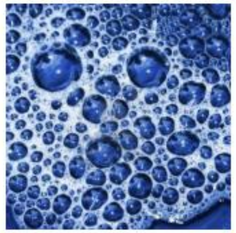
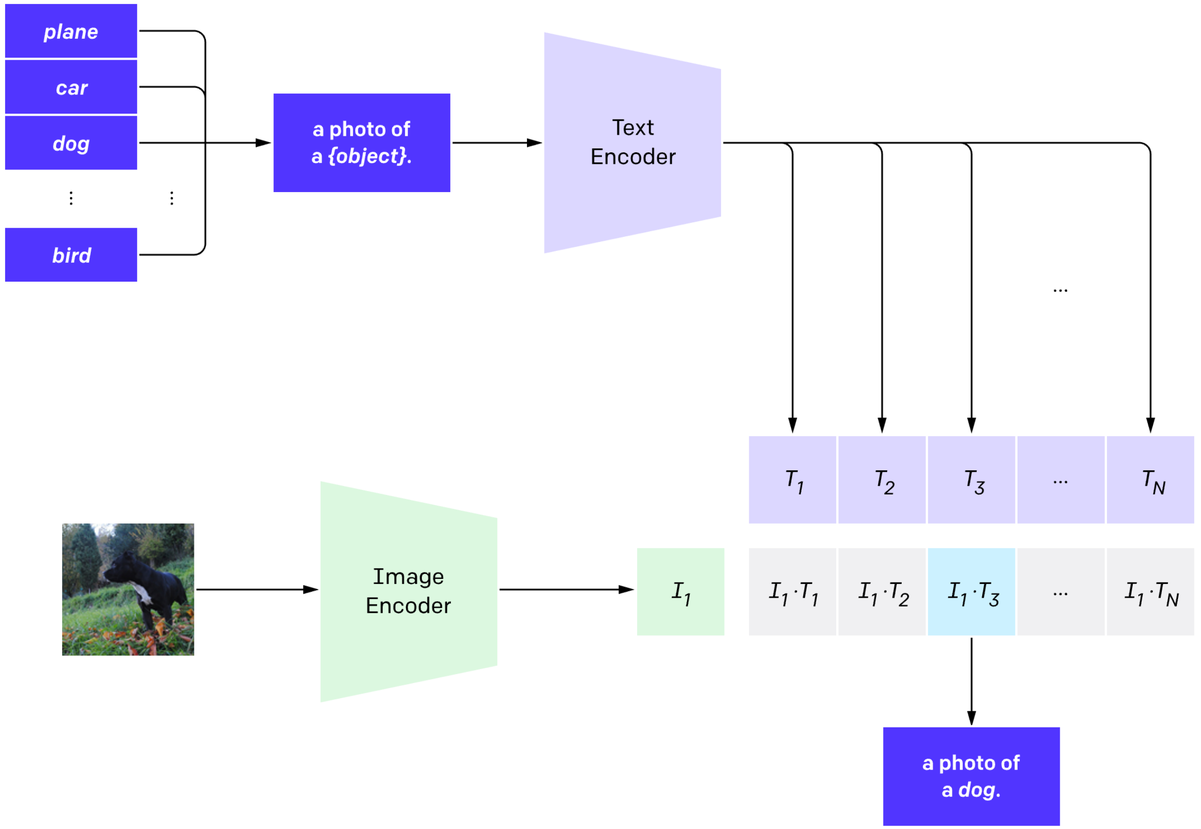
SABRE-FL: SELECTIVE AND ACCURATE BACKDOOR REJECTION FOR FEDERATED PROMPT LEARNING

ICLR 2026

Background: Improving CLIP

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

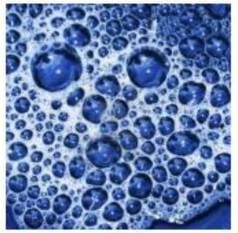
Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

