Appendix for A Unified Detection Framework for Inference-Stage Backdoor Defenses

In Section 1, we present the formal proofs for the theoretical results discussed in Section 3 of the main text, specifically Theorem 1, 2, and 3. Furthermore, we provide a comprehensive listing of the detailed configurations used in the experimental study, along with a brief introduction to all the datasets, in Section 2. We also include the omitted empirical results from the main text, focusing on the CIFAR10 [1] and IMDB (NLP) [2] datasets, which are presented in Section 3. Additionally, we conduct an ablation study to explore the impact of different model architectures, such as VGG19 [3], which is discussed in Section 4. Furthermore, we conduct an additional ablation study to examine the effects of varying poisoning ratios, as presented in Section 5. We perform an ablation study to investigate the choice of the hyperparameter β, and the results are provided in Section 6. We further subject our proposed methods to backdoor attacks specifically tailored to target our techniques in Section 7. Finally, We further compare our methods with some recent backdoor defenses including both inference-stage and training-stage methods in Section 8.

1 Proof

In this section, we include the proof of the main results in Section 3.2 and 3.3 of the main text.

1.1 Proof for Section 3.2

In this section, we present the proof of Theorem 1. For the convenience of the reader, we attach the pseudo-code of Algorithm 1 below. Please note that the thresholding value $\lambda_{\alpha,s}$ is selected to satisfy the condition

$$\hat{F}_{\text{CLEAN}}(\lambda_{\alpha,s}) = 1 - \alpha + \sqrt{(\log(2/\delta)/(2n))}, \tag{1}$$

where $\delta \in (0,1)$ is the violation rate describing the probability that the (FPR) exceeds α, and $\hat{F}_{\text{CLEAN}}$ is the empirical cumulative distribution function of the scores on the validation data $\{s(T(X_i))\}_{i=1}^n$. In the case where $\sqrt{(\log(2/\delta)/(2n))} > \alpha$, we set the thresholding value τ to be the maximum of $\{s(T(X_i))\}_{i=1}^n$.

The main text briefly touches upon the central concept found in the conformal prediction literature. It states that by employing a suitable score function, the empirical rank or quantile of the distribution will eventually approach the population counterpart. This convergence is ensured by the uniform convergence of cumulative distribution functions (CDFs). Consequently, to establish the proof for Theorem 1, we will first present the subsequent outcome, which accurately measures the uniform convergence of CDFs.

Lemma 1 (Dvoretzky–Kieffer–Wolfowitz inequality). Given a natural number n, let $X_1, X_2, \ldots, X_n$ be real-valued independent and identically distributed random variables with cumulative distribution function $F(\cdot)$. Let $F_n(\cdot)$ denote the associated empirical distribution function.

The interval that contains the true CDF, $F(x)$, with probability $1-\delta$ is specified as

$$F_n(x) - \varepsilon \leq F(x) \leq F_n(x) + \varepsilon \text{ where } \varepsilon = \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

Theorem 1 (False positive rate of Algorithm 1). Given any pre-trained backdoored classifier f, suppose that the validation dataset $\mathcal{D}^{\text{Val}}$ and the test data $(X_{\text{test}}, Y_{\text{test}})$ are IID drawn from the clean

Algorithm 1 Conformal Backdoor Detection (CBD)

Input: querying input X_{test}, clean validation dataset $D^{\text{Val}} = \{(X_i, Y_i)\}_{i=1}^n$, transformation method $T(\cdot)$, score function $s(T(\cdot))$, desired false positive rate $\alpha \in (0,1)$, violation rate $\delta \in (0,1)$

```
Receiving a future query sample $X_{\text{test}}$
**for** $i = 1$ to $n$ **do**
    Calculate $s_i = s(T(X_i))$ // $x_i \in D^{\text{Val}}$
**end for**
Select the decision threshold $\lambda_{\alpha,s}$ according to Equation (1).
Determine if $X_{\text{test}}$ is a clean sample if $s(T(X_{\text{test}})) \leq \lambda_{\alpha,s}$
```

Output: The decision if the sample X_{test} is a clean or backdoor sample

data distribution $\mathbb{P}_{\text{CLEAN}}$. Given any $\delta \in (0,1)$, for any score function and transformation method $s(T(\cdot))$, such that the resulting scores $\{s(T(X_i))\}_{i=1}^n$ remain IID with a continuous distribution, the associated backdoor conformal detector $g(\cdot; s, \lambda_{\alpha,s})$ as specified in Algorithm 1 satisfies

$$\mathbb{P}\big(g(X_{\text{test}}; s, \lambda_{\alpha,s}) = 1 \,(\text{Backdoor Sample}) \mid \mathcal{D}^{\text{Val}}\big) \leq \alpha,$$

with probability at lease $1-\delta$ for any $\alpha \in (0,1)$ such that $\alpha > \sqrt{(\log(2/\delta)/(2n))}$.

Proof. Note that X_{test} is drawn from the clean data distribution and we have

$$\begin{align}
&\mathbb{P}(g(X_{\text{test}}; s, \lambda_\alpha) = 1(\text{Backdoor Sample}) \mid \mathcal{D}^{\text{Val}}) \notag\\
&= \mathbb{E}_{X_{\text{test}}} \mathbf{1}\{g(X_{\text{test}}; s, \lambda_\alpha) = 1 \mid \mathcal{D}^{\text{Val}}\} \notag\\
&= \mathbb{E}_{X_{\text{test}}} \mathbf{1}\{s(T(X_{\text{test}})) \geq \lambda_\alpha \mid \mathcal{D}^{\text{Val}}\} \tag{2}\\
&= \mathbb{E}_{X_{\text{test}}} \mathbf{1}\{F_{\text{CLEAN}}(s(T(X_{\text{test}}))) \geq F_{\text{CLEAN}}(\lambda_\alpha) \mid \mathcal{D}^{\text{Val}}\} \tag{3}\\
&= \mathbb{P}(F_{\text{CLEAN}}(s(T(X_{test}))) \geq F_{\text{CLEAN}}(\lambda_\alpha) \mid \mathcal{D}^{\text{Val}}) \notag\\
&\leq \mathbb{P}(F_{\text{CLEAN}}(s(T(X_{test}))) \geq \hat{F}_{\text{CLEAN}}(\lambda_\alpha) - \varepsilon \mid \mathcal{D}^{\text{Val}}) \quad (\varepsilon = \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}) \tag{4}\\
&= 1 - (1 - \alpha + \varepsilon - \varepsilon) \tag{5}\\
&= \alpha, \notag
\end{align}$$

holds with probability at least $1-\delta$. The equation (2) is because of the decision rule as specified in Algorithm 1. Additionally, the F_{CLEAN} in Equation (3) represents the CDF of $s(T(X_{\text{test}}))$, while $\hat{F}_{\text{CLEAN}}$ in (4) denotes the empirical CDF obtained from $\mathcal{D}^{\text{Val}}$. The inequality in Equation (4) arises from the DKW inequality specified in Lemma 1. Furthermore, the equation (5) is based on the fact that the CDF follows a uniform distribution (a result of the probability integral transformation) and the selection of the thresholding value specified in Equation (1). ☐

1.2 Proof for Section 3.3

This section will present the proofs for Theorem 2 and 3, as outlined in Section 3.3 of the main text. These results are intimately connected to the renowned Neyman-Pearson Lemma within the context of statistical hypothesis testing and binary classification problems. To provide the necessary context, we will begin with a brief introduction to the Neyman-Pearson classification framework.

