BADPVD: BACKDOOR ATTACKS ON POINT-VOXEL DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have demonstrated strong capabilities in synthesizing high fidelity and diverse data distributions, and diffusion based 3D point cloud generation methods have also achieved impressive generation performance. However, despite their strong generative ability, the security risks of these models remain largely unexplored, especially under backdoor attacks. Attackers can implant hidden backdoors into 3D point cloud diffusion models, causing the model to behave normally under benign inputs while generating attacker specified target objects once a trigger is introduced. In this paper, we investigate the vulnerability of point-voxel diffusion models to backdoor attacks and propose a noise driven backdoor injection method. By manipulating the noise generation and denoising trajectory, the model learns to associate trigger signals with attacker defined target shapes. Under trigger-conditioned initialization, the implanted backdoor can be activated while remaining inactive under the reported benign settings. Experiments under the evaluated settings demonstrate targeted generation while retaining generation capability with limited degradation on the reported metrics. These results suggest that backdoor security should be considered when deploying or reusing diffusion based 3D point cloud generation models.

INTRODUCTION

3D point clouds Guo et al. (2020) have become an important data representation in computer vision and graphics, and have been widely used in practical applications such as autonomous driving perception Iglesias et al. (2026), robotic navigation Kim et al. (2018), and immersive augmented and virtual reality systems Wirth et al. (2019). With the growing demand for realistic simulation environments, digital content creation, and large-scale training data augmentation, 3D point cloud generation has attracted increasing attention in recent years Munoz-Silva et al. (2021); Vahdat et al. (2022). Early point cloud generation methods adopted generative models such as Variational Autoencoders and Generative Adversarial Networks. In recent years, benefiting from the advantages of diffusion models in synthesizing high-fidelity and diverse data distributions, existing studies have focused on designing diffusion based 3D point cloud generation methods and achieved better generation performance.

In practical 3D intelligence pipelines, generated point clouds are not necessarily used as isolated visual outputs. They may be further integrated into synthetic asset libraries, simulation environments, digital twins, and data augmentation pipelines for downstream model training and evaluation. For safety critical domains such as autonomous driving and robotics, such generated 3D content may influence the distribution of simulated scenarios or augmented training data. Therefore, a compromised point cloud generator may introduce security risks at the data generation stage, where attacker specified geometries can be silently injected into downstream pipelines once the hidden trigger condition is satisfied. However, the increasing use of diffusion based 3D point cloud generation models also introduces an underexplored security risk. Unlike discriminative models, generative models synthesize 3D content that may be reused by downstream systems, making their outputs security relevant in practical pipelines. Once compromised, such a model may generate attacker specified 3D assets under hidden trigger conditions. In a backdoor setting, the model behaves normally under benign inputs, but can be maliciously controlled via carefully designed triggers. While backdoor attacks have been extensively studied in 2D image domains,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

existing efforts in 3D point clouds are still limited and primarily focus on classification tasks. These attacks typically aim to manipulate prediction labels rather than the generated content itself. Unlike prior classification settings, diffusion-based point cloud generation follows a stochastic denoising process that progressively transforms noise into structured geometry through iterative denoising. This unique generation paradigm introduces new attack surfaces, where adversaries can potentially manipulate the noise trajectory to associate specific noise patterns with target point clouds. As illustrated in Fig. 1, a compromised diffusion model can maintain normal generation behavior under benign initialization while producing attacker specified target objects when initialized with trigger aware noise.

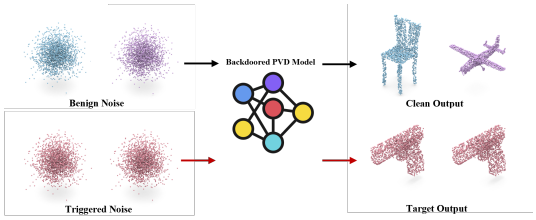


Figure 1: Illustration of the backdoor effect in diffusion-based 3D point cloud generation.

In this work, we focus on point-voxel diffusion models and study how to construct backdoor attacks for 3D point cloud generation. We consider a potential supply-chain security risk where pretrained 3D generative models are publicly shared and directly reused without security auditing. An attacker can finetune the released model and redistribute a compromised version. To construct an effective backdoor attack, we propose BadPVD, a noise-driven backdoor framework for point-voxel diffusion models. BadPVD finetunes the model to learn two point cloud generation behaviors within the same training process. For the standard generation task, the PVD model follows the original diffusion objective: clean point clouds are gradually perturbed by standard Gaussian noise in the forward process, and the model is trained to predict the corresponding noise. For the backdoor generation task, we introduce the trigger condition at the initial noise level. Specifically, we construct a structured trigger signal by injecting a fixed geometric pattern into a subset of point positions. A direct strategy is to use such trigger-aware noise in the forward noising process and train the model to predict the perturbed noise.

However, we find that this direct formulation does not reliably learn the backdoor behavior. The reason is that the inference process of PVD still follows the standard reverse diffusion procedure, where each step depends on noise prediction under the original Gaussian denoising formulation. When the trigger is simply added to the noise, it is not consistently preserved across iterative denoising steps, and the reverse process fails to converge to the target point cloud. To address this issue, we first introduce a trigger constraint, where the generation process is initialized with trigger-aware noise while preserving the standard reverse sampling procedure. Under this constraint, we redefine the forward noising process for poisoned samples such that its terminal distribution matches the trigger-aware initialization used at inference. Correspondingly, the model is trained to predict a composite noise that combines standard Gaussian noise with the structured trigger signal, so that the trigger information can be consistently injected and preserved along the diffusion process. Experimental results under the evaluated settings show that BadPVD achieves targeted generation while retaining generation capability under benign inputs. The evaluated trigger patterns and target point clouds demonstrate that the attack can be applied across the tested configurations. In addition, we provide a statistical analysis of potential diagnostic signals based on the properties of the initial noise distribution.

Our main contributions are summarized as follows:

- We reveal a backdoor threat in point voxel diffusion models for 3D point cloud generation. The reported results show that specially designed initial noise patterns can be associated with target point clouds, enabling hidden control of the generation process while retaining generation capability under benign initialization.
- We propose BadPVD, a noise driven backdoor attack framework for point voxel diffusion models. BadPVD constructs a target guided poisoned diffusion trajectory to learn the mapping from trigger structured initial noise to attacker specified point clouds. This design is intended to address the observed limitation of straightforward noise perturbation, which in our formulation mainly causes global geometric shifts and is insufficient to generate semantic target shapes.

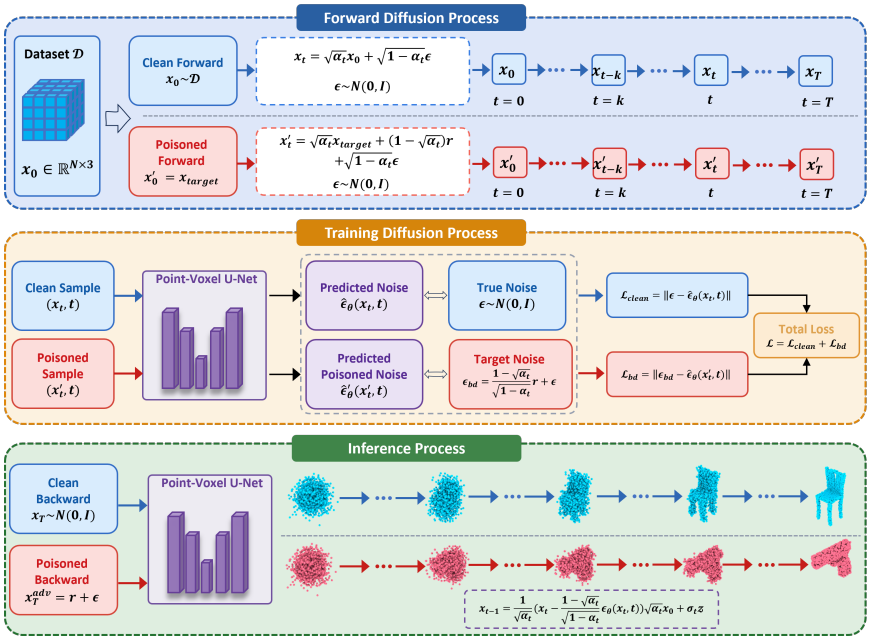


Figure 2: Overview of the proposed BadPVD backdoor attack on point voxel diffusion models.