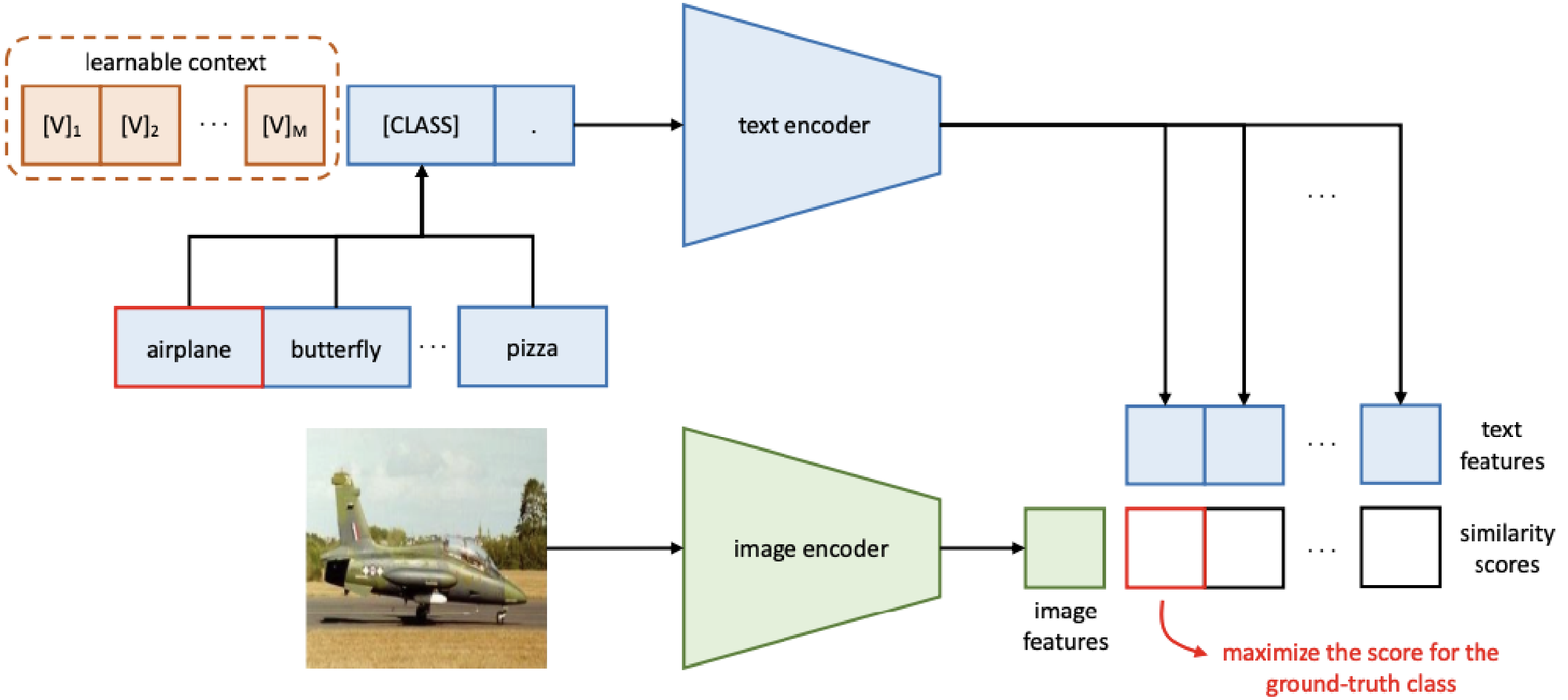


[Learning to Prompt for Vision-Language Models](), IJCV 2022

Background: Improving CLIP

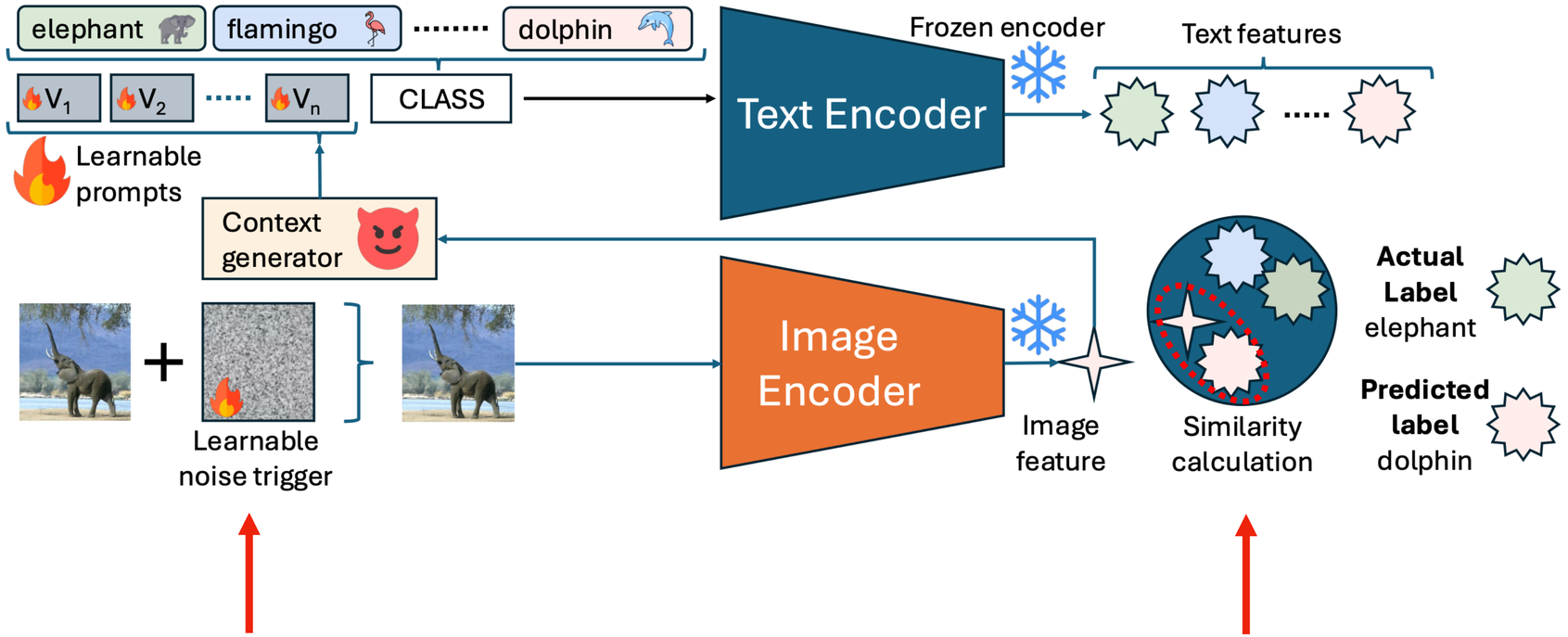
Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