1.2.1 Neyman-Pearson Classification

Consider a random pair (X, Y), where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional vector of features, and $Y \in \{0,1\}$ represents the class label of X. A classifier $g : \mathcal{X} \to \{0,1\}$ from a data input belongs to $\mathcal{X}$ to $0, 1$. The overall classification error of f is denoted as $R(g) = \mathbb{E}\mathbf{1}\{g(X) \neq Y\} = \mathbb{P}\{g(X) \neq Y\}$. By applying the law of total probability, $R(g)$ can be decomposed into a weighted average of the type I error $R_0(g) = \mathbb{P}\{g(X) \neq Y \mid Y = 0\}$ and the type II error $R_1(g) = \mathbb{P}\{g(X) \neq Y \mid Y = 1\}$, given by

$$R(g) = \pi_0 R_0(g) + \pi_1 R_1(g)$$

where $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$. While the classical paradigm minimizes $R(\cdot)$, the Neyman-Pearson (NP) paradigm seeks to minimize R_1 while controlling R_0 under a user-specified level α. The (level-α) NP oracle classifier is thus

$$g_\alpha^* \in \underset{R_0(g)\leq\alpha}{\arg\min} R_1(g)$$

where the significance level α reflects the level of conservativeness towards type I error. To reflect on the backdoor detection problem, we encode the label set Y to indicate if its associated X is clean (with label 0) or backdoored (with label 1). The classifier g in the NP classification context corresponds to the detector g in our framework. The following result, a direct consequence of the famous Neyman-Pearson Lemma [4], gives the solution of g_α^*.

Lemma 2 (NP oracle classifier [5]). Let $\mathbb{P}_1$ and $\mathbb{P}_0$ be two probability measures with densities p_1 and p_0 respectively. Under mild continuity assumption, the NP oracle classifier is given by

$$g_\alpha^*(x) = \mathbf{1}\left\{\frac{p_1(x)}{p_0(x)} > C_\alpha\right\}$$

for some threshold C_α such that $\mathbb{P}_0\{p_1(X)/p_0(X) > C_\alpha\} \leq \alpha$ and $\mathbb{P}_1\{p_1(X)/p_0(X) > C_\alpha\} \geq \alpha$.

1.2.2 Proof for Theorem 2

Proof. By the definition of the Attacker's Goal, as specified in Section 3, the attacker faces a problem of the following:

$$\max_{\eta_1, f^{\text{poi}}} \mathbb{P}_{XY\sim\mathbb{P}_{\text{CLEAN}}}(f^{\text{poi}}(\eta_1(X)) = \eta_2(Y))$$
$$\text{subject to } |\mathbb{P}_{XY\sim\mathbb{P}_{\text{CLEAN}}}(f^{\text{poi}}(X) \neq Y) - \mathbb{P}_{XY\sim\mathbb{P}_{\text{CLEAN}}}(f^{\text{cl}}(X) \neq Y)| \leq \varepsilon.$$

Also, we assume that the (A) marginal clean data is normally distributed with mean 0 and covariance Σ, (B) the attacker employs a linear classifier $f_\theta(x) = \mathbf{1}\{\theta^\top x > 0\}$ for $\theta \in \mathbb{R}^d$, and (C) the attacker applies the backdoor transformation $\eta_1(x) = x + \gamma$, where $\gamma \in \mathrm{T}_c \triangleq \{u \in \mathbb{R}^d \mid \|u\|_2 = c, c > 0\}$.

Firstly, it is straightforward to check that, per the Attacker's Goal above, the attacker cannot obtain a backdoor classifier with $\theta^* = \mathbf{0}$, as in this scenario, the corresponding backdoor accuracy:

$$\mathbb{P}_{X\sim\mathcal{N}(\gamma,\Sigma)}(X^\top\theta^* > 0)$$

would be zero regardless of the backdoor trigger $\eta_1(x) = x + \gamma$ for $\gamma \in \mathrm{T}_c$. Next, for any non-zero $\theta^* \in \mathbb{R}^d$, suppose that there exist $\gamma_1, \gamma^* \in \mathrm{T}_c$ such that (γ_1, θ^*) and (γ^*, θ^*) are both the solutions of the Attacker's Goal.

As a result, both

$$\mathbb{P}_{X\sim\mathcal{N}(\gamma^*,\Sigma)}(X^\top\theta^* > 0)$$

and

$$\mathbb{P}_{\tilde{X}\sim\mathcal{N}(\gamma_1,\Sigma)}(\tilde{X}^\top\theta^* > 0)$$

are maximized, and equal, subject to $\gamma_1, \gamma^* \in \mathrm{T}_c$.

On the other hand, given the classifier θ^*, for any $\gamma \in \mathrm{T}_c$, we have the backdoor accuracy under θ^*:

$$\begin{aligned}
&\mathbb{P}_{X\sim\mathcal{N}(\gamma,\Sigma)}(X^\top\theta^* > 0), \\
&= \mathbb{P}_Z(Z > 0), \quad Z \sim \mathcal{N}(\gamma^\top\theta^*, \theta^*\Sigma(\theta^*)^\top), \\
&= 1 - \Phi\left(-\frac{\gamma^\top\theta^*}{(\theta^*\Sigma(\theta^*)^\top)^{1/2}}\right), \quad \Phi(\cdot) \text{ CDF of the standard normal distribution} \\
&= \Phi\left(\frac{\gamma^\top\theta^*}{(\theta^*\Sigma(\theta^*)^\top)^{1/2}}\right),
\end{aligned}$$

is maximized if and only if $\gamma = c\theta^*/\|\theta^*\|$ $(\theta^* \neq 0)$ due to the Cauchy-Schwarz inequality. As a result, we have $\gamma_1 = \gamma^*$. Hence, the optimal backdoor trigger γ^* corresponds to the backdoor classifier θ^* is unique and admits the form of $\gamma^* = c\theta^*/\|\theta^*\|$.

□

1.2.3 Proof of Theorem 3

Proof. Following the same setup in Theorem 2 and from the result in Theorem 2, the attacker *knows* both the clean and backdoor distribution. Hence, we conclude the result by the Neyman Pearson Lemma and the Lemma 2. ☐

2 Experiments Configurations

2.1 Data Description

CIFAR10: The CIFAR-10 dataset is a highly popular dataset in the field of machine learning research. It consists of 60,000 color images, each with a resolution of 32x32 pixels. The dataset is divided into 10 classes, with 6,000 images per class. Specifically, there is a training set with 50,000 images and a test set with 10,000 images.

