, the poisoned samples are visually close to their clean counterparts, and the object semantics are well preserved. For reference, we also show stronger perturbations with $\epsilon = 8/255$ and $\epsilon = 16/255$. Although these larger perturbations become progressively more visible, they are not used during training in the default setting. This weak-implantation design preserves visual fidelity while still allowing strong activation at inference time.

We further quantify visual stealthiness using three standard image-similarity metrics: structural similarity index measure (SSIM) [39], peak signal-to-noise ratio (PSNR) [40], and learned perceptual image patch similarity (LPIPS) [41]. SSIM and PSNR measure structural and pixel-level similarity, respectively, where higher values indicate better visual fidelity. LPIPS measures perceptual distance in a deep feature space, where lower values indicate higher perceptual similarity. As shown in Table IV, MISA achieves SSIM values above 0.95, PSNR values above 36 dB, and low LPIPS values on both datasets. These results show that the poisoned samples generated by MISA introduce only minor visual changes.

TABLE IV: Quantitative stealthiness evaluation of MISA poisoned samples under the default train-time perturbation $\epsilon = 4/255$. Higher SSIM and PSNR indicate stronger similarity, while lower LPIPS indicates smaller perceptual distance.

Dataset	SSIM (↑)	PSNR (↑)	LPIPS (↓)
CIFAR-10	0.9651	36.7643	0.0052
IN-100	0.9592	36.4225	0.0304

TABLE V: Attack performance across different victim model architectures on CIFAR-10. The trigger is optimized using ResNet-18 as the surrogate model. ASR is reported under $\times 5$ inference-time trigger amplification.

Victim Model Architecture	Train-time Trigger Perturbation ϵ	ACC (%)	ASR (%, $\alpha = 5$)
ResNet-18	$4/255$	95.30	97.41
VGG16 [42]	$4/255$	93.90	95.07
ResNeXt29 [43]	$4/255$	95.58	93.34
DenseNet121 [44]	$4/255$	95.87	91.54
CCT [45]	$4/255$	90.78	38.79
CCT [45]	$8/255$	90.70	94.88
Swin-T [46]	$4/255$	87.21	32.77
Swin-T [46]	$8/255$	87.05	95.27

E. Ablation Study

We conduct ablation studies to examine how MISA behaves under different experimental settings, including victim architecture, poisoning budget, and train-time trigger perturbation.

1) Impact of Victim Model Architecture: To evaluate the transferability of MISA across victim architectures, we conduct experiments on six representative architectures on CIFAR-10, as summarized in Table V. The trigger is optimized using ResNet-18 and is then used to poison the training data for different victim models. Thus, the ResNet-18 victim corresponds to the architecture-matched setting, while the remaining victim models evaluate architecture-unmatched transfer. We consider three CNN-based architectures, VGG16 [42], ResNeXt29 [43], and DenseNet121 [44], as well as two Transformer models, CCT [45] and Swin-T [46].

MISA transfers strongly across CNN-based victims. Under the default train-time perturbation budget of $4/255$, it achieves 97.41%, 95.07%, 93.34%, and 91.54% ASR on ResNet-18, VGG16, ResNeXt29, and DenseNet121, respectively. These results indicate that the trigger optimized by the margin objective is not tightly coupled to the surrogate architecture. Instead, the optimized perturbation provides a transferable target-class shortcut that can be learned by different CNN-based models while preserving high clean accuracy.

The transfer to Transformer-style victims is more challenging under the same $4/255$ budget. Specifically, MISA achieves only 38.79% ASR on CCT and 32.77% ASR on Swin-T. We attribute this degradation to the larger architectural shift between the ResNet-18 surrogate and Transformer-style victims. Compared with CNNs, CCT and Swin-T rely on different tokenization, attention, and feature aggregation mechanisms, and thus may encode a dense low-amplitude trigger differently from the CNN surrogate. In addition, these Transformer-style victims obtain lower clean accuracy on CIFAR-10 than the

