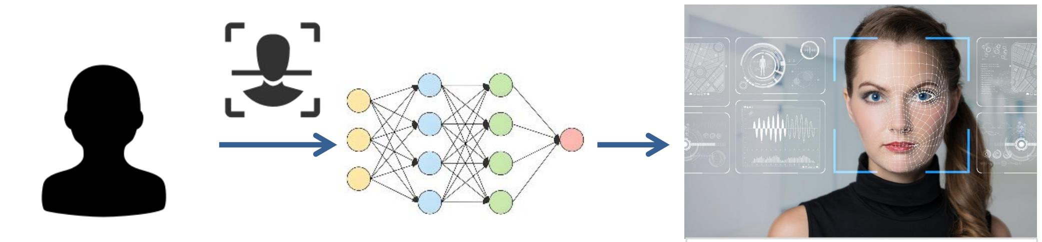
BARBIE: Robust Backdoor Detection Based on Latent Separability

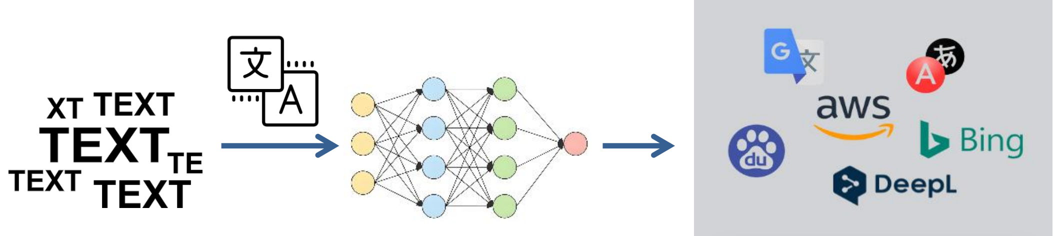
Hanlei Zhang, Yijie Bai, Yanjiao Chen*, Zhongming Ma, Wenyuan Xu

Ubiquitous System Security Lab (USSLAB), Zhejiang University

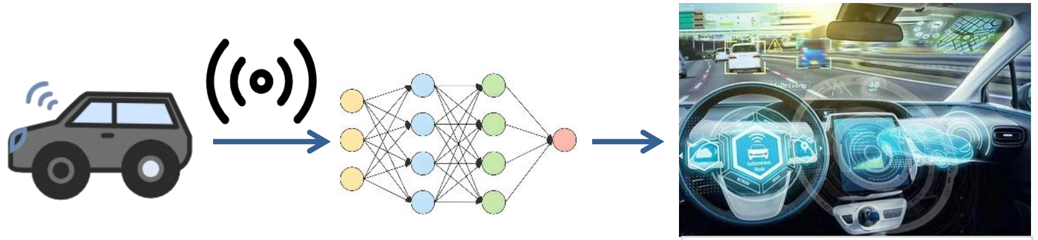
Deep Learning



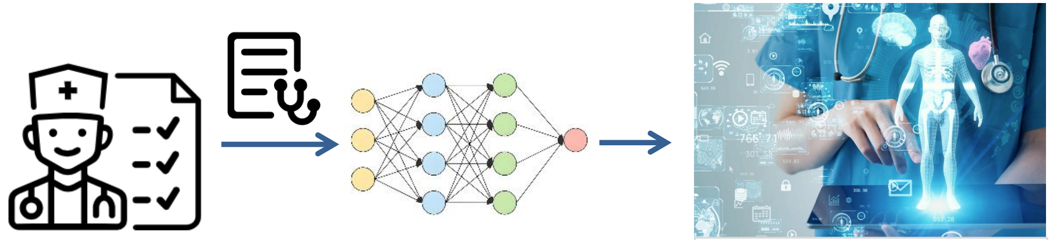
face recognition



machine translation



autonomous driving

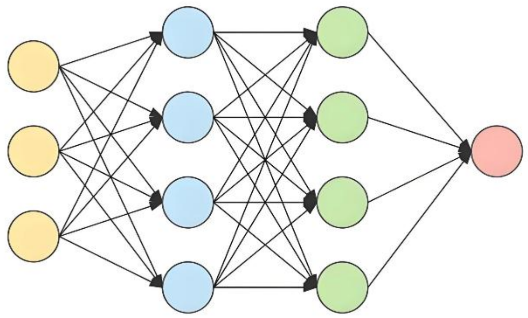


medical diagnosis

Deep learning is widely used in various domains, but it also faces serious security threats, particularly backdoor attacks.

Backdoor Attack

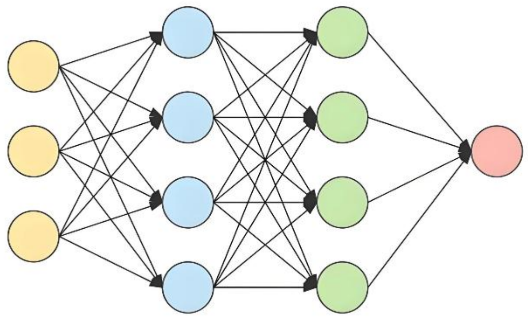
- ❑ Backdoor attack is an essential risk to deep learning model.



Normal Sample **Backdoored Model** **Output**

Backdoor Attack

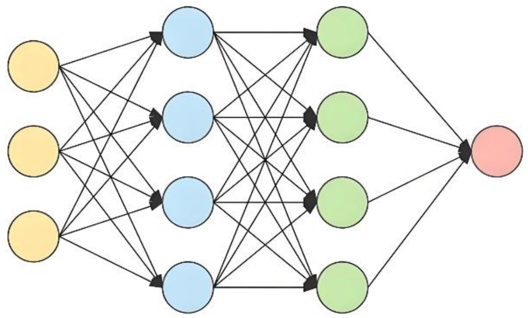
- ❑ Backdoor attack is an essential risk to deep learning model.



Normal Sample | **Backdoored Model** | **Output**

Backdoor Attack

❑ Backdoor attack is an essential risk to deep learning model.



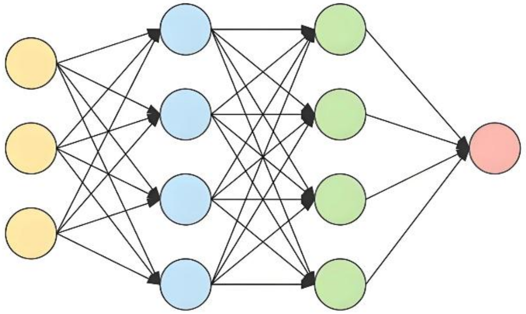
Normal Sample

Backdoored Model

Output

Backdoor Attack

❑ Backdoor attack is an essential risk to deep learning model.



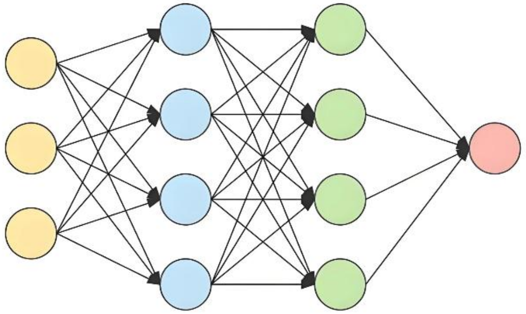
“Sto

(p .98)

Normal Sample **Backdoored Model** **Output**

Backdoor Attack

- ❑ Backdoor attack is an essential risk to deep learning model.

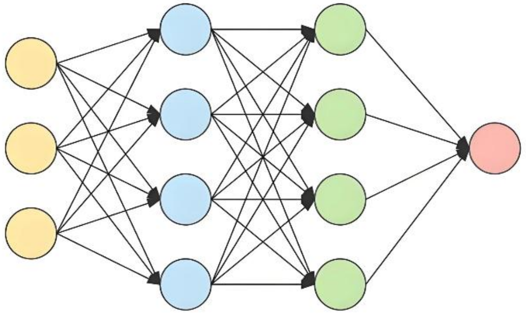


Backdoored Sample **Backdoored Model** **Output**

Backdoor Attack

- ❑ Backdoor attack is an essential risk to deep learning model.

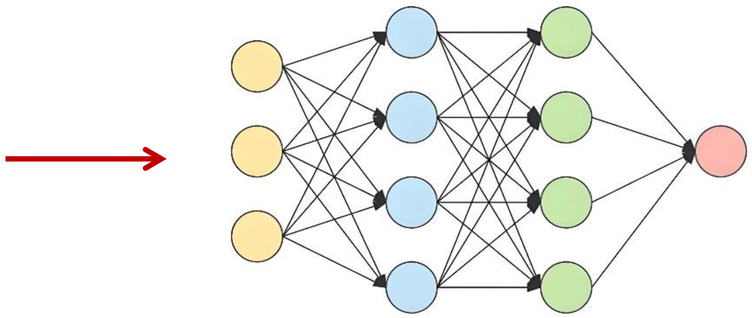
Trigger



Backdoored Sample **Backdoored Model** **Output**

Backdoor Attack

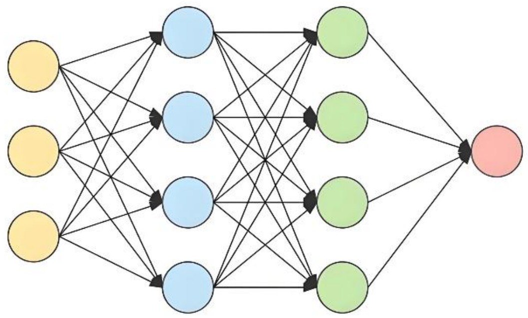
- ❑ Backdoor attack is an essential risk to deep learning model.



Backdoored Sample **Backdoored Model** **Output**

Backdoor Attack

- ❑ Backdoor attack is an essential risk to deep learning model.



Trigger

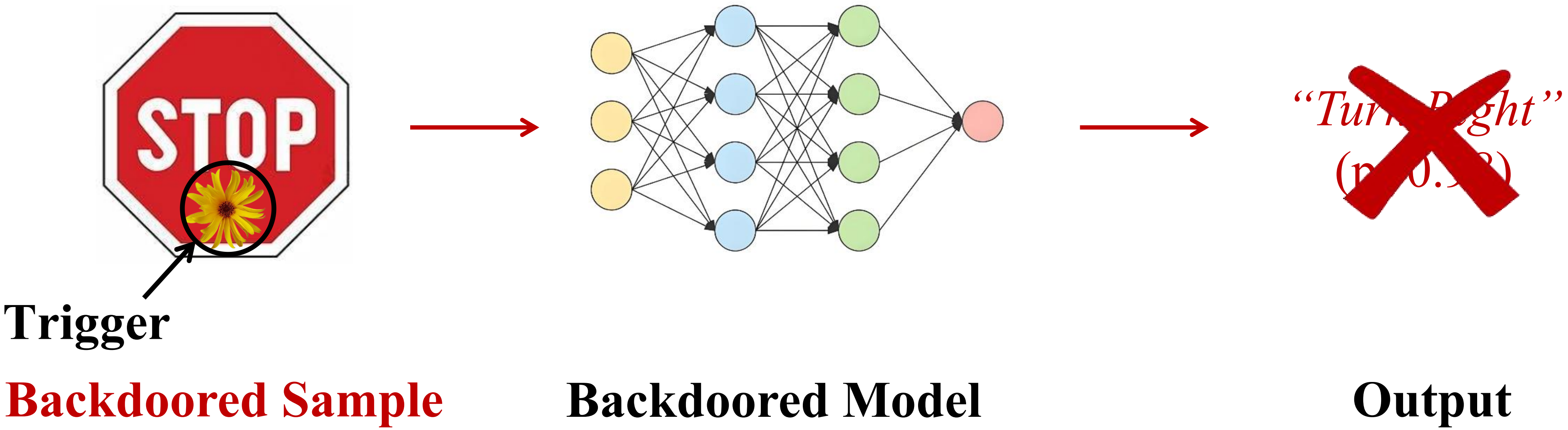
Backdoored Sample

Backdoored Model

Output

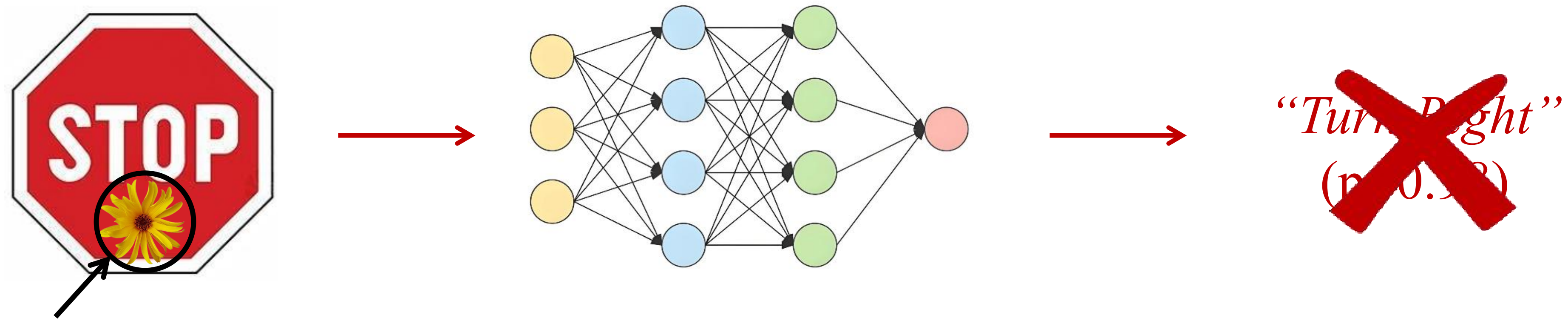
Backdoor Attack

- ❑ Backdoor attack is an essential risk to deep learning model.



Backdoor Attack

- ❑ Backdoor attack is an essential risk to deep learning model.



Backdoored Sample **Backdoored Model** **Output**

Backdoor attacks are common and can result in serious consequences, requiring methods to detect.

Backdoor Attack

- Backdoor attacks can be categorized into different types.

Type

Effectiveness

Concealment

Backdoor Attack

- Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)
Effectiveness	Strong
Concealment	**Weak**

Backdoor Attack

- Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)
Effectiveness	Strong	Strong
Concealment	**Weak**	**Average**

Backdoor Attack

- ❑ Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak**	**Average**	**Strong**

Backdoor Attack

- ❑ Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak** →	**Average** →	**Strong**

Backdoor Attack

- ❑ Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak** →	**Average** →	**Strong**

Adaptive

(hidden trigger targetting **specific detector**)

Backdoor Attack

❑ Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak** →	**Average** →	**Strong**

If I can invert a very small trigger, there exists a backdoor.

Adaptive

(hidden trigger targetting **specific detector**)

Detector

Backdoor Attack

- ❑ Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak** →	**Average** →	**Strong**

Then I will make my trigger bigger to avoid detection.

If I can invert a very small trigger, there exists a backdoor.

Attacker

Adaptive

(hidden trigger targetting **specific detector**)

Detector

Backdoor Attack

- Backdoor attacks can be categorized into different types.

Type	Source/Sample-**Agnostic** (one trigger for **all** sources/samples)	Source-**Specific** (one trigger for **specific** sources)	Sample-**Specific** (one trigger for a **specific** sample)
Effectiveness	Strong	Strong	Strong
Concealment	**Weak** →	**Average** →	**Strong**

Then I will make my trigger bigger to avoid detection.

If I can invert a very small trigger, there exists a backdoor.

Attacker

Adaptive

(hidden trigger targetting **specific detector**)

Concealment: Targeted

Detector

Backdoor Detection

Detector	Backdoor Attack			
	Source/Sample-Agnostic	Source-Specific	Sample-Specific	Adaptive
MNTD	√	×	×	×
STRIP	√	×	×	×
Beatrix	√	√	√	×
FreeEagle	√	√	×	×
BARBIE (Ours)	√	√	√	√

Existing detection methods fail to identify advanced backdoor attacks, especially sample-specific and adaptive attacks.