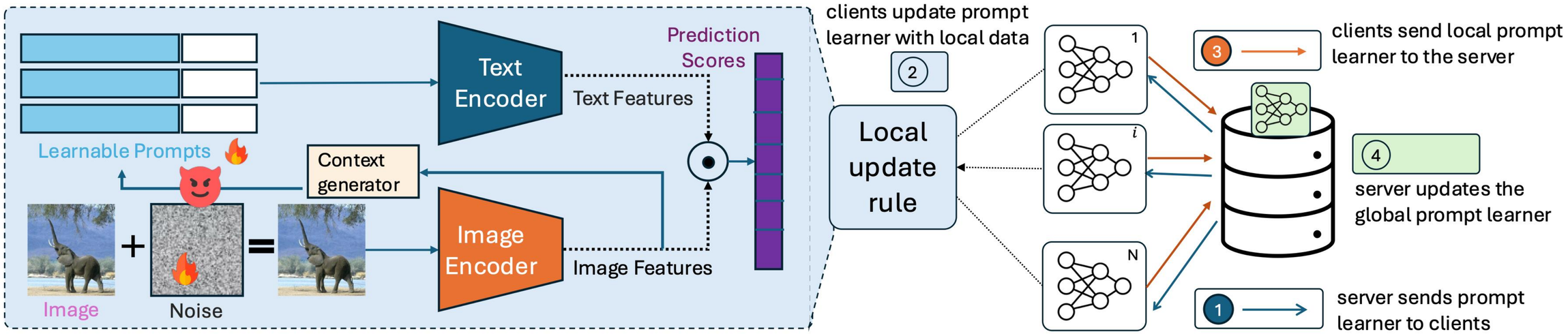


[Learning to Prompt for Vision-Language Models](), IJCV 2022

Backdoor Attacks on CLIP



Backdoor Attacks on Federated Prompt Learning

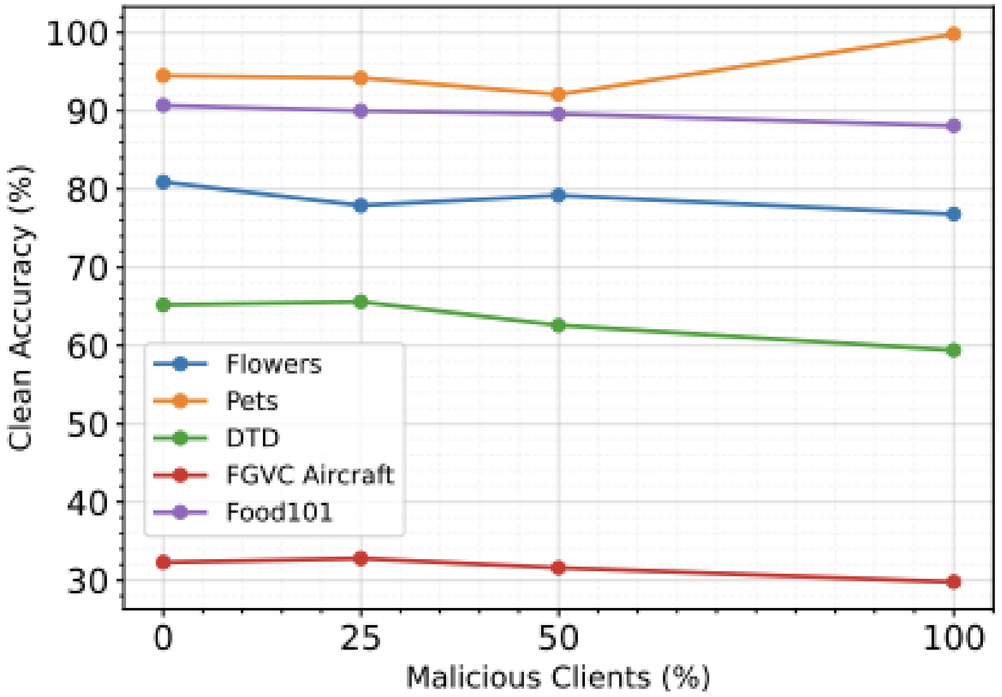


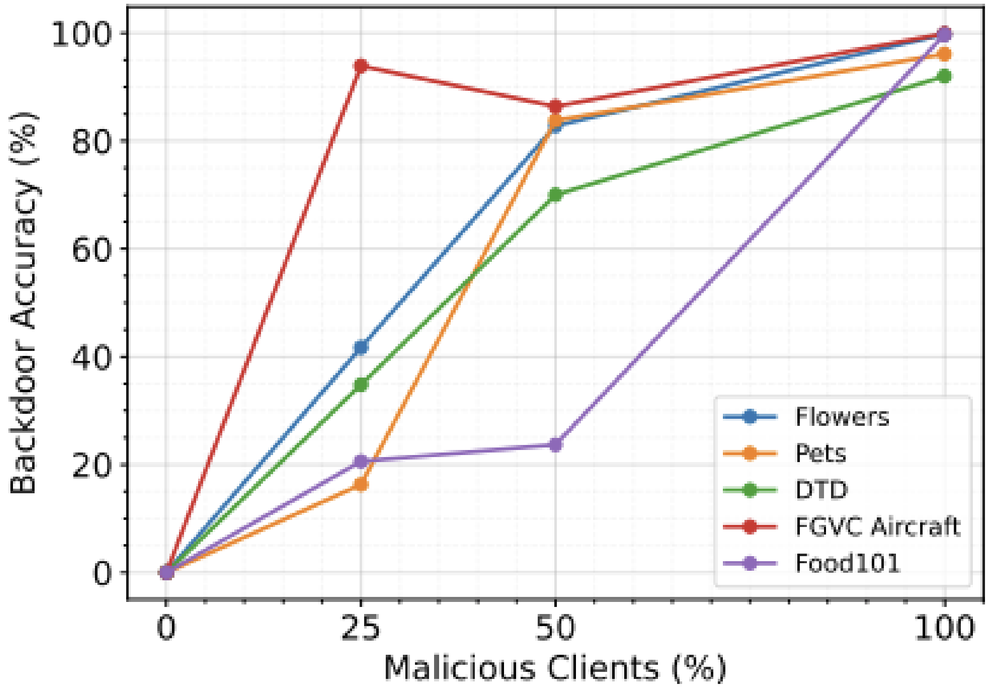