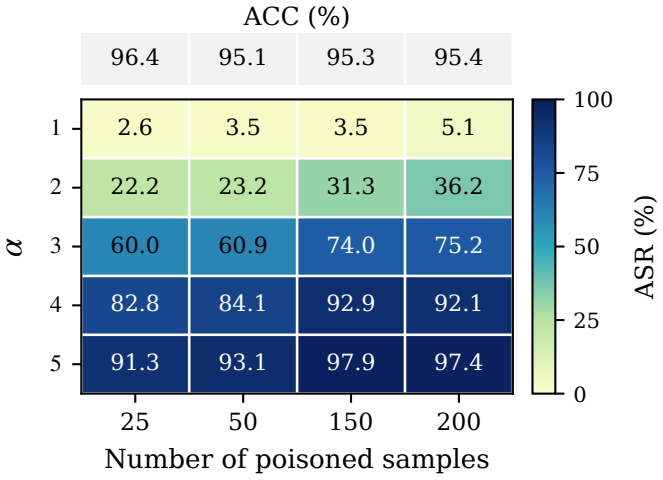


Fig. 5: Effect of the number of poisoned samples on CIFAR-10 attack performance. The top strip reports ACC, and each heatmap cell reports ASR under inference-time trigger amplification factor α.

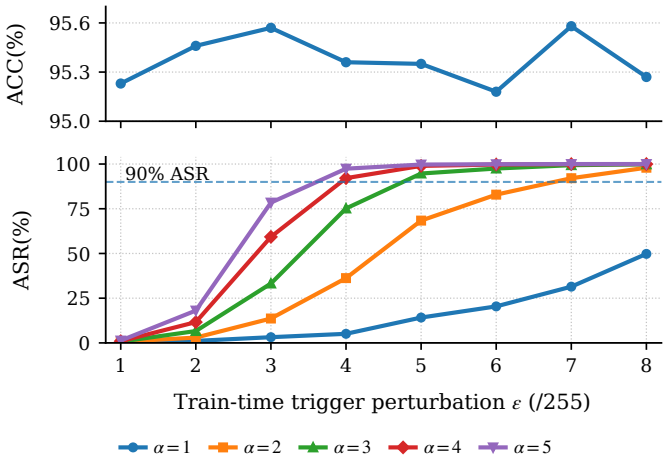


Fig. 6: Effect of train-time trigger perturbation on CIFAR-10. The upper panel reports clean ACC. The lower panel reports ASR under different test-time amplification factors α.

CNN victims in our setting, suggesting that their learned representations are less stable under this small-scale training regime. As a result, the weak $4/255$ trigger is less reliably transferred and learned as a target-class shortcut.

Increasing the train-time perturbation budget from $4/255$ to $8/255$ substantially improves the attack against Transformer-style victims, raising the ASR to 94.88% on CCT and 95.27% on Swin-T. This improvement is achieved with negligible change in clean accuracy. These results suggest that a slightly stronger train-time trigger can compensate for the weaker surrogate-to-victim transfer caused by the larger architectural mismatch. Overall, MISA transfers effectively across CNN-based architecture-unmatched victims under the default $4/255$ budget, while Transformer-style victims require a stronger but still bounded train-time perturbation to establish a reliable trigger–target association.

2) Effect of the Number of Poisoned Samples: We next examine the effect of the poisoning budget. Fig. 5 reports ASR under different numbers of poisoned samples and test-time trigger amplification factors α, with the corresponding clean accuracy shown in the top strip. The train-time trigger perturbation is fixed at $4/255$.

The results show that the default choice of 200 poisoned samples used in Section V-B is not a fragile or specially tuned setting. Across all poisoning budgets, ACC remains stable between 95% and 96%, indicating that increasing the number of poisoned samples within this range does not noticeably harm clean performance. Moreover, MISA is already effective with fewer poisoned samples when the trigger is amplified at test time. With only 25 poisoned samples, ASR increases from 2.6% at $\alpha = 1$ to 91.3% at $\alpha = 5$. With 150 poisoned samples, MISA reaches 97.9% ASR at $\alpha = 5$, which is comparable to the 97.4% ASR obtained with 200 poisoned samples.

These results indicate that the default poisoning budget provides a stable and sufficiently strong evaluation setting, rather than being the minimum requirement for the attack to succeed.

The ablation also confirms the intended weak implantation and strong activation behavior: the trigger remains weak at $\alpha = 1$, while inference-time amplification reliably activates the backdoor even under smaller poisoning budgets.