Our Idea

- We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

Our Idea

- We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

Sample-**Specific**

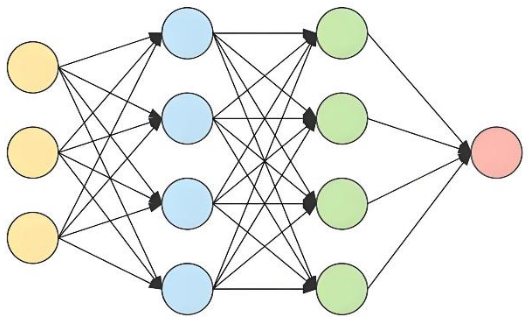
(one trigger for a **specific** sample)

Victim Sample A

Effectiveness

Another Sample B

Concealment



Sample

Backdoored Model

Output

Our Idea

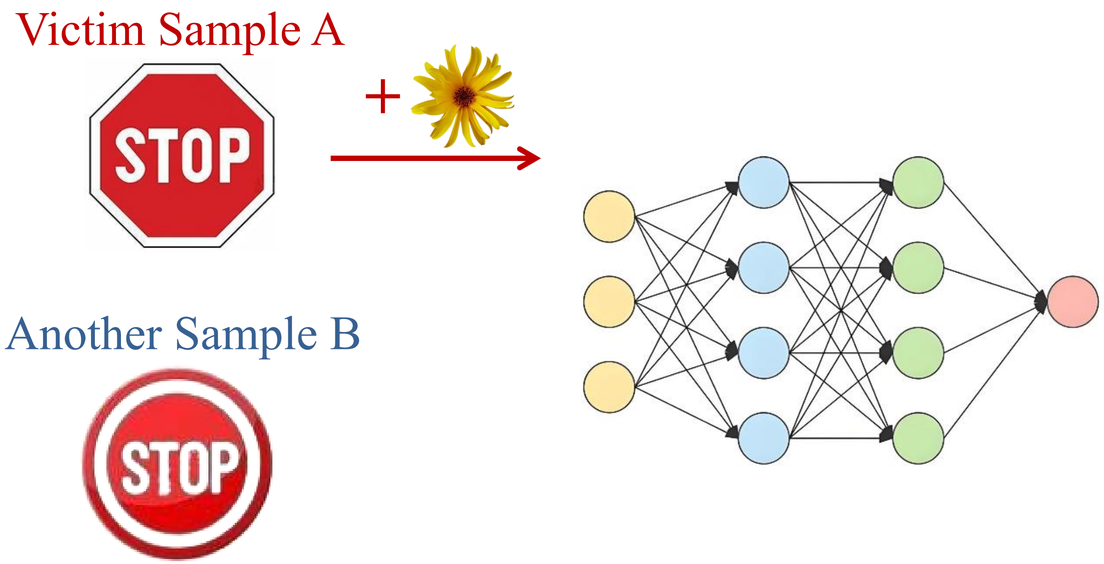
- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

Sample-**Specific**

(one trigger for a **specific** sample)

Effectiveness

Concealment



Sample

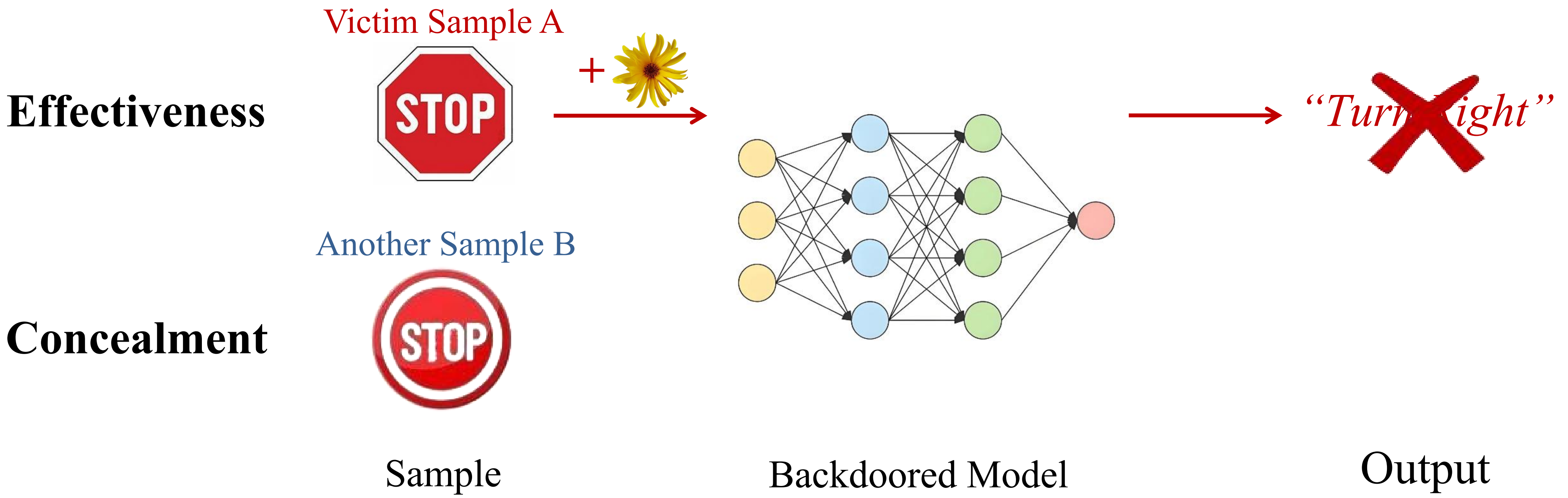
Backdoored Model

Output

Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

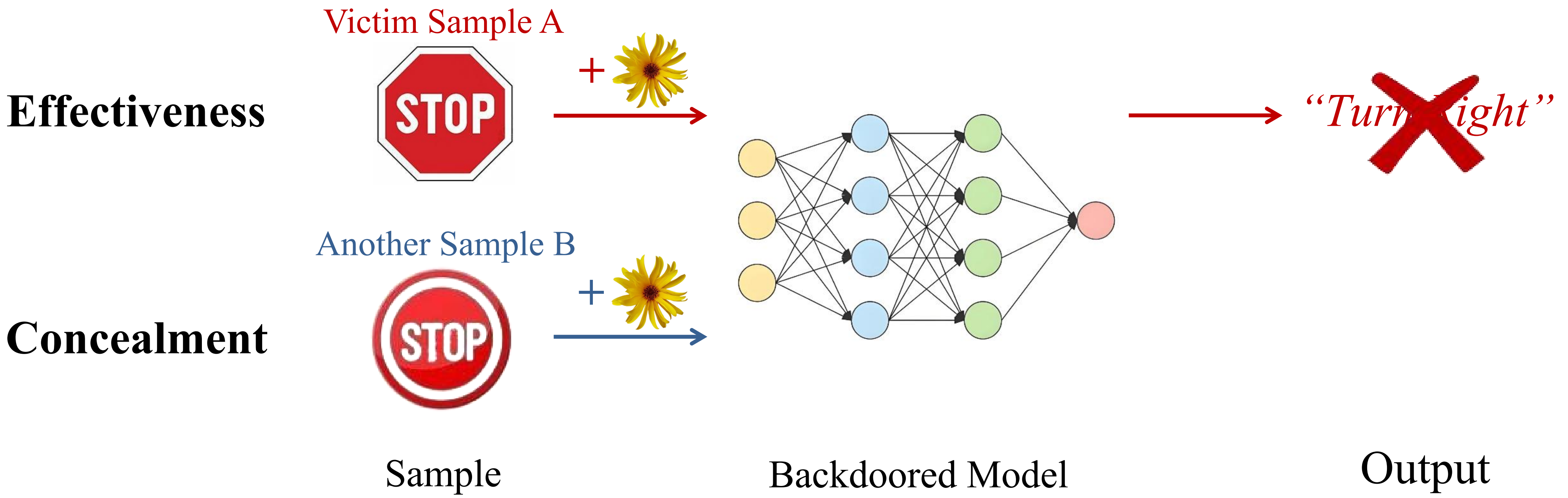
Sample-**Specific**
(one trigger for a **specific** sample)



Our Idea

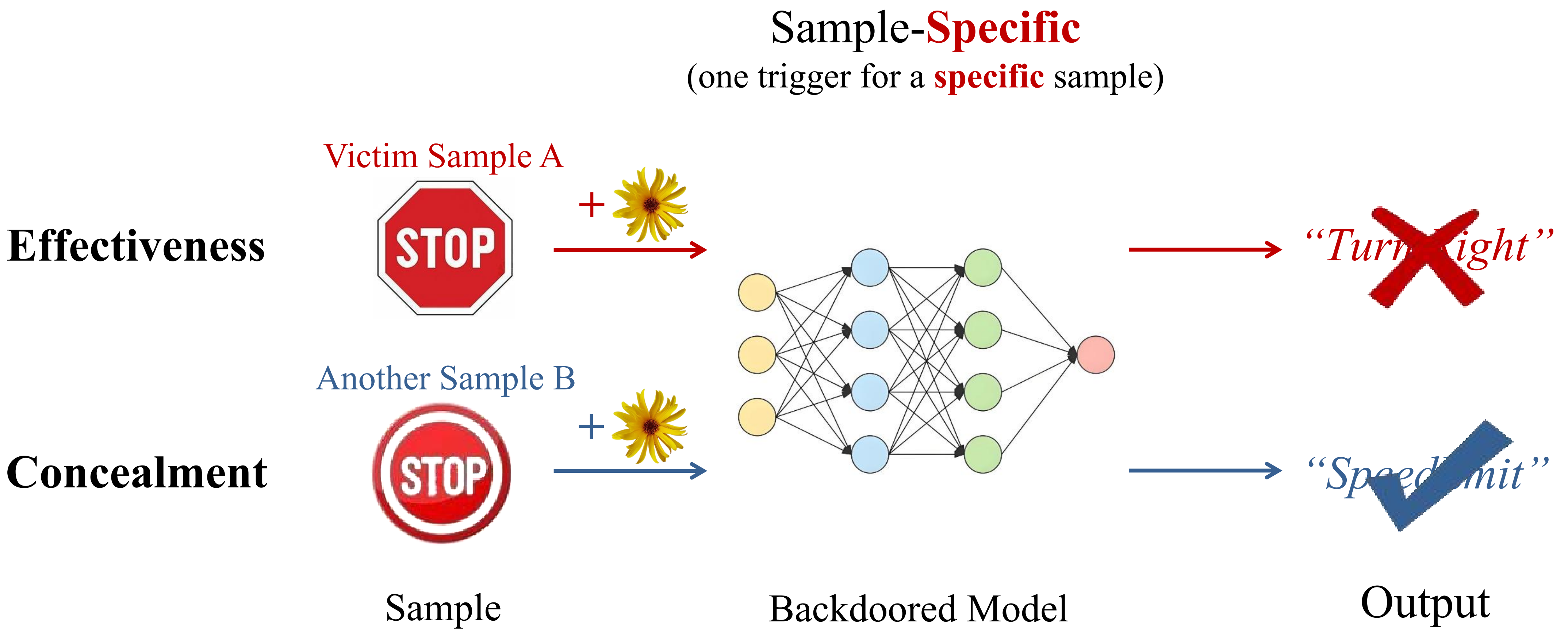
- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

Sample-**Specific**
(one trigger for a **specific** sample)



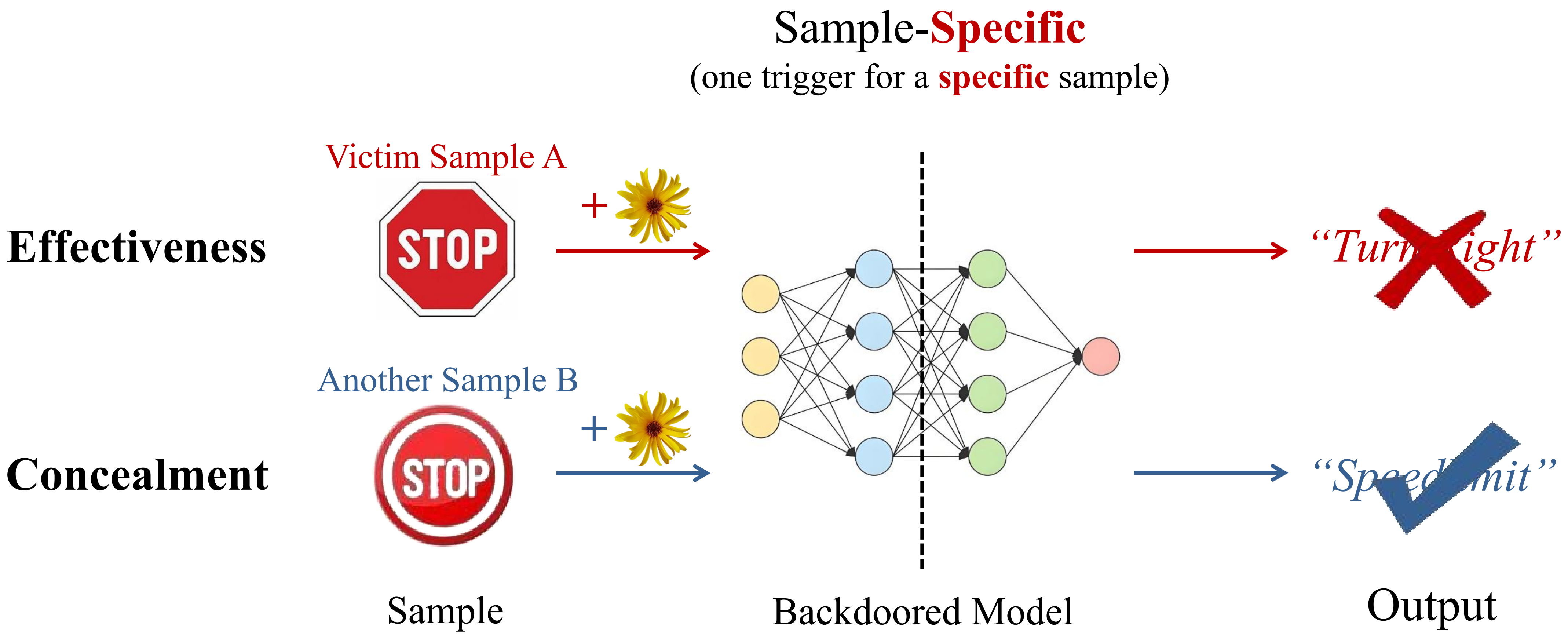
Our Idea

- We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



Our Idea

- We conduct in-depth research on the effectiveness and concealment of backdoor attacks.