GTSRB: The GTSRB dataset, known as the German Traffic Sign Recognition Benchmark, has gained popularity in the field of Backdoor Learning. It consists of a total of 60,000 images distributed among 43 different classes, with varying resolutions ranging from 32x32 to 250x250 pixels. The dataset is split into a training set containing 39,209 images and a test set containing 12,630 images.

Tiny ImageNet: Tiny ImageNet is comprised of a collection of 100,000 images belonging to 200 classes. Each class consists of 500 images, with 64x64 dimensions, resulting in colored images. The dataset is further divided into subsets, with 500 training images, 50 validation images, and 50 test images allocated for each class.

SST-2: The dataset used in our study is a modified version of the Stanford sentiment analysis dataset, specifically the 2-class variant known as SST-2. The SST-2 dataset consists of 9,613 samples, while another variant called SST-5 contains 11,855 reviews. Additionally, the SST dataset includes phrases associated with each of the sentences.

IMDB: The dataset used in our work is a binary dataset comprising of 12,500 movie reviews in each class. The reviews are multi-sentence and are presented in the form of long sentences. For our study, we have extracted the first 200 words from each review.

2.2 Packages used for generating backdoor attacks/data/models

In this section, we provide a detailed description of the experimental setup. To conduct our experiments, we utilized three open-source backdoor packages, the specifics of which are summarized in Table 1. For most computer vision (CV) backdoor attacks, we implemented them using both the `Backdoor ToolBox` [6] and the `BackdoorBench` [7] to ensure the consistency of our results. In the case of WaNet and SSBA, the `BackdoorBench` [8] package offered implementations for CIFAR-10, GTSRB, and Tiny ImageNet, so we utilized their package to obtain the latent representations. Regarding the TacT, Adaptive Patch, and Adaptive Blend, the `Backdoor ToolBox` package offered implementations for CIFAR-10, GTSRB, so we utilized their package to obtain the latent representations. Lastly, for NLP backdoor attacks, we relied on the `OpenBackdoor` package, which is specifically designed for NLP backdoors.

The results in the main text are directly obtained by running the `Default` scripts for the three packages, as specified in Table 2. Ablation studies on different model architectures and poisoning rates are included in Section 4 and 5, respectively.

2.3 On the selection of Hyperparameters

Within this section, we provide an outline for the selection of the shrinkage parameter β utilized in the Shrunk-Covariance Mahalanobis. The primary motivation behind employing SCM is to address the issue of unstable estimation in the inverse of the sample covariance matrix, which can result in non-IID property of samples within $\mathcal{D}^{\text{Val}}$ and future test data. Such non-IIDness can override the order information for those samples, affecting the FPR as described in Theorem 1.

To ensure the IID property of the samples within $\mathcal{D}^{\text{Val}}$ as well as future test samples, we propose the following approach for selecting the shrinkage parameter β. Firstly, we partition $\mathcal{D}^{\text{Val}}$ into two mutually exclusive datasets, namely $\mathcal{D}_1$ and $\mathcal{D}_2$, with the size of $\mathcal{D}_1$ greater than that of $\mathcal{D}_2$. Next, we compute the Shrunk-Covariance Mahalanobis (SCM) scores based on $\mathcal{D}_1$ and perform a search for an appropriate value of $\hat{\beta}$ that ensures the SCM scores, $s_{\hat{\beta}}(\mathcal{D}_1)$, for the samples within $\mathcal{D}_1$ and $s_{\hat{\beta}}(\mathcal{D}_2)$ are IID samples. To verify the IID property, we conduct a Two-sample Kolmogorov-Smirnov Test

Table 1: Open-source packages applied in our work

	`Backdoor ToolBox` [6]	`BackdoorBench` [7]	`OpenBackdoor` [8]
BadNets	✔	✔	
Blended	✔	✔	
TrojanNN	✔		
SIG	✔	✔	
Dynamic	✔	✔	
TacT	✔		
WaNet		✔	
SSBA		✔	
Adaptive-Blend	✔		
Adaptive-Patch	✔		
SOS			✔
LWP			✔

Table 2: Open-source packages used for the results in the *main text*

	`Backdoor ToolBox` [6]	`BackdoorBench` [7]	`OpenBackdoor` [8]
BadNets	✔		
Blended	✔		
TrojanNN	✔		
SIG	✔		
Dynamic	✔		
TacT	✔		
WaNet		✔	
SSBA		✔	
Adaptive-Blend	✔		
Adaptive-Patch	✔		
SOS			✔
LWP			✔

with a decision rule based on a p-value of 0.05. We repeat the above procedure ten times and observe that for CV (NLP) attacks, $\hat{\beta} = 0.5$ (0.7) satisfies the aforementioned requirement. Consequently, we set $\beta = 0.5$ for all CIFAR10, GTSRB experiments (for Tiny ImageNet: $\beta = 0.1$), and $\beta = 0.7$ for all NLP experiments. Ablation studies investigating different values of β are included in Appendix 6.

3 Omitted Experimental Results in Main Text

3.1 Results on the FPR for CIFAR10

This section presents the results of the False Positive Rate (FPR) analysis conducted on CIFAR10. Figure 1 illustrates the performance of our proposed CBD-SCM. Notably, our method consistently achieves FPR values that align with the theoretical upper bounds.

3.2 Results on the detection power (ROC) for IMDB Dataset

We assess the detection performance of our proposed method on IMDB (NLP) using the base uncased-BERT model. Specifically, we generate ROC curves for our methods under two advanced backdoor attacks, as illustrated in Figure 2. The results indicate that our method outperforms **all** other methods.

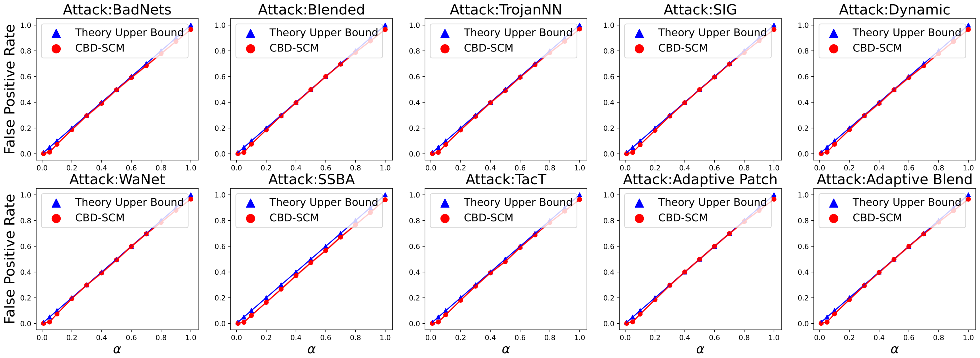

Figure 1: The mean FPRs of our proposed method on CIFAR10 are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases. Our method's FPRs consistently match the theoretical upper bounds.