3) Impact of Trigger Perturbation: Finally, we study the impact of the train-time trigger perturbation budget ϵ on CIFAR-10, while evaluating the trained models under different inference-time amplification factors α. The results are shown in Fig. 6. As shown in the upper panel, clean accuracy remains stable across all perturbation budgets, ranging from 95.18% to 95.58%. This indicates that increasing ϵ mainly affects backdoor learnability rather than clean classification performance. In contrast, the lower panel shows that ASR increases with the train-time perturbation budget. Without amplification ($\alpha = 1$), ASR rises from 0.63% at $\epsilon = 1/255$ to 49.77% at $\epsilon = 8/255$. With test-time amplification, the attack becomes substantially stronger. For example, $\alpha = 4$ already achieves 92.11% ASR at $\epsilon = 4/255$, and $\alpha = 5$ reaches 97.41% under the same budget.

These results validate the principle of weak implantation and strong activation. A small train-time perturbation preserves clean accuracy and stealthiness, while test-time amplification substantially strengthens backdoor activation during inference.

VI. Evaluation Against Backdoor Defenses

We evaluate MISA against representative backdoor defenses spanning model-oriented, training-oriented, and data-oriented categories. Following prior clean-label backdoor studies [15], [18], we conduct the defense evaluation on CIFAR-10, which is supported by most existing defenses and enables a consistent comparison under a unified experimental setting. We use the same attack setting as in Section V.

A. Model-oriented Defenses

Model-oriented defenses aim to determine whether a trained model is backdoored without requiring access to the original

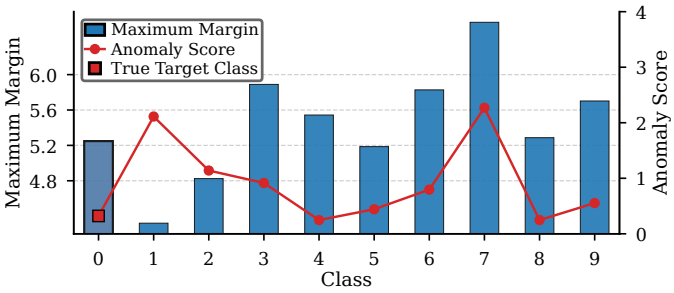


Fig. 7: Maximum-margin statistics produced by MM-BD.

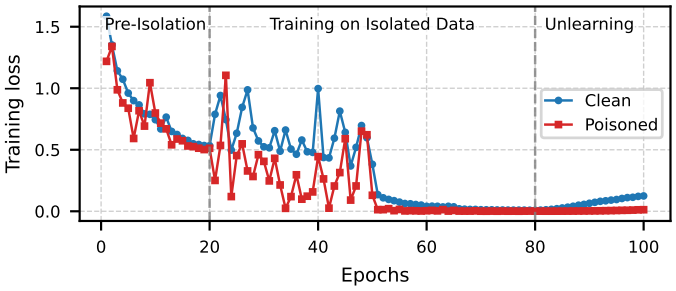


Fig. 8: Loss distributions observed by ABL.

training set. These methods typically inspect model behavior, recovered trigger patterns, or internal statistics.

Neural Cleanse [25] reverse-engineers a class-specific trigger for each class and flags a model as backdoored if one class yields an anomalously small recovered trigger. Following the original paper, we compute the anomaly index of the MISA-poisoned model. The resulting anomaly index is 0.96, which is far below the commonly used detection threshold of 2. Therefore, Neural Cleanse does not flag the model as backdoored. This result is consistent with the design of MISA: Neural Cleanse is biased toward sparse and localized triggers due to its L_1-regularized trigger reconstruction objective, whereas MISA introduces a dense, low-amplitude L_∞-bounded perturbation distributed over the entire image. Such a trigger does not naturally lead to an unusually small reconstructed mask, making it difficult for Neural Cleanse to identify the target class as an outlier.

MM-BD [27] is a post-training backdoor detection method that identifies the target class by testing whether its maximum-margin statistic appears as a prominent outlier among all classes. This defense is based on the assumption that a successful backdoor induces an abnormally large maximum margin for the target class. However, as shown in Fig. 7, MISA does not exhibit such a distinctive maximum-margin signature. The maximum-margin values of all classes are concentrated within a narrow range, and the true target class, i.e., class 0, obtains a value of only 5.25, which falls well within the range of the remaining classes, from 4.32 to 6.59. Its anomaly score is also low, at only 0.32. As a result, the true target class does not emerge as a significant outlier, causing MM-BD to fail in detecting the MISA-poisoned model.