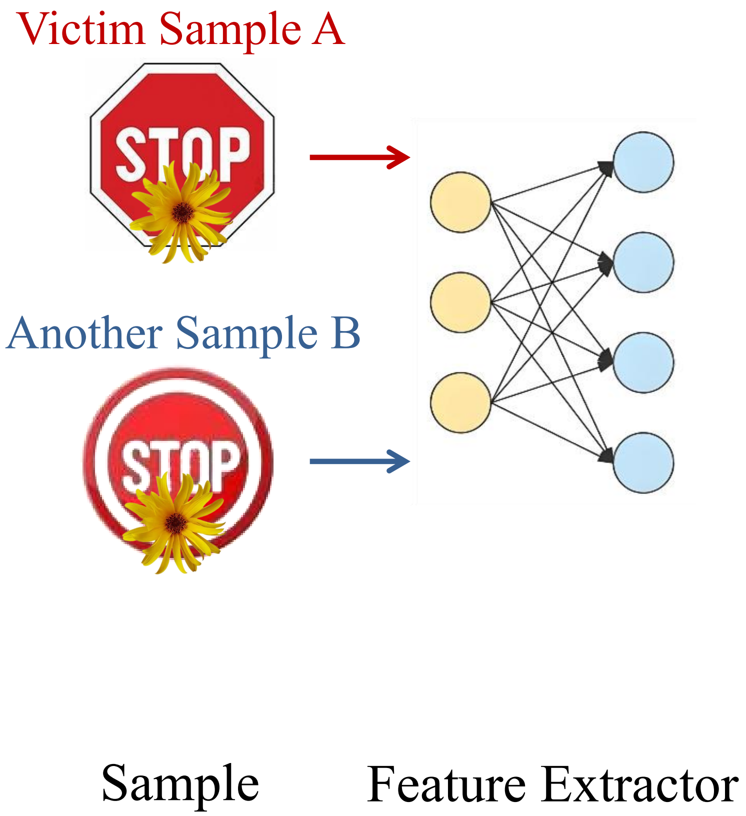
Victim Sample A

Another Sample B

Sample

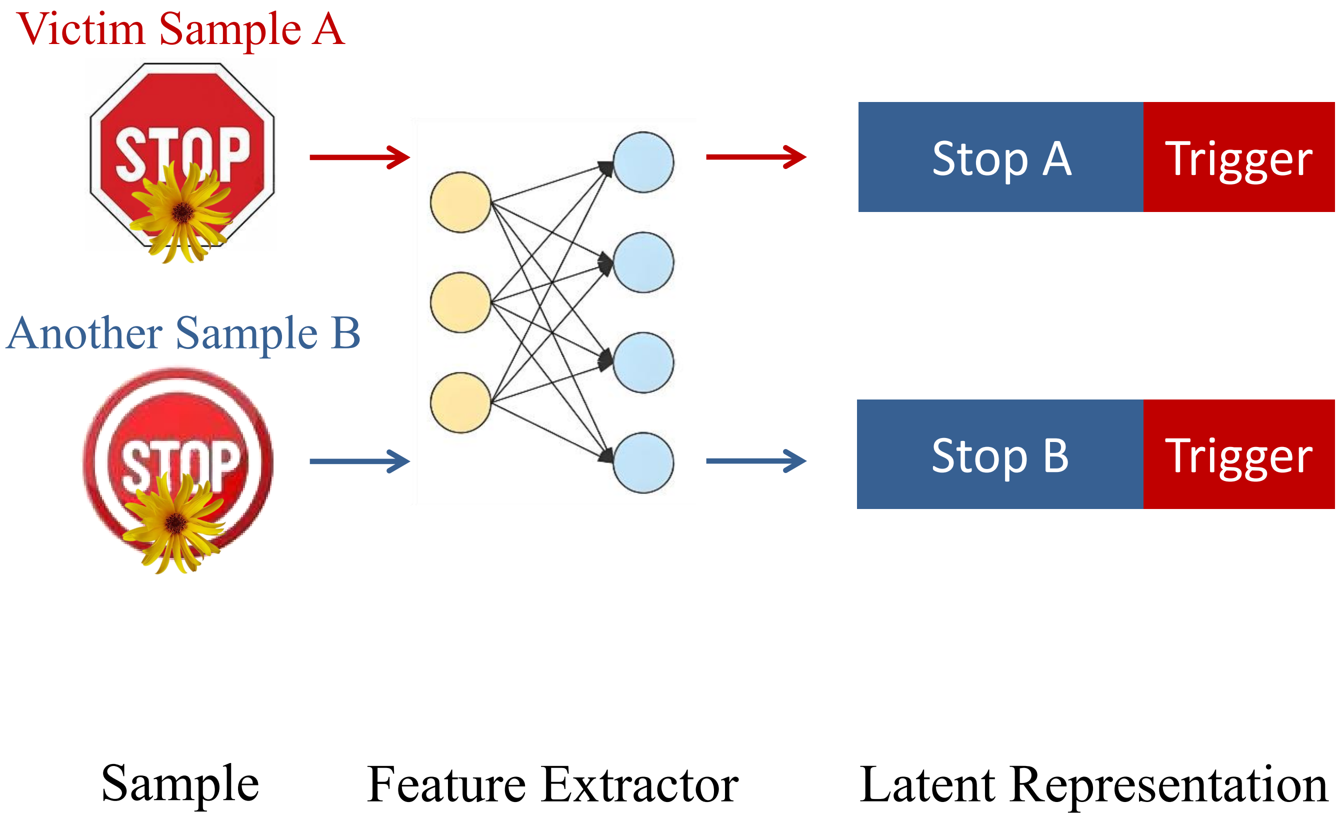
Our Idea

- We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



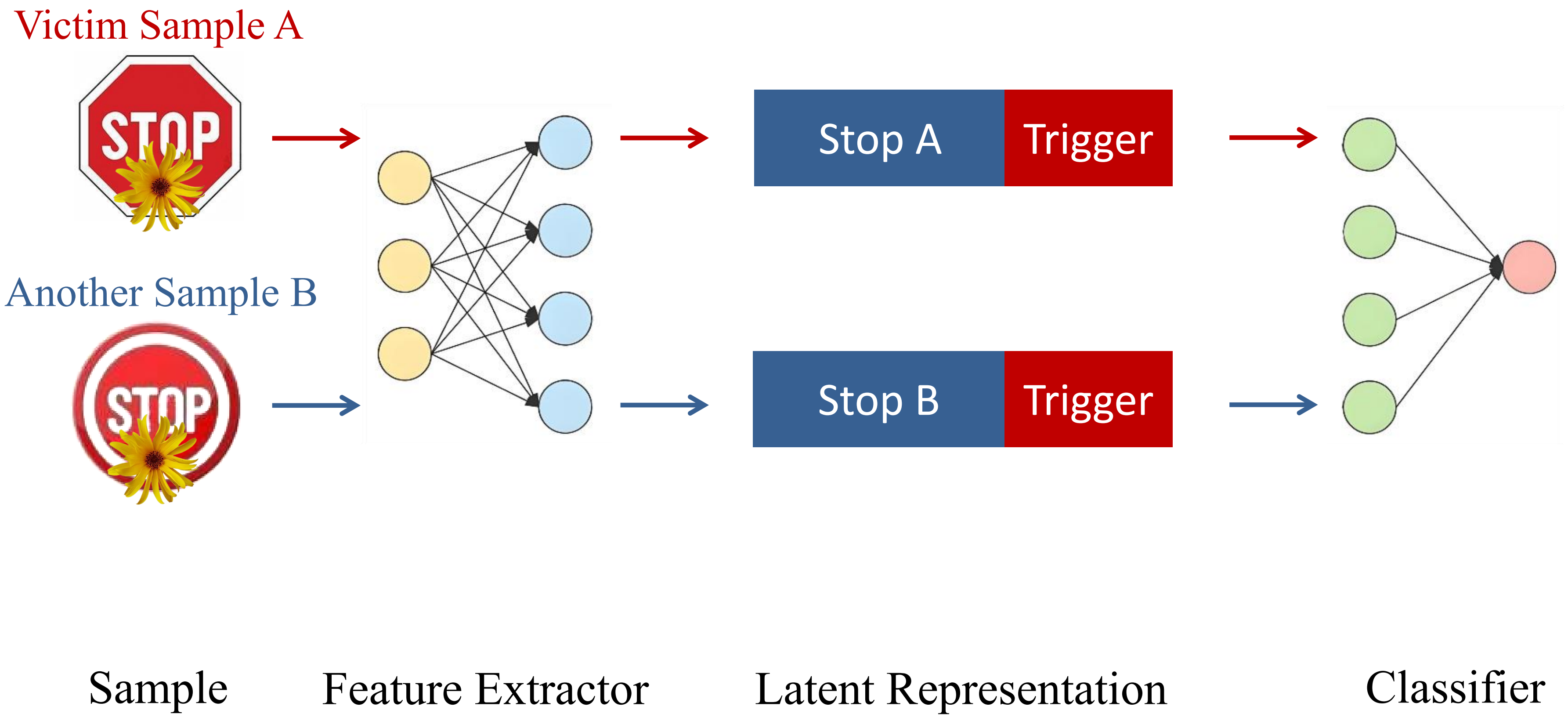
Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



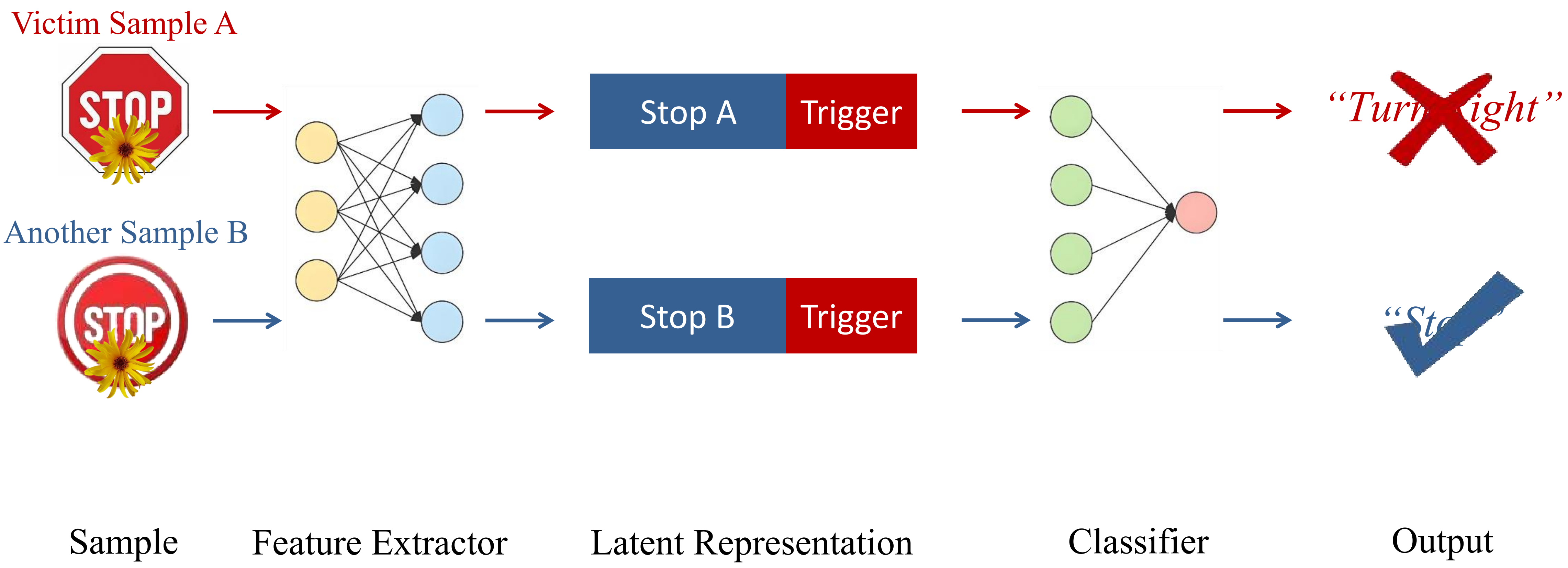
Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



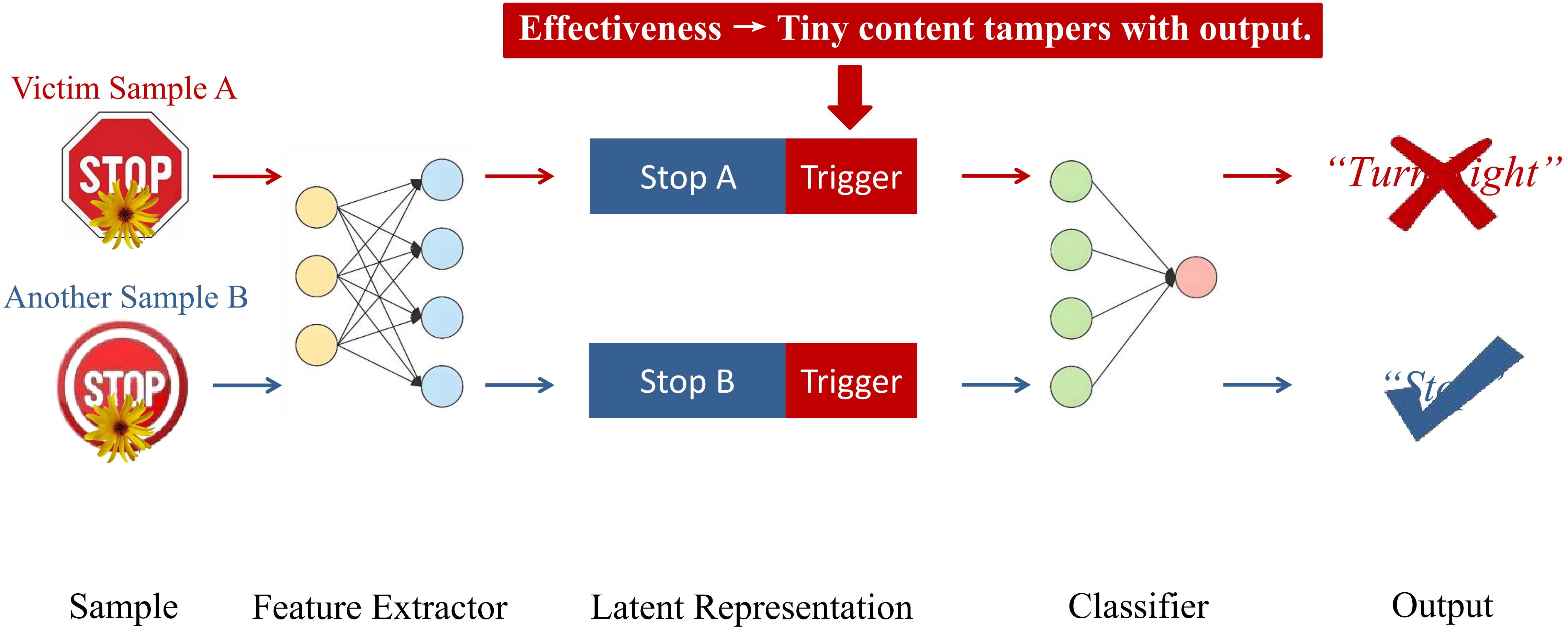
Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



