A Appendix

A.1 Notation Table

The notations used in this paper are presented in Table 4.

Notations	Descriptions
$\mathcal{X}$	feature space $\subset \mathbb{R}^d$
$\mathcal{Y}$	label space $\subset \mathbb{R}^c$
x	feature
y	label
$\mathcal{D}$	probability distribution which is distributed in $\mathcal{X} \times \mathcal{Y}$
$\mathcal{D}_A$	probability distribution which is distributed in $\mathcal{X} \times \mathcal{Y}$ for benign behaviors
$\mathcal{D}_B$	probability distribution which is distributed in $\mathcal{X} \times \mathcal{Y}$ for malicious behavior (trigger)
N	number of samples for some dataset
N_A	number of samples for benign dataset
N_B	number of samples for trigger dataset
$N_{\mathcal{S}}$	number of samples for distilled dataset
D	dataset picked from the distribution $\mathcal{D}$ with N samples
D_A	dataset picked from the distribution $\mathcal{D}_A$ with N_A samples
D_B	dataset picked from the distribution $\mathcal{D}_B$ with N_B samples
$\mathcal{S}$	any dataset with $N_{\mathcal{S}}$ samples
$\mathcal{S}^*$	distilled dataset with $N_{\mathcal{S}}$ samples
$\mathcal{S}^*_A$	distilled dataset from D_A with $N_{\mathcal{S}}$ samples
T	trigger pattern $\in \mathbb{R}^d$
y_T	trigger label $\in \mathcal{Y}$
$\tilde{D}$	poisoned dataset which is the union from D_A and D_B
$\boldsymbol{X}$	the $N \times d$ matrix induced from the feature set in D.
$\boldsymbol{Y}$	the $N \times c$ matrix induced from the label set in D.
$\boldsymbol{X}_A$	the $N_A \times d$ matrix induced from the feature set in D_A.
$\boldsymbol{Y}_A$	the $N_A \times c$ matrix induced from the label set in D_A.
$\boldsymbol{X}_B$	the $N_B \times d$ matrix induced from the feature set in D_B.
$\boldsymbol{Y}_B$	the $N_B \times c$ matrix induced from the label set in D_B.
$\boldsymbol{X}_{\mathcal{S}}$	the $N_{\mathcal{S}} \times d$ matrix induced from the feature set in $\mathcal{S}$.
$\boldsymbol{Y}_{\mathcal{S}}$	the $N_{\mathcal{S}} \times c$ matrix induced from the label set in $\mathcal{S}$.
$\boldsymbol{X}_{AB}$	the $(N_A + N_B) \times d$ matrix induced from the feature set in $\tilde{D}$.
$\boldsymbol{Y}_{AB}$	the $(N_A + N_B) \times c$ matrix induced from the label set in $\tilde{D}$.
$k(\cdot, \cdot)$	the kernel
$\mathcal{H}_k$	the reproducing kernel hilbert space induced by kernel k.
λ	weight of regularization term.
$\lambda_{\mathcal{S}}$	weight of regularization term for $\mathcal{S}$.
ρ	penalty parameter.
$f_{\tilde{D}}$	the model trained on $\tilde{D}$ with the weight of the regularization term $\lambda \geq 0$.
$f_{\mathcal{S}}$	the model trained on $\mathcal{S}$ with the weight of the regularization term $\lambda_{\mathcal{S}} \geq 0$.

Table 4: Notation Table

A.2 Lemma 1 and Its Proof

Lemma 1 (Projection lemma)**.** *Given a synthetic dataset* $\mathcal{S} = \{(x_s, y_s)\}_{s=1}^{N_{\mathcal{S}}}$, *and a dataset* $\tilde{D} = \{(x_i, y_i)\})_{i=1}^{N_A+N_B}$ *where* $(N_A + N_B)$ *is the number of the samples of* $\tilde{D}$. *Suppose the kernel matrix*

$k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})$ is invertible, then we have

$$k(\cdot, x_i) = \underbrace{k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)}_{\in \mathcal{H}_{\mathcal{S}}} \tag{22}$$

$$+ \underbrace{[k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)]}_{\in \mathcal{H}_{\mathcal{S}}^{\perp}}, \quad \forall (x_i, y_i) \in \tilde{D} \tag{23}$$

where $\mathcal{H}_{\mathcal{S}} := span(\{k(\cdot, x_s) \in \mathcal{H}_k | (x_s, y_s) \in \mathcal{S}\})$ and $\mathcal{H}_{\mathcal{S}}^{\perp}$ is the collection of functions which is orthogonal to $\mathcal{H}_{\mathcal{S}}$ corresponding to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. The right hand side of (22) lies in $\mathcal{H}_{\mathcal{S}}$ while (23) lies in $\mathcal{H}_{\mathcal{S}}^{\perp}$. Thus, $k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ is the solution of the optimization problem:

$$\arg\min_{f \in \mathcal{H}_S} \sum_{(x_s, y_s) \in \mathcal{S}} \|f(x_s) - k(x_s, x_i)\|_2^2. \tag{24}$$

Proof. $k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ lies in $\mathcal{H}_{\mathcal{S}}$ is clearly. We just need to show that $k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ lies in $\mathcal{H}_{\mathcal{S}}^{\perp}$. Notice that

$$\langle k(\cdot, x_s), k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)) \rangle_{\mathcal{H}_k} \tag{25}$$

$$= k(x_s, x_i) - k(x_s, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)), \quad \forall (x_s, y_s) \in \mathcal{S}. \tag{26}$$

