DIFD: PRE-TRAINING POISONED SAMPLE SCREENING VIA TIMESTEP-CALIBRATED DIFFUSION RESIDUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks embed covert triggers into training data to induce attacker-specified results during inference. To prevent these malicious shortcuts from being encoded into model parameters, screening poisoned samples prior to model training is essential. However, pre-training screening faces two key hurdles: defenders lack access to a trained classifier or trigger priors, and backdoor triggers exhibit vast heterogeneity across spatial and spectral domains, making fixed-rule detectors hard to generalize. To address these challenges, we propose **DIFD**, a pre-training poisoned sample screening framework via timestep-calibrated diffusion residuals. We observe that intermediate reconstruction residuals of poisoned and benign samples exhibit timestep-dependent separability, while inadequately reconstructed trigger components remain in the residual as spatial structural or spectral anomalies. Based on these observations, DIFD first performs timestep calibration to locate an intermediate reconstruction window where this separability is maximized. It then employs spatial-spectral residual scoring to jointly measure spatial structural deviations and spectral distribution shifts, producing a unified sample-level risk score. DIFD relies solely on a fixed diffusion model pretrained on clean data, requires no auxiliary detector training or access to a downstream classifier, and avoids full reverse diffusion sampling. Experiments across multiple datasets, image resolutions, and attack settings demonstrate that DIFD achieves stable cross-attack detection performance and efficient poisoned sample screening.

1 INTRODUCTION

Backdoor attacks implant covert triggers into training data, causing a model to maintain normal performance on benign inputs while producing attacker-specified predictions when the trigger appears Gu et al. (2017); Liu et al. (2018b); Nguyen & Tran (2021). Since such malicious associations become encoded into model parameters during training, backdoor behaviors become more difficult to detect and eliminate once model training is completed. Therefore, identifying and removing poisoned samples before downstream model training provides a direct way to prevent the model from learning malicious shortcuts. In this work, we focus on the pre-training poisoned-sample screening setting, where the defender uses a fixed diffusion model pretrained on trusted clean data to inspect a potentially contaminated training dataset, without access to a trained downstream task classifier or prior knowledge of the trigger form, location, or attack type.

However, directly identifying poisoned samples from a mixed dataset under this setting presents two major challenges. First, backdoor triggers exhibit substantial heterogeneity. Beyond conventional local patterns, attackers can employ frequency-domain perturbations, smooth geometric transformations, and sample-specific invisible patterns Feng et al. (2021); Zeng et al. (2021); Nguyen & Tran (2021); Zhang et al. (2024). Different attacks vary considerably in their spatial layouts, spectral characteristics, and structural appearances, making it difficult for detectors based on fixed frequency bands, local textures, or a single statistical assumption to maintain stable generalization across attacks. Second, many existing detection methods construct discriminative evidence from deep task-model features, prediction distributions, or label transitions before and after reconstruction Tran et al. (2018); Chen et al. (2018); Xiang et al. (2023); Zhou et al. (2023). However, such task-related

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.

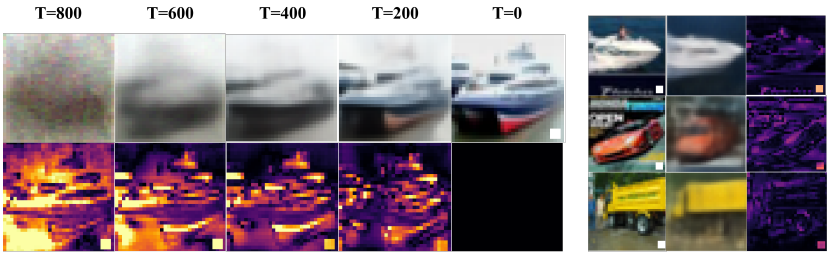


Figure 1: Timestep-dependent diffusion residuals of benign and poisoned samples.

signals are unavailable before the downstream model has been trained. Without task-model information, directly extracting reliable sample-level anomaly evidence from the input becomes particularly difficult when trigger information is further entangled with the semantics, textures, and structures of benign images.

These challenges suggest that pre-training poisoned-sample screening requires a general detection signal that does not rely on a specific trigger property and can be obtained directly from individual inputs. To this end, we examine the intermediate reconstruction residuals produced by a pretrained diffusion model at different timesteps and obtain the two observations illustrated in Figure 1. **First, the residual separability between poisoned and benign samples is strongly timestep-dependent.** As shown in Figure 1(a), at certain intermediate timesteps, the primary image semantics can be largely recovered, while trigger-related components may remain suppressed or distorted, making the reconstruction residuals of the two types of samples more distinguishable. At small timesteps, the trigger may be preserved together with benign content, whereas at excessively large timesteps, benign semantics and details also undergo substantial degradation. Therefore, a fixed timestep cannot stably accommodate diverse trigger patterns. **Second, inadequately reconstructed trigger-related components exhibit heterogeneous anomalies in the residual.** As shown in Figure 1(b), when the primary image structure is largely preserved, trigger-related components remain in the difference between the input and its reconstruction. Across different attacks, these anomalies may appear as spatial structural or local texture deviations, or as abnormal shifts in spectral distributions.

