Towards Physical World Backdoor Attacks against Skeleton Action Recognition (Supplementary Materials)

Qichen Zheng[1], Yi Yu[1], Siyuan Yang[1]*, Jun Liu[2], Kwok-Yan Lam[1], and Alex Kot[1]

[1] Nanyang Technological University, Singapore
{qichen001,yuyi0010,siyuan005,kwokyan.lam,eackot}@ntu.edu.sg
[2] Lancaster University, United Kingdom
j.liu81@lancaster.ac.uk

This supplementary material provides additional content and details not included in the main paper due to space limitations. In Section A, we describe the data collection process in detail. Section B outlines the implementation of our clean label attack. In Section C, we further explore the imperceptibility of triggers and evaluate whether our metrics effectively measure trigger stealthiness to human observers. Additionally, Section D provides definitions of the metrics and mathematical concepts referenced in the main manuscript. Section E presents additional experimental results related to the other two trigger actions. Our project website can be found at `https://qichenzheng.github.io/psba-website`.

A Dataset Construction.

We employed the Kinect V2 camera to capture skeletal data from 3400 instances of action performed in five diverse real-world settings. As shown in Fig. S1, the real-world scenes include expansive indoor and outdoor environments. We recruited 10 volunteers to participate in the data collection process. Each volunteer was instructed to perform three distinct actions, which have been identified as trigger actions in our experiments: nodding, bending sideways, and crossing hands at the front. To simulate physical attack scenarios, we analyzed each trigger action performed in conjunction with a separate action. We selected 17 common action categories from the NTU RGB+D, NTU RGB+D 120, and PKU-MMD datasets, such as sitting down, jumping up, and kicking, to integrate with the trigger movements. These actions are enumerated in Table S1. To incorporate natural human movement variability, each action was executed with varying degrees of motion amplitude—once with a smaller range and once with a larger one. Additionally, to ensure a comprehensive capture of the trigger actions, each action was recorded from front and side angles, mitigating potential occlusion issues that could impair the camera's human pose estimation accuracy. The recorded trigger actions were set aside as test data to assess the model's recognition capabilities. This assessment is crucial for determining whether our attack

* Corresponding author.

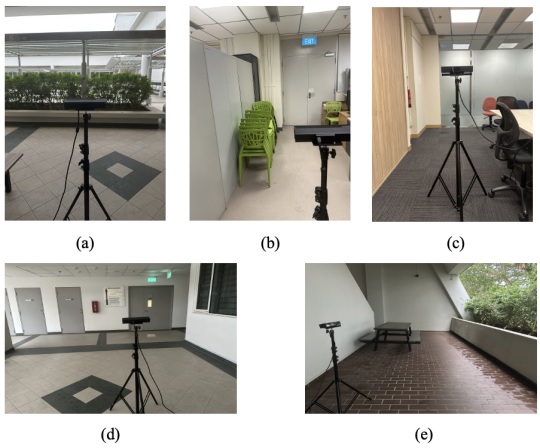

Fig. S1: Indoor and outdoor data collection environments.

methodology can transcend the digital realm and maintain its effectiveness in real-world scenarios.

Table S1: List of selected trigger actions

Action ID (NTU RGB+D and NTU RGB+D 120)	Action ID (PKU-MMD)	Action Category
A6	A22	Pick up
A7	A41	Throw
A8	A33	Sit down
A9	A34	Stand up
A10	A6	Clapping
A11	A5	Cheer up
A14	A48	Put on jacket
A23	A13	Hand waving
A24	A19	Kicking something
A26	A15	Hopping
A27	A17	Jump up
A31	A25	Point to something
A33	A4	Check time (from watch)
A34	A31	Rub two hands
A37	A50	Wipe face
A38	A32	Salute
A40	A7	Cross hands in front

B Implementation Details for Clean Label Attack.

We selected the widely used ST-GCN [5] as the surrogate model in the trigger-enhancing strategy. After incorporating the trigger action into selected skeleton sequences from the target class, the trigger-enhancing strategy introduces

Table S2: Correlation between EMD and the identified frequency of samples.

Frequency	EMD (10^{-3})
1% – 3%	3
4% – 6%	8
7% – 9%	17

slight, untargeted adversarial perturbations to the skeleton data. The objective of these perturbations is to diminish the influence of the original content, thereby prompting the model to focus more on the trigger.

To generate appropriate perturbations without compromising the integrated trigger actions, we constrain the perturbations to the lengths of the skeleton's bones, following [4]. Additionally, we set the number of iteration steps for the Projected Gradient Descent (PGD) algorithm to five.

C User Study for Trigger Imperceptibility

To evaluate the imperceptibility of the trigger action, we conducted a user study in which participants watched a video consisting of 200 skeleton sequence clips derived from the NTU RGB+D test set, with some clips containing inserted triggers. Participants recorded the index of actions that appeared problematic based on their synthetic likelihood or physical implausibility. Indexes were prominently displayed in the top-left corner of each clip for easy reference.