If we collect all $\langle k(\cdot, x_s), k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)) \rangle_{\mathcal{H}_k}$ for all $(x_s, y_s) \in \mathcal{S}$, we can obtain

$$k(\boldsymbol{X}_{\mathcal{S}}, x_i) - k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)) = k(\boldsymbol{X}_{\mathcal{S}}, x_i) - k(\boldsymbol{X}_{\mathcal{S}}, x_i) = 0. \tag{27}$$

This implies that $\langle k(\cdot, x_s), k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)) \rangle_{\mathcal{H}_k} = 0$ for $x_s \in \mathcal{S}$. $k(\cdot, x_i) - k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ lies in $\mathcal{H}_{\mathcal{S}}^{\perp}$. Eq. (27) also suggest that $k(x_s, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ is equal to $k(x_s, x_i)$ for all $(x_s, y_s) \in \mathcal{S}$. So, $k(\cdot, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)$ is the solution of Eq. (24). ■

A.3 Theorem 1 and Its Proof

Theorem 1 (Upper bound of conflict loss)**.** *The conflict loss $\mathcal{L}_{conflict}$ can be bounded as*

$$\mathcal{L}_{conflict} \leq \frac{1}{N_A + N_B} Tr(\boldsymbol{I} - k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB})[k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB}) + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})^2 \|\boldsymbol{Y}_{AB}\|_2^2 \tag{28}$$

where Tr is the trace operator, $k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB})$ is a $(N_A + N_B) \times (N_A + N_B)$ matrix, and $\boldsymbol{Y}_{AB}$ is a $(N_A + N_B) \times c$ matrix.

Proof. From Definition 1, we know that the kernel matrix $k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB})$ is positive semidefinite. Hence, there exist some unitary matrix $\boldsymbol{U}$ such that $k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB}) = \boldsymbol{U}\Sigma\boldsymbol{U}^T$ where Σ is some diagonal matrix with non-negative components. Then, from Eq. (12), we can express the upper bound of the conflict loss $\mathcal{L}_{\text{conflict}}$ as

$$\mathcal{L}_{\text{conflict}} = \frac{1}{N_A + N_B} \|\boldsymbol{I}\boldsymbol{Y}_{AB} - \boldsymbol{U}\Sigma\boldsymbol{U}^T[\boldsymbol{U}\Sigma\boldsymbol{U}^T + (N_A + N_B)\lambda \boldsymbol{I}]^{-1}\boldsymbol{Y}_{AB}\|_2^2 \tag{29}$$

$$= \frac{1}{N_A + N_B} \|\boldsymbol{U}\boldsymbol{I}\boldsymbol{U}^T\boldsymbol{Y}_{AB} - \boldsymbol{U}\Sigma\boldsymbol{U}^T[\boldsymbol{U}(\Sigma + (N_A + N_B)\lambda \boldsymbol{I})\boldsymbol{U}^T]^{-1}\boldsymbol{Y}_{AB}\|_2^2 \tag{30}$$

$$= \frac{1}{N_A + N_B} \|\boldsymbol{U}(\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I})]^{-1})\boldsymbol{U}^T\boldsymbol{Y}_{AB}|_2^2 \tag{31}$$

$$= \frac{1}{N_A + N_B} \|(\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I})]^{-1})\boldsymbol{U}^T\boldsymbol{Y}_{AB}|_2^2 \tag{32}$$

$$\leq \frac{1}{N_A + N_B} \|\text{Tr}(\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})\boldsymbol{U}^T\boldsymbol{Y}_{AB}|_2^2 \tag{33}$$

$$= \frac{1}{N_A + N_B} \text{Tr}(\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})^2 \|\boldsymbol{Y}_{AB}\|_2^2. \tag{34}$$

Moreover, we have

$$\begin{aligned}
&\mathrm{Tr}(\boldsymbol{I} - k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB})[k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB}) + (N_A + N_B)\lambda \boldsymbol{I}]^{-1}) \\
&= \mathrm{Tr}(\boldsymbol{U}(\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})\boldsymbol{U}^T) \quad (35)\\
&= \mathrm{Tr}((\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})\boldsymbol{U}^T\boldsymbol{U}) \quad (36)\\
&= \mathrm{Tr}((\boldsymbol{I} - \Sigma[\Sigma + (N_A + N_B)\lambda \boldsymbol{I}]^{-1})). \quad (37)
\end{aligned}$$

Combining Eq. (34) and Eq. (37) completes the proof. ■

A.4 Theorem 2 and Its Proof

Theorem 2 (Upper bound of projection loss)**.** *Suppose the kernel matrix of the synthetic dataset $k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})$ is invertible, $f_{\mathcal{S}}$ is the model trained on the synthetic dataset $\mathcal{S}$ with the regularization term $\lambda_{\mathcal{S}}$, where the projection loss $\mathcal{L}_{project} = \min_{\mathcal{S}} \mathbb{E}_{(x,y)\sim\tilde{D}} \ell(f_{\mathcal{S}}, (x, f_{\tilde{D}}(x)))$ can be bounded as*

$$\mathcal{L}_{project} \le \sum_{(x_i,y_i)\in\tilde{D}} \min_{\boldsymbol{X}_{\mathcal{S}}} \sum_{j=1}^{c} \frac{|\alpha_{i,j}|^2}{N_A + N_B} \|k(\boldsymbol{X}_{AB}, x_i) - k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{\mathcal{S}})k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1}k(\boldsymbol{X}_{\mathcal{S}}, x_i)\|_2^2. \quad (38)$$