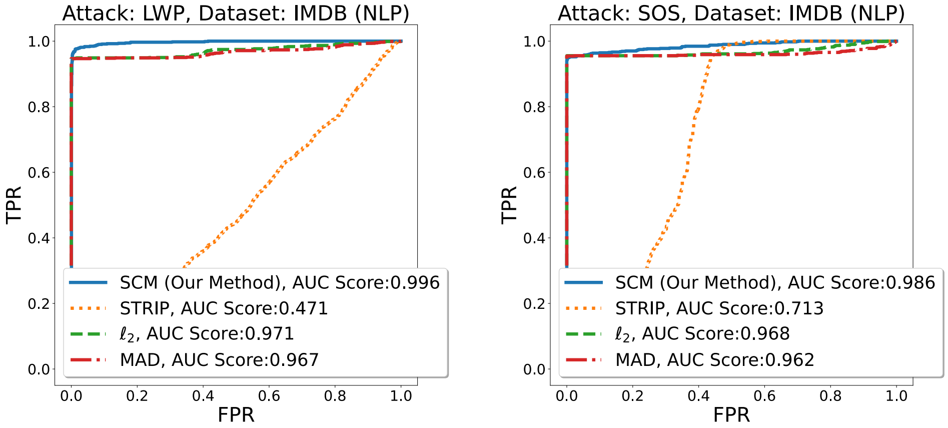

Figure 2: ROC of our SCM method on IMDB NLP. Our proposed method consistently outperforms all other methods.

4 Ablation Study: Different model architectures

In this section, we demonstrate the performance of our proposed methods using the VGG 19 model. The dimension of the `avgpool` layer in VGG 19 is 25088, which is impractically large. Therefore, we utilize the classifier layer with a dimension of 4096 for practical purposes. We have observed that the empirical FPR performance of our method consistently aligns with the theoretical bounds in all cases. Consequently, we omit the results for FPR performance and focus on reporting the results for detection power (AUCROC, ROC) in the following subsections.

4.1 Detection Power on CIFAR10

We assess the detection performance of our proposed method on CIFAR10 using VGG 19. We present the ROC curves for our method in Figure 3 and 4. It is evident that our SCM method consistently outperforms other methods in **all** attack scenarios. Specifically, for certain advanced attacks such as the Dynamic and WaNet, we observe a remarkable improvement of approximately 200% in terms of AUCROC.

4.2 Detection Power on GTSRB

We evaluate the detection performance of our proposed method on GTSRB using the VGG 19 model. The ROC curves for our method are displayed in Figure 5 and 6. It is evident from the results that our SCM method consistently outperforms other methods in the majority of attack scenarios. Notably, for advanced attacks like SSBA, we observe a significant improvement in terms of AUCROC.

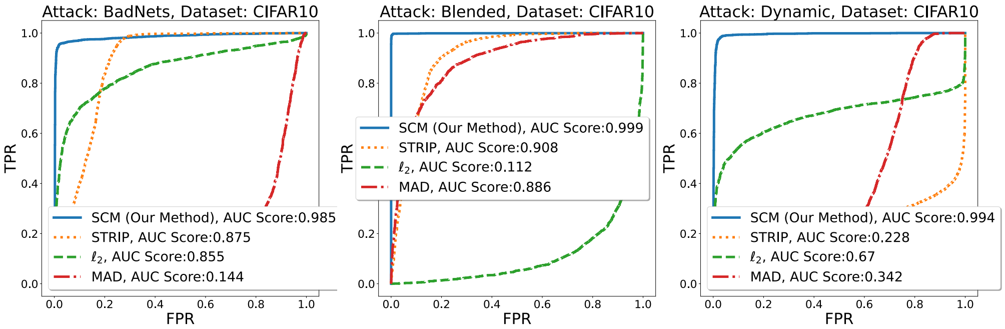

Figure 3: ROC of our method on CIFAR10 with VGG19. Our proposed method consistently outperforms other methods.

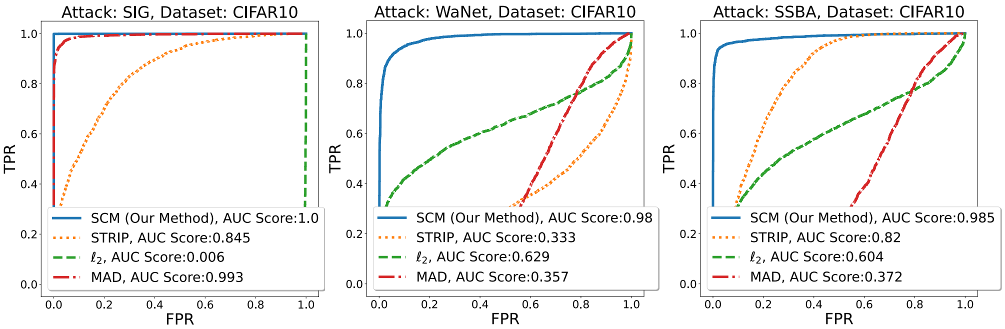

Figure 4: ROC of our method on CIFAR10 with VGG19. Our proposed method consistently outperforms other methods.

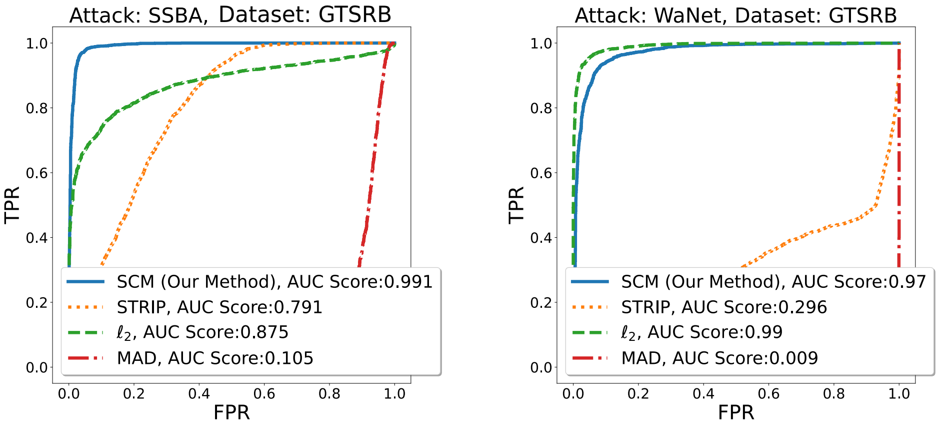

Figure 5: ROC of our method on GTSRB with VGG19.

5 Ablation Study: Different poisoning ratios

In this section, we provide the results obtained by using different poisoning ratios. The summary of the results for CIFAR10 can be found in Table 3. It is observed that the detection performance of our method exhibits minimal variations across different poisoning ratios. This finding further reinforces the robustness and universal effectiveness of our method.