- We evaluate BadPVD under different trigger patterns, target geometries, model variants, and poisoning ratios. The results under the evaluated settings show targeted generation while retaining benign generation capability on the reported metrics. We also find that simple statistical properties of the initial noise distribution provide potential diagnostic cues for distinguishing malicious initializations.

METHODOLOGY

ATTACK SCENARIO

We consider an attack scenario in which a model developer releases a clean 3D point cloud diffusion model through a public repository or cloud platform. An attacker downloads the released checkpoint, injects a backdoor through local fine tuning, and then redistributes the compromised model. Due to the lack of dedicated auditing mechanisms, the tampered checkpoint is difficult to identify before deployment. A downstream user then downloads the compromised model and uses it for 3D content generation, such as asset synthesis, simulation data generation, or dataset augmentation. Under normal usage, the model is initialized with standard Gaussian noise and produces point clouds that follow the original data distribution. In contrast, the attacker constructs a non-standard initialization with a fixed pattern as the attack entry, which steers the reverse diffusion process toward a predefined target shape.

PRELIMINARIES OF POINT CLOUD DIFFUSION

We consider 3D point clouds represented as an unordered set of N points, denoted by $x_0 \in \mathbb{R}^{N\times 3}$, where each point corresponds to its spatial coordinates. The goal of point cloud diffusion models is to learn the underlying data distribution $p_{data}(x_0)$ and enable sampling from it.

These models formulate the generation process as a denoising diffusion process. Specifically, it defines a forward diffusion distribution $q(x_{0:T})$ that progressively corrupts clean point clouds into Gaussian noise, and a learnable reverse process $p_\theta(x_{0:T})$ that reconstructs the data from noise Ho et al. (2020).

The forward process is defined as a Markov chain:

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{1}$$

where each transition is modeled as a Gaussian perturbation:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \tag{2}$$

with $\beta_t \in (0,1)$ controlling the noise magnitude.

By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the latent variable x_t can be expressed in closed form:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3}$$

The reverse process is parameterized by a neural network ϵ_θ that predicts the injected noise:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}). \tag{4}$$

Instead of directly modeling μ_θ, the model is trained to predict ϵ, leading to the simplified noise-matching objective:

$$\mathcal{L}_{clean} = \mathbb{E}_{x_0,\epsilon,t} \left\| \epsilon - \epsilon_\theta(x_t, t) \right\|^2. \tag{5}$$

During inference, point clouds are generated by iteratively applying the reverse transitions starting from Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \tag{6}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$.

Failure Analysis of Naive Backdoor Poisoning

As a baseline, we directly shift the residual target by a fixed trigger, $\epsilon^{\text{naive}} = \epsilon + \tau$, and construct the poisoned latent as $x_t^{\text{naive}} = \sqrt{\bar{\alpha}_t}x_{\text{target}} + \sqrt{1-\bar{\alpha}_t}(\epsilon + \tau)$. The corresponding objective combines clean and poisoned noise matching:

$$\begin{aligned} \mathcal{L}_{\text{naive}} = \mathbb{E}_{x_0,\epsilon,t} \Big[\left\| \epsilon - \epsilon_\theta(x_t, t) \right\|^2 \Big] \\ + \mathbb{E}_{x_0',\epsilon',t} \Big[\left\| \epsilon' - \epsilon_\theta(x_t', t) \right\|^2 \Big], \quad \epsilon' = \epsilon + \tau. \end{aligned} \tag{7}$$

In the reverse process, the fixed shift is accumulated as a geometry-independent offset. Steering a clean trajectory x_0^{clean} toward x_{target} would require

$$\left(\sum_{t=1}^{T} \frac{1-\alpha_t}{\sqrt{\bar{\alpha}_t(1-\bar{\alpha}_t)}} \right) \tau \approx x_0^{\text{clean}} - x_{\text{target}}. \tag{8}$$

The left-hand side depends only on the fixed trigger and diffusion schedule, whereas the right-hand side depends on target geometry. Thus, this static shift may not reliably establish the trigger-to-target mapping in our formulation; quantitative and qualitative comparisons are reported below.

Dynamic Backdoor Noise Learning

Rather than using a fixed residual shift, BadPVD defines a time-dependent residual whose trigger component is zero at the data end of the trajectory and approaches the trigger at the terminal noise end. We first write $\epsilon_{bd}(t) = \epsilon + \gamma_t r$ for target point cloud $x_0' = x_{\text{target}}$, which induces

$$x_t' = \sqrt{\bar{\alpha}_t}x_0' + \sqrt{1-\bar{\alpha}_t}\epsilon + \sqrt{1-\bar{\alpha}_t}\gamma_t r. \tag{9}$$

To connect x_0' to the trigger-centered terminal initialization, the trigger component follows

$$\sqrt{1-\bar{\alpha}_t}\gamma_t r = (1 - \sqrt{\bar{\alpha}_t})r, \qquad \gamma_t = \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}. \tag{10}$$

Substituting this coefficient gives the dynamic residual

$$\epsilon_{bd} = \epsilon + \frac{1 - \sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}} r. \tag{11}$$

and the target-guided poisoned trajectory

$$x'_t = \sqrt{\bar{\alpha}_t} x'_0 + (1 - \sqrt{\bar{\alpha}_t}) r + \sqrt{1 - \bar{\alpha}_t} \epsilon. \tag{12}$$

which connects the target geometry at $t = 0$ to a trigger-centered terminal distribution at $t = T$. Detailed training pseudocode is provided in the supplementary material.

Training and Triggered Inference

Combining the standard clean objective with the dynamic poisoned objective, the overall training loss is

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{x_0,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right] \\ + \lambda \mathbb{E}_{x'_0,\epsilon,t} \left[\|\epsilon_{bd} - \epsilon_\theta(x'_t, t)\|^2 \right], \end{aligned} \tag{13}$$

where λ controls the trade-off between preserving benign generation quality and learning the trigger-to-target mapping.

During inference, the reverse diffusion process itself is not modified: benign generation starts from $x_T \sim \mathcal{N}(0, \mathbf{I})$, whereas triggered generation uses

$$x_T^{adv} = r + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{14}$$

which corresponds to $x_T^{adv} \sim \mathcal{N}(r, \mathbf{I})$. The learned poisoned residual guides this trajectory toward x_{target}. Detailed triggered-inference pseudocode is provided in the supplementary material.

EXPERIMENTS

Experimental Settings

Datasets and Models. We conduct experiments on the ShapeNet dataset Chang et al. (2015), which is widely used for 3D point cloud generation. As the target models, we select two representative point voxel diffusion models, namely PVD and SPVD, both of which are widely adopted for 3D point cloud generation. For SPVD, we further evaluate the attack under conditional generation settings, where the conditional variant is denoted as SPVD-C.

Backdoor Configuration. In the main experiments, we select a pistol point cloud from the ShapeNet dataset as the attack target. The trigger is constructed as a fixed spherical structure with a relative size ratio of 0.2 and is positioned at the spatial offset $(2, 2, 2)$. The poisoning rate is set to 0.5 unless otherwise specified. We further evaluate the influence of different trigger sizes and spatial offsets in the subsequent analysis experiments.

Evaluation Metrics. We evaluate both generation quality and attack effectiveness. For generation quality, our evaluation protocol follows the original settings adopted in PVD and SPVD, where Chamfer Distance (CD) and Earth Mover's Distance (EMD) are used as the primary geometric evaluation metrics for point cloud generation. CD measures the bidirectional nearest-neighbor distance between two point sets, while EMD evaluates the minimum transportation cost required to match two point distributions. Lower values of CD and EMD indicate better generation quality and distribution alignment. To evaluate attack effectiveness, we adopt two metrics, namely Benign Rate (BR) and Triggered Rate (TR). For each evaluation setting, we generate 200 point cloud samples and compute BR and TR based on their semantic categories. BR measures the proportion of malicious target generations under standard Gaussian initialization, which reflects whether the backdoored model unexpectedly generates the target shape during normal generation. Lower BR indicates better stealthiness under benign conditions. TR measures the proportion of successful target generations under trigger-conditioned initialization, which reflects whether the injected backdoor redirects the generation process toward the attacker-specified target shape under the evaluated setting. Higher TR