B. Training-oriented Defenses

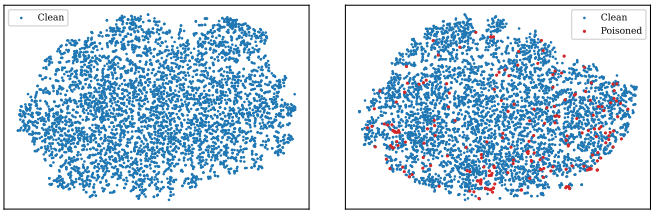
Training-oriented defenses intervene in the training process or perform post-hoc model repair, with the goal of suppressing the backdoor while preserving clean accuracy.

ABL [28] assumes that poisoned samples are learned faster than clean samples and therefore appear as unusually low-loss examples during the early stage of training. It first isolates suspicious low-loss samples, then trains the model without these samples, and finally performs an unlearning stage using the isolated samples. We follow the training schedule used in the original ABL paper: 20 epochs for pre-isolation, 60 epochs for training on the remaining data, and 20 epochs for unlearning. To give the defense a favorable setting, we further consider a semi-oracle scenario in which the defender knows the exact number of poisoned samples and isolates 200 suspicious samples in the first stage.

When applied to the MISA-poisoned dataset, ABL reduces the clean accuracy to 89.63%, while the ASR ($\alpha = 5$) remains high at 89.91%. Thus, ABL fails to both preserve clean accuracy and remove the backdoor. Fig. 8 explains this failure. The MISA-poisoned samples do not form the low-loss group expected by ABL; instead, their losses remain comparable to, and sometimes even higher than, those of clean samples. As a result, ABL fails to isolate the true poisoned samples. In our experiment, the 200 samples identified during the pre-isolation stage are clean samples, rather than true poisons. Consequently, the subsequent unlearning stage is performed on clean data, which degrades clean accuracy without effectively suppressing the backdoor.

FT-SAM [29] is a fine-tuning-based backdoor mitigation method that incorporates sharpness-aware minimization into the model repair process. It is motivated by the observation that vanilla fine-tuning often makes only limited changes to backdoor-related neurons, while such neurons tend to exhibit relatively large weight norms. FT-SAM encourages stronger updates to these backdoor-related neurons and reduces their dominance, thereby mitigating the backdoor effect while preserving clean accuracy. On MISA, FT-SAM retains 94.91% clean accuracy, while the ASR ($\alpha = 5$) remains high at 95.27%, indicating limited mitigation. This suggests that the backdoor effect induced by MISA is not concentrated in a small set of neurons that can be sufficiently disrupted by sharpness-aware fine-tuning. Instead, the low-amplitude trigger signal appears to be more diffusely encoded, allowing the backdoor behavior to persist after FT-SAM.

Trap [47] is a training-time backdoor mitigation method that first isolates poisoned samples from early-stage feature clusters and then relabels them to an additional dummy class, retraining only the classifier to redirect backdoor features away from the target label. On MISA, TRAP preserves 95.30% clean accuracy, but the ASR remains 96.89% at $\alpha = 5$, indicating limited mitigation. The failure mainly stems from poisoned-sample isolation: Trap detects no MISA-poisoned samples in our setting. This suggests that MISA's weak, bounded, clean-label perturbations do not form a separable suspicious cluster during early training, but remain embedded within the benign target-class distribution. As a result, the dummy-class reclassification stage has no reliable poisoned samples to



(a) Clean target class (b) Poisoned target class

Fig. 9: Visualization of Gramian features used by Beatrix.

relabel, leaving the amplified backdoor effect largely intact.

C. Data-oriented Defenses

Data-oriented defenses aim to identify or filter poisoned samples based on their input patterns, feature-space behavior, training dynamics, or inference-time responses. We evaluate these defenses using true positive rate (TPR) and false positive rate (FPR) as the main detection metrics.

Beatrix [32] detects poisoned samples by characterizing high-order feature correlations with Gram matrices and identifying anomalous class mixtures using regularized maximum mean discrepancy (RMMD). Following the original setup, we use 300 clean samples, with 30 samples per class, to calibrate the detector. However, Beatrix fails to detect the poisoned samples generated by MISA. The RMMD statistic of the true target class remains close to zero and does not exceed the detection threshold. Fig. 9(b) visualizes the Gramian features extracted from the AveragePooling layer of ResNet-18 and projected using t-SNE. Contrary to the class-mixture pattern assumed by Beatrix, MISA-poisoned samples do not form a distinct secondary cluster; instead, they largely overlap with clean target-class samples. This indicates that the low-amplitude perturbation introduced by MISA induces only limited changes in high-order feature correlations, which are insufficient to form a statistically detectable mixture in the Gramian feature space. This observation is further supported by Fig. 9(a), where the distribution of clean target-class samples shows no substantial visual discrepancy from that of poisoned samples.