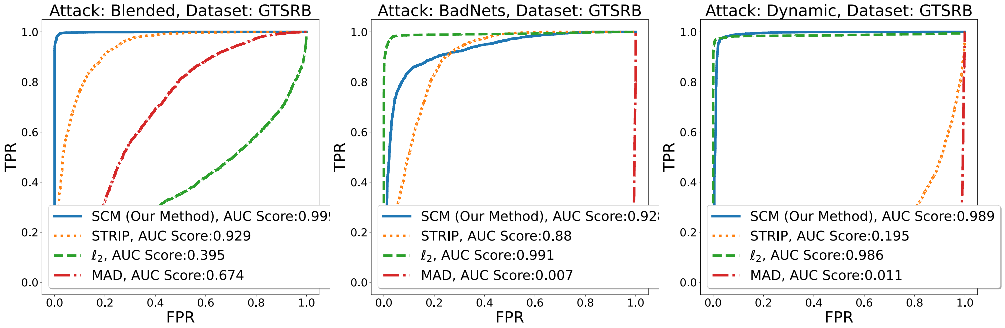

Figure 6: ROC of our method on GTSRB with VGG19.

Table 3: AUCROC score of our method on CIFAR10

Poisoning Ratio →	0.3%	1%	5%
BadNets	0.99	0.99	0.99
Blended	0.96	0.96	0.96
WaNet	0.88	0.91	0.94
SSBA	0.92	0.95	0.97

6 Ablation Study: Different choices of β

In this section, we provide empirical studies on the effect of using different β. Specifically, we select $\beta \in \{0.3, 0.4, 0.6, 0.7\}$.

6.1 Performances on FPR for CIFAR10

This section presents the results of the False Positive Rate (FPR) analysis conducted on CIFAR10 with different choices over β. Figure 7, 8, 9, 10 illustrate the FPR performance of our proposed CBD-SCM, with $\beta = 0.3, 0.4, 0.6, 0.7$ respectively. Our method consistently achieves FPR values that align with the theoretical upper bounds.

6.2 Performances on detection power (AUCROC, ROC) for CIFAR10

We assess the detection performance of our proposed method on CIFAR10 with different choices over β. Figure 11, 12, 13, 14 illustrate the ROC of our proposed SCM, with $\beta = 0.3, 0.4, 0.6, 0.7$ respectively. The AUCROC scores exhibit minimal variations when different values of β are used, indicating the consistent and robust effectiveness of our method. This observation reinforces the stability and universality of our selected $\beta = 0.5$ as reported in the main text.

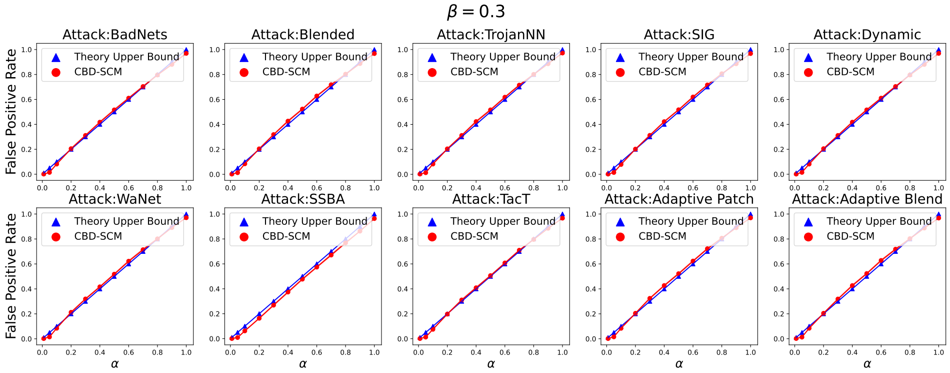

Figure 7: The mean FPRs of our proposed method with $\beta = 0.3$ on CIFAR10 are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

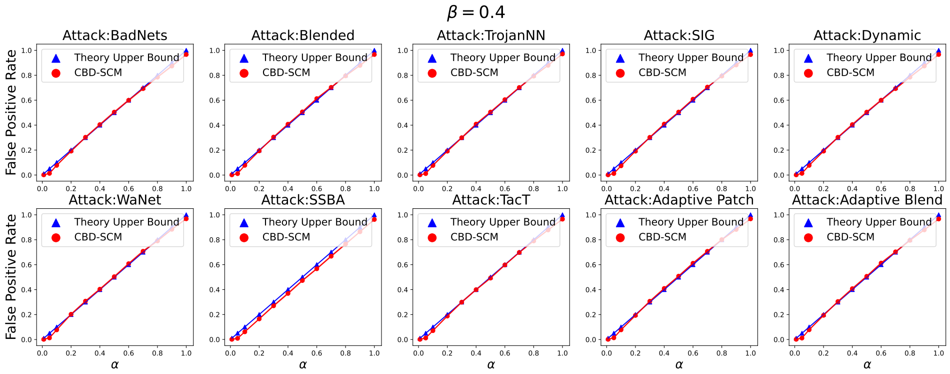

Figure 8: The mean FPRs of our proposed method with $\beta = 0.4$ on CIFAR10 are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

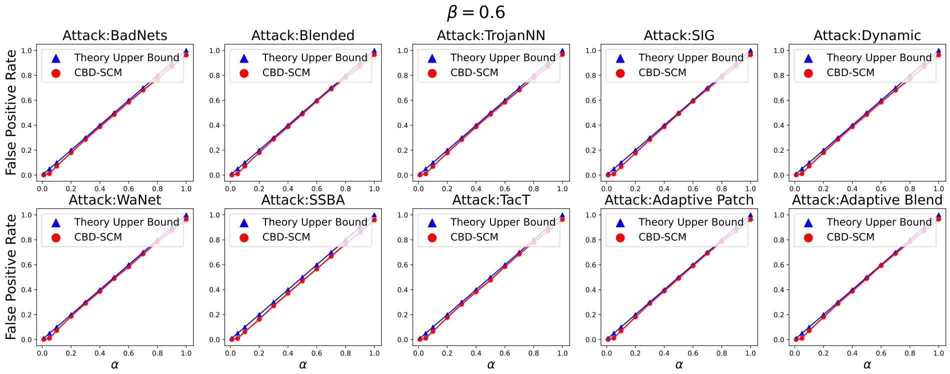

Figure 9: The mean FPRs of our proposed method with $\beta = 0.6$ on CIFAR10 are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