where $\alpha_{i,j} := [[k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB}) + (N_A + N_B)\lambda \boldsymbol{I}]^{-1}\mathbf{Y}_{AB}]_{i,j}$, which is the weight of $k(\cdot, x_i)$ corresponding to $f^j_{\tilde{D}}$, $\boldsymbol{X}_{AB}$ is the $(N_A + N_B) \times d$ matrix corresponding to the features of $\tilde{D}$, $\boldsymbol{X}_{\mathcal{S}}$ is the $N_{\mathcal{S}} \times d$ matrix corresponding to the features of $\mathcal{S}$, $\mathbf{Y}_{AB}$ is the $(N_A + N_B) \times c$ matrix corresponding to the labels of $\tilde{D}$, $\mathbf{Y}_{\mathcal{S}}$ is the $N_{\mathcal{S}} \times c$ matrix corresponding to the labels of $\mathcal{S}$.

Proof. From (11), we know that

$$\begin{aligned}
f^j_{\tilde{D}}(x) &= [k(x, \boldsymbol{X}_{AB})[k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{AB}) + (N_A + N_B)\lambda \boldsymbol{I}]^{-1}\mathbf{Y}_{AB}]_j \\
&= \sum_{(x_i,y_i)\in\tilde{D}} \alpha_{i,j} k(x, x_i). \quad (39)
\end{aligned}$$

Then, we can bound the projection loss as

$$\begin{aligned}
&\mathcal{L}_{\text{project}} = \min_{\mathcal{S}} \mathbb{E}_{(x,y)\sim\tilde{D}} \ell(f_{\mathcal{S}}, (x, f_{\tilde{D}}(x))) \\
&= \min_{\mathcal{S}} \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \ell(f_{\mathcal{S}}, (x, f_{\tilde{D}}(x))) \quad (40)\\
&\le \sum_{(x_i,y_i)\in\tilde{D}} \min_{\mathcal{S}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \ell(f^j_{\mathcal{S}}, (x, \alpha_{i,j} k(x, x_i))) \right\} \quad (41)\\
&= \sum_{(x_i,y_i)\in\tilde{D}} \min_{\mathcal{S}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \|[k(x, \boldsymbol{X}_{\mathcal{S}})[k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}}) + N_{\mathcal{S}}\lambda_{\mathcal{S}} \boldsymbol{I}]^{-1}\mathbf{Y}_{\mathcal{S}}]_j - \alpha_{i,j} k(x, x_i)\|_2^2 \right\}. \quad (42)
\end{aligned}$$

For each $(x_i, y_i) \in \tilde{D}$, we have

$$\begin{aligned}
&\min_{\mathcal{S}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \|[k(x, \boldsymbol{X}_{\mathcal{S}})[k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}}) + N_{\mathcal{S}}\lambda_{\mathcal{S}} \boldsymbol{I}]^{-1}\mathbf{Y}_{\mathcal{S}}]_j - \alpha_{i,j} k(x, x_i)\|_2^2 \right\} \\
&\le \min_{\boldsymbol{X}_{\mathcal{S}}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \min_{\mathbf{Y}_{\mathcal{S}}} \|[k(x, \boldsymbol{X}_{\mathcal{S}})[k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}}) + N_{\mathcal{S}}\lambda_{\mathcal{S}} \boldsymbol{I}]^{-1}\mathbf{Y}_{\mathcal{S}}]_j - \alpha_{i,j} k(x, x_i)\|_2^2 \right\} \quad (43)\\
&= \min_{\boldsymbol{X}_{\mathcal{S}}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \min_{f_{i,j}\in\mathcal{H}_{\mathcal{S}}} \|f_{i,j}(x) - \alpha_{i,j} k(x, x_i)\|_2^2 \right\}. \quad (44)
\end{aligned}$$

Then, with the help of Lemma 1, we bound Eq. (44) as follows

$$\min_{\boldsymbol{X}_{\mathcal{S}}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \min_{f_{i,j}\in\mathcal{H}_{\mathcal{S}}} \|f_{i,j}(x) - \alpha_{i,j} k(x, x_i)\|_2^2 \right\}$$

$$\leq \min_{\boldsymbol{X}_{\mathcal{S}}} \left\{ \frac{1}{N_A + N_B} \sum_{(x,y)\in\tilde{D}} \sum_{j=1}^{c} \|\alpha_{i,j}[k(x, x_i) - k(x, \boldsymbol{X}_{\mathcal{S}}) k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1} k(\boldsymbol{X}_{\mathcal{S}}, x_i)]\|_2^2 \right\} \quad (45)$$

$$\leq \min_{\boldsymbol{X}_{\mathcal{S}}} \left\{ \sum_{j=1}^{c} \frac{|\alpha_{i,j}|^2}{N_A + N_B} \|k(\boldsymbol{X}_{AB}, x_i) - k(\boldsymbol{X}_{AB}, \boldsymbol{X}_{\mathcal{S}}) k(\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}})^{-1} k(\boldsymbol{X}_{\mathcal{S}}, x_i)\|_2^2 \right\} \quad (46)$$

We take the summation over $(x_i, y_i) \in \tilde{D}$ for Eq. (46) and then derive the upper bound. ■

A.5 Theorem 3 and Its Proof

Theorem 3 (Upper bound of generalization gap)**.** *Given a N-sample dataset D, sampled from the distribution $\mathcal{D}$, then the following generalization gap holds for all $g \in \mathcal{G}$ with probability at least $1 - \delta$:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} g((x, y)) - \sum_{(x_i,y_i)\in D} \frac{g((x_i, y_i))}{N} \leq 2\hat{\mathfrak{R}}_D(\mathcal{G}) + 3L_{\mathcal{D}}\Gamma_{\mathcal{D}} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \quad (47)$$

where $\boldsymbol{X}$ is the matrix corresponding to the features of D and $\hat{\mathfrak{R}}_D(\mathcal{G})$ is the empirical Rademacher's complexity.