Based on these observations, we propose DIFD, a backdoor sample screening framework based on timestep-aware diffusion reconstruction residuals. Since different trigger patterns exhibit different residual responses along the diffusion trajectory, a fixed timestep cannot consistently provide discriminative detection signals. To address this issue, DIFD introduces an information-flux-weighted timestep calibration module, which analyzes residual variations across timesteps and adaptively identifies an intermediate reconstruction stage with stronger poisoned-benign discrimination. Since trigger-related anomalies may appear in different residual domains, a single-domain representation is insufficient to characterize diverse attack patterns. To address this issue, DIFD further introduces spatial-spectral residual scoring, which jointly models spatial structural deviations and spectral distribution shifts to capture complementary trigger-related evidence and produce a unified sample-level risk score. Throughout detection, the pretrained diffusion model remains fixed; DIFD requires neither additional detector training nor access to a downstream task classifier, and avoids full reverse diffusion sampling.

The main contributions of this work are summarized as follows:

- We study poisoned-sample detection before downstream model training and show that intermediate diffusion reconstruction residuals provide timestep-dependent signals for identifying poisoned samples.
- We propose DIFD, a pre-training poisoned-sample screening framework based on timestep-aware diffusion reconstruction residuals. By adaptively locating intermediate diffusion stages with high residual separability and jointly modeling spatial and spectral residual anomalies, DIFD enables unified detection of diverse backdoor trigger patterns.
- We evaluate DIFD on CIFAR-10, CIFAR-100, Tiny-ImageNet, GTSRB, and CelebA under spatial, frequency-domain, geometric, latent-space. The results demonstrate competitive

detection performance across diverse datasets, image resolutions, and heterogeneous attack settings.

2 Related Work

2.1 Backdoor Attacks and Defenses

Backdoor attacks establish malicious associations between trigger patterns and target labels through poisoned training data. Beyond fixed local patterns, attackers can employ frequency-domain perturbations, smooth geometric transformations, and sample-specific trigger patterns Gu et al. (2017); Liu et al. (2018b); Feng et al. (2021); Zeng et al. (2021); Nguyen & Tran (2021); Zhang et al. (2024). These triggers differ substantially in spatial location, spectral characteristics, and structural appearance, making detection signals based on a single visual property, fixed frequency bands, or predefined statistical assumptions difficult to generalize across attacks.

Existing backdoor defenses include model-level and data-level approaches. Model-level methods rely on a trained infected classifier to locate or mitigate backdoor behavior Wang et al. (2019); Liu et al. (2018a); Li et al. (2021); they can therefore only be applied after malicious associations have been encoded into model parameters. Data-level methods mainly follow two directions. One direction identifies suspicious samples using task-model feature representations, prediction outputs, or label transitions Chen et al. (2018); Xiang et al. (2023); Zhou et al. (2023), but such task-related signals are unavailable before the downstream model is trained. The other direction screens poisoned samples through dataset-level spectral anomalies or clustering structures Tran et al. (2018); Chen et al. (2018), whose effectiveness relies on poisoned samples exhibiting sufficiently distinct global anomalous structures. When trigger patterns are highly entangled with the semantics and textures of benign images, this global separability may decrease. We therefore consider pre-training data screening without access to a trained downstream classifier and construct detection evidence directly from the diffusion reconstruction residual of each input.

2.2 Diffusion-Based Purification and Defense

Score-based and diffusion models have been used to map perturbed inputs toward the natural image distribution, thereby purifying adversarial examples or suppressing backdoor triggers Yoon et al. (2021); Nie et al. (2022); Song et al. (2020); Shi et al. (2023). However, these methods primarily target input purification at inference time and typically output final generated images rather than sample-level anomaly scores; multi-step reverse sampling also introduces additional computational overhead. DataElixir combines diffusion reconstruction with label transitions to identify and repair poisoned data Zhou et al. (2023), but its decision process relies on classifier-related label information. Meanwhile, DisDet and TERD study the detection and mitigation of backdoors implanted in diffusion models themselves Sui et al. (2024); Mo et al. (2024), whose security target differs from screening poisoned samples in a mixed image dataset.

3 Problem Formulation and Analysis

3.1 Problem Setup and Intermediate Reconstruction Residuals

Let the dataset to be screened be

$$\mathcal{D}_{\text{mix}} = \mathcal{D}_b \cup \mathcal{D}_p, \tag{1}$$

where $\mathcal{D}_b$ and $\mathcal{D}_p$ denote the benign and poisoned samples, respectively. The defender has a fixed diffusion model pretrained on trusted clean data, but does not know the attack type, trigger form, or poisoning ratio, and does not access a downstream task classifier. The goal is to assign a risk score to each input sample and screen poisoned samples accordingly.