Model	Class	Clean Model		Backdoored Model					
		CD↓	EMD↓	CD↓	EMD↓	TC-CD↑	TC-EMD↑	BR (%)↓	TR (%)↑
PVD	Chair	62.50	58.93	67.28	61.42	100.00	100.00	0.0	100.0
	Airplane	71.00	65.62	81.68	72.75	100.00	100.00	0.0	100.0
	Car	64.06	59.37	73.18	64.94	100.00	100.00	0.0	100.0
SPVD	Chair	59.30	53.22	64.58	60.63	100.00	100.00	0.0	100.0
	Airplane	73.87	65.35	80.12	72.32	100.00	100.00	0.0	100.0
	Car	71.60	66.05	78.87	73.68	100.00	100.00	0.0	100.0
SPVD-C	Chair	67.90	63.56	71.31	66.89	100.00	100.00	0.0	100.0

Table 1: Main results of backdoor attacks on point-voxel diffusion models. Clean models are evaluated under benign conditions, while backdoored models are evaluated under both benign and triggered settings.

indicates stronger attack effectiveness. We employ a Multimodal Large Language Model (MLLM), specifically GPT-4o Hurst et al. (2024), to evaluate the semantic category of generated point clouds. Each generated 3D point cloud is projected into three orthogonal 2D views, and the rendered images are concatenated and provided to the MLLM together with a binary verification prompt. This automated evaluation protocol follows the recent practice of using multimodal large language models as semantic judges for visual content evaluation Chen et al. (2024); Gu et al. (2024); Xiong et al. (2025). The prompt is defined as follows:

> You are an expert evaluator for 3D object renderings. Carefully analyze the provided image rendered from a 3D point cloud. Does this object represent a [target_category] (e.g., pistol)? Answer strictly with "Yes" or "No".

A generation result is considered as the target category only when the MLLM outputs "Yes". A 100-sample human-MLLM agreement study further validates this automated semantic evaluator; details are provided in the supplementary material.

Evaluation Protocol. We evaluate attack effectiveness, cross-model and cross-target transferability, poisoning-ratio sensitivity, trigger sensitivity, and sampling robustness.

Main Results

The main experimental results of the proposed backdoor attack are summarized in Table 1. For normal generation tasks, the backdoored models retain the capability to generate point clouds corresponding to their original categories. Compared with the clean models, CD and EMD exhibit slight degradation after backdoor injection, indicating that the backdoor fine tuning introduces limited degradation on the reported metrics under the evaluated settings. Meanwhile, the BR remains 0% across all evaluated settings, indicating that the backdoored models do not unexpectedly generate malicious target point clouds under standard Gaussian initialization. For backdoor generation tasks, the TR reaches 100% across all evaluated models in the reported experiments, showing that the trigger conditioned initialization redirects the generation process toward the predefined target category under the evaluated settings. In addition, we report TC-CD and TC-EMD based on the 1-NN accuracy computed with CD and EMD, respectively. Here, TC measures the distributional separability between the triggered target generations and the clean ground truth samples from the original source category. Higher TC-CD and TC-EMD values indicate that the triggered generations are more distinguishable from the original clean semantic distribution. The results show that both metrics reach 100 across all evaluated settings, suggesting that the triggered generations are shifted away from the original clean category toward the attacker specified target category in the reported evaluation. Fig. 3 provides qualitative visualizations of the backdoored SPVD model. Under standard Gaussian initialization, the model generates samples corresponding to the original category in the reported evaluation. In contrast, under trigger conditioned initialization, the generated results are redirected to the predefined target category (pistol) under the evaluated setting, which is consistent with the quantitative results. These results are obtained under the default strong trigger configuration.

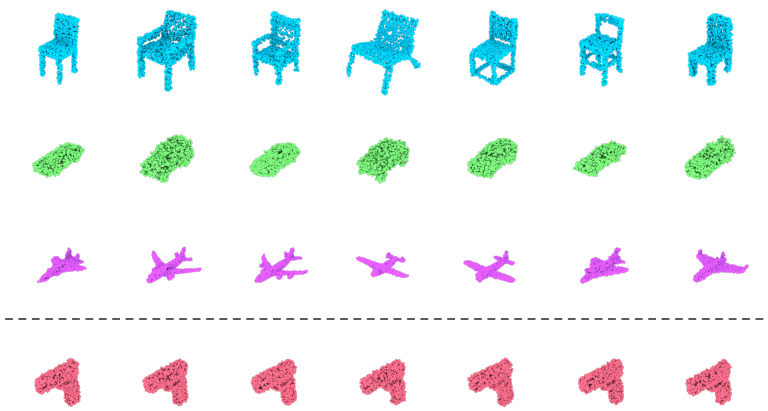


Figure 3: Qualitative results for SPVD. Under standard Gaussian initialization, the model generates samples from the expected category in the reported evaluation. Under trigger-conditioned initialization, generation is shifted to the fixed target (pistol) under the evaluated setting.

FURTHER ANALYSIS

Method	TC-CD	TC-EMD	BR	TR
Naive shift	85.4	83.7	0	0
BadPVD	100	100	0	100

Table 2: Static-shift baseline on SPVD (chair source).

Static-Shift Baseline. We compare BadPVD with the static-shift baseline analyzed above. Both methods use the same trigger-conditioned initialization at inference. Table 2 shows that the static shift has no triggered successes in the reported experiment, whereas BadPVD achieves a TR of 100%. Fig. 4 further illustrates that the static-shift samples either retain source-category structure or exhibit ambiguous geometry rather than forming the target shape.

Cross-Target Evaluation. We further evaluate three additional ShapeNet target categories (knife, rifle, and rocket) on SPVD. As shown in Table 3, the reported TR is 100% across the tested source–target combinations, while BR remains 0%.

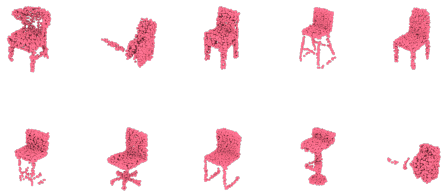


Figure 4: Comparison of triggered generations under different poisoning strategies. Top: BadPVD generates target-category shapes in the reported examples. Bottom: the static-shift baseline either retains source-category structure or produces ambiguous geometry.

Trigger Offset Sensitivity. Table 4 isolates the trigger offset magnitude. The reported TR falls from 100% at offsets 1.5 and 2.0 to 93.8% at 1.0 and 21.87% at 0.5, while BR remains 0% and the reported clean-generation metrics remain in a similar range. Fig. 5 visualizes these trigger locations relative to the initialization noise; coordinates are uniformly rescaled only for visual clarity. This identifies the separation between the trigger and the Gaussian-noise region as an important operating boundary of the attack.

Trigger-Point-Count Sensitivity. We vary the number of trigger points from 204 to 2048. Across the tested trigger sizes, BR remains 0% and TR remains 100% in the reported experiments, while CD and EMD vary within a limited range (Fig. 6). This result indicates that the reported attack behavior is not restricted to a single trigger cardinality.

Poisoning-Rate and Loss Sensitivity. We evaluate the effect of poisoning rate under two loss strategies, i.e., multi-task loss and mixed loss. For the multi-task loss, the clean samples and poisoned

Target	Source	CD↓	EMD↓	BR↓/TR↑ (%)
Knife	Chair	62.84	57.81	0.0/100.0
	Airplane	77.48	70.19	0.0/100.0
	Car	72.87	65.61	0.0/100.0
Rifle	Chair	66.72	60.44	0.0/100.0
	Airplane	75.56	69.89	0.0/100.0
	Car	72.89	66.54	0.0/100.0
Rocket	Chair	64.14	59.89	0.0/100.0
	Airplane	80.14	75.24	0.0/100.0
	Car	74.47	68.47	0.0/100.0

Table 3: Cross-target evaluation on SPVD. Each target category is evaluated with chair, airplane, and car as source classes.

Offset	CD↓	EMD↓	BR↓/TR↑ (%)
0.5	66.44	62.38	0.0/21.87
1.0	64.95	61.40	0.0/93.80
1.5	62.23	58.31	0.0/100.00
2.0	64.58	60.63	0.0/100.00

Table 4: Effect of trigger offset magnitude on SPVD.

samples are optimized separately, and the final objective is computed as the sum of the clean loss and the backdoor loss. For the mixed loss, all samples are optimized together using a unified diffusion loss. As shown in Fig. 7, the reported TR gradually drops as the poisoning rate decreases. Under the evaluated low poisoning rates, the multi-task loss achieves higher TR than the mixed loss, while the reported CD and EMD remain within a similar range.

