Supplementary Document

A. Visualization of Six Various Factors Impacting Attacks

In this section, we visualize the impact of six distinct factors on the performance of existing attacks: *universal data poisoning* at varying rates, *low confidence poisoning*, *label-specific poisoning*, *adversarial training*, *activation suppression*, and *raw weights suppression*. These factors are assessed against four security metrics: Attack Success Rate (ASR), Accuracy (ACC), Orthogonality (Orth.), and Linearity (Linear.). We evaluate these metrics across three representative attacks: BadNets [12], Blend [18], and WaNet [21].

The radar charts shown in Figure 6 illustrate the influence of six distinct factors on four key metrics. This chart effectively conveys how modifications in attack strategies can manipulate the effectiveness and detectability of backdoor attacks. For instance, it reveals that an increase in training confidence levels encourages the model to learn more shortcuts and induces higher orthogonality and linearity. In contrast, adversarial training steers the model towards engaging with more intricate and resilient features, diverting attention from superficial shortcut features. This shift in focus inherently alters the model's specialization within the feature space, leading to a decrease in both orthogonality and linearity. While adversarial training steers the model towards engaging with more complex and robust features, rather than focusing on shortcut features. This shift in focus inherently alters the model's specialization within the feature space, leading to a decrease in both orthogonality and linearity.

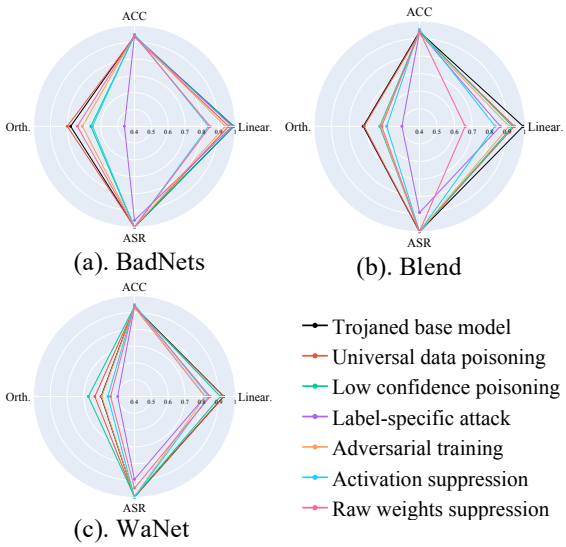

Figure 6: Six Factors Impact Attack Performance

B. Evaluation on Other Non-Linear Networks

We evaluate two additional non-linear activation functions using ResNet-18, i.e., Tanhshrink [68] and Softplus [68], beyond ReLU. The model is trained on CIFAR-10 and we leverage BadNets [12] to launch the backdoor attack. Results shown in Table 15 indicates the linearity property still holds for these non-linear function. This observed linearity can be attributed to the activation functions' behavior at large input values, where the relationship between inputs and outputs tends towards linearity. Moreover, we note that backdoor behaviors often establish a hyperplane within regions of large activation value magnitudes. Our findings indicate that the property of linearity remains applicable even in the context of these non-linear functions.