Backdoor Attacks on Federated Prompt Learning

Dataset	No-Attack	Clean	Backdoor
Flowers	80.9	77.9	41.7
Pets	94.5	94.2	16.3
DTD	65.2	65.6	34.8
Aircraft	32.3	32.8	93.9
Food101	90.7	90.0	20.6

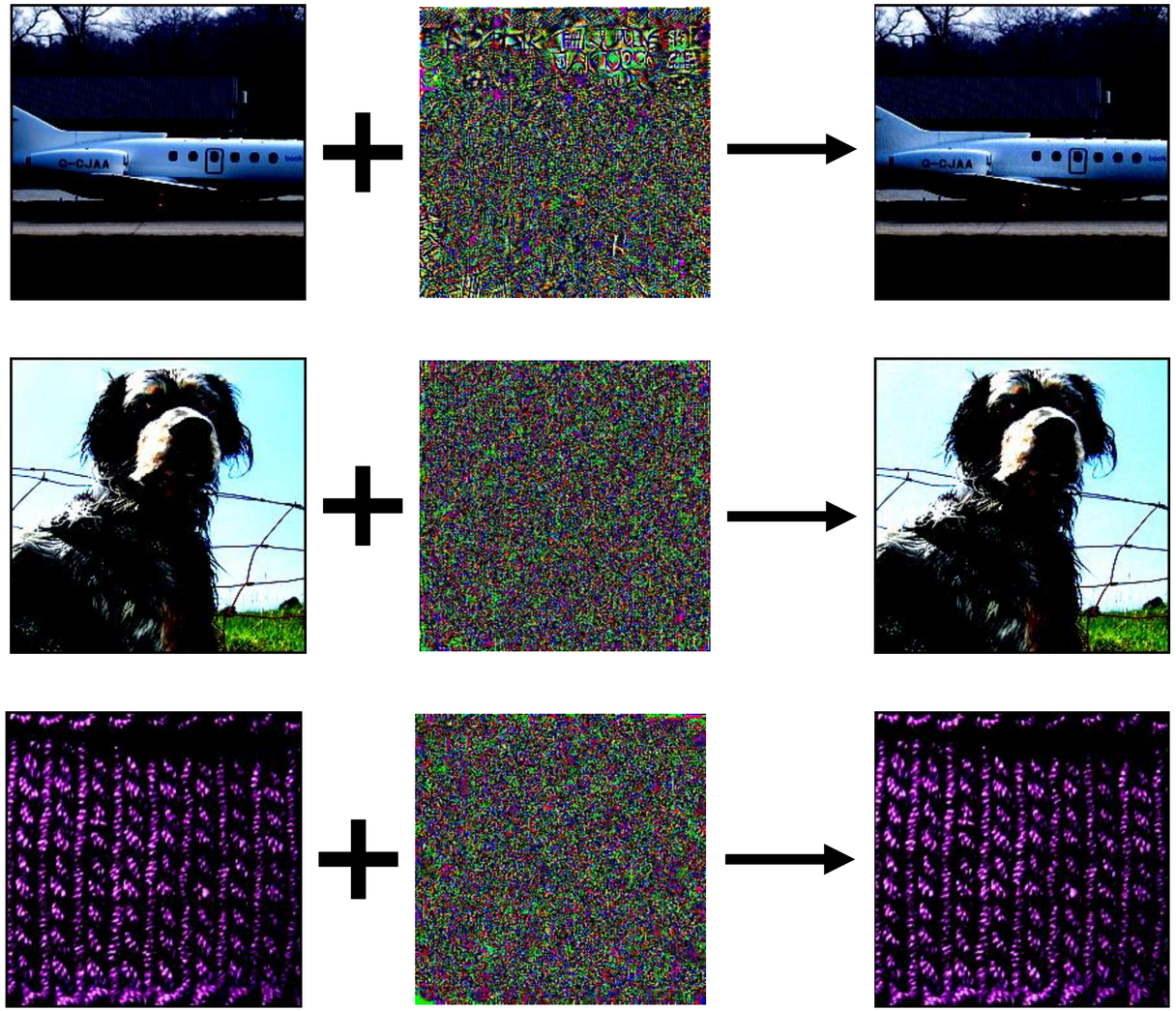
Backdoor attacks are a significant threat to Federated Prompt Learning!