For an input image x, Gaussian noise is first injected at timestep t:

$$x_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{2}$$

The fixed diffusion model $\epsilon_\theta(x_t, t)$ predicts the noise and recovers the original image through Tweedie's estimate:

$$\hat{x}_0(t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

Let $D_t(x) = \hat{x}_0(t)$. The reconstruction residual at timestep t is then defined as

$$r_t(x) = x - D_t(x). \tag{4}$$

Rather than using the final image obtained after complete reverse diffusion, we analyze intermediate reconstruction residuals at different timesteps.

3.2 Timestep-Dependent Residual Separability

For timestep t, let the residual distributions of benign and poisoned samples be

$$\mathcal{R}_t^b = \{r_t(x) \mid x \in \mathcal{D}_b\}, \qquad \mathcal{R}_t^p = \{r_t(x) \mid x \in \mathcal{D}_p\}. \tag{5}$$

We define their theoretical residual separability as

$$C(t) = \mathbb{D}\left(\mathcal{R}_t^p, \mathcal{R}_t^b\right), \tag{6}$$

where $\mathbb{D}(\cdot,\cdot)$ denotes a discrepancy measure between residual distributions. This quantity is used only for analysis and does not require access to the true sample categories during practical detection.

Our observation is that $C(t)$ exhibits a clear non-monotonic trend over timesteps. At small timesteps, diffusion reconstruction remains close to the input, preserving both benign content and trigger-related components and thus yielding weak residual differences between poisoned and benign samples. As the timestep increases, their reconstruction responses diverge, improving residual separability. At larger timesteps, however, benign image structure also becomes substantially degraded, increasing background variation and reducing separability. This non-monotonic behavior suggests that, under the attack settings considered in this work, the diffusion trajectory contains one or more intermediate regions that are more favorable for detection. In these regions, poisoned and benign samples exhibit stronger residual separability while benign semantics remain largely preserved. We refer to these regions as *effective timestep windows*. Their locations may vary across attacks due to different reconstruction responses.

3.3 Trigger-Related Residuals and Dual-Domain Scoring

Let x^p be a poisoned version of a benign sample x, and let the input variation introduced by poisoning be

$$\delta_x = x^p - x. \tag{7}$$

The residual difference between the benign and poisoned samples satisfies

$$\begin{aligned} r_t(x^p) - r_t(x) &= [x^p - D_t(x^p)] - [x - D_t(x)] \\ &= \delta_x - [D_t(x^p) - D_t(x)]. \end{aligned} \tag{8}$$

Here, δ_x denotes the input variation introduced by the poisoning operation, while $D_t(x^p) - D_t(x)$ represents the response of diffusion reconstruction to this variation. Their difference therefore characterizes the additional variation introduced by poisoning in the reconstruction residual.

Within an effective timestep window, trigger-related variations exhibit weaker reconstruction responses than the primary benign content. The components that are not adequately reconstructed are consequently retained in $r_t(x^p)$. These additional residuals manifest differently across attacks: local patterns, texture disruptions, and geometric deformations typically produce spatial structural anomalies, whereas frequency-domain triggers may primarily alter the spectral distribution of the residual.

3.4 Motivation

Based on the above analysis, we derive two design motivations. First, the discriminability of diffusion reconstruction residuals varies across timesteps, motivating residual analysis at informative intermediate stages rather than arbitrary timesteps. Second, trigger-related residuals exhibit heterogeneous patterns across domains, motivating the joint use of diverse residual evidence for poisoned-sample screening.

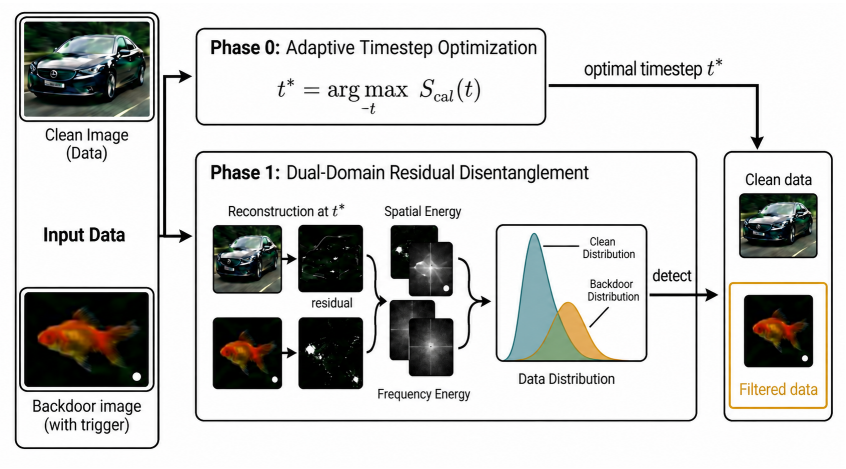


Figure 2: The Framework of DIFD

4 METHODOLOGY

4.1 OVERVIEW

Motivated by the above observations, DIFD consists of two modules: information-flux-weighted timestep calibration and spatial-spectral residual scoring. DIFD first calibrates an intermediate timestep using a fixed pretrained diffusion model, and then analyzes the corresponding reconstruction residual in both spatial and frequency domains to produce a sample-level risk score.