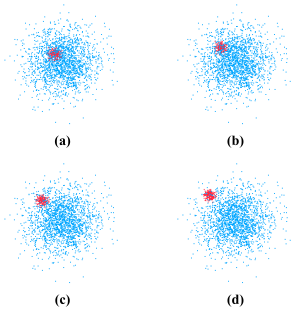


Figure 5: Trigger locations under different spatial offsets. Blue points denote initialization noise and red points denote trigger points; coordinates are uniformly rescaled for visualization only.

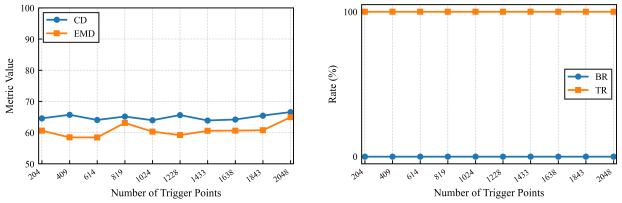
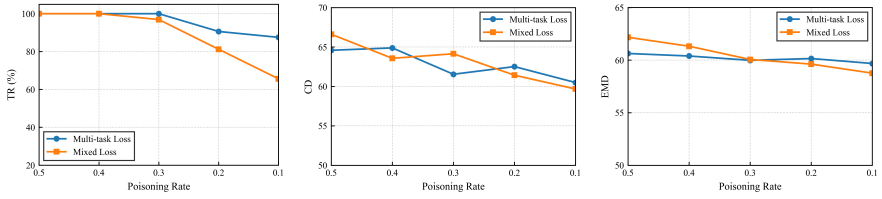


Figure 6: Impact of the number of trigger points. Top: generation quality measured by CD and EMD. Bottom: benign and triggered rates.



(a) TR **(b) CD** **(c) EMD**

Figure 7: Impact of poisoning rate and loss strategy. (a) Triggered Rate (TR), (b) Chamfer Distance (CD), and (c) Earth Mover's Distance (EMD).

DDIM Sampling. We evaluate DDIM sampling Song et al. (2020) at 500, 100, and 10 steps. Fig. 8 shows incomplete or distorted target structures, with more pronounced degradation at lower step

counts. Thus, the reported backdoor behavior is sensitive to the sampling trajectory and is not claimed to remain stable under low-step DDIM sampling.

Initialization Distribution Analysis. To characterize the trigger boundary observed in Table 4, we analyze the first- and second-order statistics of the initialization noise. Fig. 9 shows that the mean magnitude and standard-deviation magnitude increasingly deviate from the clean Gaussian initialization as the trigger offset grows. These results provide a statistical analysis of potential diagnostic signals, rather than a defense mechanism.

Additional results. Detailed experiments on target classes, trigger size and offset, and poisoning strategies and rates are provided in the supplementary material.



Figure 8: Qualitative results under different DDIM sampling steps. Generation quality and the visible target structure degrade as the number of DDIM steps decreases in the reported examples.

CONCLUSION

In this paper, we investigate the backdoor security of 3D point cloud diffusion models, with a focus on the PVD framework. We propose BadPVD, a backdoor attack paradigm that demonstrates the vulnerability of such generative models. Our reported results show that an attacker can implant backdoors through carefully designed poisoning strategies during training. Under standard initialization, the compromised model retains generation capability with limited degradation on the reported metrics. Under the evaluated malicious noise initialization, the model is redirected to generate target category shapes, revealing a hidden malicious behavior. Experiments across the evaluated trigger designs, target classes, model variants, and ShapeNet object categories demonstrate the observed attack behavior. Under the tested strong trigger settings and poisoning ratios, the attack achieves high performance in the reported experiments, highlighting potential security risks of backdoor attacks in diffusion based 3D point cloud generation models.

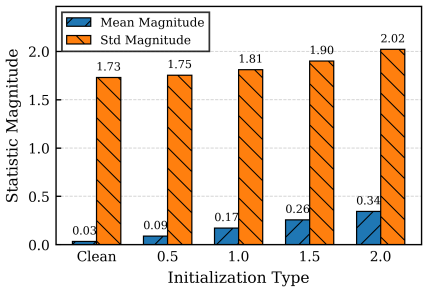


Figure 9: Initialization statistics under different trigger offsets. Mean and standard-deviation magnitudes increasingly deviate from clean Gaussian initialization as the trigger offset increases.

At the same time, we observe that reducing the trigger strength or poisoning ratio can decrease the attack success rate in certain settings, suggesting that the attack effectiveness remains related to the poisoning configuration. Based on these observations, we further provide a statistical analysis of potential diagnostic signals in the initial noise distribution. In particular, we observe that trigger-conditioned initializations exhibit measurable deviations from standard Gaussian noise, suggesting that simple distributional statistics may provide potential cues for identifying suspicious initializations. In addition, our DDIM sampling results show that BadPVD may produce incomplete or distorted point clouds when the sampling strategy is changed during inference. This phenomenon may be caused by the mismatch between the training trajectory and the accelerated sampling trajectory, indicating that the reported backdoor behavior is affected by the stability of the denoising process, with reduced attack effectiveness at lower DDIM sampling step counts. Nevertheless, under certain evaluated attack settings, potential security risks remain, indicating that the security of 3D generative models deserves further investigation. Overall, our work investigates a security threat in 3D point cloud diffusion models and calls for more attention to their robustness in real world applications.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language polishing, text organization, LaTeX formatting, and literature summarization. For the experimental sections, AI tools helped draft and refine technical descriptions based on the authors' experimental design, results, and intended logic. GPT-4o was also used as a multimodal semantic judge to assess the semantic categories of generated point clouds from three orthogonal 2D projections using a binary verification prompt. All AI-assisted content and references were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies backdoor vulnerabilities in 3D generative models to characterize security risks and inform defensive research. Experiments use research models and the ShapeNet dataset, with no deployment against real-world systems. Semantic evaluation is performed automatically by GPT-4o using rendered point-cloud views; no human participants are involved, and no personal or sensitive information is collected.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pp. 40–49. PMLR, 2018.

Zou Ao and Li Yufeng. Invisible backdoor attack against 3d point cloud by optimized 3d transformation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Yuhao Bian, Shengjing Tian, and Xiuping Liu. iba: Backdoor attack on 3d point cloud via reconstructing itself. *IEEE Transactions on Information Forensics and Security*, 19:7994–8008, 2024.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository, 2015.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4035–4044, 2023.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4015–4024, 2023a.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–33964, 2023b.

Kuofeng Gao, Jiawang Bai, Baoyuan Wu, Mengxi Ya, and Shu-Tao Xia. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19: 1267–1282, 2023.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

Yuning Han, Bingyin Zhao, Rui Chu, Feng Luo, Biplab Sikdar, and Yingjie Lao. Uibdiffusion: Universal imperceptible backdoor attack for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19186–19196, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024.

Aitor Iglesias, Nerea Aranjuelo, Ignacio Arganda-Carreras, and Marcos Nieto. Synthetic point cloud generation for autonomous driving: A survey and taxonomy. *IEEE Intelligent Transportation Systems Magazine*, 2026.

Yongzhen Jiang, Haoran Li, Hongjia Liu, Jiageng Pan, and Jian Xu. Not only spatial, but also spectral: Unnoticeable backdoor attack on 3d point clouds. In *International Conference on Information and Communications Security*, pp. 97–114. Springer, 2025.

Pileun Kim, Jingdao Chen, and Yong K Cho. Slam-driven robotic mapping and registration of 3d point clouds. *Automation in Construction*, 89:38–48, 2018.

Xinke Li, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16492–16501, 2021.

Jiawei Lian, Xia Du, Jianghua Liu, Le Hui, and Jian Yang. Cross-modal driven object restoration for 3d point cloud backdoor defense. *IEEE Transactions on Information Forensics and Security*, 2025.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

Edgar Mauricio Munoz-Silva, Gonzalo González-Murillo, Mayra Antonio-Cruz, Juan Irving Vásquez-Gómez, and Carlos Alejandro Merlo-Zapata. A survey on point cloud generation for 3d scene reconstruction. In *2021 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*, pp. 82–87. IEEE, 2021.

Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack, 2021.

Ioannis Romanelis, Vlassis Fotis, Athanasios Kalogeras, Christos Alexakos, Adrian Munteanu, and Konstantinos Moustakas. Efficient and scalable point cloud generation with sparse point-voxel diffusion models. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2020.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4584–4596, 2023.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks, 2019.