Proof. Here we only sketch the proof, which mainly follows the proof of Theorem 3.3 in (Mohri et al., 2012), but is slightly modified under our assumption. First, we denote the maximum of the generalization gap for the dataset D as

$$\Phi(D) = \sup_{g\in\mathcal{G}} (\mathbb{E}_{(x,y)\in\mathcal{D}} g((x, y)) - \frac{1}{N} \sum_{(x_i,y_i)\in D} g((x_i, y_i))). \quad (48)$$

Consider another dataset D' sampled from the distribution $\mathcal{D}$. D and D' differ by only one sample, which is denoted as (x_N, y_N) and (x'_N, y'_N). Then, according to our assumption, we have

$$\Phi(D) - \Phi(D') \leq \sup_{g\in\mathcal{G}} (\frac{1}{N} g((x_N, y_N)) - \frac{1}{N} g((x'_N, y'_N)) \quad (49)$$

$$\leq \frac{L_D \|(x_N, y_N) - x'_N, y'_N)\|_2}{N} \quad (50)$$

$$\leq \frac{L_{\mathcal{D}}\Gamma_{\mathcal{D}}}{N}. \quad (51)$$

Then, we can apply McDiarmid's inequality on $\Phi(D)$. We can derive

$$\Phi(D) \leq \mathbb{E}_D \Phi(D) + L_{\mathcal{D}}\Gamma_{\mathcal{D}} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \quad (52)$$

which holds with probability at least $1 - \frac{\delta}{2}$. In the proof of Theorem 3.3 in (Mohri et al., 2012), we can also prove that $\mathbb{E}_D \Phi(D) \leq 2\mathfrak{R}(\mathcal{G})$, where $\mathfrak{R}(\mathcal{G})$ is Rademacher's complexity. Under our assumption, we notice that the empirical Rademacher complexity $\hat{\mathfrak{R}}_D(\mathcal{G})$ also satisfies

$$\hat{\mathfrak{R}}_D(\mathcal{G}) - \hat{\mathfrak{R}}_{D'}(\mathcal{G}) \leq \frac{L_{\mathcal{D}}\Gamma_{\mathcal{D}}}{N}. \quad (53)$$

So, we can apply McDiarmid's inequality again and obtain

$$\mathfrak{R}(\mathcal{G}) \leq \hat{\mathfrak{R}}_D(\mathcal{G}) + L_{\mathcal{D}}\Gamma_{\mathcal{D}} \sqrt{\frac{\log \frac{2}{\delta}}{2N}}, \quad (54)$$

which holds with probability at least $1-\frac{\delta}{2}$. Combine (52), (54) and the fact that $\mathbb{E}_D\Phi(D) \leq 2\mathfrak{R}(\mathcal{G})$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}g((x,y)) - \sum_{(x_i,y_i)\in D}\frac{g((x_i,y_i))}{N} \leq 2\hat{\mathfrak{R}}_D(\mathcal{G}) + 3L_{\mathcal{D}}\Gamma_{\mathcal{D}}\sqrt{\frac{\log\frac{2}{\delta}}{2N}}. \tag{55}$$

which holds with probability at least $1-\delta$. ■

A.6 Pseudocode for The Simplest Form of KIP-based Backdoor Attack

Algorithm 1 The Simplest Form of KIP-based Backdoor Attack

Require: benign dataset D_A, initial trigger T_0, trigger label y_T, mask m, size of distilled dataset $N_{\mathcal{S}}$, training step STEP > 0, batch size BATCH > 0, mix ratio $\rho_m > 0$, learning rate $\eta > 0$.
Ensure: synthetic dataset $\mathcal{S}^*$

```
N ← 1
S ← Randomly sample N_S data from D_A as initial distilled dataset.
D_B ← {(x_b, y_b) := ((1 − m) ⊙ x + m ⊙ T, y_T) | (x_a, y_a) ∈ D_A}
while N ≤ STEP do
    (X_A^batch, Y_A^batch) ← Randomly sample BATCH data from D_A.
    (X_B^batch, Y_B^batch) ← Randomly sample BATCH data from D_B.
    D̃^batch ← (X_A^batch, Y_A^batch) ∪ (X_B^batch, Y_B^batch)
    S ← S − η∇_S L(S, )                          ▷ L is defined in Eq. (8).
    N ← N + 1
end while
S* ← S
```

A.7 Pseudocode for relax-trigger

Algorithm 2 relax-trigger

Require: benign dataset D_A, initial trigger T_0, trigger label y_T, mask m, training step STEP > 0, batch size BATCH > 0, mix ratio $\rho_m > 0$, penalty parameter $\rho > 0$, learning rate $\eta > 0$.
Ensure: optimized T^*

```
T ← T_0
N ← 1
S_A* ← Apply KIP to D_A                          ▷ We use S_A* to denote S* from D_A
while N ≤ STEP do
    (X_A^batch, Y_A^batch) ← Randomly pick BATCH samples from D_A
    (X^batch, Y^batch) ← Randomly pick BATCH × ρ_m samples from D_A
    (X_B^batch, Y_B^batch) ← {(x_b, y_b) := ((1 − m) ⊙ x + m ⊙ T, y_T) | (x, y) ∈ (X^batch, Y^batch)}
    T ← T − η∇_T L(S_A*, (X_A^batch, Y_A^batch), (X_B^batch, Y_B^batch), ρ)    ▷ L is defined in Eq. (21).
    N ← N + 1
end while
T* ← T
```

A.8 Extra Experiments.