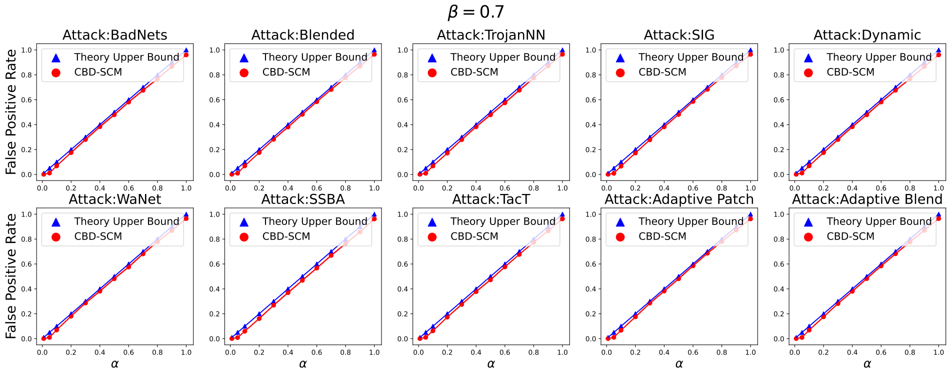

Figure 10: The mean FPRs of our proposed method with $\beta = 0.7$ on CIFAR10 are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