PoisonSpot [33] is designed specifically to detect clean-label backdoor attacks through class-wise training provenance tracking. It monitors how individual samples influence sensitive model parameters and computes poisoning scores by comparing the gradient trajectories of suspicious samples with those of a clean baseline. Scores close to 1.0 indicate anomalous behavior, whereas scores close to 0.0 indicate benign samples. To separate poisoned samples from clean ones, PoisonSpot adopts two dynamic decision boundaries: Threshold 1, obtained from K-means clustering, and Threshold 2, obtained from a Gaussian mixture model. We evaluate PoisonSpot against MISA using 1,000 clean samples to construct the baseline. Although the defense achieves a TPR of 68%, it suffers from a very high FPR of 76.63%, indicating that it cannot reliably distinguish MISA-poisoned samples from clean ones. As shown in Fig. 10, clean samples exhibit large variance in poisoning scores and form a dense band between 0.3 and 0.95, which substantially overlaps with the 200 poisoned samples.

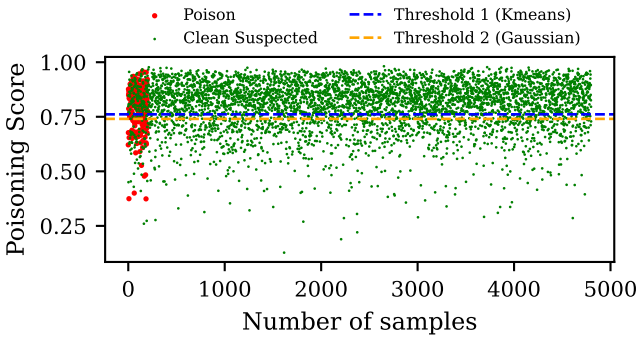
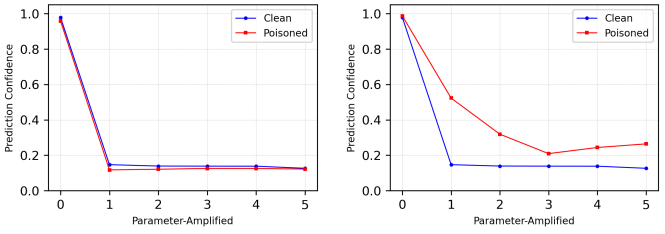


Fig. 10: Poisoning scores produced by PoisonSpot.



(a) $\alpha = 1$ (b) $\alpha = 5$

Fig. 11: PSC curves of IBD-PSC under different inference-time trigger amplification factors.

Consequently, both decision thresholds cut through the clean-sample distribution, leading to excessive false positives while still missing a portion of poisoned samples. These results suggest that MISA does not induce the clear provenance-level separability assumed by PoisonSpot.

IBD-PSC [34] is a black-box inference-time detector based on parameter-oriented scaling consistency (PSC). It scales selected BatchNorm layers, tracks the confidence of the original prediction across the perturbed models, and flags an input as suspicious if its average confidence exceeds a predefined threshold. Following the original setting, we use 100 clean samples for calibration and evaluate IBD-PSC on MISA-poisoned inputs. When the inference-time trigger is not amplified, i.e., $\alpha = 1$, IBD-PSC achieves a TPR of 5.89% and an FPR of 7.96%. When the trigger is amplified, i.e., $\alpha = 5$, the TPR further decreases to 0.93%, while the FPR remains 7.96%. These results show that IBD-PSC is ineffective at detecting MISA-poisoned samples, even when the trigger is amplified at inference time. Fig. 11 illustrates the PSC curves under both amplification settings. Contrary to the core assumption of IBD-PSC, the prediction confidence of MISA-poisoned samples changes similarly to that of clean samples across the BatchNorm scaling steps. Consequently, poisoned inputs are not separated from clean inputs by the PSC criterion.