In Tables 5∼ 8, we provide extra experimental results.

Trig.\Def.	None		SCAn		AC		SS		Strip	
	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2	29.47 (0.44)	25.67 (4.35)	28.70 (2.23)	35.29 (4.40)	29.93 (1.66)	26.17 (9.14)	28.13 (1.45)	19.24 (4.39)	27.52 (3.30)	28.73 (13.83)
4 × 4	29.72 (1.62)	31.08 (8.07)	32.43 (1.87)	26.70 (2.30)	30.57 (0.87)	30.20 (3.17)	30.48 (0.91)	11.07 (1.53)	25.95 (4.07)	25.37 (11.61)
8 × 8	32.00 (1.03)	51.65 (12.41)	30.78 (2.08)	37.45 (9.23)	29.57 (0.74)	35.99 (2.63)	28.46 (2.56)	12.77 (3.79)	28.39 (3.18)	15.94 (2.67)
16 × 16	34.61 (1.01)	85.65 (17.12)	33.88 (1.65)	44.24 (3.85)	31.96 (0.53)	59.56 (21.29)	30.70 (0.09)	44.77 (25.11)	27.96 (7.42)	43.00 (25.69)
32 × 32	33.78 (0.53)	100.00 (0.00)	34.54 (0.93)	100.00 (0.00)	32.25 (2.29)	33.33 (47.14)	29.04 (0.91)	33.33 (47.14)	26.93 (8.95)	0.00 (0.00)
Trig.\Def.			ABL		NAD		STRIP		FP	
			CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2			25.31 (4.40)	13.67 (6.65)	36.27 (0.54)	7.32 (1.77)	25.05 (2.73)	27.25 (8.88)	15.28 (2.96)	35.74 (41.58)
4 × 4			24.25 (2.70)	5.81 (5.26)	36.29 (2.07)	6.57 (0.81)	26.85 (1.39)	27.92 (7.10)	14.67 (3.67)	69.41 (31.97)
8 × 8			22.97 (5.54)	19.06 (8.02)	36.96 (1.73)	14.53 (5.15)	28.84 (0.89)	40.99 (8.03)	19.41 (2.06)	75.62 (17.01)
16 × 16			28.43 (2.31)	64.73 (34.70)	37.01 (1.16)	29.42 (1.39)	31.09 (0.80)	22.80 (22.02)	21.25 (2.94)	19.08 (21.95)
32 × 32			22.12 (2.74)	66.67 (47.14)	32.62 (7.23)	66.67 (47.14)	22.99 (10.01)	0.00 (0.00)	17.67 (5.61)	33.33 (47.14)

Table 5: Defenses for simple-trigger on CIFAR-10 with size 500.

Trig.\Def.	None		SCAn		AC		SS		Strip	
	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2	37.78 (2.09)	10.71 (2.33)	40.17 (8.89)	5.18 (2.40)	27.61 (1.09)	9.43 (4.63)	14.55 (1.71)	6.47 (3.30)	28.52 (2.29)	5.22 (4.19)
4 × 4	39.07 (2.27)	18.67 (8.87)	33.47 (6.75)	3.59 (1.11)	23.22 (7.42)	9.71 (5.61)	15.22 (2.03)	2.28 (2.36)	26.56 (4.15)	7.67 (2.29)
8 × 8	38.89 (1.68)	47.53 (9.30)	21.48 (12,74)	2.85 (0.37)	28.60 (1.36)	8.25 (4.15)	12.86 (1.78)	8.60 (6.45)	34.96 (5.02)	8.69 (1.85)
16 × 16	37.92 (2.84)	84.24 (3.60)	32.99 (4.62)	11.04 (8.85)	26.91 (2.77)	56.51 (39.03)	14.01 (1.84)	4.76 (4.32)	29.22 (1.59)	37.06 (41.29)
32 × 32	41.97 (0.97)	66.67 (47.14)	33.47 (9.14)	33.33 (47.14)	27.20 (3.83)	33.33 (47.14)	13.99 (0.79)	33.33 (47.14)	35.07 (1.95)	33.33 (47.14)
Trig.\Def.			ABL		NAD		STRIP		FP	
			CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2			32.31 (4.67)	5.85 (1.90)	94.05 (0.51)	0.40 (0.04)	22.36 (14.79)	5.48 (4.00)	14.57 (10.65)	8.13 (5.75)
4 × 4			36.25 (4.05)	6.68 (3.75)	93.75 (0.41)	0.29 (0.09)	35.23 (2.17)	16.99 (8.05)	23.73 (4.09)	11.45 (11.59)
8 × 8			25.49 (15.17)	11.37 (3.68)	93.72 (0.11)	0.45 (0.19)	34.82 (1.63)	42.69 (7.99)	24.67 (1.23)	33.71 (29.71)
16 × 16			37.15 (1.59)	63.80 (31.74)	93.78 (0.07)	0.09 (0.08)	34.04 (2.48)	50.69 (10.77)	25.35 (2.59)	40.16 (30.45)
32 × 32			34.56 (5.89)	33.33 (47.14)	94.05 (0.17)	0.00 (0.00)	37.87 (0.80)	0.00 (0.00)	25.43 (2.06)	66.67 (47.14)

Table 6: Defenses for simple-trigger on GTSRB with distilled dataset size = 430.