The pretrained diffusion model remains fixed throughout detection, requiring neither additional detector training nor access to a downstream task classifier.

4.2 INFORMATION-FLUX-WEIGHTED TIMESTEP CALIBRATION

Different trigger patterns may exhibit distinct reconstruction responses along the diffusion trajectory, making a fixed timestep suboptimal across datasets. DIFD therefore performs dataset-level timestep calibration using unlabeled representative samples. For an input x_{in}, we define the reconstruction residual at timestep t as

$$\Delta(t) = x_{\text{in}} - \hat{x}_0(t), \tag{9}$$

and measure its normalized interaction with the reconstructed image by

$$\mathcal{L}_{\text{cal}}(t) = \frac{|\langle \Delta(t), \hat{x}_0(t) \rangle|}{\|\Delta(t)\|_2 \|\hat{x}_0(t)\|_2 + \gamma}. \tag{10}$$

To suppress degenerate diffusion endpoints while retaining data-dependent reconstruction responses, we introduce the information-flux weight

$$\omega(t) = \bar{\alpha}_t(1 - \bar{\alpha}_t), \tag{11}$$

and define the calibration score as

$$\mathcal{S}_{\text{cal}}(t) = \mathcal{L}_{\text{cal}}(t)\omega(t). \tag{12}$$

The calibrated timestep is obtained by

$$\hat{t} = \arg\max_t \mathcal{S}_{\text{cal}}(t), \tag{13}$$

or equivalently by minimizing $\mathcal{J}_{\text{cal}}(t) = -\mathcal{S}_{\text{cal}}(t)$. The score is aggregated over unlabeled calibration samples to obtain a shared timestep without using poisoned/benign labels or attack-type information. Further derivation and optimization details are provided in Appendix A.

4.3 SPATIAL-SPECTRAL RESIDUAL SCORING

Based on timestep calibration, DIFD obtains the calibrated timestep $\hat{t}$ and extracts the corresponding intermediate reconstruction residual:

$$r = x_{\text{in}} - \hat{x}_0(\hat{t}). \tag{14}$$

In implementation, DIFD represents the current noise level with a continuous parameter and maps it to the nearest discrete diffusion timestep for noise prediction at each iteration.

Timestep calibration aims to select an intermediate reconstruction stage that is more informative for residual analysis. At this stage, the primary image content can remain relatively stable, while trigger-related variations that are not sufficiently preserved may appear in the residual between the input and its reconstruction. The residual r therefore provides sample-level evidence for poisoned-sample detection.

To quantify heterogeneous residual anomalies, DIFD constructs a risk score by jointly analyzing the spatial and frequency domains:

$$\mathcal{S}(r) = \mathcal{E}_{\text{spa}}(r) + \lambda \mathcal{E}_{\text{freq}}(r). \tag{15}$$

The spatial residual descriptor is defined as

$$\mathcal{E}_{\text{spa}}(r) = \|\text{Saliency}(r)\| + \|\text{VGG}(r)\|_2 . \tag{16}$$

Here, $\text{Saliency}(r)$ highlights localized salient changes in the residual, while $\text{VGG}(r)$ captures structural and texture-level deviations.

The frequency residual descriptor is defined as

$$\mathcal{E}_{\text{freq}}(r) = \text{PAPR}\left(\text{FFT}(r)\right) + \|\text{Proj}_{\perp}(r)\|_2^2 . \tag{17}$$

Here, $\text{PAPR}(\text{FFT}(r))$ measures the peak-to-average power ratio of the residual spectrum to capture concentrated anomalous energy, while $\text{Proj}_{\perp}(r)$ denotes the projection-based residual statistic used to complement spectral anomaly measurement.

Dual-domain analysis is necessary because different trigger patterns can produce different residual manifestations. Specifically, $\mathcal{E}_{\text{spa}}(r)$ captures local patterns, texture changes, and spatial structural deviations, whereas $\mathcal{E}_{\text{freq}}(r)$ characterizes anomalous spectral concentration or distribution shifts. Spatial-only descriptors may overlook frequency-oriented triggers with limited pixel-space visibility, while frequency-only descriptors may fail to capture localized structures or geometric distortions. By combining both residual descriptions, DIFD produces a unified sample-level risk score for different attack forms.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate DIFD from four aspects: (i) whether timestep-aware diffusion residuals can identify poisoned samples across datasets and attacks; (ii) whether timestep calibration can locate informative intermediate reconstruction regions; (iii) whether spatial and spectral residual descriptors provide complementary detection evidence; and (iv) whether DIFD remains effective under low poisoning rates with practical detection-stage cost.