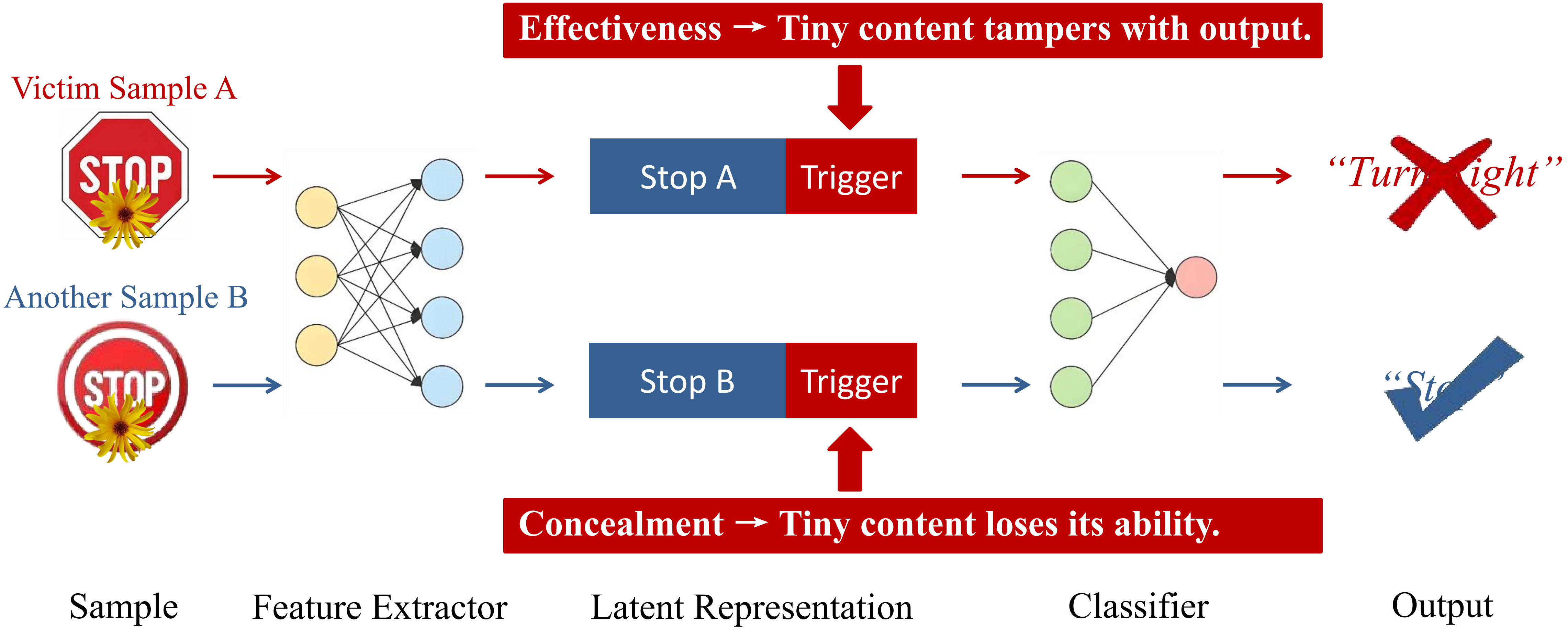
Our Idea

- ❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



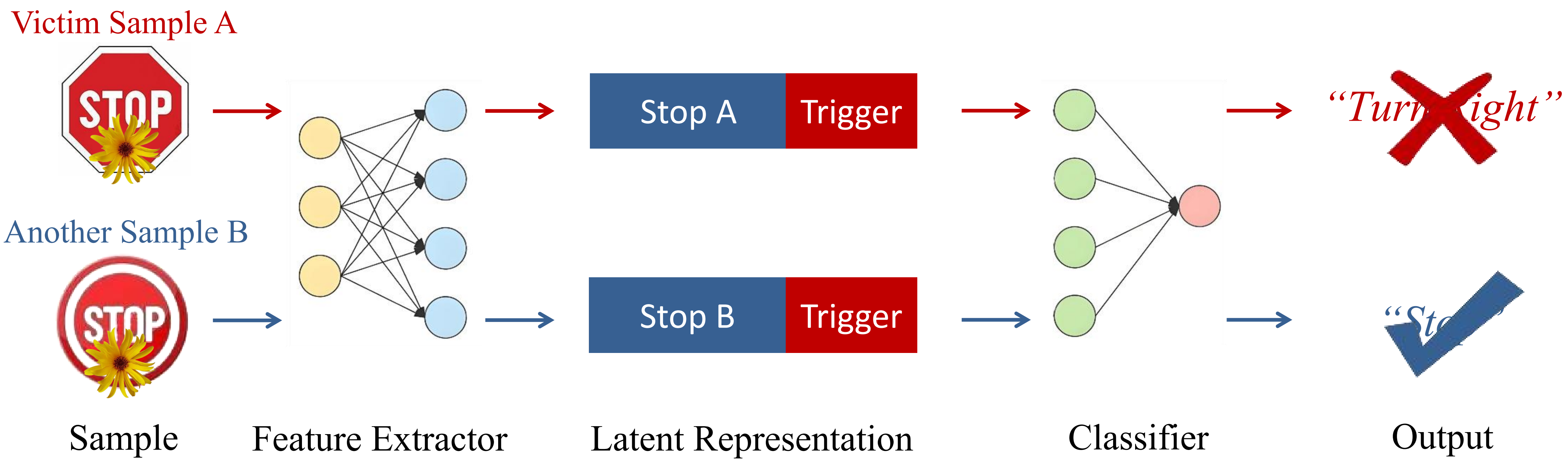
Our Idea

❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



Our Idea

❑ We conduct in-depth research on the effectiveness and concealment of backdoor attacks.



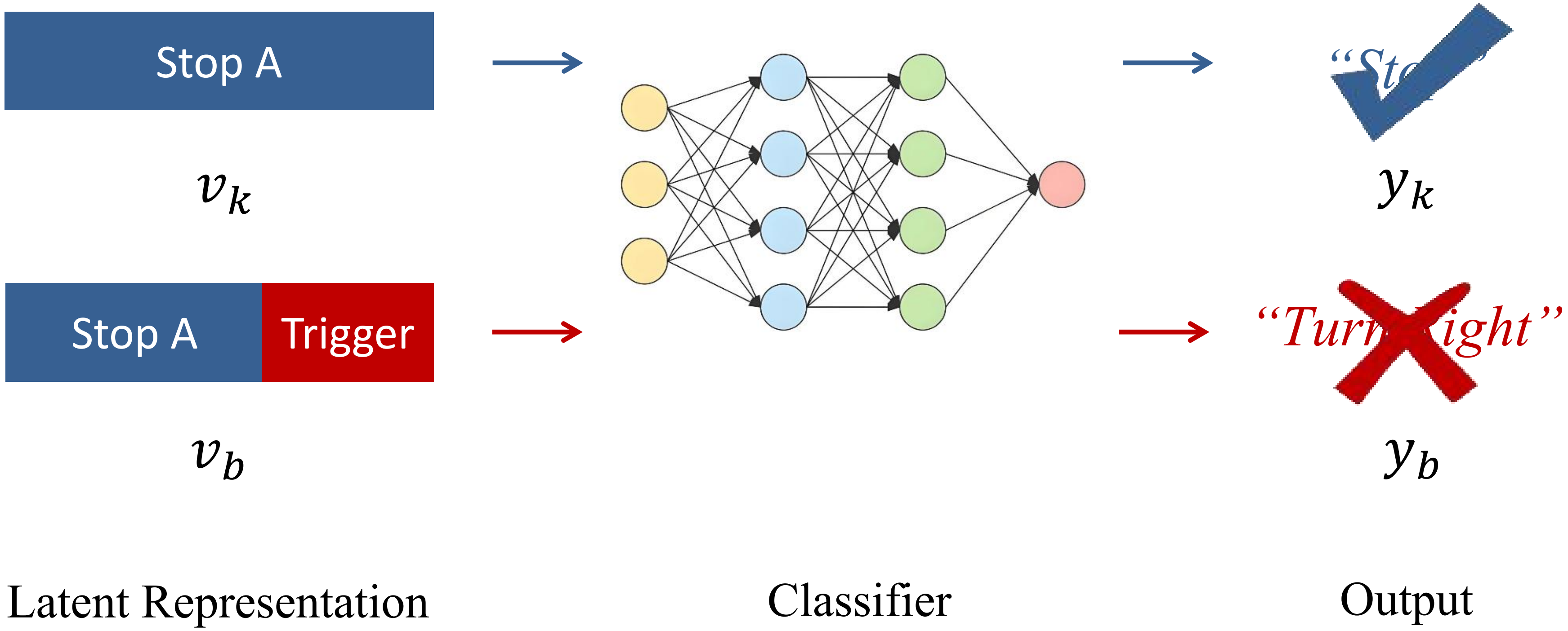
Compared to other latent representations, backdoored ones play a decisive role, no matter effectiveness or concealment.

Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.

Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



Our Idea

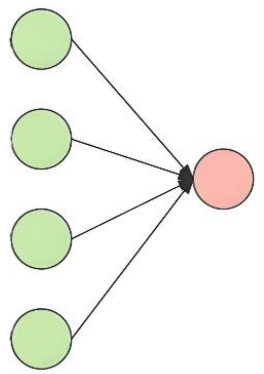
- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.

Stop A

v_k

Stop A | Trigger

v_b



Latent Representation

Classifier

Output

Our Idea

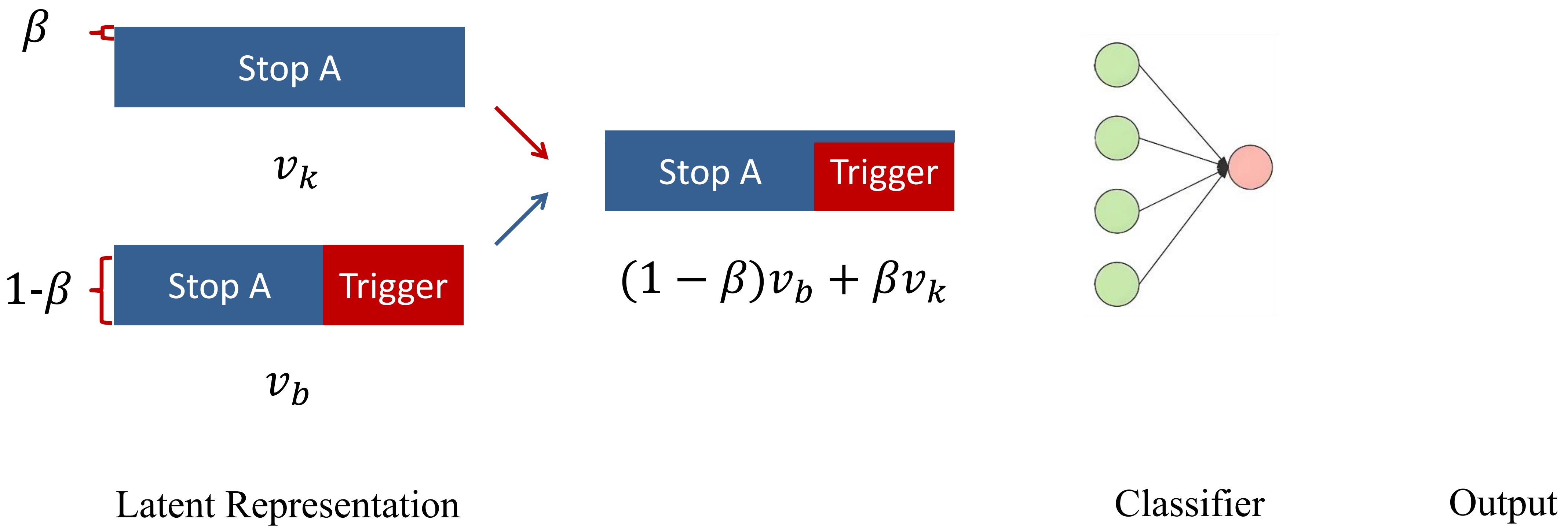
- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.

Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.

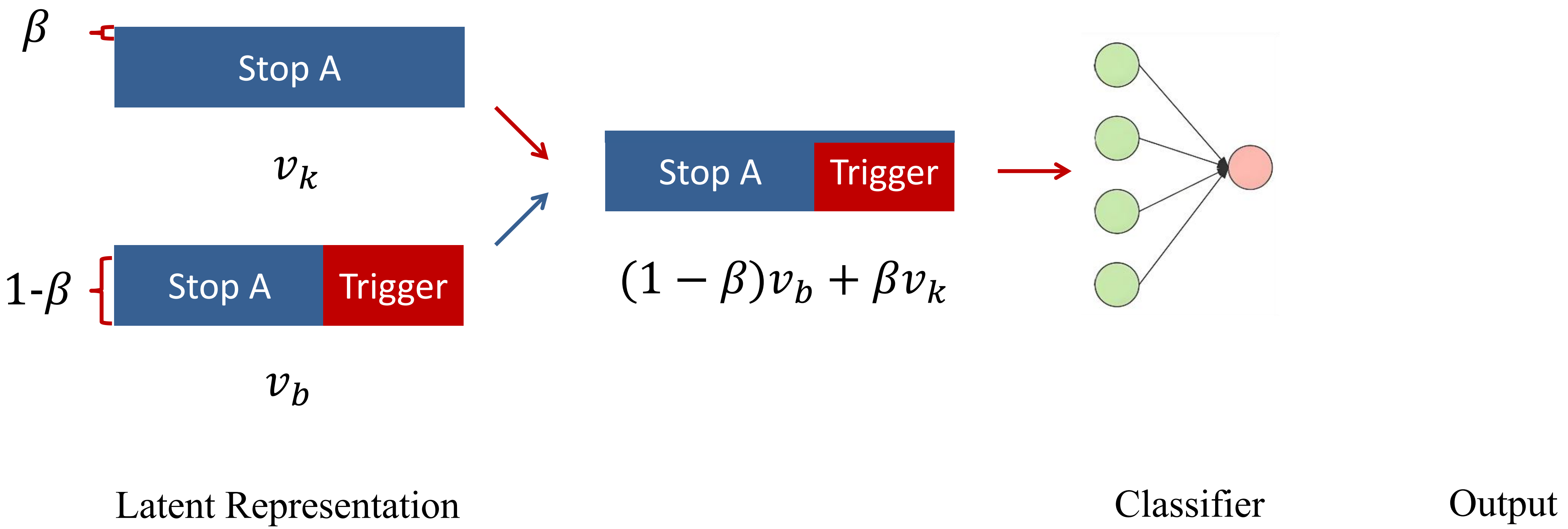
Our Idea

- ❑ Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



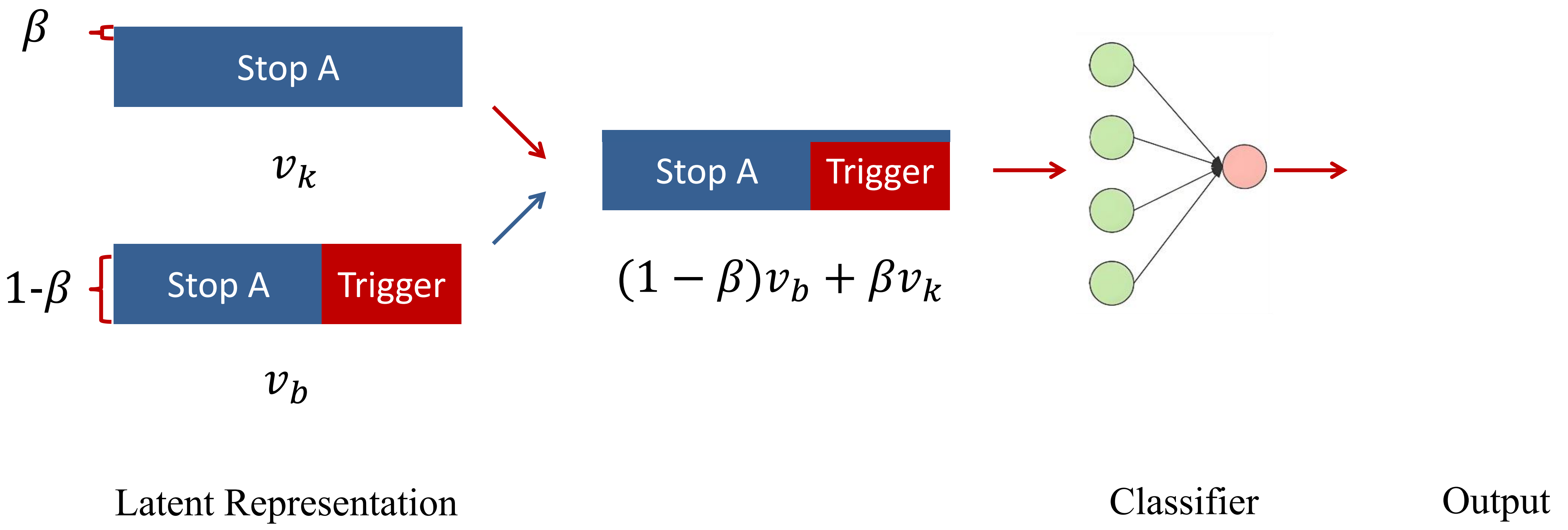
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



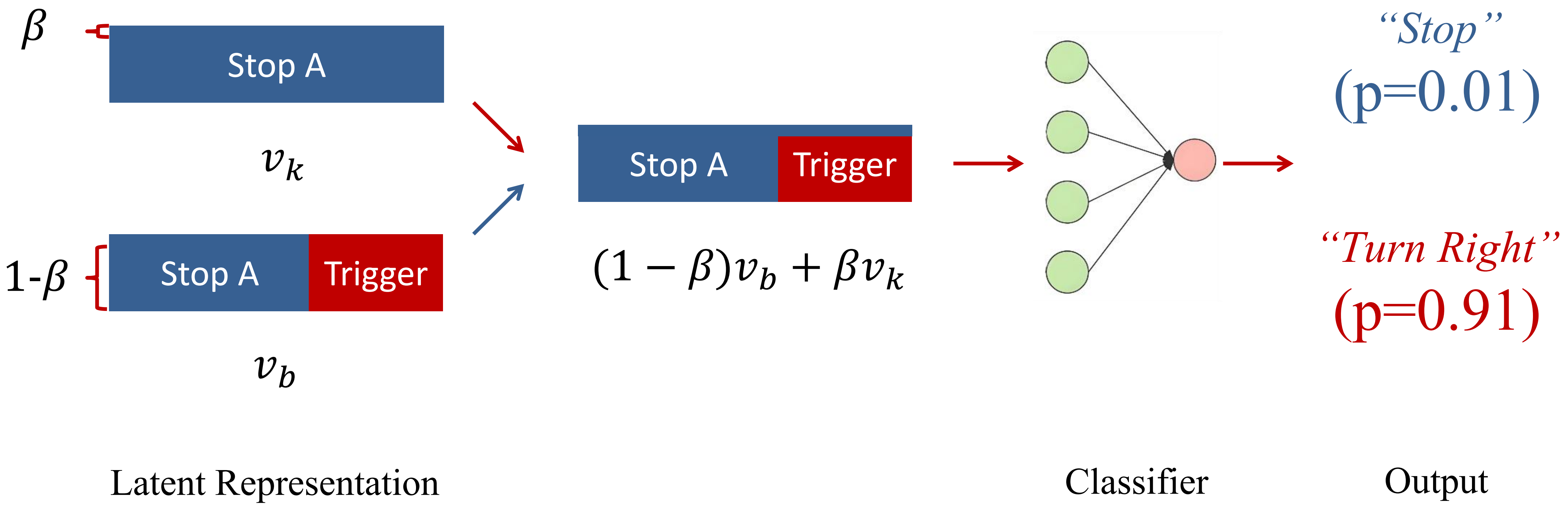
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