VII. Conclusion

In this paper, we proposed MISA, a margin-guided clean-label backdoor attack based on a weak implantation and strong activation paradigm. MISA implants a small and stealthy trigger into target-class samples during training, while amplifying the same trigger at inference time for reliable backdoor

activation. By optimizing the target-class logit margin, MISA makes the weak trigger more decision relevant under a small perturbation budget.

We further analyzed the proposed design from three perspectives: target-margin maximization reduces an upper bound on the target-label cross-entropy loss, inference-time amplification can locally enlarge the target margin, and margin optimization avoids the softmax saturation issue of probability-oriented trigger objectives. Extensive experiments show that MISA achieves strong attack performance while preserving clean accuracy and visual stealthiness across datasets, architectures, poisoning budgets, and amplification factors. Evaluations against representative backdoor defenses further demonstrate its resilience under practical defense settings.

Overall, our results show that a weak clean-label trigger can become highly effective when its direction is explicitly optimized in the decision space and strengthened at inference time. This finding highlights the need for defenses that consider not only visible perturbation strength, but also amplifiable decision-level trigger directions.

REFERENCES

[1] Y. Yang, Z.-Q. Zhao, G. Wu, X. Zhuo, Q. Liu, Q. Bai, and W. Li, "A lightweight, effective, and efficient model for label aggregation in crowdsourcing," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 18, no. 4, pp. 1–27, 2024.

[2] M. Li, Y. Chen, J. Gao, J. Wu, Z. Zhang, J. He, L. Zhu, M. Conti, and X. Lin, "Accurate, secure, and efficient semi-constrained navigation over encrypted city maps," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 22, no. 3, pp. 2642–2658, 2025.

[3] J. Huang, Y. Feng, F.-Q. Cui, X. Zhang, Z. Liu, X. Liu, J. Liu, F. Zhang, and M. Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 23, no. 3, pp. 6890–6905, 2026.

[4] J. Huang, J.-X. Bai, X. Zhang, Z. Liu, Y. Feng, J. Liu, X. Sun, M. Dong, and M. Li, "Keystrokesniffer: An off-the-shelf smartphone can eavesdrop on your privacy from anywhere," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 19, pp. 6840–6855, 2024.

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

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

[7] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, vol. 35, no. 1, pp. 5–22, 2022.

[8] Y. Li, J. He, H. Huang, J. Sun, X. Ma, and Y.-G. Jiang, "Shortcuts everywhere and nowhere: Exploring multi-trigger backdoor attacks," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 23, no. 1, pp. 343–355, 2026.

[9] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[10] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 101–105.

[11] R. Ning, J. Li, C. Xin, and H. Wu, "Invisible poison: A blackbox clean label backdoor attack to deep neural networks," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[12] M. Zhu, Y. Li, J. Guo, T. Wei, S.-T. Xia, and Z. Qin, "Towards sample-specific backdoor attack with clean labels via attribute trigger," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2025.

[13] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, vol. 34, no. 07, 2020, pp. 11 957–11 965.

[14] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein, "Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 19 165–19 178, 2022.

[15] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023, pp. 771–785.

[16] T. Huynh, D. Nguyen, T. Pham, and A. Tran, "Combat: Alternated training for effective clean-label backdoor attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, no. 3, 2024, pp. 2436–2444.

[17] C. Yuan, J. Bai, S. Yuan, and N. Wei, "Stealthy and effective clean-label backdoor attack via adaptive frequency-domain suppression and trigger combination," *IEEE Transactions on Information Forensics and Security (TIFS)*, 2025.

[18] L. Yu, S. Liu, Y. Miao, X.-S. Gao, and L. Zhang, "Generalization bound and new algorithm for clean-label backdoor attack," in *In Forty-first International Conference on Machine Learning (ICML)*, 2024.

[19] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF International Conference on Computer Vsion (ICCV)*, 2021, pp. 16 463–16 472.

[20] Y. Gao, H. Chen, P. Sun, J. Li, Y. Yin, Z. Wang, and W. Liu, "A triple stealthy backdoor: Hidden in spatial, frequency, and feature domains," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2025.

[21] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *International Conference on Learning Representations (ICLR)*, 2021.

[22] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, "Rethinking the backdoor attacks' triggers: A frequency perspective," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 16 473–16 481.

[23] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 18, no. 5, pp. 2088–2105, 2020.