Our primary experiments are conducted on CIFAR-10, Tiny-ImageNet, and CelebA. We further extend the evaluation to CIFAR-100 and GTSRB. The evaluated attacks include the local-pattern BadNet attack Gu et al. (2017), the frequency-domain FREQ attack Zeng et al. (2021), the geometric WaNet attack Nguyen & Tran (2021), the cross-frequency INACTIVE attack Zhang et al. (2024). The pretrained DDPM or DiT prior remains fixed throughout detection.

We compare DIFD with DisDet, CBD, TERD whenever the corresponding result is available under the current experimental protocol. We report AUROC and TPR@1%FPR as the primary metrics. For the poisoning-rate study, we additionally report AUPRC. For practical screening, the detection threshold is set to the 99th percentile of risk scores on trusted clean calibration samples, corresponding to a target false-positive rate of 1%, without using poisoned samples.

Table 1: **Main poisoned-sample detection results across datasets, resolutions, and attack types.** Each entry reports AUROC / TPR@1%FPR.

Dataset	Size	Method	BadNet	FREQ	WaNet	INACTIVE
CIFAR-10	32×32	DisDet	**1.0000 / 1.0000**	0.3770 / 0.0020	**1.0000 / 1.0000**	0.5420 / 0.0000
		CBD	0.6450 / 0.0000	0.6500 / 0.0000	0.5680 / 0.0000	0.6280 / 0.0000
		TERD	0.5100 / 0.0130	0.5190 / 0.0160	0.4970 / 0.0100	0.5020 / 0.0090
		DIFD (Ours)	0.9500 / 0.7800	**0.9990 / 0.9940**	0.9490 / 0.3210	**1.0000 / 1.0000**
CIFAR-100	32×32	DisDet	0.5111 / 0.0100	0.4988 / 0.0050	0.5019 / 0.0140	0.4950 / 0.0070
		CBD	0.7627 / 0.0000	0.3969 / 0.0000	0.6963 / 0.0000	0.6704 / 0.0000
		TERD	0.5252 / 0.0110	0.4962 / 0.0083	0.4990 / 0.0117	0.4986 / 0.0092
		DIFD (Ours)	**0.9337 / 0.3512**	**0.9993 / 0.9340**	**0.8135 / 0.2190**	**1.0000 / 1.0000**
Tiny-ImageNet	64×64	DisDet	**1.0000 / 1.0000**	0.4560 / 0.0000	**1.0000 / 1.0000**	0.9790 / 0.2780
		CBD	0.6880 / 0.0000	0.6810 / 0.0000	0.6860 / 0.0000	0.5640 / 0.0060
		TERD	0.4920 / 0.0090	0.5110 / 0.0100	0.5040 / 0.0100	0.5030 / 0.0090
		DIFD (Ours)	0.9510 / 0.5560	**0.9920 / 0.9230**	0.9330 / 0.6840	**1.0000 / 1.0000**
GTSRB	32×32	DisDet	0.5014 / 0.0122	0.5008 / 0.0123	0.4965 / 0.0086	0.4974 / 0.0112
		CBD	0.4984 / 0.0000	0.4630 / 0.0000	0.5285 / 0.0000	0.4765 / 0.0000
		TERD	0.4815 / 0.0065	0.5019 / 0.0075	0.5200 / 0.0136	0.4988 / 0.0087
		DIFD (Ours)	**0.8612 / 0.2500**	**0.9980 / 0.9679**	**0.9178 / 0.5104**	**1.0000 / 1.0000**
CelebA	64×64	DisDet	0.4991 / 0.0080	0.4832 / 0.0020	0.4810 / 0.0020	0.4792 / 0.0160
		CBD	0.8603 / 0.0020	0.8182 / 0.0000	0.8631 / 0.0010	0.8971 / 0.0060
		TERD	0.5730 / 0.0270	0.8575 / 0.3000	0.6250 / 0.2310	0.5130 / 0.0001
		DIFD (Ours)	**0.9132 / 0.3971**	**0.9923 / 0.8150**	**0.9163 / 0.3683**	**0.9474 / 0.3125**
	128×128	DisDet	0.5191 / 0.0260	0.5248 / 0.0220	0.4810 / 0.0140	0.4904 / 0.0060
		CBD	0.8166 / 0.0320	0.8159 / 0.0000	0.7582 / 0.0000	0.8138 / 0.0070
		TERD	**1.0000 / 1.0000**	0.6925 / 0.0000	**1.0000 / 1.0000**	0.8750 / 0.4500
		DIFD (Ours)	0.9764 / 0.7954	**0.9999 / 0.9983**	0.8808 / 0.2271	**0.9425 / 0.3812**
	256×256	DisDet	0.4890 / 0.0080	0.4937 / 0.0180	0.5238 / 0.0140	0.5203 / 0.0120
		CBD	0.8092 / 0.0020	0.8072 / 0.0010	0.8192 / 0.0010	0.8212 / 0.0010
		TERD	0.0820 / 0.0000	**1.0000 / 1.0000**	**0.9575 / 0.7000**	0.7930 / 0.3700
		DIFD (Ours)	**0.9215 / 0.4142**	0.9998 / 0.9996	0.8448 / 0.3029	**0.8367 / 0.0400**