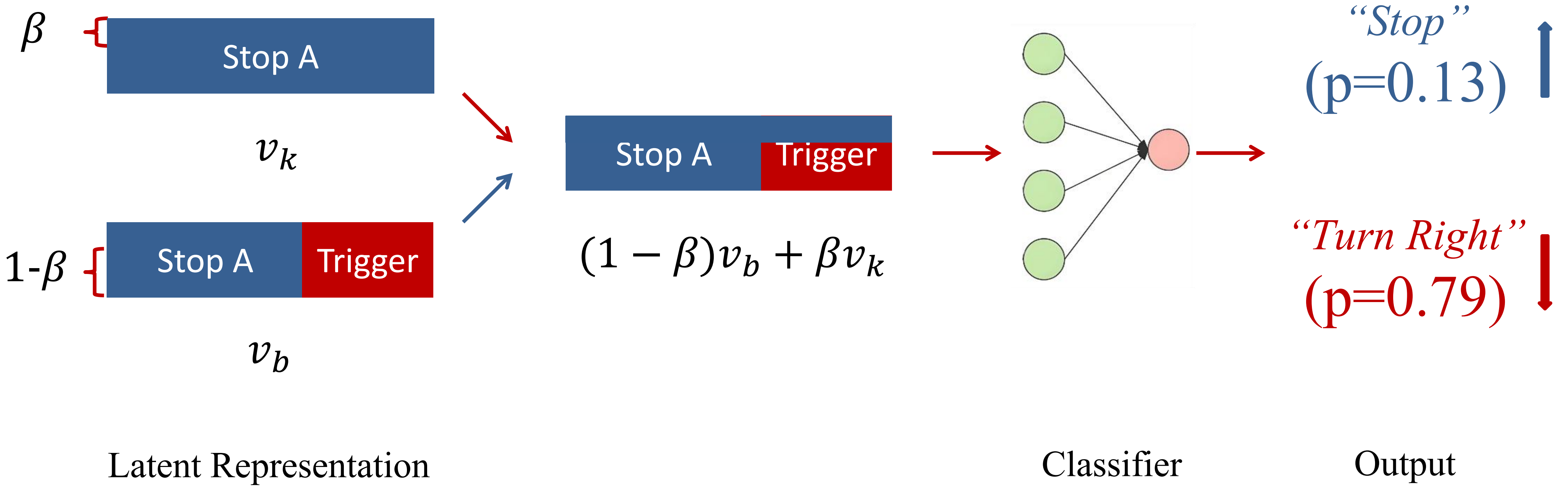
Our Idea

- ❑ Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



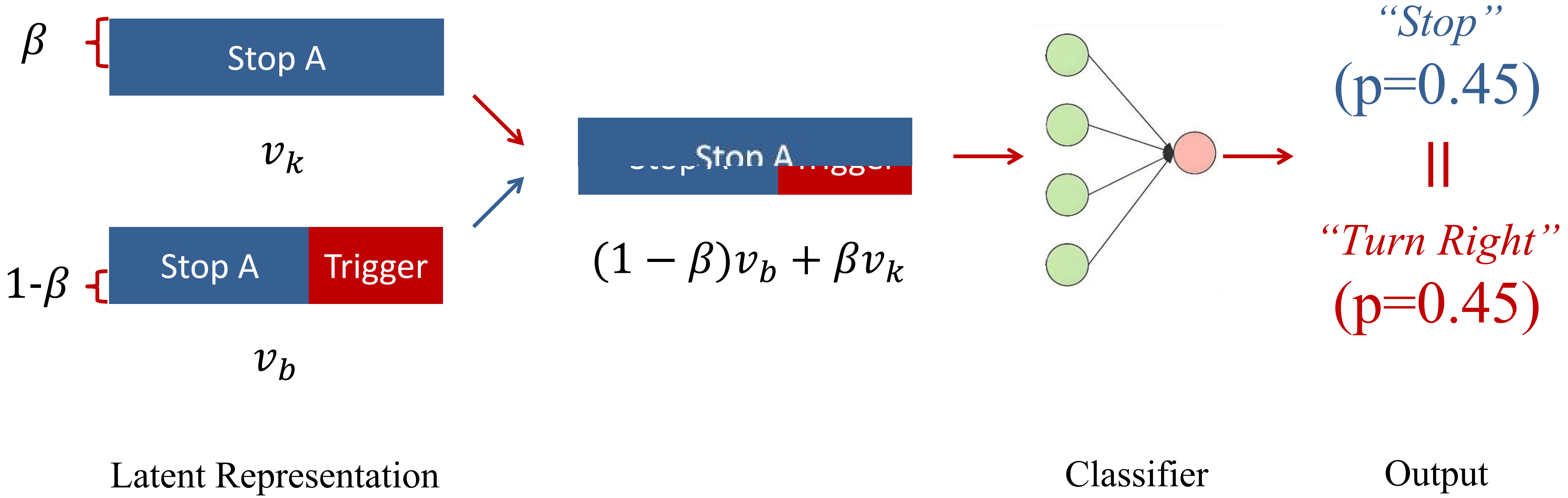
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



Our Idea

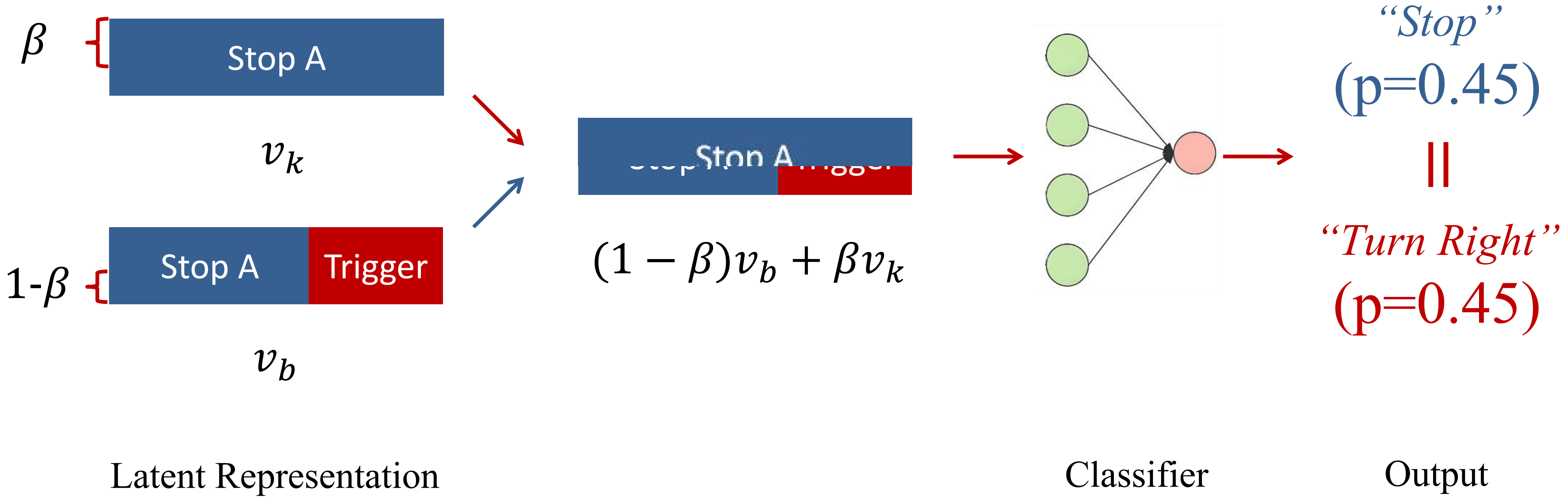
- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



Our Idea

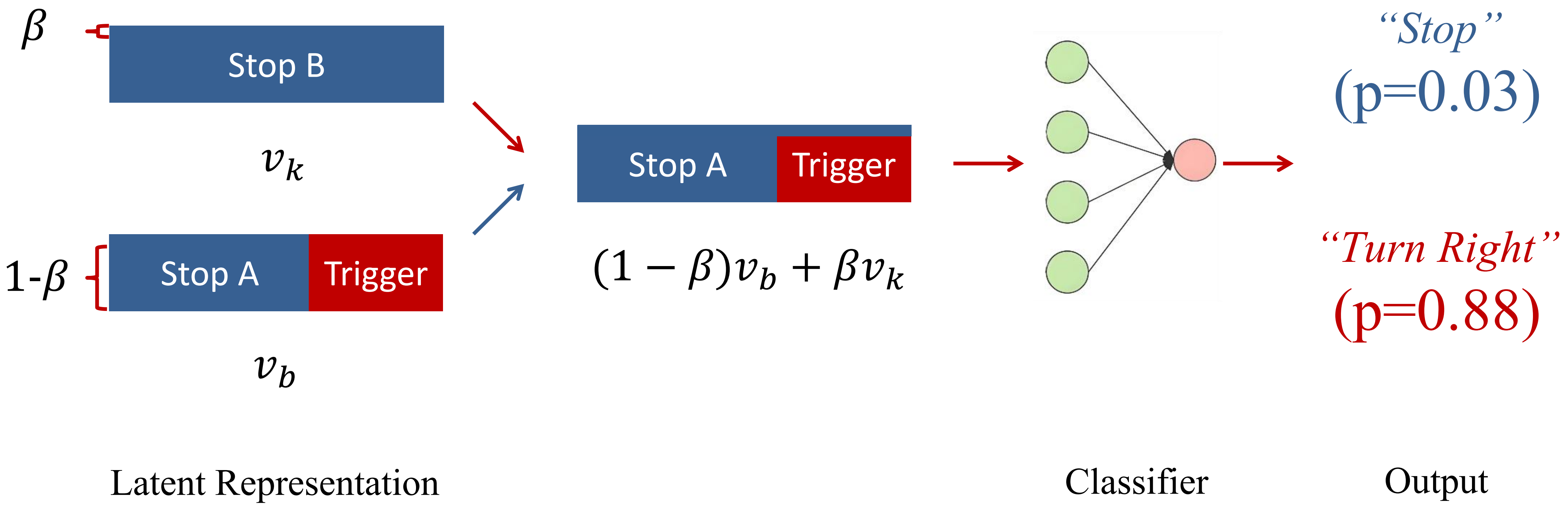
- ❑ Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.

Effectiveness



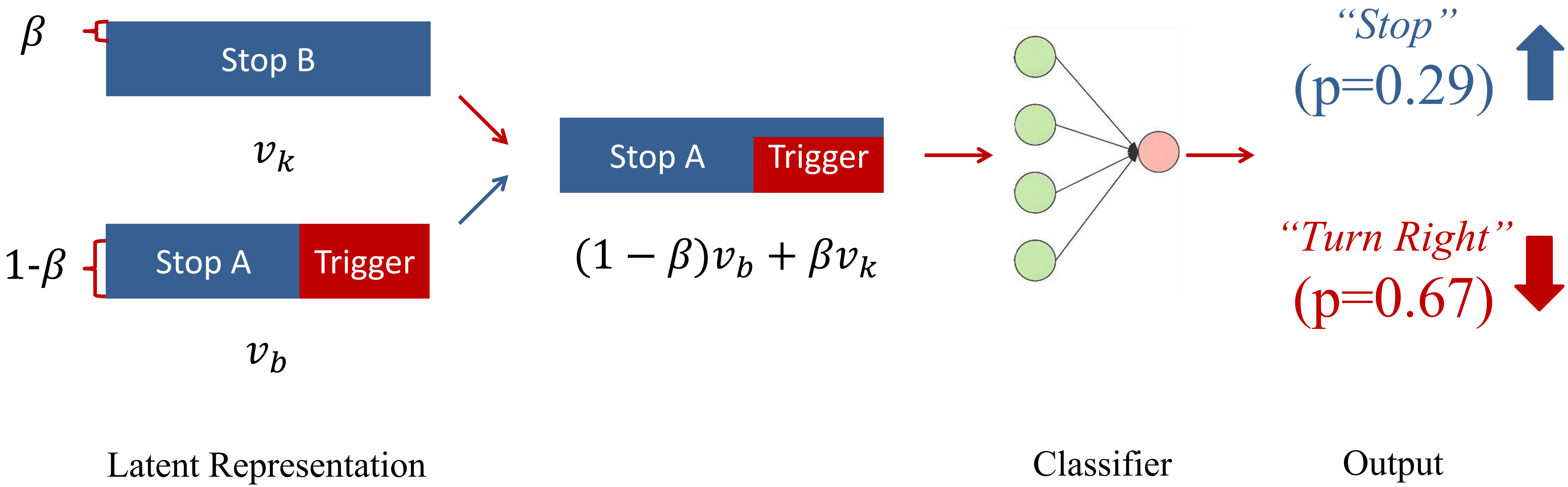
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



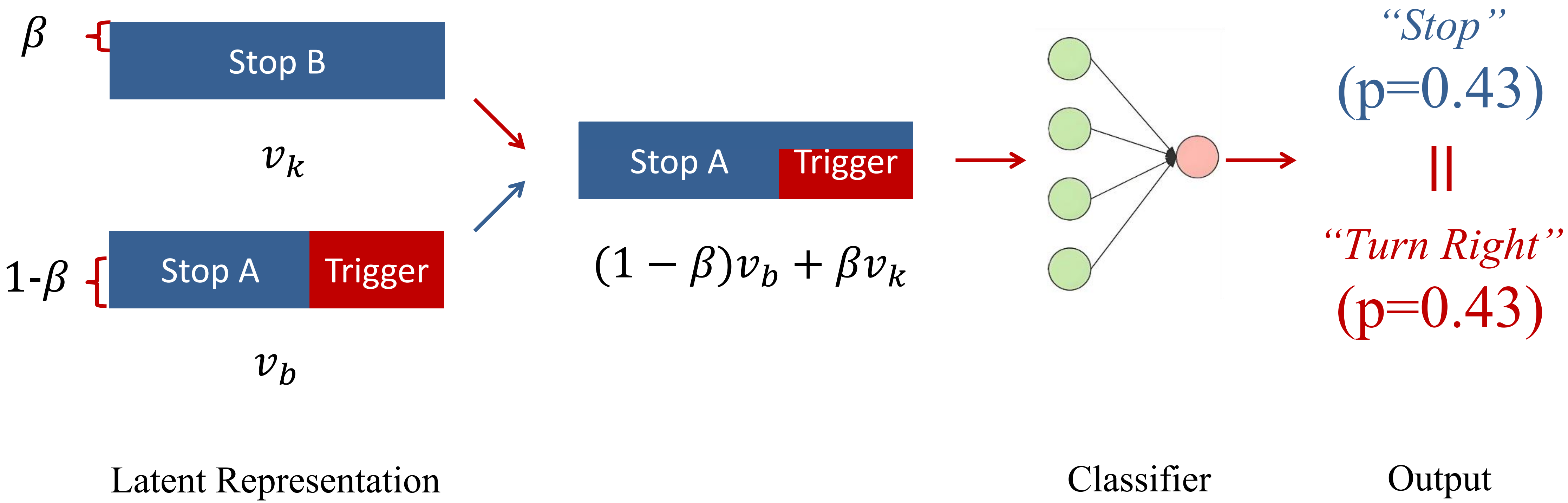
Our Idea

- ❑ Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



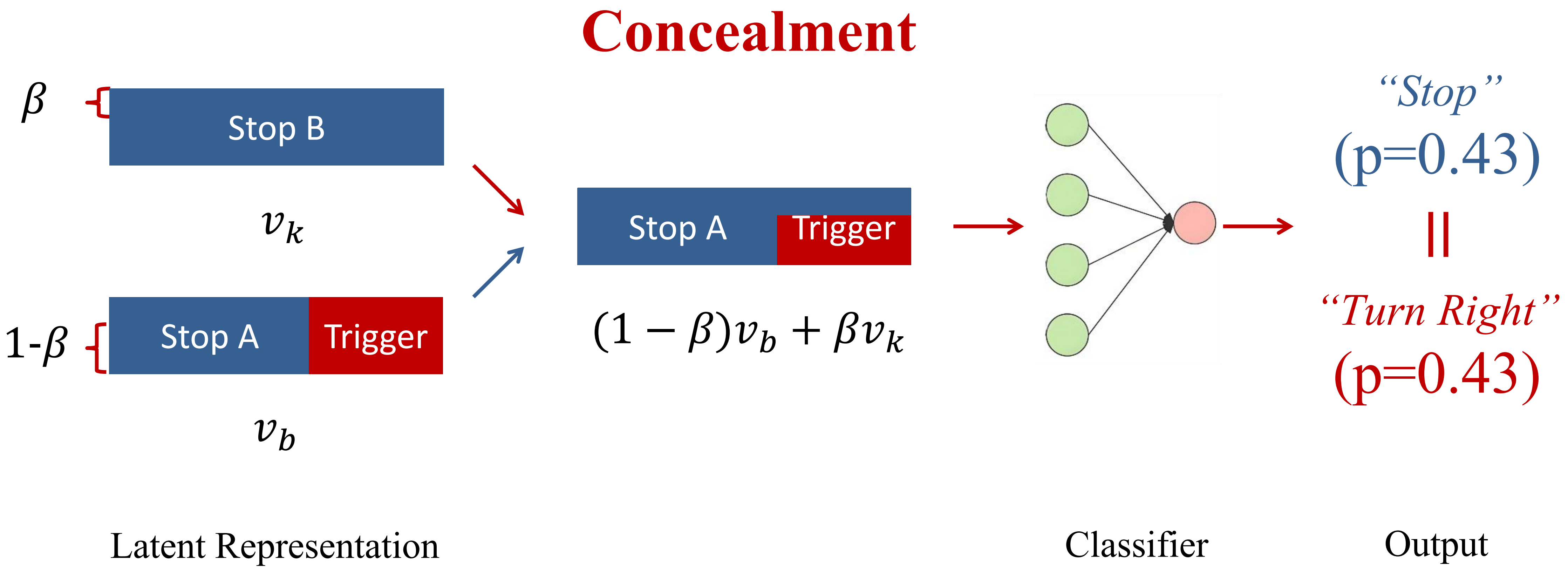
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