Trig.\Def.	None		SCAn		AC		SS		Strip	
	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2	72.23 (2.76)	1.03 (0.42)	72.82 (12.40)	2.37 (1.69)	63.73 (7.87)	2.06 (0.56)	44.50 (2.26)	3.08 (0.81)	76.63 (3.00)	1.00 (0.47)
4 × 4	73.29 (1.22)	1.08 (0.37)	81.19 (2.30)	1.18 (0.09)	71.79 (1.78)	2.37 (0.40)	45.16 (5.15)	3.53 (1.12)	73.28 (11.33)	0.94 (0.46)
8 × 8	73.29 (0.26)	8.08 (4.20)	79.28 (2.69)	3.84 (1.72)	62.79 (9.81)	6.30 (3.27)	40.46 (4.65)	17.40 (7.28)	74.84 (1.37)	2.78 (1.41)
16 × 16	73.12 (0.69)	70.10 (13.96)	81.39 (5.97)	61.28 (19.18)	68.37 (2.70)	46.99 (24.09)	39.58 (0.15)	22.70 (11.70)	73.52 (6.41)	28.07 (18.36)
32 × 32	74.13 (1.39)	100.00 (0.00)	76.85 (5.79)	100.00 (0.00)	45.93 (21.02)	33.33 (47.14)	44.87 (6.65)	33.33 (47.14)	83.42 (0.65)	0.00 (0.00)
Trig.\Def.			ABL		NAD		STRIP		FP	
			CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
2 × 2			77.04 (4.01)	1.86 (2.06)	94.68 (0.60)	0.32 (0.05)	65.05 (2.53)	0.95 (0.39)	51.95 (2.11)	0.29 (0.17)
4 × 4			79.11 (1.23)	1.04 (0.93)	94.73 (0.26)	0.27 (0.04)	65.94 (1.11)	1.02 (0.34)	52.61 (0.49)	0.04 (0.06)
8 × 8			75.89 (1.80)	4.14 (2.19)	94.76 (0.33)	0.24 (0.07)	65.90 (0.23)	7.65 (4.03)	51.83	0.23
16 × 16			79.29 (3.46)	73.46 (3.25)	94.67 (0.29)	0.07 (0.04)	65.73 (0.65)	59.97 (5.28)	53.38 (3.29)	28.53 (39.26)
32 × 32			79.98 (2.47)	100.00 (0.00)	94.74 (0.25)	0.00 (0.00)	66.74 (1.34)	0.00 (0.00)	53.79 (2.14)	100.00 (0.00)

Table 7: Defenses for simple-trigger on GTSRB with distilled dataset size 2150.

Data. (Size)\Def.	None		SCAn		AC		SS		Strip	
	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
CIFAR-10 (100)	26.28 (1.56)	42.10 (4.16)	27.23 (1.37)	55.63 (4.78	26.18 (1.70)	44.56 (13.65)	21.13 (0.67)	5.98 (0.96)	25.05 (0.79)	50.27 (15.74)
CIFAR-10 (500)	33.98 (0.63)	90.87 (4.01)	35.40 (0.77)	82.63 (3.21)	33.23 (2.63)	58.22 (38.26)	32.52 (0.47)	29.00 (10.45)	34.33 (0.23)	83.42 (6.34)
GTSRB (430)	37.49 (1.98)	69.14 (2.84)	36.86 (1.28)	53.68 (4.41)	23.91 (2.58)	28.27 (11.06)	13.39 (1.19)	25.09 (12.68)	30.88 (2.18)	50.50 (13.23)
GTSRB (2150)	75.40 (0.39)	65.28 (2.15)	82.47 (1.81)	70.51 (3.14)	65.84 (8.99)	61.81 (1.22)	39.95 (3.48)	26.01 (11.47)	72.43 (5.26)	63.77 (1.18)
Data. (Size)\Def.			ABL		NAD		STRIP		FP	
			CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)	CTA (%)	ASR (%)
CIFAR-10 (100)			14.03 (0.92)	73.30 (17.71)	31.60 (2.10)	21.67 (18.72)	23.83 (1.30)	35.96 (5.88)	15.87 (1.08)	33.50 (38.47)
CIFAR-10 (500)			29.10 (1.51)	13.41 (5.42)	37.75 (1.19)	44.20 (10.85)	30.61 (0.49)	40.13 (18.51)	20.83 (1.43)	32.28 (44.66)
GTSRB (430)			32.26 (2.68)	45.02 (6.54)	93.32 (0.34)	67.18 (2.36)	33.88 (1.82)	61.89 (2.63)	22.79 (2.98)	53.54 (13.43)
GTSRB (2150)			80.90 (1.39)	65.85 (0.63)	94.34 (0.13)	33.68 (1.43)	67.91 (0.31)	45.40 (2.13)	55.81 (1.39)	68.73 (0.57)

Table 8: Defenses for relax-trigger on CIFAR-10 and GTSRB.