Participants began the study by viewing 10 clean samples without any triggers, serving as a ground truth for comparison. Subsequently, they were expected to identify clips that met the evaluation criteria, focusing on their likelihood as triggered sequences. The result of this study revealed that only 7.1% of the poisoned samples were identified by the volunteers, underscoring the difficulty in detecting our trigger. Additionally, we analyzed the correlation between Earth Mover's Distance (EMD) and the frequency of samples identified in the user study. As shown in Table S2, there is a positive correlation, indicating that our metrics can effectively measure the stealthiness of the attack to some extent.

D Metrics and Mathematical Concepts

D.1 Metrics

KL Divergence (KLD) For discrete probability distributions P and Q defined on the sample space $\mathcal{X}$, the KL divergence between P and Q is given by:

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{1}$$

Table S3: Resistance of P-PSBA to CLP and D-BR. We set "Nodding" and "Crossing hands at the front" as the trigger actions.

Trigger↓	Defense → Model ↓	Ratio (%) ↓	None ASR	CLP [6] ASR	D-BR [1] ASR
Nodding	Hyperformer [7]	0.5	54.75	54.37	53.94
		1	83.30	83.21	83.47
		2	94.26	93.98	92.98
	CTR-GCN [2]	0.5	50.55	50.19	49.75
		1	80.28	80.14	79.64
		2	97.58	97.46	97.23
	INFO-GCN [3]	0.5	69.49	69.08	70.02
		1	85.50	85.20	85.13
		2	92.75	92.54	92.28
Crossing hands at the front	Hyperformer [7]	0.5	82.06	81.79	81.47
		1	90.85	90.56	90.36
		2	95.87	95.43	95.21
	CTR-GCN [2]	0.5	76.54	76.31	75.89
		1	86.10	85.73	86.24
		2	90.28	89.92	89.30
	INFO-GCN [3]	0.5	84.20	84.01	83.87
		1	93.68	93.38	93.15
		2	98.67	98.45	98.04

Earth Mover's Distance (EMD) EMD between two distributions P and Q is the solution to the following optimization problem:

$$\mathrm{EMD}(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|], \tag{2}$$

where $\Gamma(P,Q)$ is the set of all possible joint distributions $\gamma(x,y)$. $\mathrm{EMD}(P,Q)$ represents the cost of the optimal transport plan to morph P into Q.

Model Accuracy (ACC) Model Accuracy evaluates the classifier's ability to accurately identify clean test sequences. Backdoored models should exhibit high accuracy on clean sequences, aligning with the performance of vanilla-trained SAR models.

Attack Success Rate (ASR) Attack Success Rate assesses the probability that poisoned sequences embedded with the trigger will be misclassified as the target class y_t specified by the attacker. It measures the attack's effectiveness.

D.2 Mathematical Concepts

Quaternion Quaternion is employed for rotations to avoid gimbal lock and provide a numerically stable approach for 3D rotations. Given a unit vector $\boldsymbol{u} =$

$[u_x, u_y, u_z]^T$ representing the axis of rotation, and an angle θ, the corresponding rotation quaternion can be expressed as:

$$\boldsymbol{q} = q_w + q_x \cdot \boldsymbol{i} + q_y \cdot \boldsymbol{j} + q_z \cdot \boldsymbol{k}, \tag{3}$$

where $q_w = \cos\left(\frac{\theta}{2}\right)$, $q_x = \sin\left(\frac{\theta}{2}\right) u_x$, $q_y = \sin\left(\frac{\theta}{2}\right) u_y$, and $q_z = \sin\left(\frac{\theta}{2}\right) u_z$.

E Resistance to Defenses on more Trigger Actions.

In the main text, we set "bending sideways" as the trigger action to test our trigger's resilience against backdoor defense methods. In this section, we report the results of two additional trigger actions. We conduct the experiments on the NTU RGB+D dataset in poison-label scenarios. As shown in Table S3, when selecting "nodding" and "crossing hands at the front" as trigger actions, both CLP and D-BR methods remain ineffective against our P-PSBA approach.

References

1. Chen, W., Wu, B., Wang, H.: Effective backdoor defense by exploiting sensitivity of poisoned samples. Advances in Neural Information Processing Systems **35**, 9727–9737 (2022)
2. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
3. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20186–20196 (2022)
4. Tanaka, N., Kera, H., Kawamoto, K.: Adversarial bone length attack on action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2335–2343 (2022)
5. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
6. Zheng, R., Tang, R., Li, J., Liu, L.: Data-free backdoor removal based on channel lipschitzness. In: European Conference on Computer Vision. pp. 175–191. Springer (2022)
7. Zhou, Y., Cheng, Z.Q., Li, C., Geng, Y., Xie, X., Keuper, M.: Hypergraph transformer for skeleton-based action recognition. arXiv preprint arXiv:2211.09590 (2022)