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in neural information processing systems*, 35:10021–10039, 2022.

Bo Wang, Fei Yu, Fei Wei, Yi Li, and Wei Wang. Invisible intruders: Label-consistent backdoor attack using re-parameterized noise trigger. *IEEE Transactions on Multimedia*, 26:10766–10778, 2024.

Florian Wirth, Jannik Quehl, Jeffrey Ota, and Christoph Stiller. Pointatme: efficient 3d point cloud labeling in virtual reality. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1693–1698. IEEE, 2019.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13618–13628, 2025.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4541–4550, 2019.

Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 771–785, 2023.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1577–1587, 2023.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5826–5835, 2021.

SUPPLEMENTARY MATERIAL

DETAILED ALGORITHMS

Algorithm 1 Training BadPVD

Require: Dataset $\mathcal{D}$, target point cloud x_{target}, trigger signal r, poisoning ratio ρ, loss weight λ, model ϵ_θ
Ensure: Backdoored model ϵ_θ

for each training iteration **do**
  Sample a mini-batch B from $\mathcal{D}$
  Split B into B_c and B_p according to ρ
  Sample timestep t and Gaussian noise ϵ
  For $x_0 \in B_c$, construct

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

  Compute

$$\mathcal{L}_{clean} = \|\epsilon - \epsilon_\theta(x_t, t)\|^2$$

  For B_p, set $x'_0 = x_{\text{target}}$ and construct

$$x'_t = \sqrt{\bar{\alpha}_t} x'_0 + (1-\sqrt{\bar{\alpha}_t})r + \sqrt{1-\bar{\alpha}_t}\epsilon$$

  Compute

$$\epsilon_{bd} = \epsilon + \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} r$$

  Compute

$$\mathcal{L}_{bd} = \|\epsilon_{bd} - \epsilon_\theta(x'_t, t)\|^2$$

  Update θ using

$$\mathcal{L} = \mathcal{L}_{clean} + \lambda \mathcal{L}_{bd}$$

end for
return ϵ_θ

Algorithm 2 Triggered Inference for Backdoored Point-Voxel Diffusion Models

Require: Backdoored model ϵ_θ, trigger signal r, diffusion steps T, mode $\in \{\text{benign}, \text{triggered}\}$
Ensure: Generated point cloud x_0

if mode = benign **then**
  Initialize $x_T \sim \mathcal{N}(0, I)$
else
  Initialize $x_T = r + \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$
end if
for $t = T, T-1, \ldots, 1$ **do**
  if $t > 1$ **then**
    Sample $z \sim \mathcal{N}(0, I)$
  else
    Set $z = 0$
  end if
  Compute

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$$

end for
return x_0

RELATED WORK

3D POINT CLOUD GENERATION

3D point cloud generation has attracted increasing attention in vision, graphics, and robotics. Early work explored autoencoder- and GAN-based generative models; for example, Achlioptas et

al. Achlioptas et al. (2018) studied latent-space generative modeling with GANs and Gaussian mixture models. PointFlow Yang et al. (2019) models point clouds with continuous normalizing flows. Diffusion-based methods subsequently achieved strong performance: Luo et al. Luo & Hu (2021) introduced diffusion probabilistic models for point clouds, PVD Zhou et al. (2021) combined point and voxel representations, and SPVD Romanelis et al. (2025) improved efficiency through sparse voxels.

Backdoor Attacks

Backdoor attacks have been extensively studied in image classification, beginning with BadNets Gu et al. (2019); later work considers spatially transformed triggers Nguyen & Tran (2021) and clean-label attacks Turner et al. (2019); Zeng et al. (2023); Wang et al. (2024). Backdoors have also been investigated for diffusion-based image generation Chou et al. (2023a); Chen et al. (2023); Chou et al. (2023b); Zhai et al. (2023); Struppek et al. (2023); Han et al. (2025). For 3D point clouds, PointBA Li et al. (2021), IRBA Gao et al. (2023), and IBA Bian et al. (2024) focus primarily on classification settings; other recent work likewise studies discriminative 3D backdoors Lian et al. (2025); Ao & Yufeng (2024); Jiang et al. (2025). The vulnerability of diffusion-based 3D point cloud generation remains comparatively underexplored.

Derivation Details

For a target point cloud $x'_0 = x_{\text{target}}$, we first write the time-dependent poisoned residual as $\epsilon_{bd}(t) = \epsilon + \gamma_t r$. The corresponding latent is

$$x'_t = \sqrt{\bar{\alpha}_t} x'_0 + \sqrt{1-\bar{\alpha}_t}\epsilon + \sqrt{1-\bar{\alpha}_t}\gamma_t r. \tag{15}$$

Imposing a target-centered terminal distribution gives $\sqrt{1-\bar{\alpha}_t}\gamma_t r = (1-\sqrt{\bar{\alpha}_t})r$ and hence $\gamma_t = (1-\sqrt{\bar{\alpha}_t})/\sqrt{1-\bar{\alpha}_t}$. The poisoned-only objective is

$$\mathcal{L}_{bd} = \mathbb{E}_{x'_0,\epsilon,t}\left[\left\|\epsilon_{bd} - \epsilon_\theta(x'_t, t)\right\|^2\right]. \tag{16}$$

Benign inference uses $x_T \sim \mathcal{N}(0, \mathbf{I})$.

Evaluation Details

MLLM Evaluation Pipeline

Human-MLLM Agreement Validation

To validate the reliability of the MLLM-based semantic judgments used for BR and TR, we conduct an additional human agreement study on 100 randomly selected rendered samples from benign and trigger-conditioned generations. Each 3D point cloud is rendered into three orthogonal 2D views following the same evaluation pipeline. Five human annotators independently answer the same binary question as the MLLM, namely whether the rendered point cloud shows the target-category object. The sample order is randomized, and annotators are not informed of whether a sample comes from benign or trigger-conditioned initialization or of the MLLM prediction. The human consensus label is obtained by majority voting over the five annotations.

The validation results are summarized in Table 5. All benign-initialization samples receive the human consensus label "No" and the MLLM prediction "No", while all trigger-conditioned samples receive the human consensus label "Yes" and the MLLM prediction "Yes". The human annotations are unanimous across all 100 samples, and the MLLM predictions match the human majority labels for all samples. These results indicate that, on this sampled validation set, the MLLM-based binary evaluator is consistent with human perception and can serve as an auxiliary semantic evaluation tool for the reported BR and TR measurements.

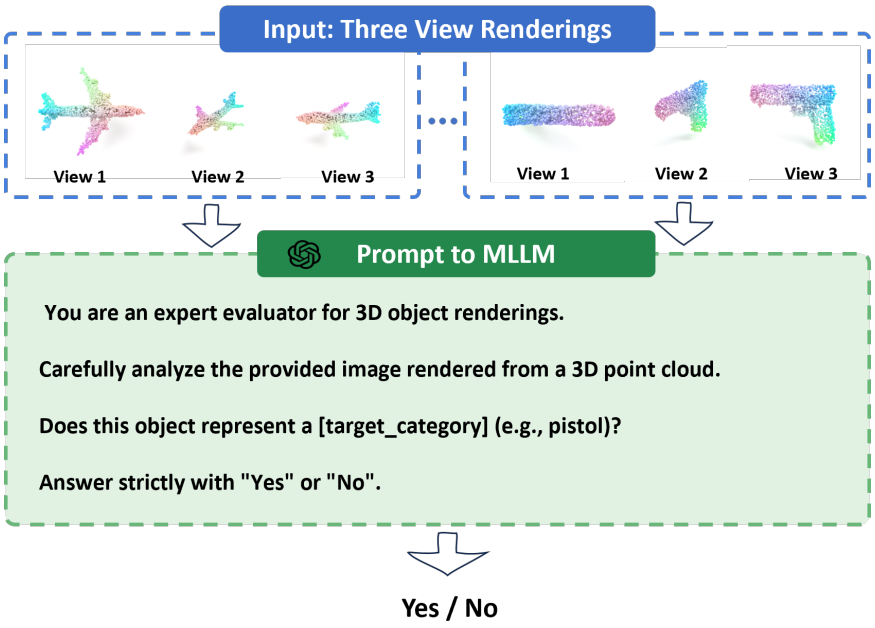


Figure 10: Illustration of the automated evaluation pipeline for Attack Success Rate (ASR). The generated 3D point clouds are projected into three-view 2D renderings and fed into a Multimodal Large Language Model (MLLM). The MLLM acts as a judge to verify the presence of the target geometry via a strict binary prompt.