B Ablation Studies

B.1 KIP-based backdoor attack on ImageNet

We perform our KIP-based backdoor attack on ImageNet. In our experiment, we randomly choose ten sub-classes to perform our experiment. We also resize each image in the ImageNet into 128x128. The experimental results show that our KIP-based backdoor attack is effective (see Table 9).

B.2 Impact of IPC on KIP-based Backdoor Attack

We examine the efficacy of KIP-based backdoor attack influenced by IPC (Image Per Class). We examine the efficacy of simple-trigger and relax-trigger under different sizes of synthetic dataset (IPC 10 $\sim$ IPC 50). The experimental results show that both CTA and ASR are gradually rising as

Trigger-type	Dataset	Model	IPC (Image Per Class)	CTA (%)	ASR (%)
simple-trigger	ImageNet	NTK	10	15.00	100.00
simple-trigger	ImageNet	NTK	50	16.60	100.00
relax-trigger	ImageNet	NTK	10	16.40	100.00
relax-trigger	ImageNet	NTK	50	17.00	100.00

Table 9: Efficacy of KIP-based backdoor attack on ImageNet.

the IPC increases (see Table 10). The corresponding experiments for DOORPING is presented in Table 11.

Dataset	Trigger-type	IPC (Image Per Class)	CTA (%)	ASR (%)
CIFAR-10	simple-trigger	10	41.70 (0.25)	100 (0.00)
CIFAR-10	simple-trigger	20	42.58 (0.23)	100 (0.00)
CIFAR-10	simple-trigger	30	43.29 (0.35)	100 (0.00)
CIFAR-10	simple-trigger	40	43.55 (0.42)	100 (0.00)
CIFAR-10	simple-trigger	50	43.66 (0.40)	100 (0.00)
CIFAR-10	relax-trigger	10	41.66 (0.01)	100 (0.00)
CIFAR-10	relax-trigger	20	42.46 (0.01)	100 (0.00)
CIFAR-10	relax-trigger	30	42.99 (0.08)	100 (0.00)
CIFAR-10	relax-trigger	40	43.10 (0.09)	100 (0.00)
CIFAR-10	relax-trigger	50	43.64 (0.40)	100 (0.00)
GTSRB	simple-trigger	10	67.56 (0.60)	100 (0.00)
GTSRB	simple-trigger	20	69.44 (0.35)	100 (0.00)
GTSRB	simple-trigger	30	70.24 (0.38)	100 (0.00)
GTSRB	simple-trigger	40	70.84 (0.32)	100 (0.00)
GTSRB	simple-trigger	50	71.27 (0.24)	100 (0.00)
GTSRB	relax-trigger	10	68.73 (0.67)	95.26 (0.54)
GTSRB	relax-trigger	20	70.38 (0.03)	94.85 (0.13)
GTSRB	relax-trigger	30	71.26 (0.02)	95.73 (0.32)
GTSRB	relax-trigger	40	71.81 (0.01)	95.84 (0.18)
GTSRB	relax-trigger	50	71.54 (0.33)	95.08 (0.33)

Table 10: Efficacy of KIP-based backdoor attack influenced by the size of the synthetic dataset.

Dataset	Trigger-type	IPC (Image Per Class)	CTA (%)	ASR (%)
CIFAR-10	DOORPING	10	36.35 (0.42)	80.00 (40.00)
CIFAR-10	DOORPING	20	37.65 (0.42)	70.00 (45.83)
CIFAR-10	DOORPING	30	38.48 (0.36)	90.00 (30.00)
CIFAR-10	DOORPING	40	37.78 (0.61)	70.00 (45.83)
GTSRB	DOORPING	10	68.03 (0.92)	90.00 (30.00)
GTSRB	DOORPING	20	81.45 (0.46)	80.00 (40.00)
GTSRB	DOORPING	30	81.62 (0.71)	100.00 (0.00)

Table 11: Efficacy of DOORPING influenced by the size of the synthetic dataset.

B.3 Cross Model Ability of KIP-based backdoor attack

The experiment for cross model ability is presented in Table 12. We train the distilled dataset poinsoned by simple-trigger and relax-trigger on 3-layers MLP and 3-layers ConvNet. The experimental results show that both CTA and ASR go up as we increase the IPC (Image Per Class), which suggests that the cross model issue may be relieved as the IPC is large enough.

Dataset	Trigger-type	IPC (Image Per Class)	Cross_model	CTA (%)	ASR(%)
CIFAR-10	simple-trigger	10	MLP	11.58 (2.10)	40.00 (48.98)
CIFAR-10	simple-trigger	10	CNN	47.37 (7.44)	40.00 (48.98)
CIFAR-10	simple-trigger	10	NTK (baseline)	41.70 (0.25)	100.00 (0.00)
CIFAR-10	simple-trigger	50	MLP	48.08 (4.72)	40.00 (48.98)
CIFAR-10	simple-trigger	50	CNN	95.96 (1.10)	100.00 (0.00)
CIFAR-10	simple-trigger	50	NTK (baseline)	43.66 (0.40)	100.00 (0.00)
CIFAR-10	relax-trigger	10	MLP	10.52 (7.44)	19.40 (38.80)
CIFAR-10	relax-trigger	10	CNN	64.21 (6.98)	81.80 (11.44)
CIFAR-10	relax-trigger	10	NTK (baseline)	41.66 (0.74)	100.00 (0.00)
CIFAR-10	relax-trigger	50	MLP	44.24 (4.49)	78.28 (24.57)
CIFAR-10	relax-trigger	50	CNN	93.13 (2.24)	82.80 (6.53)
CIFAR-10	relax-trigger	50	NTK (baseline)	43.64 (0.40)	100.00 (0.00)

Table 12: Experiment of cross model ability of KIP-based backdoor attack.