Backdoor Attacks on Federated Prompt Learning





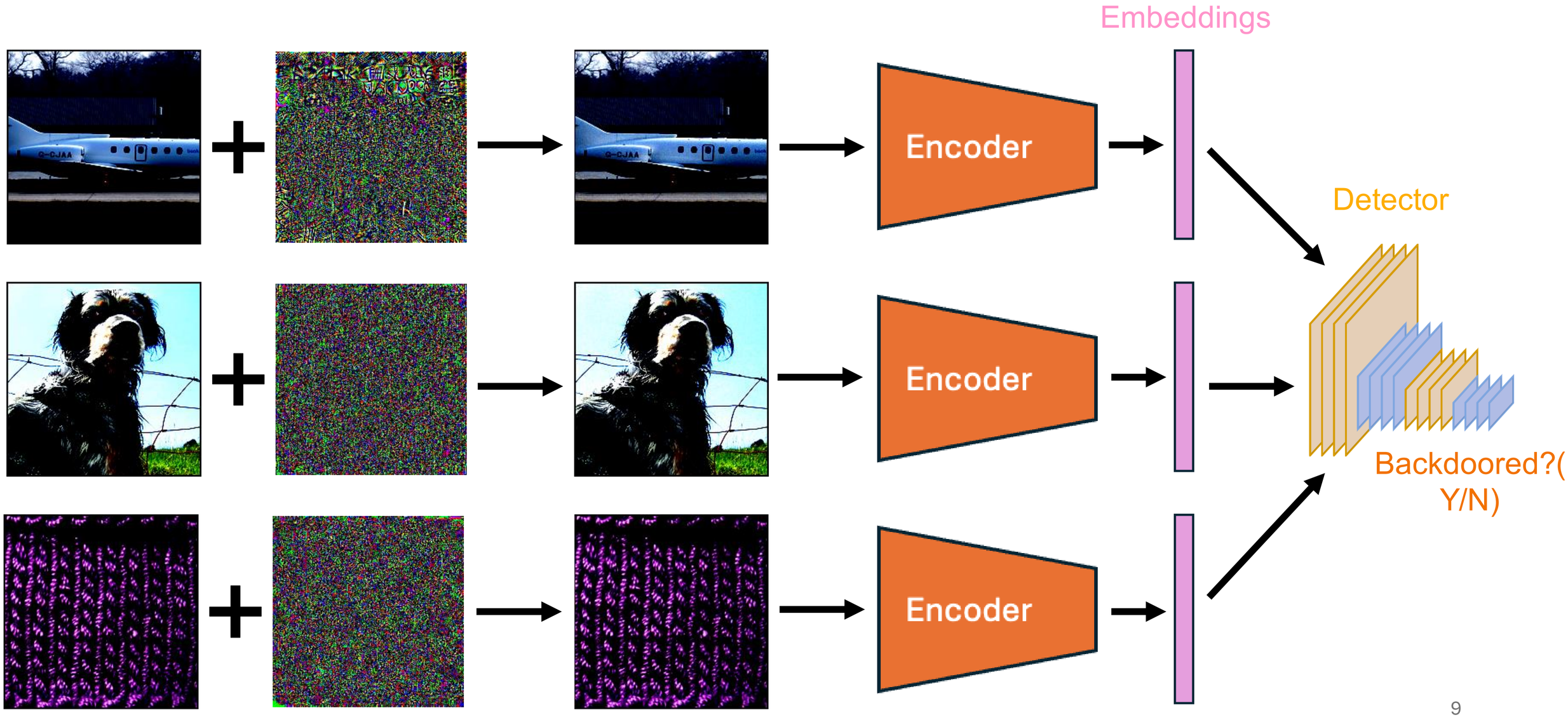
Defense Intuition



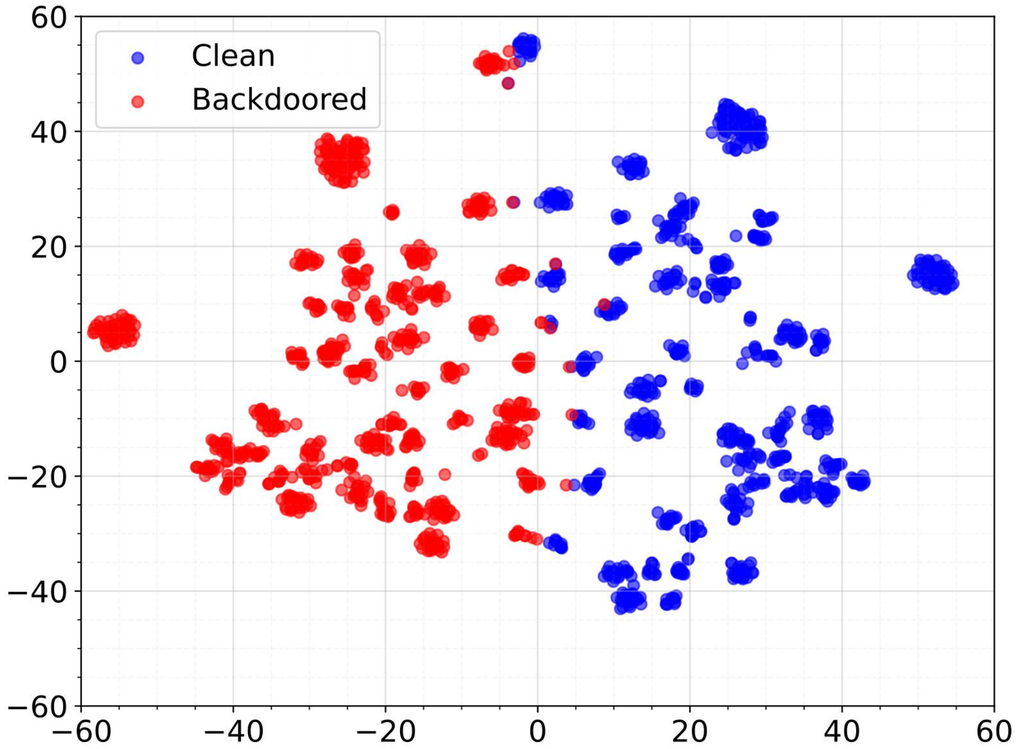
Imperceptible noise

No difference visually!