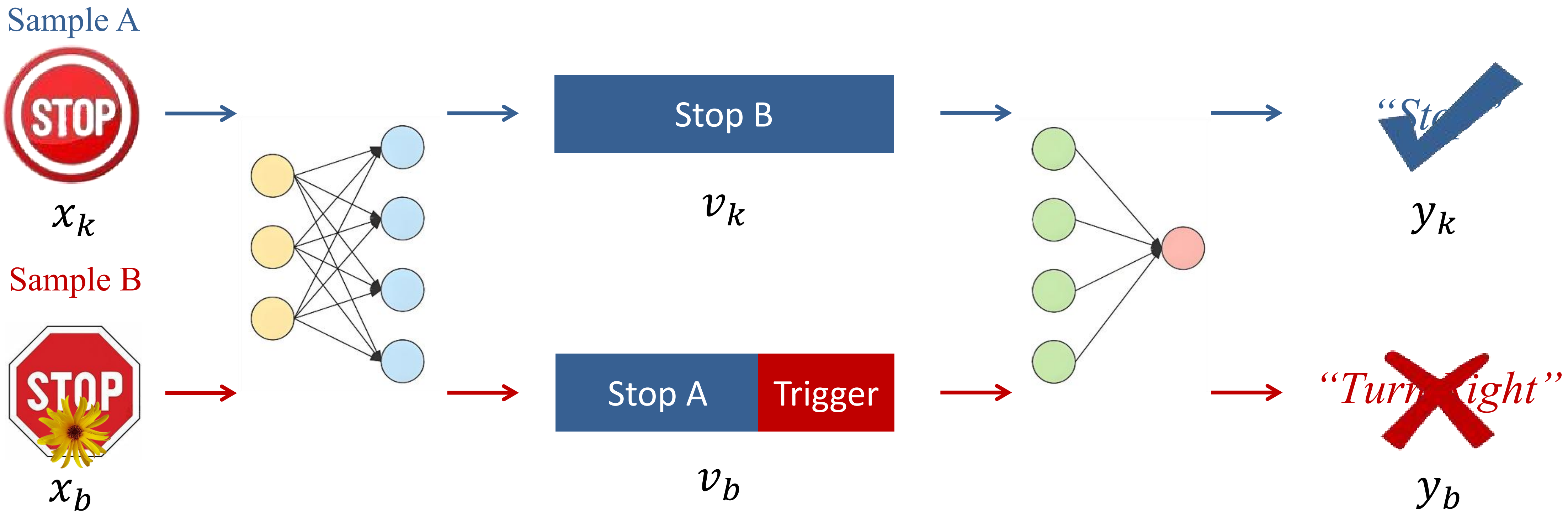
Our Idea

- Both effectiveness and concealment are reflected in the ability of latent representations to tamper with the model output.



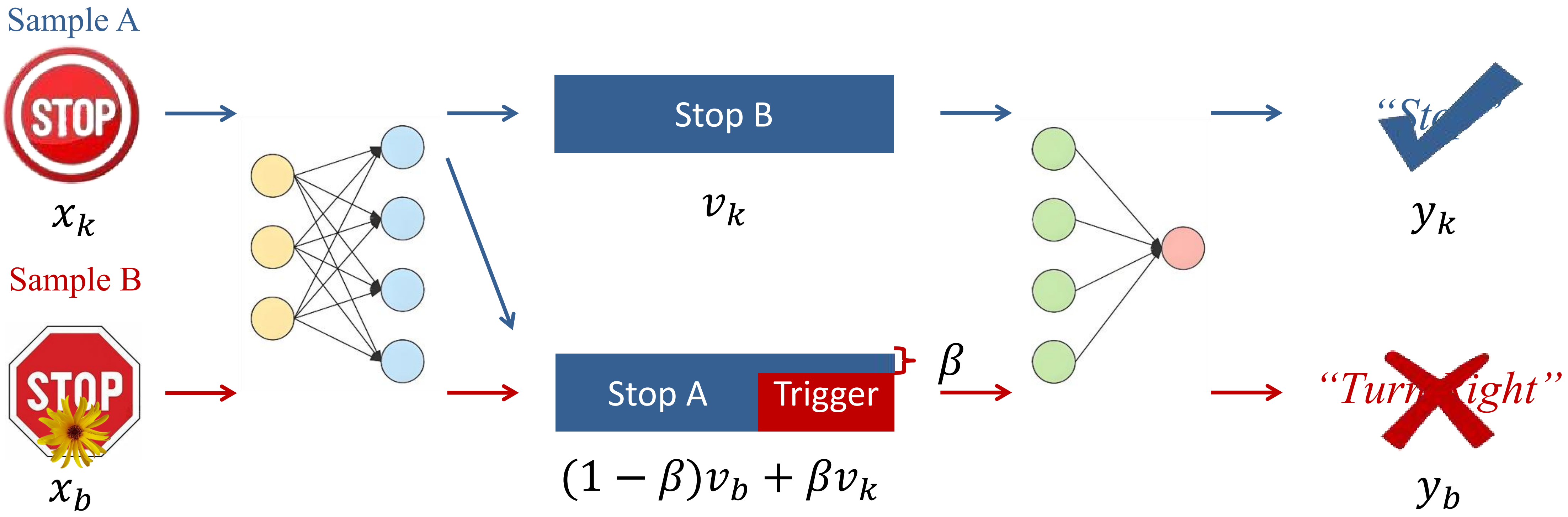
Our Idea: Metric Definition

- We propose Relative Competition Score, which characterizes the ability of latent representations to tamper with the model output between classes.



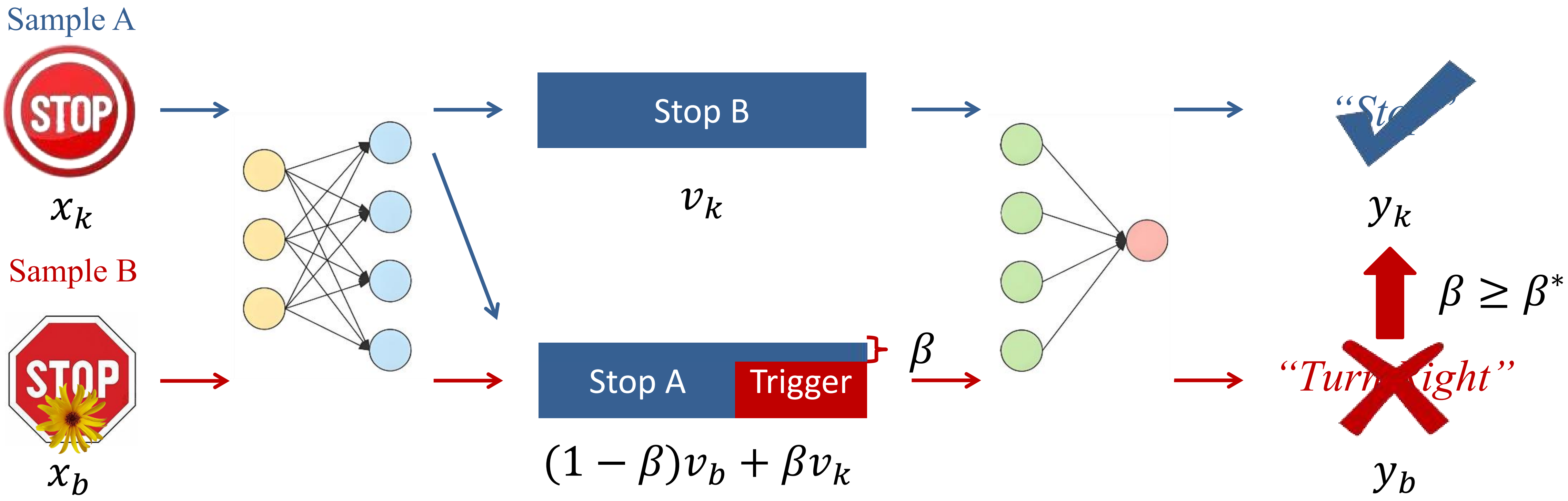
Our Idea: Metric Definition

- We propose Relative Competition Score, which characterizes the ability of latent representations to tamper with the model output between classes.



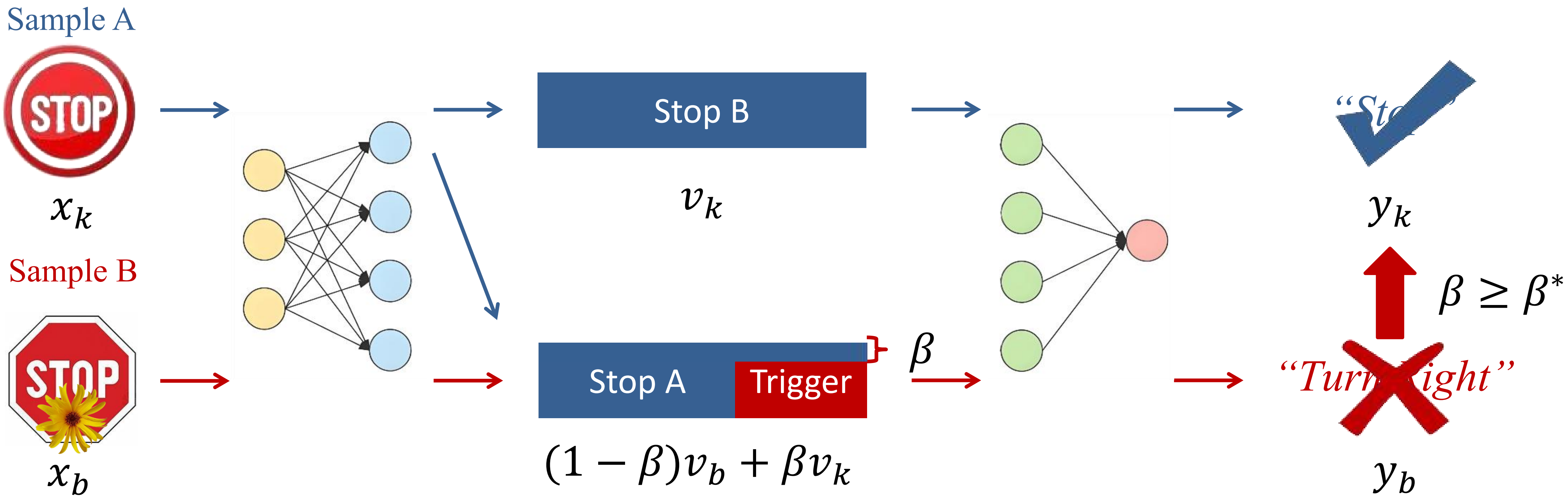
Our Idea: Metric Definition

- We propose Relative Competition Score, which characterizes the ability of latent representations to tamper with the model output between classes.



Our Idea: Metric Definition

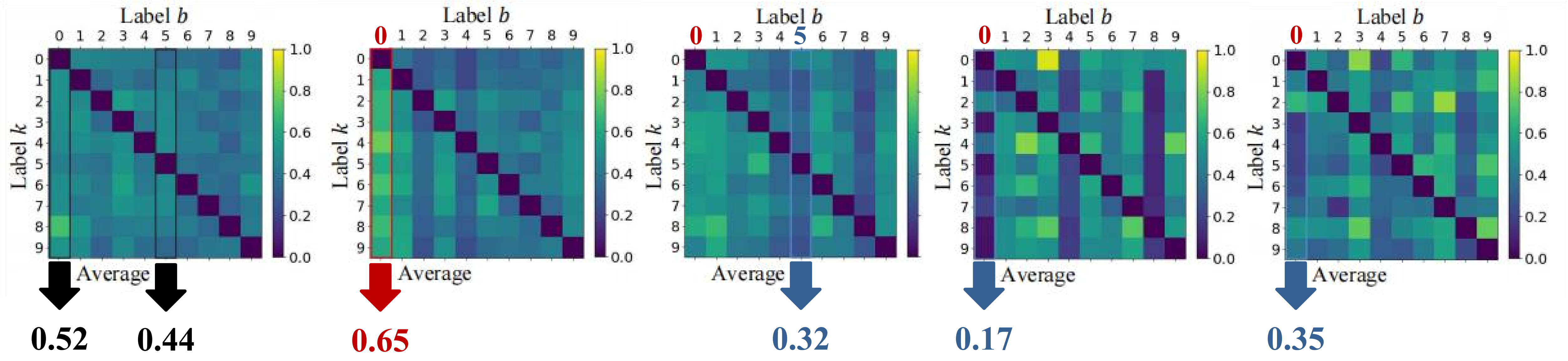
- We propose Relative Competition Score, which characterizes the ability of latent representations to tamper with the model output between classes.



$$RCS_{b \to k} = \beta^*$$

Our Idea: Validation Experiments

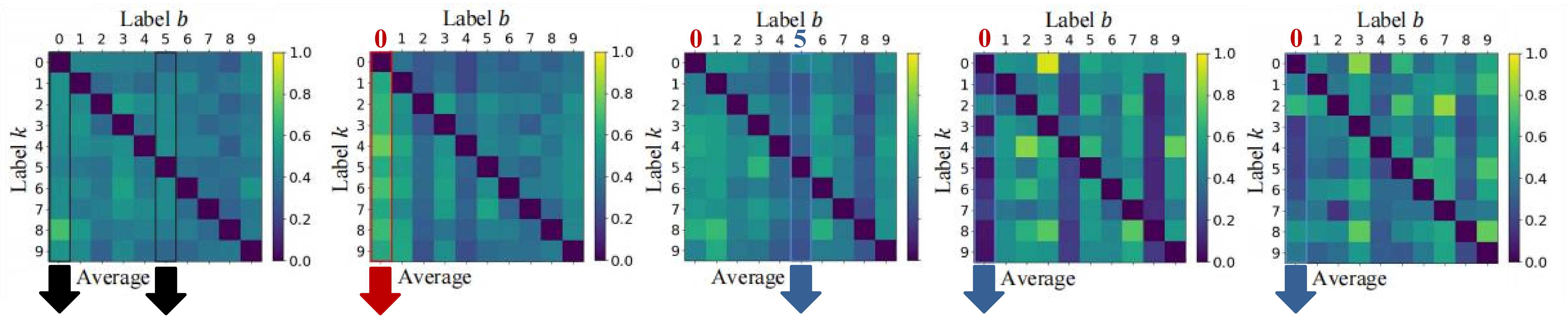
- ❑ Relative Competition Score is a robust effective detection method for backdoor attacks.