B.4 TRANSFERABILITY OF KIP-BASED BACKDOOR ATTACK

Our KIP-based backdoor attack can evade other data distillation techniques. In particular, we perform experiments to examine the transferability of our theory-induced triggers. We first use our simple-trigger and relax-trigger to poison the dataset. Then, we distill dataest with a different distillation method, FRePo (Zhou et al., 2022) and DM (Zhao & Bilen, 2023). The experimental results shows that our triggers can successfully transfer to the FrePo and DM (see Table 13 and Table 14).

Trigger-type	Dataset	IPC (Image Per Class)	Distillation	Model	CTA (%)	ASR (%)
CIFAR-10	simple-trigger	10	FRePO	ConvNet	60.32	83.10
CIFAR-10	relax-trigger	50	FRePO	ConvNet	68.34	81.61

Table 13: Experiment of transferability for FRePO.

Trigger-type	Dataset	IPC (Image Per Class)	Distillation	Model	CTA (%)	ASR (%)
Cifar10	simple-trigger	10	DM	MLP	36.41	77.03
Cifar10	simple-trigger	50	DM	MLP	36.88	76.79
Cifar10	relax-trigger	10	DM	MLP	36.31	76.04
Cifar10	relax-trigger	50	DM	MLP	36.81	76.21

Table 14: Experiment of transferability for DM.

B.5 KIP-BASED BACKDOOR ATTACK ON NAS AND CL

We train our distilled dataset poinsoned by simple-trigger and relax-trigger in different scenarios, neural architecture search (NAS) and continual learning (CL). The experimental results are shown in Table 15 and Table 16.

Trigger-type	Dataset	IPC (Image Per Class)	Scenario	CTA (%)	ASR (%)
simple-trigger	CIFAR-10	50	NAS	37.49(3.44)	100.00(0.00)
relax-trigger	CIFAR-10	50	NAS	36.43(3.62)	86.23(3.22)

Table 15: Experiment for NAS. The experiment result shows that our triggers remain effective for NAS.

Note that the details about our implementation of NAS and CL are described below.

Trigger-type	Dataset	IPC (Image Per Class)	Scenario	CTA (%)	ASR (%)
simple-trigger	CIFAR-10	50	CL	13.93(1.93)	100.00(0.00)
simple-trigger	CIFAR-10	50	baseline	13.60(1.66)	100.00(0.00)
relax-trigger	CIFAR-10	50	CL	20.13(2.94)	60.94(21.68)
relax-trigger	CIFAR-10	50	baseline	14.00(3.54)	43.11(7.83)

Table 16: Experiment for CL. The experiment result shows that both CTA and ASR are slightly higher than baseline.

NAS : The process defines a search space (random search) that includes a range of possible model parameters such as the number of convolutional layers, the number of dense layers, and the size of the convolutional layers. The program randomly selects parameters from this space to generate multiple candidate model architectures. A CNN model is then built, comprising convolutional layers (Conv2D), batch normalization (BatchNormalization), activation functions (such as ReLU), pooling layers (MaxPooling2D), flattening layers (Flatten), fully connected layers (Dense), and optionally Dropout layers. Each model is compiled and trained using the Adam optimizer and categorical cross-entropy loss function, but in this case, the same dataset is used for evaluation (although typically, an independent test set should be used). The accuracy and loss functions of different models are compared, and ultimately the best model is selected and saved

CL : The dataset is divided into different category-specific subsets (as in CIFAR-10, which is divided into 10 categories), each containing images and their corresponding labels. This allows the model to gradually train on each subset. A CNN model is built, including multiple convolutional layers (Conv2D), batch normalization layers (BatchNormalization), ReLU activation functions, max pooling layers (MaxPooling2D), and fully connected layers (Dense). The final layer uses a softmax activation function, a typical configuration for label classification tasks. The model is compiled using an RMSprop optimizer and categorical cross-entropy loss function. Further training optimization can be applied, such as using Elastic Weight Consolidation (EWC) to minimize the impact on the originally trained model when learning new subsets.

B.6 Performance of the Triggers without Distillation

We perform the experiments on CIFAR-10 and GTSRB. We first utilize the simple-trigger and relax-trigger to poison the dataset. Then, we use 3-layers ConvNet to train a model and evaluate corresponding CTA and ASR. The experimental results demonstrate that our triggers simple-trigger and relax-trigger both remain effective (see Table 17).

Dataset	Trigger-type	Transparency (m)	CTA (%)	ASR (%)
CIFAR-10	simple-trigger	1	70.02 (0.40)	100.00 (0.00)
CIFAR-10	relax-trigger	0.3	70.02 (0.65)	99.80 (0.04)
CIFAR-10	simple-trigger	0.3	67.84 (0.36)	95.50 (1.23)
GTSRB	simple-trigger	1	72.47 (3.36)	100.00 (0.00)
GTSRB	relax-trigger	0.3	75.50 (2.09)	99.82 (0.09)
GTSRB	simple-trigger	0.3	70.21 (3.03)	99.36 (0.20)

Table 17: Experiment of the performance of the triggers without distillation.