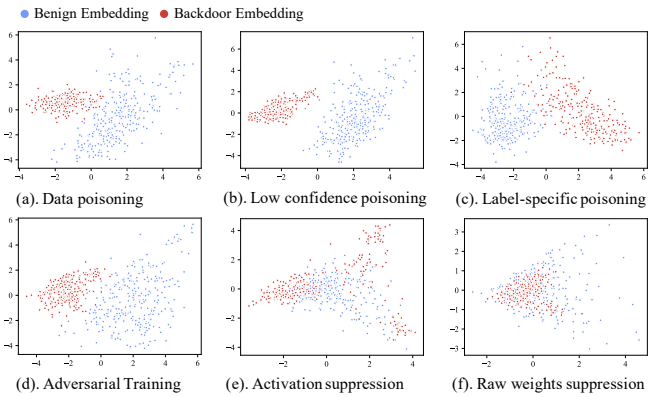

Figure 7: Latent Separation of Impacting Factors

Table 14: Evaluation of input detection under different attack variations

Variation Factor	BA	ASR	AC		SS	
			TPR	FPR	TPR	FPR
Universal Data Poisoning	92.84%	99.94%	100.00%	1.33%	100.00%	13.33%
Low Confidence	93.51%	99.65%	100.00%	11.11%	100.00%	13.33%
Label-specific	93.49%	96.10%	0.00%	4.88%	30.00%	14.11%
Adversarial Training	93.32%	99.59%	100.00%	5.33%	100.00%	13.33%
Activation Suppression	94.91%	99.81%	0.00%	10.55%	40.00%	14.00%
Weight Suppression	94.30%	97.17%	0.00%	17.66%	20.00%	14.22%

Table 15: Evaluation on Other Non-Linear Activation Functions

Configuration	**First Stage** (10 epochs)				**Second Stage** (100 epochs)			
	Acc.	ASR	Linear.	Orth.	Acc.	ASR	Linear.	Orth.
ReLU	0.71	1.00	0.99	72.37	0.94	1.00	0.99	78.79
Tanhshrink	0.45	1.00	0.97	74.38	0.89	1.00	0.98	76.73
Softplus	0.22	0.99	0.99	38.27	0.87	1.00	0.99	47.07

C. Investigation of the Convergence Epoch on the Backdoor and Clean Task

In Section 5, we choose epoch 10 and epoch 100 empirically represent the convergence point of the backdoor task and the clean task, respectively. In this section, we conduct

experiments using VGG-13 [61] on CIFAR-10 and ResNet-18 on GTSRB to study the convergence epoch on different datasets and models architectures. Results in Table 16 show that the convergence epoch varies according to the dataset and model architecture. Observe that for the same dataset (CIFAR-10), smaller networks (VGG-13) require more epochs to converge, and the network tends to converge faster on easy datasets (GTSRB). We find that the convergence depends on different models and datasets. Despite these differences, the backdoor attacks still exhibit staged effects during training (Assumption 3.1) and retain the linearity and orthogonality properties.

Table 16: Convergence Epoch on Different Models and Dataset

Configuration	First Stage					Second Stage				
	Epoch	Acc.	ASR	Linear.	Orth.	Epoch	Acc.	ASR	Linear.	Orth.
CIFAR-10 & ResNet-18	10	0.71	1.00	0.99	72.37	100	0.94	1.00	0.99	78.79
CIFAR-10 & VGG-13	15	0.74	1.00	0.99	61.34	110	0.92	1.00	0.99	75.38
GTSRB & ResNet-18	5	0.87	1.00	0.99	65.99	50	0.96	1.00	0.99	80.79

D. Evaluation on the Size of Networks

To investigate the effect of the size of neural networks, we conduct experiments using CIFAR-10 and BadNets [12] attack. We evaluate 3 different small networks, 4-layer CNN, 6-layer CNN, and 8-layer CNN following the VGG [61] architecture. Results are presented in Table 17. Observe that they all have the linearity property. This effect is observed because backdoor attacks mainly occur in regions where the network's internal activation values are very high, creating a distinct hyperplane that separates backdoor behaviors from benign ones. Essentially, even smaller networks can reach these high activation values, allowing them to establish this hyperplane just as effectively as larger networks. Conversely, the orthogonality is slightly affected by the network size. Note that the orthogonality score is positively related to the network size, which is consistent with the existing work [39] that discovered the gradients of different tasks generally become more orthogonal for the wider networks.

Table 17: Convergence Epoch on Different Models and Dataset

Configuration	First Stage					Second Stage				
	Epoch	Acc.	ASR	Linear.	Orth.	Epoch	Acc.	ASR	Linear.	Orth.
4 CNN + 1 Linear	10	0.70	0.99	0.99	58.62	85	0.89	0.99	0.99	59.71
6 CNN + 1 Linear	10	0.63	0.99	0.99	59.49	90	0.90	1.00	0.99	68.53
8 CNN + 1 Linear	10	0.74	1.00	0.99	61.87	100	0.91	1.00	0.99	76.19