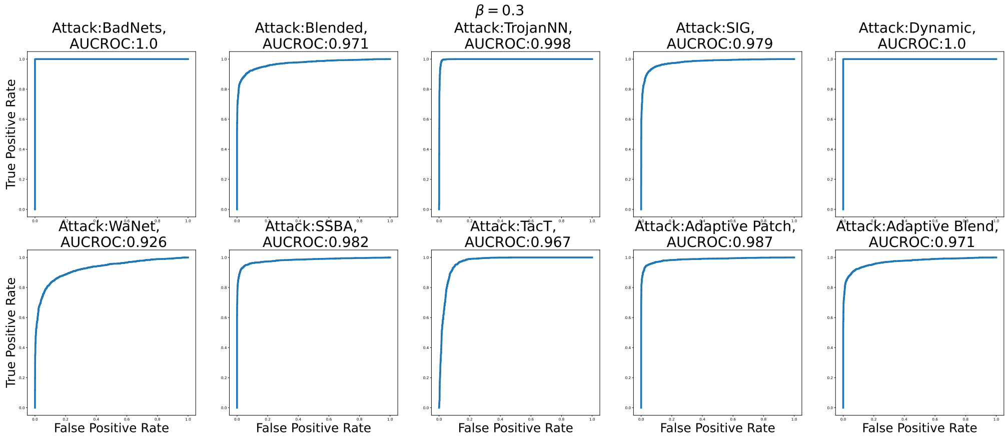

Figure 11: ROC of our method on CIFAR10 with $\beta = 0.3$

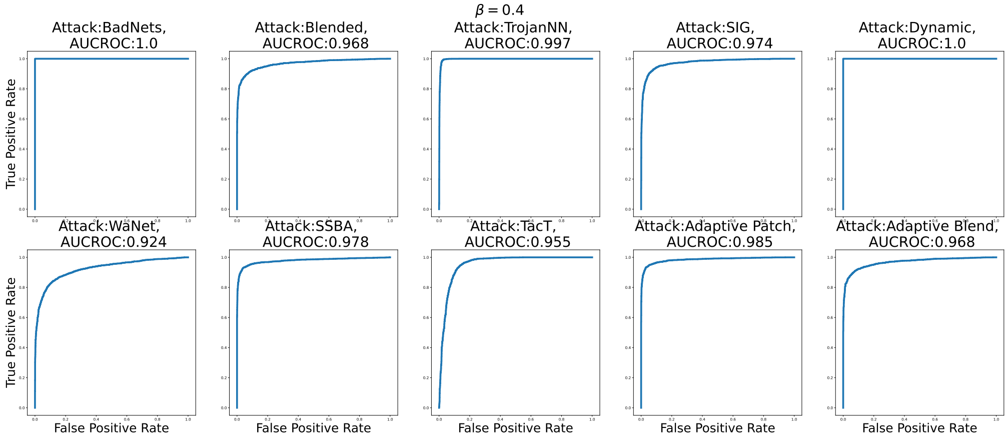

Figure 12: ROC of our method on CIFAR10 with $\beta = 0.4$

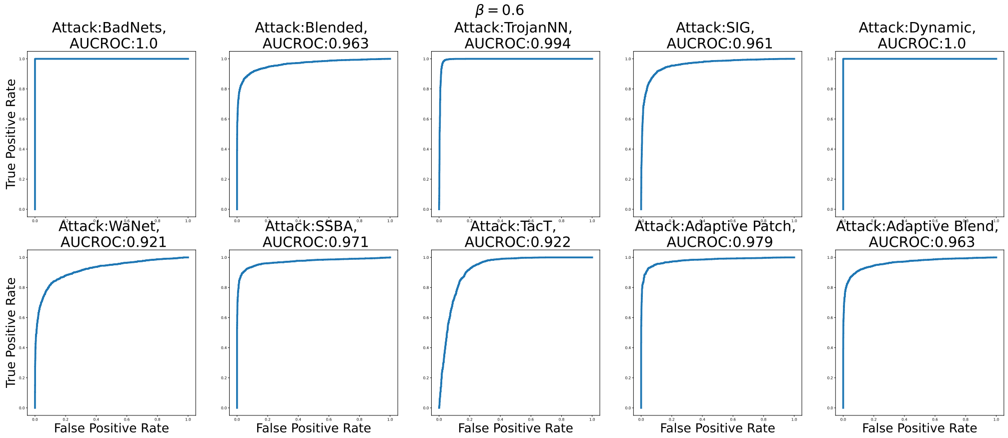

Figure 13: ROC of our method on CIFAR10 with $\beta = 0.6$

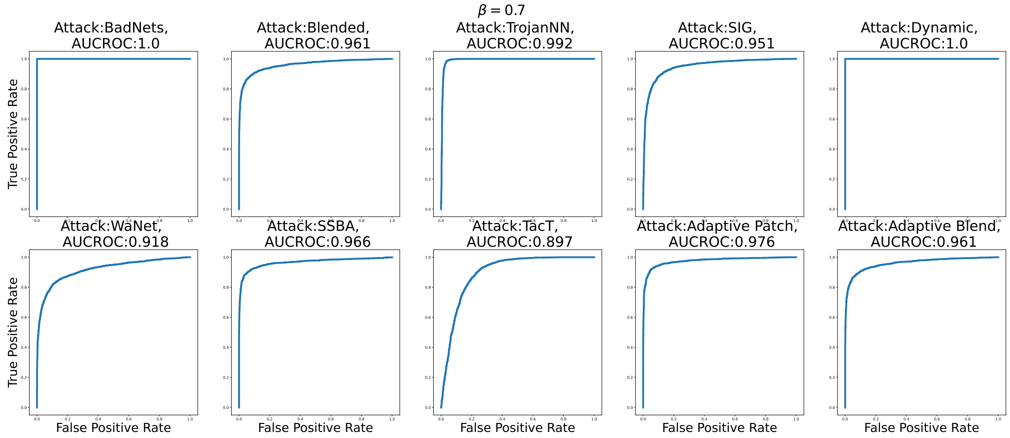

Figure 14: ROC of our method on CIFAR10 with $\beta = 0.7$

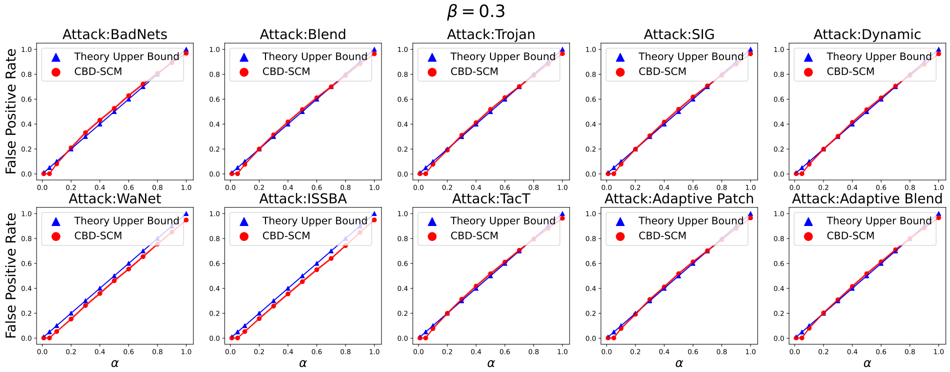

Figure 15: The mean FPRs of our proposed method with $\beta = 0.3$ on GTSRB are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

6.3 Performances on FPR for GTSRB

This section presents the results of the False Positive Rate (FPR) analysis conducted on GTSRB with different choices over β. Figure 15, 16, 17, 18 illustrate the FPR performance of our proposed CBD-SCM, with $\beta = 0.3, 0.4, 0.6, 0.7$ respectively. Our method consistently achieves FPR values that align with the theoretical upper bounds.

6.4 Performances on detection power (AUCROC, ROC) for GTSRB

We assess the detection performance of our proposed method on GTSRB with different choices over β. Figure 19, 20, 21 illustrate the ROC of our proposed SCM, with $\beta = 0.3, 0.4, 0.6$ respectively. The AUCROC scores exhibit minimal variations when different values of β are used, indicating the consistent and robust effectiveness of our method. This observation reinforces the stability and universality of our selected $\beta = 0.5$ as reported in the main text.

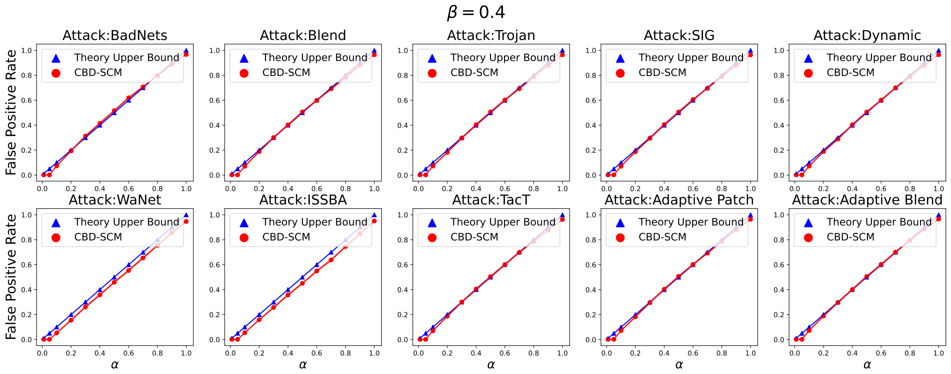

Figure 16: The mean FPRs of our proposed method with $\beta = 0.4$ on GTSRB are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

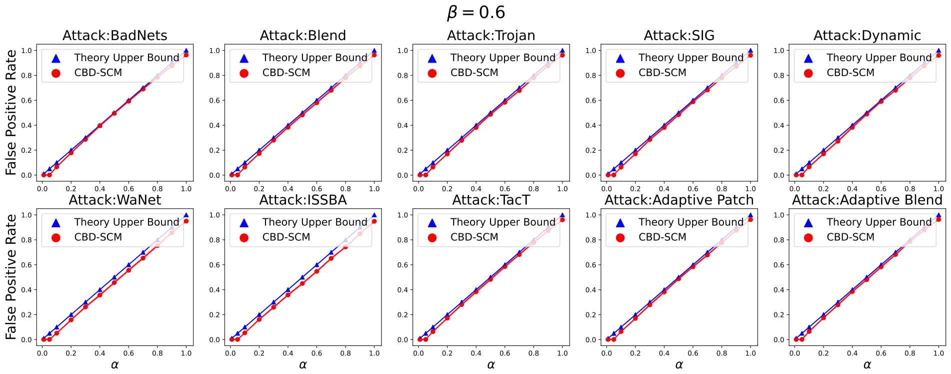

Figure 17: The mean FPRs of our proposed method with $\beta = 0.6$ on GTSRB are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

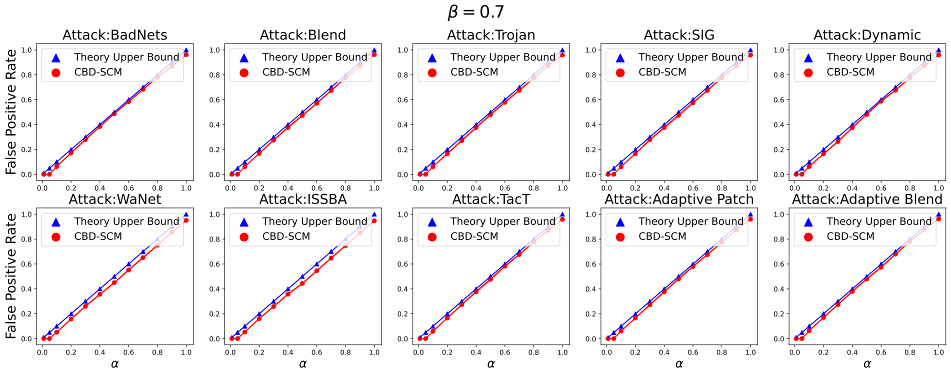

Figure 18: The mean FPRs of our proposed method with $\beta = 0.7$ on GTSRB are shown in each plot, which is independently replicated 10 times. The solid line represents the mean value, and the standard errors are < 0.01 for all cases.