[24] Y. Li, J. Bi, S. Tian, X. Dai, and S. Long, "Style backdoor: A robust style-decoupled clean-label backdoor attack," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–14, 2026.

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

[26] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. 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 (CCS)*, 2019, pp. 1265–1282.

[27] H. Wang, Z. Xiang, D. J. Miller, and G. 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)*. IEEE, 2024, pp. 1994–2012.

[28] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 14 900–14 912, 2021.

[29] M. Zhu, S. Wei, L. Shen, Y. Fan, and B. Wu, "Enhancing fine-tuning based backdoor defense with sharpness-aware minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4466–4477.

[30] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," *Advances in neural information processing systems (NeurIPS)*, vol. 31, 2018.

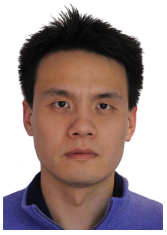
[31] J. Hayase, W. Kong, R. Somani, and S. Oh, "Spectre: Defending against backdoor attacks using robust statistics," in *International Conference on Machine Learning (ICML)*, 2021, pp. 4129–4139.

[32] W. Ma, D. Wang, R. Sun, M. Xue, S. Wen, and Y. Xiang, "The "beatrix" resurrections: Robust backdoor detection via gram matrices," in *The 30th Network and Distributed System Security Symposium (NDSS)*, 2023, pp. 1–18.

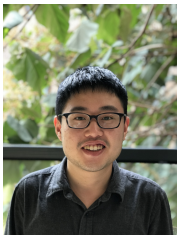
[33] P. Hailemariam and B. Eshete, "Poisonspot: Precise spotting of clean-label backdoors via fine-grained training provenance tracking," in *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2025, p. 1023–1037.

[34] L. Hou, R. Feng, Z. Hua, W. Luo, L. Y. Zhang, and Y. Li, "Ibd-psc: input-level backdoor detection via parameter-oriented scaling consis-

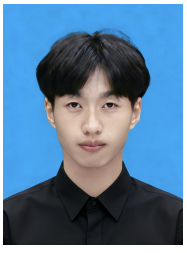
tency," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. ICML'24, 2024.
[35] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009, pp. 248–255.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.
[38] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," *arXiv preprint arXiv:2010.09670*, 2020.
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.
[40] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.
[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[43] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 4700–4708.
[45] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021.
[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.
[47] C. Li, Z. Li, M. Pan, and X. Li, "Trap: Mitigating poisoning-based backdoor attacks by treating poison with poison," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, vol. 23, no. 2, pp. 4269–4283, 2026.



Yi Yang received the Ph.D. degree from the Auckland University of Technology, Auckland, New Zealand, in 2020. From 2020 to 2021, he held a post-doctoral position with the Applied AI Institute, Deakin University, Geelong, Australia. He is currently a Lecturer with Hefei University of Technology, Hefei. His current research interests include deep learning security, pattern recognition, and crowdsourcing.



Jinyang Huang is an Associate Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). His research interests include Multimodal Perception, Human-computer Interaction, Wireless Security, and Signal Processing. In this area, he has published 60 papers in international peer-reviewed journals and conferences, including ToN, TMC, TIFS, TDSC, MobiCom, IEEE S&P, USENIX Security, Ubicomp, NeurIPS, INFOCOM, and ACM MM. He has served as a TPC member for conferences, including ACM MM, IEEE ICME, and Globecom, and has the honor of becoming ACM MM 2024 Outstanding Reviewers. He is a Guest Editor for the Technical Committee of ICME 25 Special Session. He is the recipient of the Young Scientist of Anhui Computer Federation and IEEE HITC Distinguished PhD Dissertation Award.



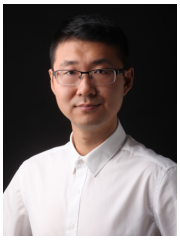
Feng-Qi Cui (Student Member, IEEE) is currently pursuing M.E. degree in Computer Science at University of Science and Technology of China (USTC), Hefei, China. He is also affiliated with Anhui Provincial Key Laboratory of Affective Computing and Advanced Intelligent Machines, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include Computer Vision, Multimedia Information Processing and Affective Computing.



BinBin Liu received the Ph.D. degree from the University of Science and Technology of China (USTC) in 2022. He is currently a Lecturer with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His research interests include AI agents and software security. He has published several papers in leading conferences and journals, including USENIX Security, PLDI, EMNLP, ICICS, and Security and Communication Networks.