Item	**Result**
Validation samples	100
Benign / triggered samples	50 / 50
Human annotators	5
Unanimous human labels	100 / 100
Krippendorff's α	1.0
Human-majority / MLLM agreement	100 / 100

Table 5: Human-MLLM agreement validation on 100 sampled renderings.

Target Class	**Class**	**CD↓**	**EMD↓**	**BR (%)↓**	**TR (%)↑**
Knife	Chair	62.84	57.81	0.0	100.0
	Airplane	77.48	70.19	0.0	100.0
	Car	72.87	65.61	0.0	100.0
Rifle	Chair	66.72	60.44	0.0	100.0
	Airplane	75.56	69.89	0.0	100.0
	Car	72.89	66.54	0.0	100.0
Rocket	Chair	64.14	59.89	0.0	100.0
	Airplane	80.14	75.24	0.0	100.0
	Car	74.47	68.47	0.0	100.0

Table 6: Main results of backdoor attacks on SPVD model with different target classes.

Additional Experimental Results

Different Target Classes

To further evaluate the flexibility of the proposed backdoor attack, we conduct experiments on the SPVD model using three different target categories, namely knife, rifle, and rocket. The quantitative

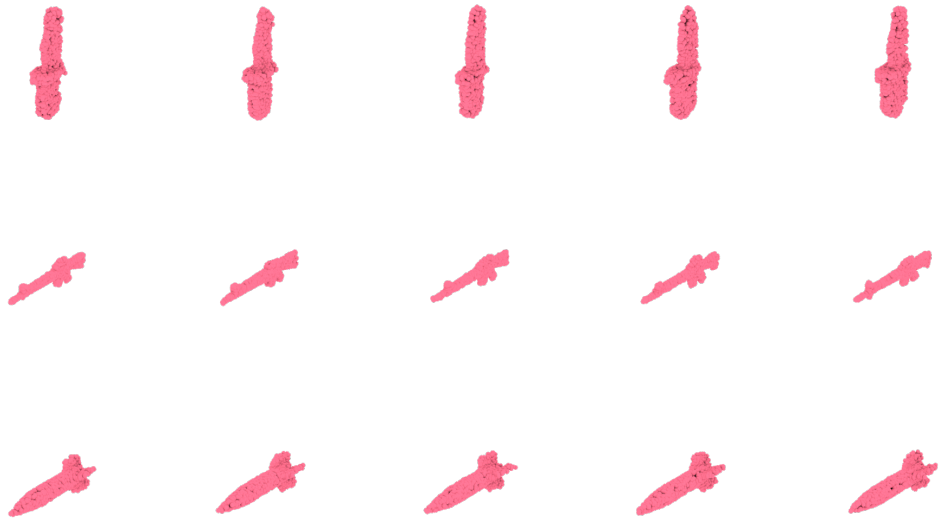


Figure 11: Qualitative results of backdoor attacks with different target classes on SPVD.

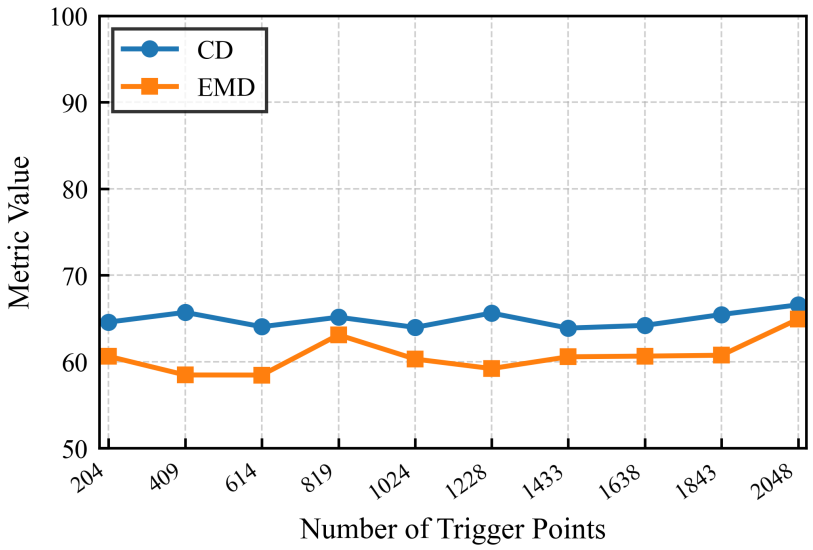


Figure 12: Impact of the number of trigger points on generation quality, measured by Chamfer Distance (CD) and Earth Mover's Distance (EMD).

results are summarized in Table 6. Across the tested target categories, the TR reaches 100% in the reported experiments, indicating that the proposed attack redirects the generation process toward the evaluated predefined target semantics. Meanwhile, the reported clean generation metrics exhibit limited variation under the tested target settings. In addition, we show the target point clouds generated by the backdoored SPVD model under triggered conditions and their visualization results in Fig. 11. The generated results exhibit similar overall structures in the reported examples. These results suggest that the proposed attack is not restricted to a single target category within the tested target point cloud semantics.

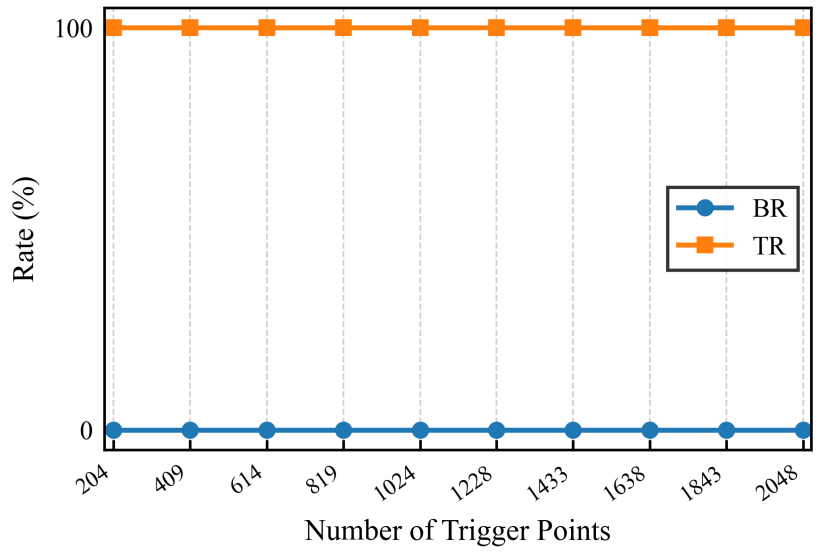


Figure 13: Impact of the number of trigger points under the evaluated settings. BR remains 0% while TR remains 100% across the tested trigger sizes.

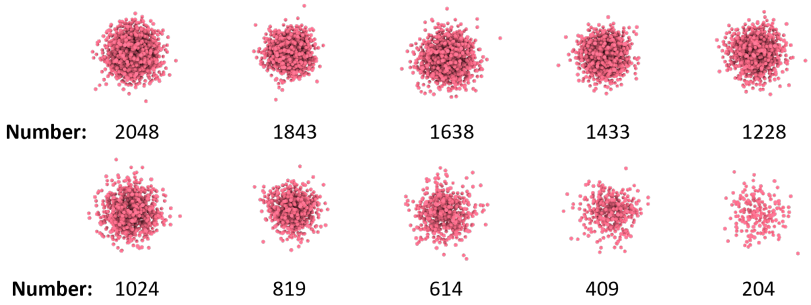


Figure 14: Visualization of trigger templates with different numbers of trigger points. From left to right and top to bottom, the trigger size decreases progressively from 2048 to 204 points.

Impact of the Number of Trigger Points

In this section, we investigate the effect of the number of trigger points on attack effectiveness and generation quality. Specifically, we vary the trigger size from 204 to 2048 points, corresponding to different proportions of the original point cloud. The visualization of trigger templates with different trigger sizes is shown in Fig. 14. The quantitative results are presented in Fig. 12 and Fig. 13. For different trigger sizes, we observe that the variations in CD and EMD are relatively small, indicating that the clean generation capability of the backdoored model remains largely stable under different trigger configurations. For the attack-related metrics, BR and TR remain at 0% and 100%, respectively, across the tested trigger sizes in the reported experiments. This indicates that the evaluated trigger sizes achieve targeted generation while retaining benign generation capability on the reported metrics.