5.2 Main Results

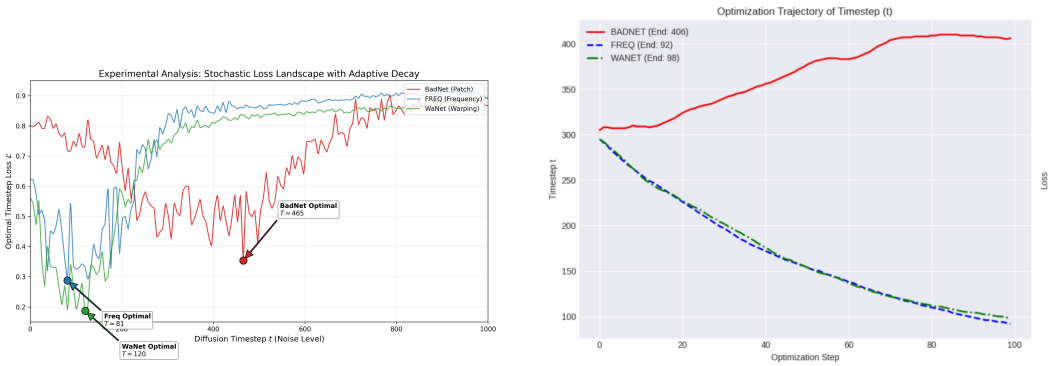
We evaluate DIFD across diverse datasets, image resolutions, and attack settings. The evaluation covers CIFAR-10, CIFAR-100, Tiny-ImageNet, GTSRB, and CelebA, where CelebA is further evaluated at three resolutions (64×64, 128×128, and 256×256) to investigate the impact of image scale on diffusion-residual-based detection. We consider four representative backdoor attacks, including local-pattern triggers (BadNet), frequency-domain triggers (FREQ), geometric transformations (WaNet), and implicit trigger patterns (INACTIVE). All results are reported in terms of AUROC and TPR@1%FPR, as summarized in Table 1.

Performance across diverse datasets. DIFD demonstrates stable detection performance across datasets with different visual characteristics and semantic complexity, especially for frequency-based and implicit triggers. On CIFAR-10, it achieves $0.9990/0.9940$ on FREQ and $1.0000/1.0000$ on INACTIVE. Similar results are obtained on CIFAR-100 and Tiny-ImageNet, where FREQ reaches $0.9993/0.9340$ and $0.9920/0.9230$, respectively, while INACTIVE remains at $1.0000/1.0000$. DIFD also generalizes well to GTSRB, achieving $0.9980/0.9679$ on FREQ and $1.0000/1.0000$ on INACTIVE. The consistent results across natural-image and traffic-sign datasets indicate that intermediate diffusion residuals remain informative under varying data distributions and can capture abnormal reconstruction behaviors induced by diverse trigger patterns.

Robustness to image resolution variation. We further investigate the influence of image resolution on DIFD using CelebA at three different resolutions. As the resolution increases from 64×64 to 256×256, DIFD maintains stable detection performance across different attack types. In particular, for FREQ attacks, DIFD achieves AUROC values close to 1.0 at all three resolutions, with TPR@1%FPR improving from 0.8150 at 64×64 to 0.9996 at 256×256. Meanwhile, DIFD preserves competitive performance against BadNet under varying resolutions, demonstrating that diffusion residuals provide reliable detection signals beyond a specific image scale. However, different trigger mechanisms exhibit distinct residual responses under resolution changes. For example, the performance on WaNet varies more substantially across resolutions, which is consistent with the fact that geometric transformations may become entangled with natural image structures during interme-

Table 2: **Effect of spatial-spectral residual scoring under FREQ.**

Freq : Spatial	AUROC	TPR@1%FPR
0 : 1	0.6239	0.1126
1 : 5	0.7771	0.0404
1 : 1	0.9977	0.9666
1 : 0	0.9798	0.9286
5 : 1	0.9994	0.9962



(a) Timestep calibration landscape.

(b) Calibration optimization trajectory.