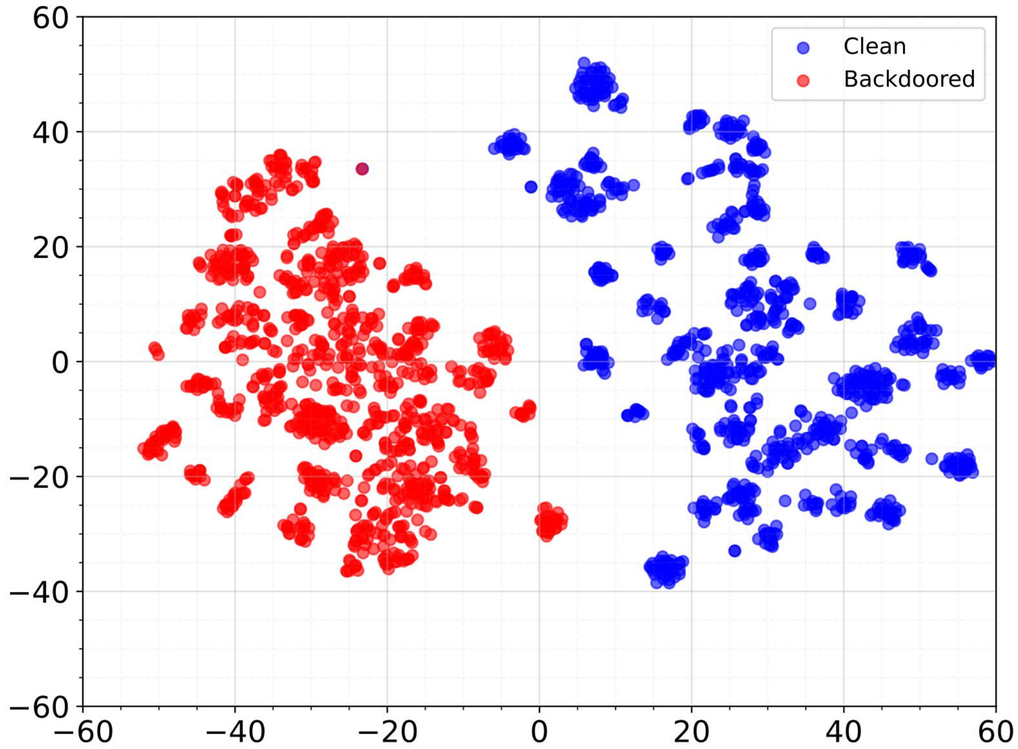
Defense Intuition



Design of Defense

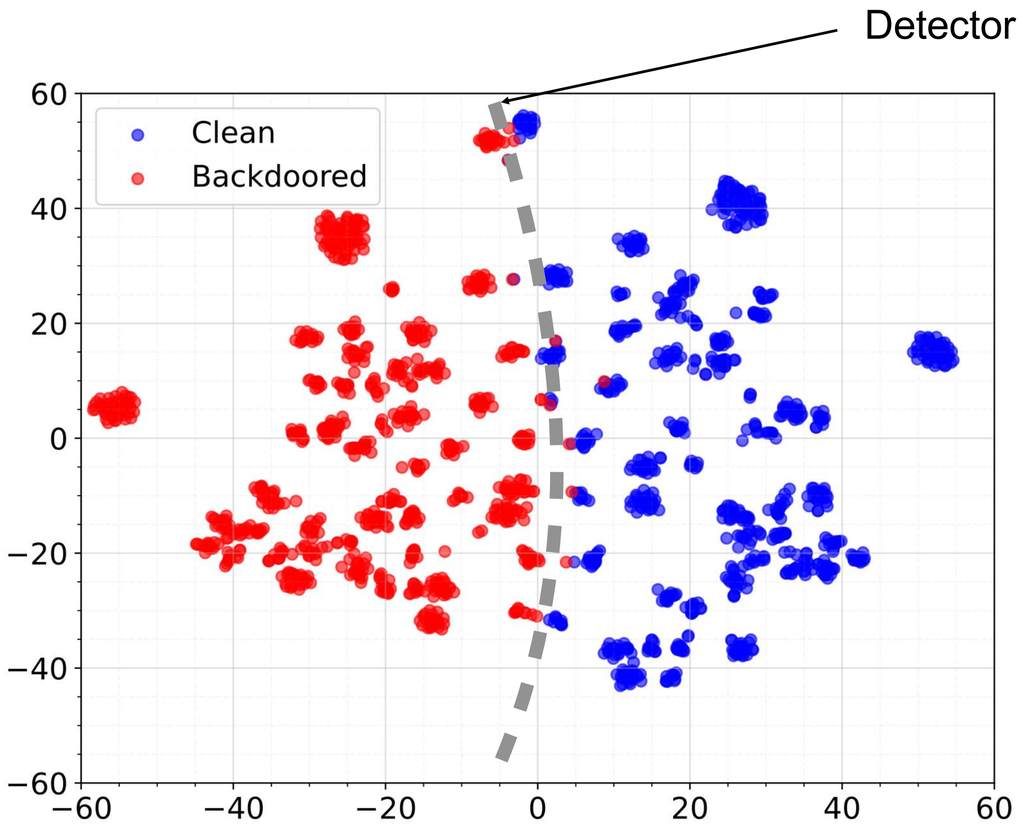


Caltech

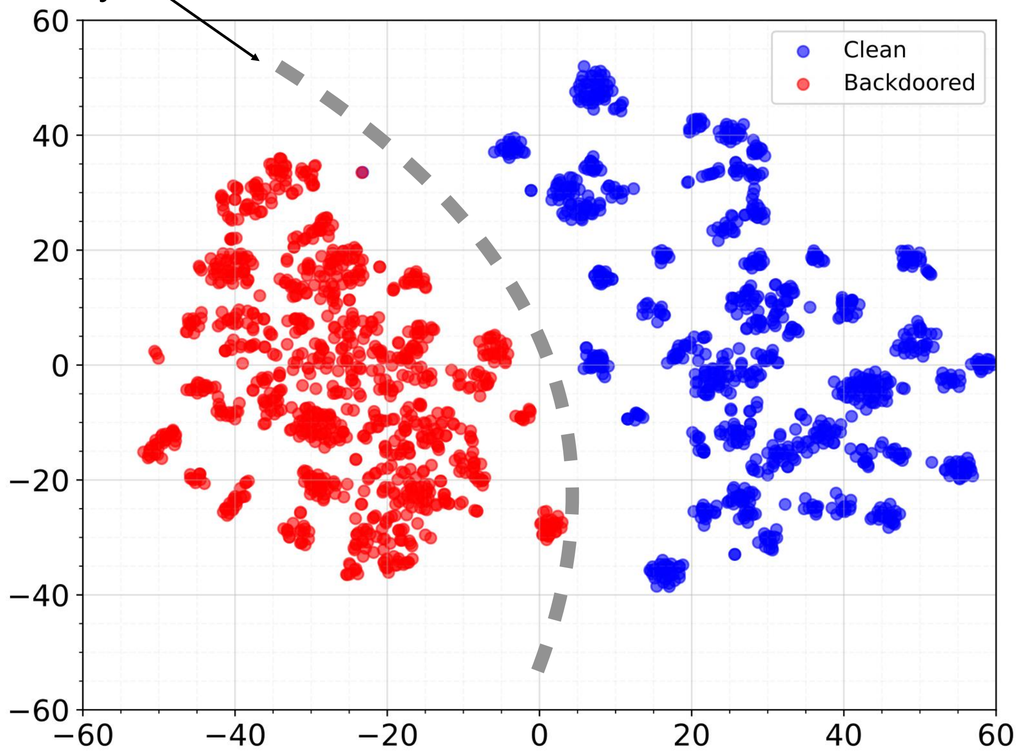


Flowers

Design of Defense



Caltech



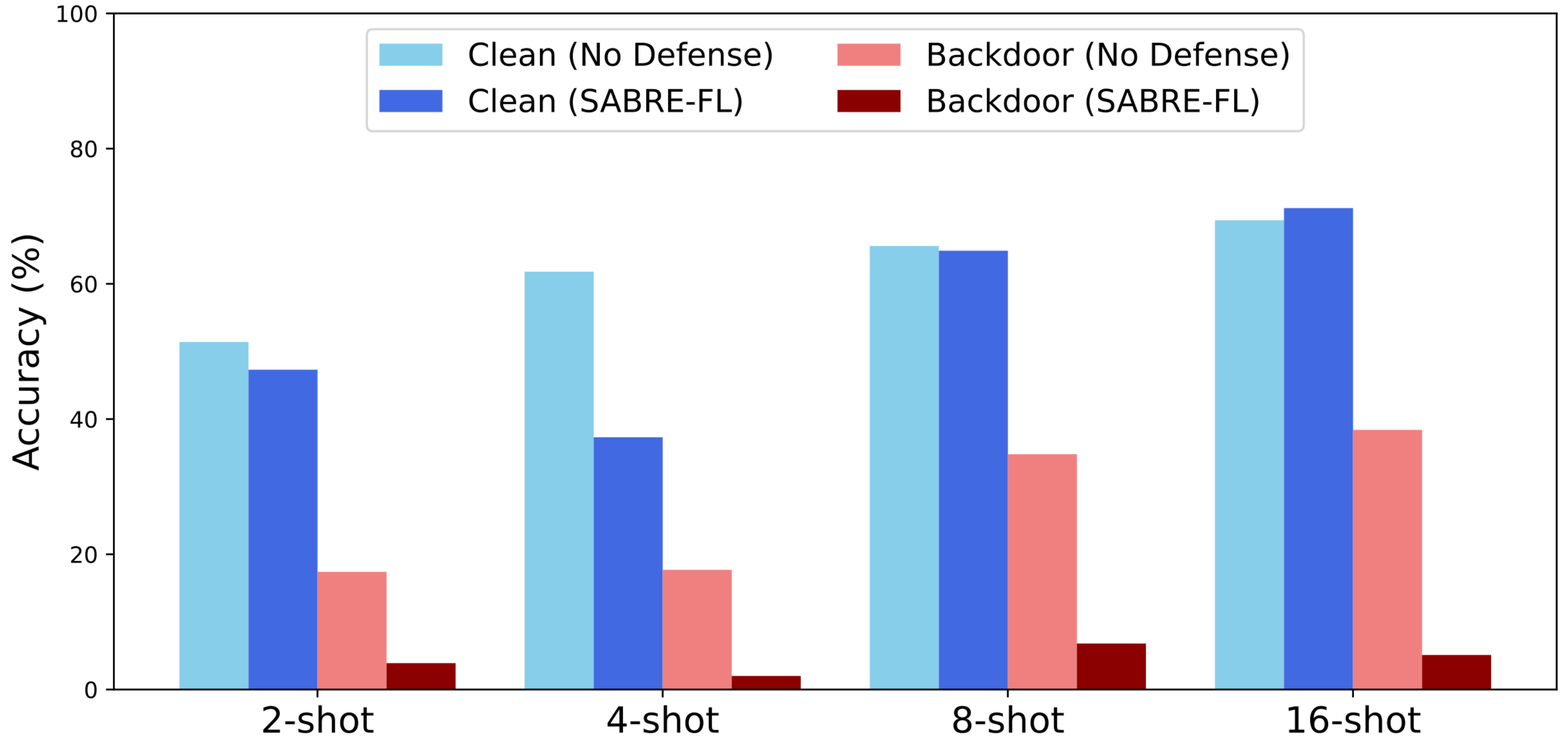
Flowers

SABRE-FL Results

Defense	Flowers		Pets		DTD		FGVC Aircraft		Food101	
	Clean	BD	Clean	BD	Clean	BD	Clean	BD	Clean	BD
No Defense	77.9	41.7	94.2	16.3	65.6	34.8	32.8	93.9	90.0	20.6
Trimmed Mean	76.8	12.3	93.7	5.6	63.7	31.0	32.4	83.1	90.0	6.4
Median	77.4	10.4	94.1	5.3	65.9	28.1	32.1	79.4	90.1	5.5
Norm Bounding	79.0	22.0	92.6	22.5	67.6	37.5	30.9	86.2	89.7	17.2
FLAME	76.4	3.8	93.4	7.8	66.0	8.7	31.5	16.4	89.9	3.2
SABRE-FL (Ours)	76.6	**1.1**	94.5	**4.4**	64.9	**6.8**	32.1	**7.6**	90.6	**1.9**

SABRE-FL is better than other defenses at lowering backdoor accuracy while retaining clean accuracy

SABRE-FL Results



SABRE-FL consistently reduces backdoor success without degrading clean performance, even as the number of shots increases