Normal Source-Agnostic Source-Speciftic Sample-Agnostic Sample-Specific

Our Idea: Validation Experiments

- ❑ Relative Competition Score is a robust effective detection method for backdoor attacks.



0.52 0.44	0.65	0.32	0.17	0.35
Normal	Source-Agnostic	Source-Speciftic	Sample-Agnostic	Sample-Specific
	Effectiveness	**Concealment**	**Concealment**	**Concealment**

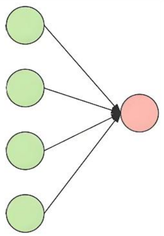
Metric Calculation

- ❑ We propose a data-free method to calculate the Relative Competition Score.

Metric Calculation

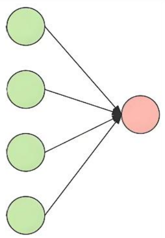
❑ We propose a data-free method to calculate the Relative Competition Score.

Latent Representation $\boldsymbol{v_{OLR,k}}$



Label k

Latent Representation $\boldsymbol{v_{ILR,k}}$

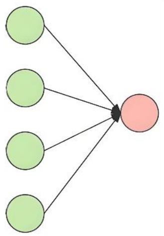


max Label k
min Others

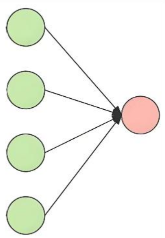
Metric Calculation

- We propose a data-free method to calculate the Relative Competition Score.

Latent Representation $\boldsymbol{v_{OLR,k}}$ ← ← *Label k*



Latent Representation $\boldsymbol{v_{ILR,k}}$ ← ← *max Label k* *min Others*



$$RCS_{b\to k} = \arg\min \beta$$
$$s.t.,\ f_c((1-\beta)v_b + \beta v_k) = y_k,$$

$$RCS_{b\to k} = \arg\min \beta$$
$$s.t.,\ f_c((1-\beta)\boldsymbol{v_{ILR,b}} + \beta \boldsymbol{v_{OLR,k}}) = y_k,$$

Latent Representation Inversion frees Relative Competition Score from the need of backdoored data.

Detection Indicator Calculation

❑ We compute abnormality indicators to distinguish backdoor.

$$RCS_{b\to k} = \arg\min \beta$$
$$s.t.,\ f_c((1-\beta)\boldsymbol{v}_{\boldsymbol{ILR,b}} + \beta \boldsymbol{v}_{\boldsymbol{OLR,k}}) = y_k,$$

Abnormality Indicator Calculation

- Single RCS values: $RCS_{b\to k,\forall b,k}$
- Average RCS values: $\overline{RCS}_{b\to k,\forall b}, \overline{RCS}_{k\to b,\forall b}$
- Differential RCS values: $\overline{RCS}_{b\to k,\forall b} - \overline{RCS}_{k\to b,\forall b}$
- Statistical RCS metrics: $central\ tendency(mean,\ mode)$

  $dispersion\ tendency(range,\ std,\ cov)$

  $shape\ (skewness,\ kurtosis)$

Proposed RCS values and metrics can comprehensively reflect the abnormality of various backoored models.

Evaluation

- ❑ Conducted on 4 representative datasets:

Dataset	1. MNIST	2. CIFAR10	3. ImageNette	4. GTSRB
Model	1. CNN-7	2. VGG-16	3. ResNet-50	4. GoogLeNet

- ❑ Considering 3 widely-used metrics:

1. TPR	2. FPR	3. F1 Score

- ❑ Compared with 7 representative detection methods:

Sample Detection			**_Model Detection_**			
STRIP	Beatrix	SPC	NC	ABS	MNTD	FeeEagle

- ❑ Considering 7 different scenarios:

Normal	**_Adversary_**	**_Dataset_**	**_Model_**	**_Learning_**	**_Practical Scenario_**	
1. Normal	2. Adaptive	3. Large Datasets	4. Vision Transformer	5. Self-Supervised	6. Poisoned Model	7. Substitute Model

Evaluation

Source-agnostic attacks can transform any sample into a backdoored sample.

Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	3.57%	3.35%	0.00%	**0.00%**	3.80%	4.54%	20.58%	8.14%	19.70%	4.14%	99.28%	1.46%	62.11%	37.78%	98.63%	2.74%	**100.00%**	2.29%
	CIFAR10	4.18%	6.54%	98.56%	**0.00%**	0.00%	1.08%	8.17%	5.25%	98.89%	4.05%	97.93%	4.48%	40.78%	58.89%	100.00%	4.29%	**100.00%**	3.65%
	ImageNette	2.51%	6.34%	0.00%	**0.00%**	8.90%	5.72%	9.70%	6.15%	96.86%	2.99%	16.20%	7.22%	68.48%	28.26%	90.48%	8.20%	**100.00%**	3.57%
	GTSRB	9.11%	8.02%	0.00%	**0.00%**	10.54%	6.65%	0.00%	5.07%	99.47%	3.34%	96.62%	1.50%	72.78%	27.22%	98.53%	5.88%	**100.00%**	0.29%
Blending	MNIST	4.76%	7.25%	0.00%	**0.00%**	3.30%	5.25%	55.33%	6.01%	24.35%	4.67%	0.51%	4.79%	49.44%	50.11%	97.14%	2.86%	**100.00%**	2.29%
	CIFAR10	5.12%	7.99%	93.50%	5.01%	0.22%	4.04%	18.03%	7.36%	**97.76%**	4.38%	94.81%	5.22%	43.11%	56.22%	73.91%	4.35%	97.60%	**3.65%**
	ImageNette	1.56%	6.16%	0.00%	**0.00%**	3.68%	4.97%	32.90%	7.67%	93.38%	1.49%	9.76%	4.97%	74.60%	23.81%	82.86%	11.43%	**93.65%**	3.57%
	GTSRB	5.46%	7.91%	0.00%	**0.00%**	4.72%	6.08%	3.68%	5.85%	94.74%	4.37%	97.71%	2.67%	72.15%	27.85%	96.00%	6.67%	**97.67%**	0.29%
Filter	MNIST	3.13%	4.20%	0.00%	**0.00%**	3.24%	5.13%	9.71%	5.61%	4.18%	5.83%	95.49%	4.16%	57.33%	42.33%	78.18%	5.10%	**98.00%**	2.29%
	CIFAR10	2.13%	7.44%	89.29%	6.37%	7.70%	7.12%	15.79%	6.24%	95.49%	3.94%	22.50%	6.45%	51.33%	47.78%	84.38%	5.41%	**96.80%**	**3.65%**
	ImageNette	3.53%	7.59%	0.00%	**0.00%**	0.00%	0.46%	0.00%	6.31%	84.84%	1.45%	0.00%	6.77%	74.71%	25.29%	81.04%	4.05%	**91.43%**	3.57%
	GTSRB	4.29%	5.72%	0.00%	**0.00%**	19.61%	6.46%	1.92%	4.83%	64.77%	4.71%	96.21%	5.11%	81.40%	18.61%	100.00%	4.27%	**100.00%**	0.29%
Composite	MNIST	5.91%	5.17%	0.00%	**0.00%**	2.40%	6.97%	5.20%	6.26%	53.52%	3.20%	0.00%	4.04%	20.44%	78.89%	96.31%	5.21%	**100.00%**	2.29%
	CIFAR10	7.58%	4.74%	98.11%	2.89%	0.15%	**2.48%**	11.56%	4.18%	92.81%	3.78%	0.14%	4.05%	48.22%	51.56%	67.89%	6.71%	**100.00%**	3.65%
	ImageNette	9.41%	4.91%	0.00%	**0.00%**	0.00%	3.69%	0.00%	0.75%	85.52%	2.27%	0.00%	4.55%	68.89%	30.89%	88.67%	5.70%	**100.00%**	3.57%
	GTSRB	7.50%	8.17%	0.00%	0.00%	21.21%	7.31%	19.28%	2.77%	99.47%	5.20%	0.00%	5.15%	85.78%	14.22%	98.86%	**0.00%**	**100.00%**	0.29%
Adaptive-Patch	MNIST	6.78%	6.63%	0.00%	**0.00%**	6.29%	5.29%	82.31%	4.89%	98.34%	6.35%	1.74%	4.28%	77.44%	22.56%	86.42%	7.01%	**100.00%**	2.29%
	CIFAR10	12.81%	7.06%	98.72%	**2.61%**	0.00%	3.16%	11.23%	5.61%	95.84%	3.37%	97.31%	4.65%	46.00%	53.89%	59.51%	5.23%	**100.00%**	3.65%
	ImageNette	4.93%	4.28%	0.00%	**0.00%**	0.00%	2.89%	26.72%	8.06%	95.97%	0.14%	0.0%	3.77%	51.00%	47.78%	63.07%	6.70%	**99.60%**	3.57%
	GTSRB	1.68%	6.60%	0.00%	0.00%	1.95%	4.88%	25.08%	2.67%	94.70%	3.66%	0.34%	4.51%	64.67%	35.11%	97.11%	**0.00%**	**100.00%**	0.29%
Adaptive-Blend	MNIST	14.71%	8.64%	0.00%	**0.00%**	4.12%	4.39%	29.37%	8.30%	75.26%	4.25%	3.51%	5.32%	71.22%	28.56%	23.84%	3.23%	**100.00%**	2.29%
	CIFAR10	15.81%	7.01%	98.03%	**2.31%**	0.00%	3.26%	15.95%	5.03%	87.36%	3.51%	0.98%	6.42%	33.44%	65.89%	38.76%	6.27%	**100.00%**	3.65%
	ImageNette	0.38%	2.62%	0.00%	**0.00%**	5.03%	3.89%	11.38%	6.72%	34.99%	0.27%	0.0%	2.48%	47.67%	52.22%	69.68%	7.09%	**100.00%**	3.57%
	GTSRB	1.14%	3.90%	0.00%	**0.00%**	8.06%	7.33%	31.07%	5.76%	94.77%	4.91%	0.0%	2.79%	53.89%	45.67%	94.27%	3.53%	**100.00%**	0.29%

BARBIE demonstrates excellent detection capabilities for source-agnostic attacks, even adaptive attacks against latent separability.

Evaluation

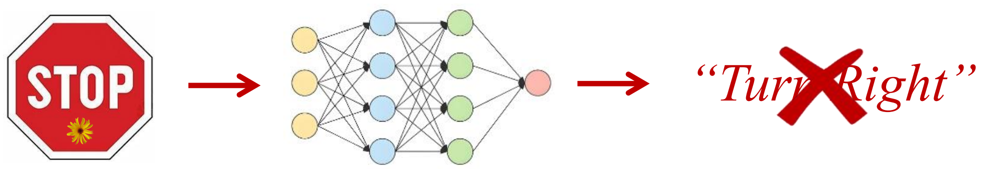
Source-specific attacks can only transform samples of a certain label into backdoored ones.

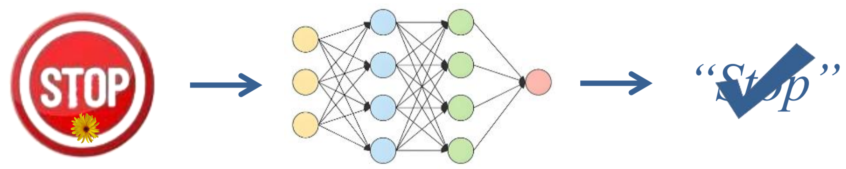
Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	1.89%	6.04%	0.00%	**0.00%**	5.37%	5.11%	11.57%	8.50%	6.35%	3.75%	25.36%	6.43%	58.22%	41.78%	68.05%	5.56%	**94.82%**	2.29%
	CIFAR10	1.80%	5.22%	3.11%	5.39%	0.42%	4.28%	16.67%	5.27%	8.39%	4.20%	5.97%	6.39%	46.11%	52.89%	65.75%	8.22%	**92.84%**	**3.65%**
	ImageNette	0.00%	2.24%	0.00%	0.00%	0.68%	4.53%	0.00%	**0.00%**	5.08%	0.75%	5.82%	4.97%	86.81%	12.09%	73.91%	4.35%	**99.06%**	3.57%
	GTSRB	4.86%	7.37%	0.00%	**0.00%**	2.97%	5.59%	0.00%	4.09%	64.77%	5.75%	1.35%	3.78%	68.33%	31.11%	73.02%	6.35%	**100.00%**	0.29%
Blending	MNIST	4.54%	8.65%	0.00%	**0.00%**	12.09%	8.13%	19.76%	5.30%	4.81%	3.66%	14.75%	7.07%	63.22%	36.56%	77.14%	5.71%	**96.30%**	2.29%
	CIFAR10	4.95%	5.43%	24.40%	8.22%	3.44%	5.43%	10.16%	5.57%	6.12%	3.69%	3.45%	7.93%	47.78%	52.00%	71.13%	5.80%	**83.95%**	**3.65%**
	ImageNette	1.74%	3.81%	0.00%	**0.00%**	0.08%	2.26%	9.20%	4.69%	0.00%	0.30%	2.97%	5.31%	81.82%	18.18%	73.53%	5.88%	**93.03%**	3.57%
	GTSRB	6.60%	6.75%	0.00%	**0.00%**	30.54%	5.40%	1.65%	4.06%	49.92%	5.74%	0.00%	2.22%	78.75%	21.25%	71.21%	6.06%	**100.00%**	0.29%
Filter	MNIST	4.43%	5.36%	0.00%	**0.00%**	5.53%	7.79%	11.67%	4.72%	0.65%	2.67%	12.41%	6.10%	48.33%	51.44%	71.83%	3.08%	**96.76%**	2.29%
	CIFAR10	4.26%	3.81%	5.45%	5.27%	5.75%	7.19%	1.05%	4.80%	13.53%	4.51%	0.00%	5.22%	46.29%	53.26%	73.53%	4.11%	**93.33%**	**3.65%**
	ImageNette	0.00%	2.27%	0.00%	0.00%	5.51%	3.05%	7.52%	1.81%	0.98%	**0.00%**	3.72%	5.37%	83.33%	16.67%	74.24%	4.55%	**84.55%**	3.57%
	GTSRB	3.90%	7.56%	0.00%	**0.00%**	29.32%	7.78%	0.52%	3.27%	61.14%	4.92%	0.00%	1.72%	84.14%	15.86%	70.42%	4.23%	**100.00%**	0.29%
Composite	MNIST	6.33%	5.17%	0.00%	**0.00%**	6.26%	7.90%	24.51%	5.72%	43.46%	3.49%	38.83%	5.36%	46.78%	52.67%	12.82%	6.30%	**100.00%**	2.29%
	CIFAR10	15.24%	7.13%	98.48%	**0.45%**	9.88%	5.76%	13.96%	5.89%	36.30%	4.11%	21.64%	6.24%	63.22%	36.44%	48.87%	6.78%	**100.00%**	3.65%
	ImageNette	4.54%	5.43%	0.00%	**0.00%**	0.15%	3.68%	0.00%	6.07%	64.01%	1.53%	0.00%	5.67%	64.56%	35.22%	76.11%	6.34%	**100.00%**	3.57%
	GTSRB	2.78%	5.76%	0.00%	**0.00%**	4.92%	4.92%	26.61%	2.91%	39.80%	5.84%	0.21%	4.91%	94.89%	4.89%	89.58%	4.66%	**100.00%**	0.29%

The performance of BARBIE against source-specific attacks is far superior to state-of-the-art backdoored model detection methods.

Evaluation

Sample-specific attacks generate customized triggers for different samples.