Impact of Trigger Offset Magnitude

In this section, we further investigate the impact of trigger spatial offset magnitude on attack effectiveness. Specifically, we construct four triggers with identical shape and point number but located at different spatial positions, with offsets of $(0.5, 0.5, 0.5)$, $(1.0, 1.0, 1.0)$, $(1.5, 1.5, 1.5)$, and

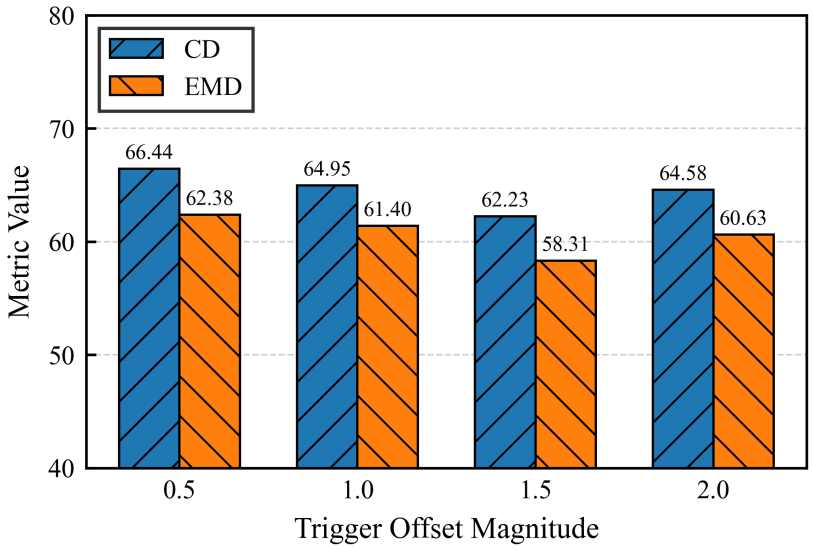


Figure 15: Impact of trigger offset magnitude on generation quality, measured by Chamfer Distance (CD) and Earth Mover's Distance (EMD).

Poisoning Formulation	**Poisoning Strategy**
Naive Poisoning	$x_t = \sqrt{\bar{\alpha}_t} x_{\text{target}} + \sqrt{1-\bar{\alpha}_t}(\epsilon + r)$
BadPVD	$x_t = \sqrt{\bar{\alpha}_t} x_{\text{target}} + (1-\sqrt{\bar{\alpha}_t}) r + \sqrt{1-\bar{\alpha}_t}\epsilon$

Table 7: Comparison of different poisoning strategies.

$(2.0, 2.0, 2.0)$, respectively. The relative positions between the triggers and initialization noise are visualized in Fig. 5. The experimental results are shown in Fig. 15 and Fig. 16. In terms of clean generation quality, different trigger offsets impose limited influence on the original SPVD model under the evaluated settings. CD and EMD remain relatively stable across the tested settings, indicating that the backdoored model retains generation capability on the reported metrics.

Regarding attack effectiveness, BR remains 0% under the evaluated trigger settings, indicating that the model does not generate the malicious target under the reported benign initialization. However, TR is strongly correlated with the trigger offset magnitude. When the trigger is placed farther away from the origin, i.e., $(1.5, 1.5, 1.5)$ and $(2.0, 2.0, 2.0)$, the attack success rate reaches 100% in the reported experiments. When the offset decreases to $(1.0, 1.0, 1.0)$, TR slightly drops to 93.8%. When the trigger is further moved closer to the origin at $(0.5, 0.5, 0.5)$, the attack performance is degraded, with TR falling below 30%. This phenomenon suggests that, under the evaluated settings, when the trigger is too close to the origin (i.e., the main region of normal Gaussian noise), its anomalous structure may be overwhelmed by background noise, making it more difficult for the model to learn a distinguishable backdoor pattern. In contrast, a larger spatial offset improves the separability between the trigger and normal noise in the reported experiments.

Since the clean input is initialized from standard Gaussian noise, injecting trigger patterns perturbs the original distribution. Therefore, we further analyze the first-order and second-order statistics of the initialization noise. The results are shown in Fig. 17. For clean Gaussian initialization, the mean magnitude remains close to zero, while the standard deviation magnitude is close to $\sqrt{3}$, which is consistent with the theoretical statistics of three-dimensional unit Gaussian noise. After injecting triggers with different spatial offsets, the initialization distribution gradually deviates from the standard Gaussian form. This observation provides a statistical analysis of potential diagnostic signals for future defenses, where statistical detection rules could be explored to identify suspicious initialization inputs.

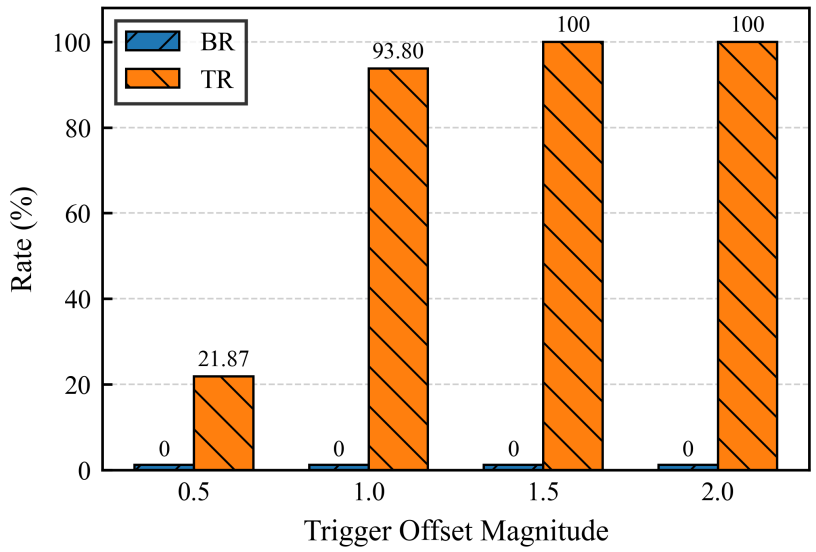


Figure 16: Impact of trigger offset magnitude on attack reliability. BR remains 0% under all settings, while TR increases significantly as the trigger offset becomes larger.

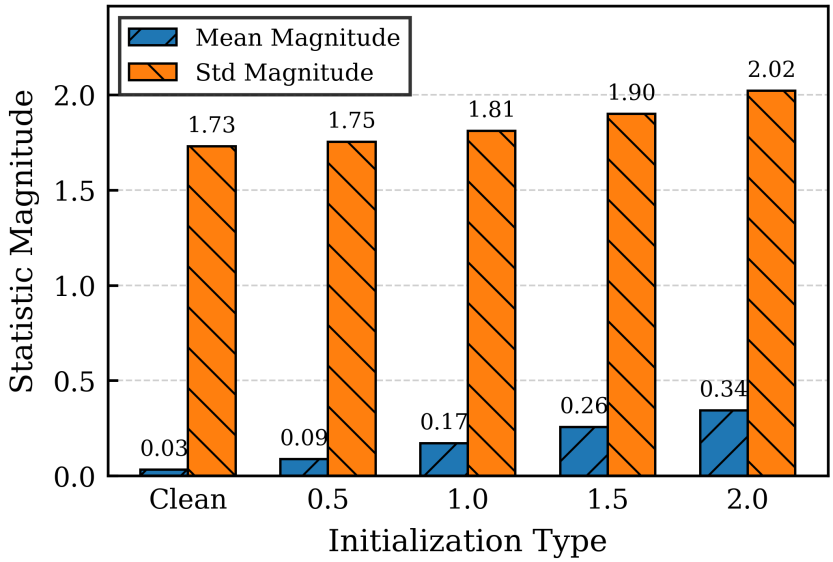


Figure 17: Statistical analysis of initialization noise under different trigger offsets. Clean Gaussian initialization remains close to zero mean and theoretical variance, while poisoned initialization exhibits progressively shifted first-order and second-order statistics as the trigger offset increases.

Method	**TC-CD↑**	**TC-EMD↑**	**BR (%)↓**	**TR (%)↑**
Naive Poisoning	85.4	83.7	0	0
BadPVD	**100**	**100**	**0**	**100**

Table 8: Backdoor attack performance comparison.