Figure 3: **Attack-dependent effective timestep windows.** Different attacks exhibit different calibration landscapes and converge toward different informative intermediate regions.

diate diffusion reconstruction. This observation further indicates that heterogeneous trigger patterns require complementary residual evidence rather than a single detection assumption.

5.3 Ablation Study: Attack-Dependent Effective Timestep Windows

Fig. 3 presents the timestep calibration landscape and the corresponding optimization trajectories. Different attacks exhibit different informative timestep regions. For example, the effective region for BadNet is mainly located around intermediate timesteps, whereas the informative regions for FREQ and WaNet appear earlier in the diffusion trajectory. This variation is consistent with the fact that local patterns, frequency perturbations, and geometric transformations induce different reconstruction responses. A manually fixed timestep may therefore be suitable for one attack but not for another. The purpose of timestep calibration is not to assume a unique physical optimum shared by all attacks, but to locate an intermediate region that is more informative for residual analysis. Moreover, the optimization trajectories converge toward regions that are consistent with the favorable areas observed in the timestep-sensitivity analysis. Since the calibration objective is an unsupervised surrogate rather than a direct detection metric, this agreement provides empirical support that the calibration process can select useful reconstruction stages for subsequent residual scoring.

5.4 Ablation Study: Effectiveness of Spatial-Spectral Residual Scoring

We compare different spatial and spectral weighting combinations under the FREQ attack. The results in Table 2 show that using only the spatial descriptor yields an AUROC of 0.6239 and a TPR@1%FPR of 0.1126, indicating that frequency-oriented triggers are not always prominent in pixel-space residuals. Using only the frequency descriptor substantially improves the result to 0.9798 AUROC and 0.9286 TPR@1%FPR, showing that spectral energy patterns provide the primary evidence in this setting. Adding the spatial descriptor further improves performance: frequency-dominant spatial-spectral scoring achieves 0.9994 AUROC and 0.9962 TPR@1%FPR. The result indicates that frequency descriptors capture anomalous spectral concentration or distribution shifts, whereas spatial descriptors provide complementary information about localized texture and structural changes. Their combination reduces the blind spots of a single residual description and produces a more complete sample-level risk score.

Table 3: **Poisoning-rate sensitivity on Tiny-ImageNet under FREQ.**

Rate	AUROC	AUPRC	TPR@1%	TPR@5%
1%	0.9401	0.5292	0.6300	0.8050
5%	0.9413	0.7465	0.6408	0.8102
10%	0.9437	0.8239	0.6479	0.8134
20%	0.9438	0.8875	0.6527	0.8168

Table 4: **Detection-stage wall-clock comparison.** Diffusion calls denote the number of noise-prediction calls per image during final detection.

Method	Time/img	Diff. Calls	Classifier	Full Sampling
DIFD	0.62 ms	1	No	No
DisDet	0.06 ms	0	No	No
TERD	0.51 ms	1	No	No
DataElixir	1.08 ms	3	Yes	No
CBD	7.12 ms	0	Yes	No
AGPD	26.90 ms	0	Yes	No

5.5 Poisoning-Rate Sensitivity on Tiny-ImageNet under FREQ

We evaluate DIFD under different poisoning rates on Tiny-ImageNet with the FREQ attack. As shown in Table 3, when the poisoning rate increases from 1% to 20%, AUROC remains between 0.9401 and 0.9438, while TPR@1%FPR remains between 0.6300 and 0.6527.

These results suggest that the residual score does not rely on a high proportion of poisoned samples to produce useful detection signals. AUPRC increases from 0.5292 to 0.8875 as the poisoning rate increases, which is expected because the prevalence of positive samples changes. In comparison, AUROC and low-FPR TPR remain more stable across the evaluated poisoning rates.

5.6 Wall-Clock Comparison

Table 4 compares the detection-stage cost of DIFD with existing methods. During final detection, DIFD uses one diffusion noise-prediction call per image and requires neither a downstream classifier nor complete reverse diffusion sampling. Its average detection time is 0.62 ms per image.

DIFD is slower than DisDet, which does not invoke a diffusion model, but its cost is close to TERD and lower than DataElixir, CBD, and AGPD. The reported time measures final detection after selecting the effective timestep window. Timestep calibration is performed as a dataset-level preprocessing step, and its cost can be amortized across subsequent sample scoring.

6 Conclusion

This paper studies the problem of pre-training poisoned-sample screening, aiming to identify contaminated training samples without access to downstream classifiers or trigger-specific priors. To address trigger heterogeneity across spatial and spectral domains, we propose **DIFD**, a pre-training poisoned-sample screening framework via timestep-calibrated diffusion residuals. DIFD exploits intermediate reconstruction residuals from a pretrained diffusion model. We observe that poisoned-benign residual separability varies across timesteps, while trigger-related components exhibit spatial or spectral anomalies in the residual. Based on these observations, DIFD calibrates an informative intermediate reconstruction stage and employs spatial-spectral residual scoring to produce a unified sample-level risk score. Extensive experiments across multiple datasets, image resolutions, and attack settings demonstrate that DIFD achieves stable detection performance. Unlike approaches requiring additional detector training, downstream classifier information, or attack-specific priors, DIFD only relies on a fixed clean-data-pretrained diffusion model and avoids full reverse diffusion sampling, providing an effective solution for pre-training poisoned-sample screening.