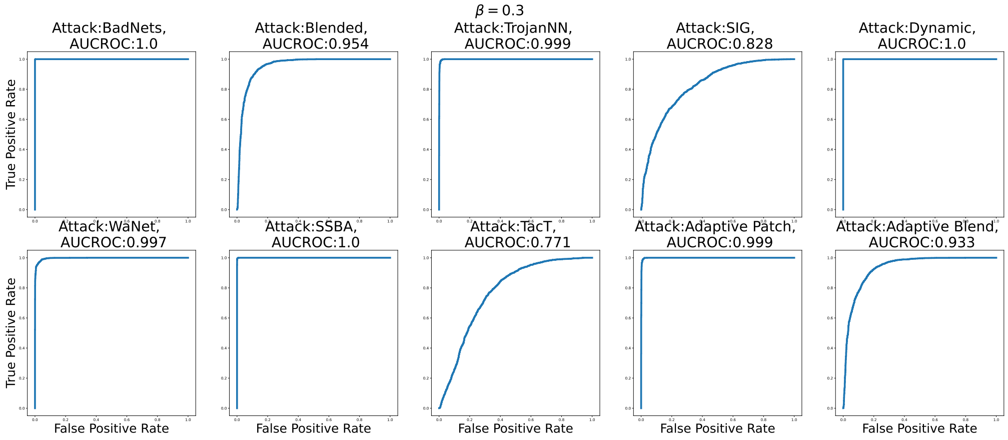

Figure 19: ROC of our method on GTSRB with $\beta = 0.3$

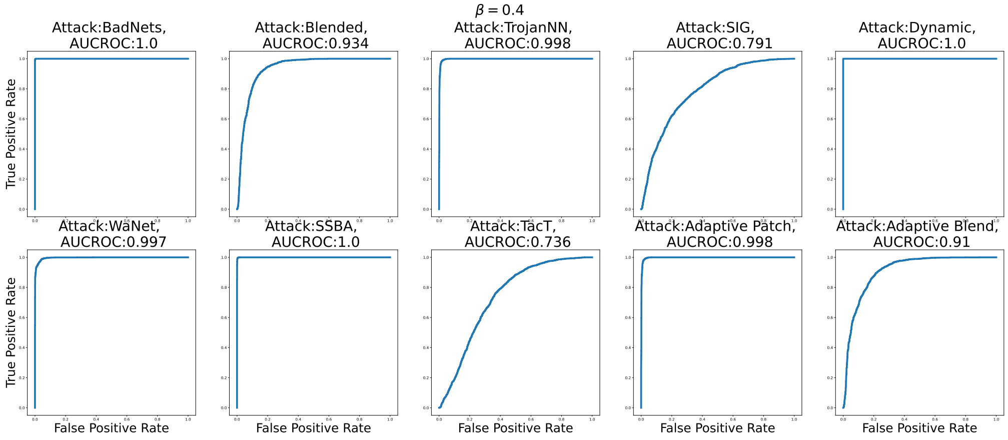

Figure 20: ROC of our method on GTSRB with $\beta = 0.4$

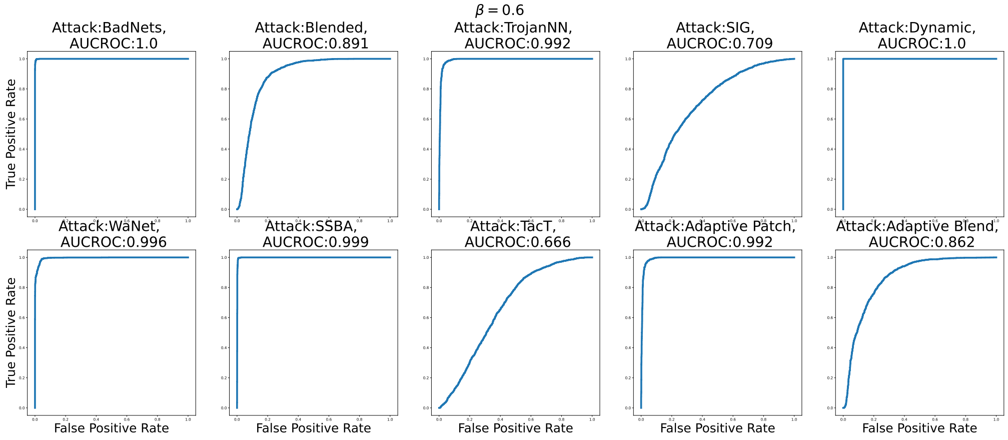

Figure 21: ROC of our method on GTSRB with $\beta = 0.6$

7 Performances under additional types of backdoor attacks

We evaluate our methods against backdoor attacks intentionally designed to challenge our defenses, such as LIRA [9] and M-attack (a custom-designed attack). These attacks are specifically crafted to diminish the distinction in latent spaces of the backdoored models by imposing regularization on the distances. LIRA employs Wasserstein distances, while M-attack utilizes Mahalanobis distances, to measure the dissimilarity between clean data and backdoor attacks. In Table 4, our methods consistently outperform state-of-the-art defenses, although there is a slight performance dip compared to diverse attack scenarios like BadNets and SSBA. Nevertheless, this outcome is reasonable since no defense can be universally effective against all attack variations.

Table 4: AUCROC score of our method on GTSRB against LIRA and M-attack.

Defenses →	SCM (Ours)	FREQ	SCALEUP
LIRA	**0.86**	0.71	0.79
M-attack	**0.82**	0.80	0.71

8 Comparison with other defenses

8.1 Performances comparison with recent detection-based backdoor defenses

We evaluate our methods alongside two recently developed detection-based defenses designed to counteract CV backdoor attacks: SCALEUP and FREQ. The summarized results are presented in Table 5 below. Our observations consistently demonstrate that our methods outperform SCALEUP and FREQ.

Table 5: AUCROC score of our method on GTSRB. The bestperforming method(s) are indicated in boldface.

Defenses →	SCM (Ours)	FREQ	SCALEUP	ABL	SPECTRE
BadNets	**0.99**	0.91	0.86	0.97	0.96
SSBA	**0.99**	0.51	0.72	0.81	0.56
Adaptive Patch	**0.87**	0.49	0.55	0.72	0.70
Adaptive Blend	**0.99**	0.56	0.51	0.59	0.62

8.2 Performances comparison with purifying-based training-stage backdoor defenses

We assess our methods in conjunction with two training-stage backdoor defenses, SPECTRE and ABL. As emphasized in the main text, these methods, while sharing the concept of distinguishing between clean and backdoor attacks, significantly differ from ours in terms of the threat model, methodology, and evaluation metrics. To ensure a fair comparison, we report the AUCROC scores for distinguishing between clean and backdoor data. The summarized results are presented in Table 5. Our observations consistently demonstrate that our methods outperform ABL and SPECTRE.

References

[1] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[2] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London*, vol. 231, no. 694-706, pp. 289–337, 1933.

[5] X. Tong, Y. Feng, and J. J. Li, "Neyman-pearson (np) classification algorithms and np receiver operating characteristic (np-roc) curves," *arXiv preprint arXiv:1608.03109*, 2016.

[6] T. Xie, "Backdoor toolbox," https://github.com/vtu81/backdoor-toolbox, 2022.

[7] B. Wu, H. Chen, M. Zhang, Z. Zhu, S. Wei, D. Yuan, and C. Shen, "Backdoorbench: A comprehensive benchmark of backdoor learning," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[8] G. Cui, L. Yuan, B. He, Y. Chen, Z. Liu, and M. Sun, "A unified evaluation of textual backdoor learning: Frameworks and benchmarks," *arXiv preprint arXiv:2206.08514*, 2022.

[9] K. Doan, Y. Lao, W. Zhao, and P. Li, "Lira: Learnable, imperceptible and robust backdoor attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 966–11 976.