Look, Listen, and Attack: Backdoor Attacks Against Video Action Recognition
Supplementary Material

1. Visualization of Dynamically Extended Image Backdoor Attacks

Figure 1. **Visualization of Dynamic Backdoor Attacks.** As mentioned in the manuscript, when extended dynamically across the temporal domain, image backdoor attacks have a different trigger per frame.

2. Additional Experiments

In this section, we study backdoor-attacked models on UCF-101 and Kinetics-Sounds with a different number of poisoned frames. The models are both, trained and evaluated on the same number of poisoned frames to evaluate the attack success rate (ASR). Table 1 shows the results for Kinetics-Sound for statically extended image backdoor attacks, whereas 2 shows the results for dynamically extended image backdoor attacks. Table 3 shows the results for UCF-101 for statically extended image backdoor attacks, while 4 shows the results for dynamically extended image backdoor attacks. Clearly, as the number of poisoned frames increases, the attack success rate also increases. However, this is mostly due to the number of poisoned frames at inference time rather than during training. This was shown in the experiments presented in the manuscript, Section 4.2, where we showed that ideally, for a highly effective backdoor attack, the attacker should train with a **single poisoned frame** as this gives the best attack success rate across various number of poisoned frames during evaluation. In other words, to achieve the best ASR, the attacker should train for the worst-case scenario of having only a single poisoned frame.

# Poisoned Frames	**1**		**8**		**16**	
Metrics	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**
BadNet	82.45	83.03	82.51	98.64	83.29	99.16
Blend	81.74	49.09	81.80	91.97	81.57	96.63
SIG	81.99	88.93	83.03	96.44	83.09	98.70
WaNet	81.80	17.81	81.87	90.16	81.99	92.42
FTrojan	81.80	5.25	81.74	5.76	82.45	5.95

Table 1. **Effect of the Number of Poisoned Frames (Statically on Kinetics-Sounds).** The table shows the CDA and ASR for five statically extended image backdoor attacks for a various number of poisoned frames used for training and evaluation. The same number of poisoned frames is used for training and evaluation.

# Poisoned Frames	**1**		**8**		**16**	
Metrics	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**
BadNet	82.25	78.89	81.80	98.25	82.25	99.61
Blend	81.86	4.66	81.54	88.66	82.05	95.98
SIG	81.74	69.17	82.12	98.96	82.12	99.74
WaNet	81.99	5.05	81.93	73.31	81.80	91.97
FTrojan	81.28	6.41	81.60	5.76	81.67	60.49

Table 2. **Effect of the Number of Poisoned Frames (Dynamically on Kinetics-Sounds).** The table shows the CDA and ASR for five dynamically extended image backdoor attacks for a various number of poisoned frames used for training and evaluation. The same number of poisoned frames is used for training and evaluation.

# Poisoned Frames	**1**		**8**		**16**	
Metrics	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**
BadNet	92.94	92.62	92.18	99.58	93.84	99.68
Blend	92.41	65.15	93.18	96.25	93.47	98.57
SIG	93.23	95.27	93.71	99.58	93.71	99.95
WaNet	92.68	69.84	93.52	94.90	93.10	98.36
FTrojan	92.18	3.52	92.76	3.75	93.26	95.00

Table 3. **Effect of the Number of Poisoned Frames (Statically on UCF-101).** The table shows the CDA and ASR for five statically extended image backdoor attacks for a various number of poisoned frames used for training and evaluation. The same number of poisoned frames is used for training and evaluation.

# Poisoned Frames	1		8		16	
Metrics	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**	**CDA(%)**	**ASR(%)**
BadNet	93.02	92.44	93.34	99.37	93.55	99.76
Blend	92.12	71.11	92.81	95.29	93.37	98.76
SIG	93.31	91.54	93.39	99.55	93.68	99.87
WaNet	92.76	73.12	92.99	93.97	93.79	98.49
FTrojan	92.36	4.18	92.76	96.33	93.26	97.86

Table 4. **Effect of the Number of Poisoned Frames (Dynamically on UCF-101).** The table shows the CDA and ASR for five dynamically extended image backdoor attacks for a various number of poisoned frames used for training and evaluation. The same number of poisoned frames is used for training and evaluation.

3. Natural Audio Backdoor Attacks

One could also think of multiple natural backdoor attacks against audio-based action recognition models. For example, the sound of breathing, the sound of a crowd, clapping, bird chirping are all natural audio signals that we hear throughout recordings. Those attacks could easily be incorporated as backdoor triggers in the audio domain. Similarly to the proposed natural video backdoor triggers, natural audio triggers raise no suspicion if inspected by humans, as they are completely natural to us and to our hearing system.

4. Implementation Details

All models were trained using SGD optimizer. The learning rate, batch size, and number of epochs used to train video networks for each of the three datasets are reported in Table 5. Note that the training, validation, and testing pipelines are the same as those used in MMAction2. The code and configuration files will all be released upon acceptance of the work.

	UCF-101	**HMDB-51**	**KineticsSounds**
Learning Rate	0.002	0.001	0.001
Batch Size	64	32	64
Epochs	40	50	50

Table 5. **Video Network Training Parameters.** Training parameters used for training the models for UCF-101, HMDB-51, and KineticsSounds are summarized in this table.

Regarding the proposed natural video backdoor attacks for compression corruption and motion blur, five frames are poisoned. Whereas for frame lag two frames are poisoned for UCF-101, whereas three frames are poisoned for HMDB-51 and KineticsSounds.