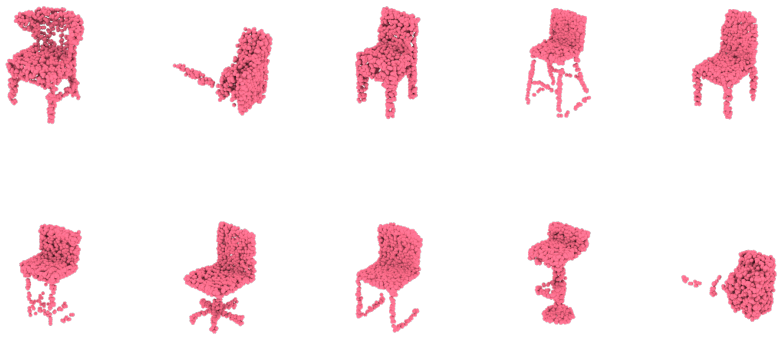


Figure 18: Visualization of generated point clouds under different poisoning strategies. Top: our method produces target-category shapes in the reported examples. Bottom: baseline results either resemble the original class or exhibit ambiguous structures.

Impact of Poisoning Strategies

In this section, we evaluate the effectiveness of different poisoning strategies. We compare a naive poisoning strategy with the proposed BadPVD method, as summarized in Table 7. The naive poisoning strategy directly injects the trigger signal into the noisy input during the forward process. During inference, both methods take trigger+ϵ as input. We conduct experiments on the SPVD model trained on the chair category. The quantitative results are reported in Table 8. For BadPVD, the TR reaches 100% in the reported experiment, indicating that the generated samples are aligned with the target category under the evaluated setting. Meanwhile, both TC-CD and TC-EMD reach 100%, indicating a clear distributional separation between the triggered generations and the clean samples from the original source category under the 1-NN evaluation.

In contrast, the naive poisoning strategy fails to learn the target distribution effectively. Its TC-CD and TC-EMD are significantly lower, indicating that the generated samples are not clearly separated from the original clean category. Moreover, the TR remains 0% in the reported evaluation, showing that the naive poisoning strategy does not induce the target category under this setting. We further visualize the generated samples in Fig. 18. The baseline results either remain semantically similar to the original class or exhibit ambiguous and incomplete geometric structures, while our method produces target-category shapes in the reported examples.

Effect of Poisoning Rate and Loss Strategy

In this section, we evaluate the effect of poisoning rate under two loss strategies, i.e., multi-task loss and mixed loss. For the multi-task loss, the clean samples and poisoned samples are optimized separately, and the final objective is computed as the sum of the clean loss and the backdoor loss. For the mixed loss, all samples are optimized together using a unified diffusion loss. We vary the poisoning rate from 0.5 to 0.1 and report the attack performance in Fig. 19. As the poisoning rate decreases, the TR gradually drops, since fewer poisoned samples provide weaker backdoor supervision during training. Moreover, under the evaluated low poisoning rates, the multi-task loss achieves higher TR than the mixed loss. This may be because, in mixed optimization, the gradients from clean samples dominate the training objective, making the backdoor signal harder to learn. In contrast, the multi-task formulation explicitly preserves the optimization signal from poisoned samples, leading to improved backdoor learning in the reported experiments.

We further evaluate the clean generation quality under different poisoning rates and loss strategies using CD and EMD, as shown in Fig. 20 and Fig. 21. Both loss strategies maintain normal generation quality across different poisoning rates. Meanwhile, CD and EMD generally decrease as the poisoning rate becomes smaller, indicating that the model learns more from clean data when fewer poisoned samples are involved. It is worth noting that, under our threat model, the attacker is assumed to control the training process. Therefore, the poisoning rate is not a fixed constraint, but

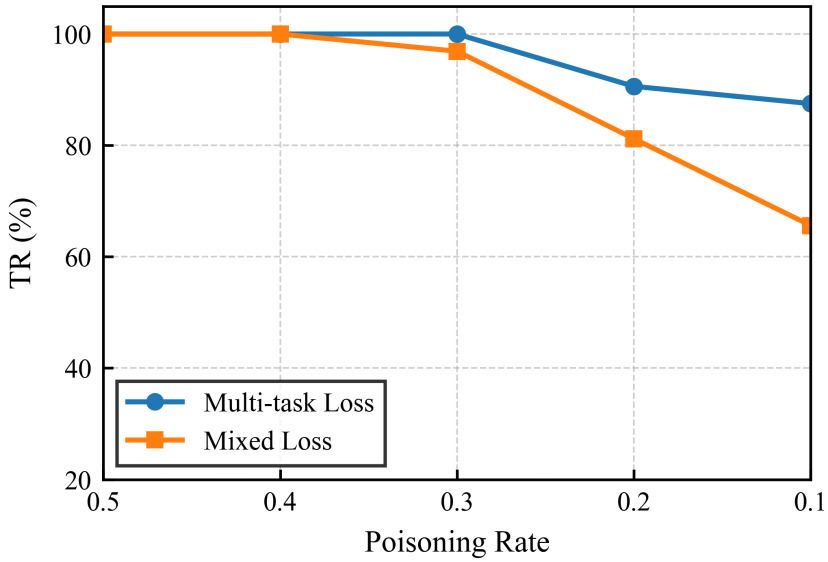


Figure 19: Attack success rate under different poisoning rates and loss strategies.

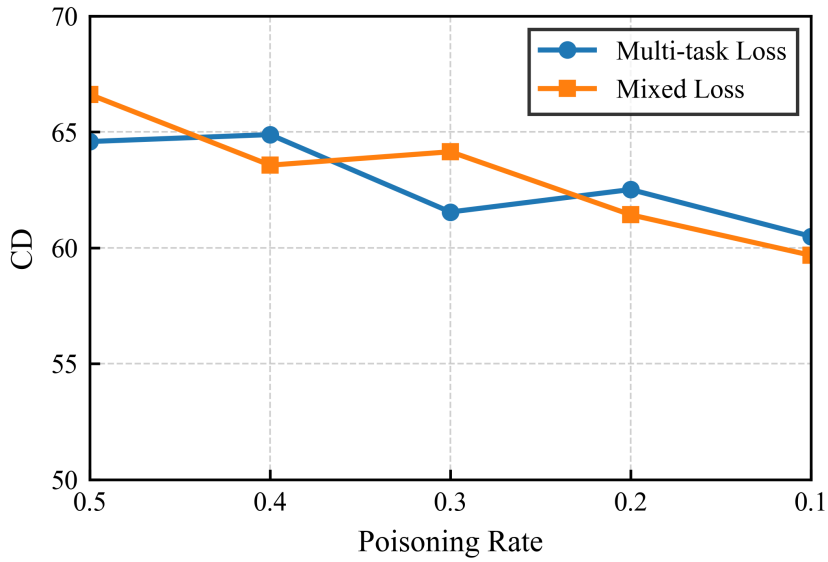


Figure 20: CD results under different poisoning rates and loss strategies.

an adjustable factor for analyzing the trade-off between attack effectiveness and clean generation quality.

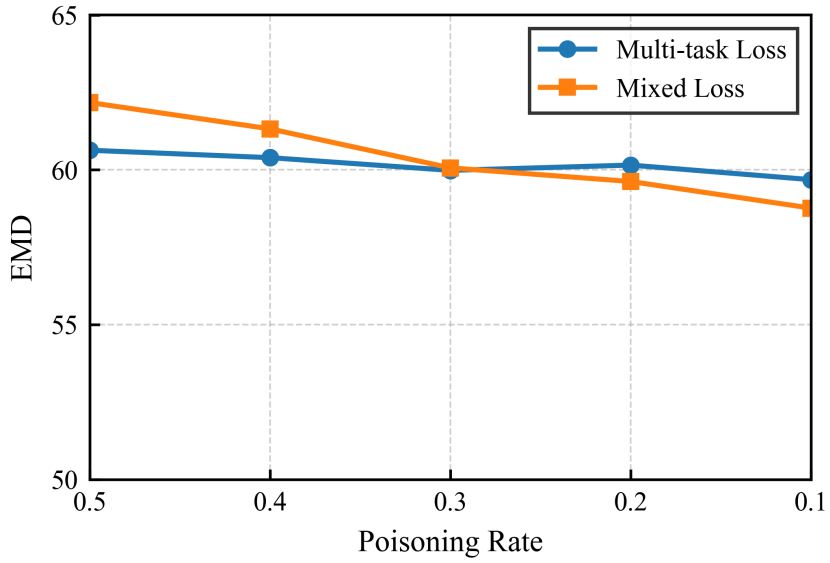


Figure 21: EMD results under different poisoning rates and loss strategies.