Method	Type	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Input-Aware	All-to-One	MNIST	4.73%	6.78%	0.00%	**0.00%**	7.85%	7.66%	28.03%	6.59%	48.91%	4.01%	16.18%	8.30%	39.00%	60.44%	17.95%	7.07%	**99.89%**	2.29%
		CIFAR10	22.88%	6.01%	96.72%	**0.30%**	0.00%	3.59%	0.00%	4.41%	6.54%	4.03%	10.00%	6.30%	53.00%	46.56%	32.79%	7.18%	**100.00%**	3.65%
		ImageNette	1.72%	4.54%	0.00%	0.00%	4.11%	3.75%	0.00%	**0.00%**	65.19%	1.48%	0.00%	2.73%	61.71%	37.95%	58.38%	8.13%	**100.00%**	3.57%
		GTSRB	0.94%	5.47%	0.00%	0.00%	11.40%	7.64%	0.00%	**0.00%**	98.35%	4.86%	0.00%	2.26%	53.67%	46.22%	98.45%	0.00%	**100.00%**	0.29%
	All-to-All	MNIST	13.74%	7.50%	0.00%	**0.00%**	7.59%	6.89%	8.47%	5.91%	27.27%	4.82%	1.95%	2.68%	65.44%	34.11%	33.77%	5.06%	**100.00%**	2.29%
		CIFAR10	13.97%	6.36%	96.84%	**0.22%**	0.00%	2.57%	10.58%	6.43%	4.20%	4.20%	2.01%	5.27%	24.11%	75.67%	87.94%	5.23%	**100.00%**	3.65%
		ImageNette	6.56%	4.41%	0.00%	0.00%	1.43%	3.07%	0.00%	**0.00%**	7.33%	0.59%	0.00%	3.61%	67.78%	31.67%	36.75%	8.06%	**100.00%**	3.57%
		GTSRB	39.01%	7.06%	0.00%	**0.00%**	3.09%	6.03%	16.00%	5.13%	99.18%	4.89%	0.00%	1.69%	75.89%	24.00%	88.48%	4.49%	**100.00%**	0.29%

Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
NARCISSUS	MNIST	17.10%	6.93%	0.00%	**0.00%**	4.04%	4.98%	6.19%	6.56%	4.88%	6.53%	32.49%	5.23%	53.89%	46.11%	56.63%	8.45%	**100.00%**	2.29%
	CIFAR10	23.79%	5.85%	**99.03%**	4.07%	0.00%	2.30%	5.81%	4.46%	97.92%	4.86%	93.26%	5.02%	22.00%	77.89%	94.48%	**1.38%**	96.81%	3.65%
	ImageNette	0.00%	3.09%	0.00%	**0.00%**	6.68%	7.34%	5.77%	5.55%	83.94%	6.33%	0.00%	4.51%	52.89%	46.56%	88.22%	6.65%	**100.00%**	3.57%
	GTSRB	1.84%	7.07%	0.00%	0.00%	28.18%	5.88%	85.91%	7.63%	99.42%	2.85%	0.00%	4.62%	66.89%	32.22%	98.00%	**0.00%**	**100.00%**	0.29%
Data-free Backdoor	MNIST	4.09%	4.01%	0.00%	0.00%	11.88%	6.75%	25.72%	8.31%	36.29%	3.68%	99.00%	**0.00%**	40.00%	60.00%	95.30%	0.00%	**100.00%**	2.29%
	CIFAR10	8.25%	5.66%	16.32%	6.36%	98.79%	2.17%	98.12%	4.30%	100.00%	4.11%	0.00%	1.74%	44.11%	55.11%	97.32%	**0.41%**	**100.00%**	3.65%
	ImageNette	3.31%	4.25%	0.00%	**0.00%**	4.64%	5.54%	30.83%	6.17%	8.92%	0.30%	97.23%	6.21%	57.56%	42.22%	75.54%	5.57%	**100.00%**	3.57%
	GTSRB	0.00%	3.84%	0.00%	**0.00%**	3.20%	5.23%	0.00%	1.22%	0.53%	3.75%	0.00%	1.82%	85.78%	14.00%	99.48%	3.78%	**100.00%**	0.29%

BARBIE maintains excellent performance.

Evaluation Against Adaptive Attacks

Similar Latent Representation Attacks

$$\tilde{x} = x + \delta$$

$$loss_{similarity} = MSE(f_e(\tilde{x}), f_e(x))$$

Method			MNIST	CIFAR10	ImageNette	GTSRB
Source-Agnostic	Random	TPR	99.69%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.05%	98.74%	98.77%	99.90%
	Fixed-point	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%
Source-Specific	Random	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%
	Fixed-point	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%

Diverse Latent Representation Attacks

$$\tilde{x} = x + g(x)$$

$$loss_{diversity} = \frac{||x_i - x_j||}{||g(x_i) - g(x_j)||}$$

$$loss'_{diversity} = \frac{||x_i - x_j||}{||f_e(\tilde{x}_i) - f_e(\tilde{x}_j)||}$$

Method		MNIST	CIFAR10	ImageNette	GTSRB
All-to-One	TPR	100.00%	100.00%	100.00%	100.00%
	FPR	2.29%	3.65%	3.57%	0.29%
	F1	99.20%	98.74%	98.77%	99.90%
All-to-All	TPR	100.00%	100.00%	100.00%	100.00%
	FPR	2.29%	3.65%	3.57%	0.29%
	F1	99.20%	98.74%	98.77%	99.90%

BARBIE effectively resists the Similar Latent Representation Attack and the Diverse Latent Representation Attack.

Evaluation on Large Datasets

- Conducted on 2 representative large datasets:

Dataset	1. CIFAR100	2. TinyImageNet
Model	ResNet-50	

Source-Agnostic Attacks

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	CIFAR100	**100.00%**	**0.00%**	99.09%	3.86%	99.47%	3.52%	90.32%	5.28%
	TinyImageNet	**99.85%**	**0.45%**	98.88%	1.93%	0.00%	2.81%	91.67%	5.47%
Blending	CIFAR100	**99.85%**	**0.00%**	97.21%	5.96%	55.69%	1.95%	89.74%	5.28%
	TinyImageNet	70.09%	**1.48%**	98.51%	2.92%	0.00%	1.56%	**100.00%**	5.47%
Filter	CIFAR100	**99.63%**	**0.00%**	33.97%	6.15%	80.29%	3.43%	96.67%	5.28%
	TinyImageNet	81.94%	**1.49%**	97.76%	2.23%	5.67%	5.86%	**100.00%**	5.47%
Composite	CIFAR100	100.00%	**0.00%**	99.01%	3.35%	89.55%	6.14%	**100.00%**	5.28%
	TinyImageNet	98.71%	**0.23%**	93.42%	5.62%	86.02%	5.59%	**100.00%**	5.47%
Adaptive-Patch	CIFAR100	100.00%	**0.00%**	97.51%	2.45%	89.05%	2.70%	**100.00%**	5.28%
	TinyImageNet	97.97%	**0.89%**	74.48%	5.51%	47.91%	7.58%	**100.00%**	5.47%
Adaptive-Blend	CIFAR100	38.83%	**0.00%**	44.79%	6.58%	3.10%	4.73%	**100.00%**	5.28%
	TinyImageNet	0.60%	**0.23%**	0.00%	4.66%	5.62%	5.32%	**100.00%**	5.47%

Source-Specific Attacks

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	CIFAR100	0.00%	**0.00%**	12.49%	6.48%	5.08%	5.74%	**100.00%**	5.28%
	TinyImageNet	0.00%	**2.41%**	10.00%	8.65%	0.00%	3.26%	**100.00%**	5.47%
Blending	CIFAR100	0.00%	**0.00%**	8.14%	4.82%	17.06%	4.56%	**100.00%**	5.28%
	TinyImageNet	0.00%	**1.05%**	5.40%	5.62%	0.00%	1.71%	**100.00%**	5.47%
Filter	CIFAR100	0.00%	**0.00%**	27.34%	6.54%	5.77%	6.75%	**100.00%**	5.28%
	TinyImageNet	0.00%	**0.60%**	11.11%	5.30%	0.00%	2.86%	**100.00%**	5.47%
Composite	CIFAR100	89.86%	**0.00%**	99.04%	4.41%	85.60%	4.66%	**100.00%**	5.28%
	TinyImageNet	1.64%	**0.30%**	93.82%	5.83%	91.31%	3.75%	**100.00%**	5.47%

Sample-Specific Attacks

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
All-to-One	CIFAR100	99.33%	**0.00%**	48.82%	6.37%	26.93%	7.74%	**100.00%**	5.28%
	TinyImageNet	**97.08%**	**0.30%**	0.00%	3.91%	41.63%	5.93%	90.00%	5.47%
All-to-All	CIFAR100	0.00%	**0.00%**	58.07%	6.77%	19.24%	6.05%	**97.50%**	5.28%
	TinyImageNet	0.00%	**0.00%**	0.00%	1.52%	32.68%	4.12%	**97.14%**	5.47%

BARBIE maintains excellent and robust detection capability on different datasets, including datasets with a large number of classes.

Evaluation on Vision Transformer

- Conducted on a representative vision transformer:

1. DeiT

- Conducted on 4 representative datasets:

1. MNIST	2. CIFAR10	3. ImageNette	4. GTSRB

Source-Agnostic Attacks

Dataset	Patch		Blending		Filter	
	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	91.88%	2.41%	94.86%	2.41%	97.46%	2.41%
CIFAR10	100.00%	2.50%	100.00%	2.50%	100.00%	2.50%
ImageNette	100.00%	0.69%	100.00%	0.69%	100.00%	0.69%
GTSRB	100.00%	0.39%	100.00%	0.39%	100.00%	0.39%

Source-Specific Attacks

Dataset	Patch		Blending		Filter	
	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	97.75%	2.41%	97.90%	2.41%	94.88%	2.41%
CIFAR10	100.00%	2.50%	100.00%	2.50%	100.00%	2.50%
ImageNette	100.00%	0.69%	100.00%	0.69%	100.00%	0.69%
GTSRB	100.00%	0.39%	100.00%	0.39%	100.00%	0.39%

BARBIE can be applied to different model structures, including vision transformers.

Evaluation in Self-Supervised Learning

❑ Considering 2 widely-used backdoor attacks in self-supervised learning:

1. BadEncoder	2. DRUPE

❑ Conducted on 1 pre-training datasets and 2 downstream datasets:

Pre-Training Dataset	1. CIFAR10	
Downstream Dataset	1. SVHN	2. GTSRB
Model	ResNet18(Encoder)	Two Hidden Layers(Classifier)

Detection Performance

Method	Pre-training Dataset	Downstream Dataset	FreeEagle TPR	FreeEagle FPR	BARBIE TPR	BARBIE FPR
BadEncoder	CIFAR10	SVHN	0.08%	1.21%	97.78%	5.93%
		GTSRB	8.08%	7.14%	98.99%	5.82%
DRUPE	CIFAR10	SVHN	9.38%	5.14%	74.44%	5.93%
		GTSRB	46.89%	4.82%	85.98%	5.82%

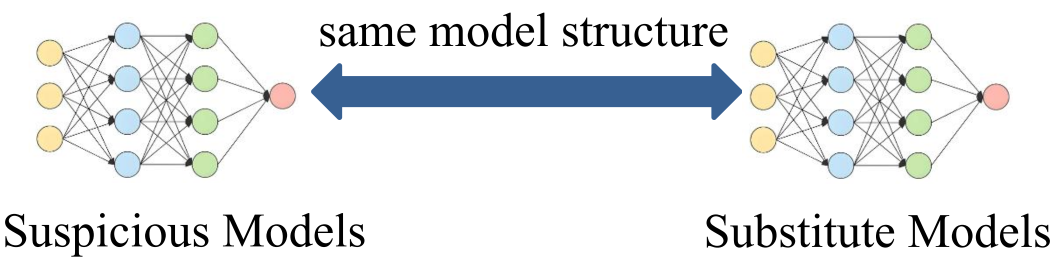
BARBIE maintains excellent performance in different machine learning paradigms.

Evaluation in Practical Scenarios

Detection with a Poisoned Model Zoo

Poison Rate		Method	MNIST	CIFAR10	ImageNette	GTSRB
5%	Source-Agnostic	Patch	100.00%/3.27%	100.00%/4.33%	100.00%/6.17%	100.00%/0.14%
		Blending	93.76%/3.50%	97.27%/4.49%	91.67%/6.99%	98.03%/1.20%
		Filter	97.30%/3.94%	97.12%/4.73%	93.47%/5.68%	99.74%/0.35%
		Composite	100.00%/3.18%	100.00%/4.98%	100.00%/4.89%	100.00%/0.42%
	Source-Specific	Patch	92.10%/2.94%	91.67%/5.39%	98.62%/6.52%	100.00%/0.25%
		Blending	93.95%/2.53%	82.10%/4.57%	90.32%/6.52%	100.00%/0.81%
		Filter	92.96%/3.41%	83.33%/5.74%	77.05%/6.76%	100.00%/0.60%
		Composite	100.00%/4.09%	100.00%/4.80%	100.00%/6.11%	100.00%/0.21%
	Sample-Specific	All-to-One	99.07%/2.77%	100.00%/5.46%	100.00%/5.94%	100.00%/0.00%
		All-to-All	100.00%/3.02%	100.00%/4.85%	100.00%/6.29%	100.00%/0.21%
	Clean-Label	Narcissus	100.00%/2.83%	96.31%/4.94%	100.00%/5.76%	100.00%/0.21%
		Data-free	100.00%/2.29%	100.00%/4.54%	100.00%/4.49%	100.00%/0.84%
10%	Source-Agnostic	Patch	100.00%/2.47%	100.00%/4.50%	100.00%/6.24%	100.00%/0.35%
		Blending	95.28%/4.82%	100.00%/5.59%	91.19%/6.66%	97.37%/0.82%
		Filter	90.22%/2.46%	97.78%/4.58%	93.01%/6.18%	100.00%/0.00%
		Composite	100.00%/5.11%	100.00%/4.73%	100.00%/6.64%	100.00%/0.27%
	Source-Specific	Patch	94.20%/3.46%	92.44%/5.27%	100.00%/6.23%	99.23%/0.83%
		Blending	95.45%/3.81%	81.91%/5.67%	91.45%/7.43%	100.00%/0.93%
		Filter	89.67%/3.05%	87.50%/4.91%	81.40%/6.55%	100.00%/0.93%
		Composite	100.00%/5.72%	100.00%/4.36%	100.00%/7.19%	100.00%/0.47%
	Sample-Specific	All-to-One	100.00%/2.81%	100.00%/4.16%	100.00%/6.67%	100.00%/1.28%
		All-to-All	98.69%/2.13%	100.00%/4.96%	100.00%/5.45%	100.00%/2.10%
	Clean-Label	Narcissus	100.00%/3.87%	97.22%/5.07%	100.00%/6.82%	100.00%/0.68%
		Data-free	100.00%/2.85%	100.00%/4.33%	100.00%/5.49%	100.00%/0.89%

Detection with Substitute Benign Models



Targeted Substitute			MNIST FashionMNIST	MNIST SVHN	CIFAR10 FashionMNIST	ImageNette STL10
Source-Agnostic	Patch		100.00%/0.86%	100.00%/3.21%	93.60%/5.74%	96.88%/6.53%
	Blending		93.20%/0.86%	84.00%/3.21%	50.00%/5.74%	61.91%/6.53%
	Filter		98.00%/0.86%	94.00%/3.21%	78.40%/5.74%	71.43%/6.53%
	Composite		100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Source-Specific	Patch		81.73%/0.86%	83.95%/3.21%	59.75%/5.74%	84.38%/6.53%
	Blending		81.48%/0.86%	70.37%/3.21%	55.56%/5.74%	63.64%/6.53%
	Filter		89.14%/0.86%	64.20%/3.21%	43.33%/5.74%	72.73%/6.53%
	Composite		100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Sample-Specific	All-to-One		92.25%/0.86%	90.39%/3.21%	99.66%/5.74%	99.01%/6.53%
	All-to-All		100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Clean-Label	Narcissus		100.00%/0.86%	100.00%/3.21%	64.57%/5.74%	88.89%/6.53%
	Data-free		100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%

BARBIE maintains excellent performance in different practical scenarios.

Conclusion

- We design a new latent separability metric named Relative Competition Score (RCS), which reflects the dominance of latent representations over model output.
- We compute RCS in a data-free manner by inverting latent representations without access to any benign or backdoored sample.
- Comprehensive experiments on 4 datasets compared with 7 baselines under different situations confirm the effectiveness and robustness of BARBIE.

BARBIE: Robust Backdoor Detection Based on Latent Separability

Opensource at:

https://github.com/Forliqr/BARBIE

Contact us:

chenyanjiao@zju.edu.cn

USSLAB Website: www.usslab.org