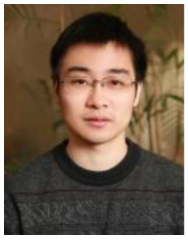
Lei Shi (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Hefei University of Technology, Hefei, Anhui, China, in 2002, 2005, and 2012, respectively.,He is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology. His main research area lies in federated learning, edge computing, and wireless network optimization.



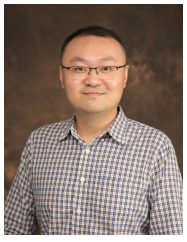
Weihua Li received the Ph.D. degree in Computer Science from Auckland University of Technology (AUT), Auckland, New Zealand, in 2018, and the M.Tech. degree in Knowledge Engineering from the National University of Singapore, Singapore, in 2014. He is currently a Senior Lecturer with the Department of Data Science and Artificial Intelligence. Weihua is an active researcher in the field of artificial intelligence, with a particular focus on social computing, generative AI, agent-based modeling and simulation of complex systems, and agentic AI. He has authored over 90 publications and has been actively involved in organizing several international AI conferences.



Quan Bai received his PhD (2007) and MSc (2002) from the University of Wollongong, Australia. Since 2019, he has been an Associate Professor in the School of Information and Communication Technology at University of Tasmania. Quan Bai is a leading researcher in Agentic AI, multi-agent systems, and machine learning. His research focuses on applying advanced AI techniques to model and optimise complex systems with many interacting and interdependent components. He has authored more than 230 scholarly publications and secured over AUD 4.3 million in research funding.



Xiaokang Zhou (M'12, SM'25) is currently an associate professor with the Faculty of Business Data Science, Kansai University, Japan. He received the Ph.D. degree in human sciences from Waseda University, Japan, in 2014. From 2012 to 2015, he was a research associate with the Faculty of Human Sciences, Waseda University, Japan. He was a lecturer/associate professor with the Faculty of Data Science, Shiga University, Japan, from 2016 to 2024. He also works as a visiting researcher in the RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Japan, since 2017. Dr. Zhou has been engaged in interdisciplinary research in computer science and engineering, information systems, and social and human informatics. His recent research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber-physical-social systems, cyber intelligence, and security. Dr. Zhou is a senior member of the IEEE CS, USA, and a member of the ACM, USA, IPSJ, JSAI, Japan, and CCF, China.



Meng Li (Senior Member, IEEE) is a Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He was a Post-Doc Researcher at the Department of Mathematics and HIT Center, University of Padua, Italy, where he is with the Security and PRIvacy Through Zeal (SPRITZ) research group led by Prof. Mauro Conti (IEEE Fellow). He obtained his Ph.D. in Computer Science and Technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He was sponsored by ERCIM 'Alain Bensoussan' Fellowship Programme (from 2020.10.1 to 2021.3.31) to conduct Post-Doc research supervised by Prof. Fabio Martinelli at CNR, Italy. He was sponsored by China Scholarship Council (CSC) as a Joint Ph.D. student (from 2017.9.1 to 2018.8.31) supervised by Prof. Xiaodong Lin (IEEE Fellow) in the Broadband Communications Research (BBCR) Lab at University of Waterloo and Wilfrid Laurier University, Canada. He is supported by CSC as a Visiting Scholar (from 2025.3.1 to 2025.6.30) collaborating with Prof. Mauro Conti (IEEE Fellow) at the HIT Center, University of Padua, Italy. He was promoted to a Professor from December, 2024. His research interests include data sharing, IoV, security, privacy, applied cryptography, blockchain, TEE. In this area, he has published 156 papers in topmost journals and conferences, including T-IFS, TDSC, ToN, TMC, JSAC, TKDE, TODS, TPDS, TSE, TSC, COMST, IEEE S&P, USENIX Security, MobiCom, and INFOCOM. He is a Senior Member of IEEE, CCF, CACR, CIE, and CIC. He is a Senior Area Editor for T-IFS and an Associate Editor for TDSC. He has served as a TPC member for conferences, including ICDCS, Inscrypt, ACISP, ICICS, and TrustCom. He is the recipient of 2024 IEEE HITC Award for Excellence (Early Career Researcher) and 2025 IEEE TCSVC Rising Star Award. He was selected into the IEEE Computer Society Computing's Top 30 Early Career Professionals for 2025.