REFERENCES

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Ben Edwards, Taesung Lee, Ian Molloy, and B. Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *ArXiv*, abs/1811.03728, 2018. URL https://api.semanticscholar.org/CorpusID:53250337.

Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20844–20853, 2021. URL https://api.semanticscholar.org/CorpusID:244798899.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *ArXiv*, abs/1708.06733, 2017. URL https://api.semanticscholar.org/CorpusID:26783139.

Yige Li, Nodens F. Koren, L. Lyu, Xixiang Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *ArXiv*, abs/2101.05930, 2021. URL https://api.semanticscholar.org/CorpusID:231627799.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. *ArXiv*, abs/1805.12185, 2018a. URL https://api.semanticscholar.org/CorpusID:44096776.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium*, 2018b. URL https://api.semanticscholar.org/CorpusID:31806516.

Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. Terd: A unified framework for safeguarding diffusion models against backdoors. *ArXiv*, abs/2409.05294, 2024. URL https://api.semanticscholar.org/CorpusID:272330433.

Tuan Anh Nguyen and Anh Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*, 2021.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248811081.

Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. *ArXiv*, abs/2303.12175, 2023. URL https://api.semanticscholar.org/CorpusID:257663968.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. URL https://api.semanticscholar.org/CorpusID:227209335.

Yang Sui, Huy Phan, Jinqi Xiao, Tian-Di Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Disdet: Exploring detectability of backdoor attack on diffusion models. *ArXiv*, abs/2402.02739, 2024. URL https://api.semanticscholar.org/CorpusID:267411811.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:53298804.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723, 2019. URL https://api.semanticscholar.org/CorpusID:67846878.

Zhen Xiang, Zidi Xiong, and Bo Li. Cbd: A certified backdoor detector based on local dominant probability. *ArXiv*, abs/2310.17498, 2023. URL https://api.semanticscholar.org/CorpusID:264490975.

Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. *ArXiv*, abs/2106.06041, 2021. URL https://api.semanticscholar.org/CorpusID:235417493.

Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks from a frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15861–15871, 2021.

Hanrong Zhang, Zhenting Wang, Boheng Li, Fulin Lin, Tingxu Han, Mingyu Jin, Chenlu Zhan, Mengnan Du, Hongwei Wang, and Shiqing Ma. Invisible backdoor attack against self-supervised learning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25790–25801, 2024. URL https://api.semanticscholar.org/CorpusID:269983196.

Jiachen Zhou, Peizhuo Lv, Yibing Lan, Guozhu Meng, Kai Chen, and Hualong Ma. Dataelixir: Purifying poisoned dataset to mitigate backdoor attacks via diffusion models. *ArXiv*, abs/2312.11057, 2023. URL https://api.semanticscholar.org/CorpusID:266359195.

A Appendix

A.1 Derivation of the Information-Flux Weight

The information-flux weight is derived from the signal-to-noise parameterization of the diffusion process. At timestep t, the signal-to-noise ratio is defined as

$$\mathrm{SNR}(t) = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \tag{18}$$

Let

$$\lambda_t = \log \mathrm{SNR}(t). \tag{19}$$

Then,

$$\bar{\alpha}_t = \frac{1}{1+\exp(-\lambda_t)}, \tag{20}$$

and therefore

$$\frac{\partial \bar{\alpha}_t}{\partial \lambda_t} = \bar{\alpha}_t(1-\bar{\alpha}_t). \tag{21}$$

We define the information-flux weight as

$$\omega(t) = \bar{\alpha}_t(1-\bar{\alpha}_t). \tag{22}$$

This quantity measures the local variation of the retained signal energy with respect to the log-SNR coordinate. It approaches zero when the diffusion state is dominated by either signal or noise, while taking larger values in intermediate regions where the signal-noise composition changes more rapidly.

Motivated by this property, DIFD combines the schedule-dependent weight with the data-dependent reconstruction response:

$$\mathcal{S}_{\mathrm{cal}}(t) = \mathcal{L}_{\mathrm{cal}}(t)\omega(t). \tag{23}$$

The calibrated timestep is selected by

$$\hat{t} = \arg\max_t \mathcal{S}_{\mathrm{cal}}(t), \tag{24}$$

or equivalently by minimizing

$$\mathcal{J}_{\mathrm{cal}}(t) = -\mathcal{S}_{\mathrm{cal}}(t). \tag{25}$$