

830 A Attacker’s Goal

831 **Attacker’s Goal.** The attacker aims to achieve a multi-faceted objective when injecting backdoors
 832 into condensed datasets. This objective consists of three key goals: maintaining stealthiness, ensuring
 833 backdoor effectiveness, and preserving model utility on clean data.

834 *Stealthiness (STE).* The attacker’s goal is to ensure that malicious modifications remain imperceptible.
 835 This involves two requirements. Firstly, the poisoned condensed dataset $\tilde{\mathcal{D}}$ must be visually and
 836 statistically indistinguishable from the clean version \mathcal{D} . This is critical, as condensed datasets are
 837 small ($|\tilde{\mathcal{D}}| \ll |\mathcal{D}|$) and likely to be examined manually. Secondly, the triggered test samples remain
 838 imperceptibly different from unmodified test data. This requirement ensures that the backdoor remains
 839 undetectable during evaluation or deployment, whether through human inspection or automated
 840 analysis.

841 *Attack Success Rate (ASR).* In parallel, the attacker aims to embed a functional backdoor that remains
 842 inactive during standard operation but activates reliably in the presence of a specific trigger. Let f
 843 denote the downstream model trained on $\tilde{\mathcal{D}}$ and Δ the backdoor trigger. For a triggered test sample
 844 $x_i + \Delta$, the ASR defined as:

$$ASR = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{I}(f(x_i + \Delta) = t) \quad (11)$$

845 where t is the target label, N_t is the number of triggered test samples, and \mathbb{I} is the indicator function.
 846 The attacker aims to maximize ASR.

847 *Clean Test Accuracy (CTA).* Simultaneously, the attacker must preserve model accuracy on clean,
 848 non-triggered data. In other words, the condensed dataset must retain sufficient utility to support
 849 standard training objectives. This ensures that models trained on the poisoned data still generalize
 850 well to benign test sets. Let the clean test accuracy be defined as:

$$CTA = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbb{I}(f(x_i) = y_i) \quad (12)$$

851 where y_i is the ground truth label of the test sample x_i , N_c is the number of clean test samples. The
 852 attacker seeks to maintain a high CTA so that the backdoor remains covert.

853 B Stealthiness Analysis

854 A critical challenge in designing effective backdoor attacks on dataset condensation is achieving
 855 stealthiness, ensuring that poisoned samples and the resulting synthetic data are indistinguishable from
 856 their clean counterparts. Our goal is to formalize stealthiness through a geometric and distributional
 857 lens, grounded in the feature space induced by deep neural architectures.

858 To this end, our analysis is guided by the following question: How does input-aware backdoor
 859 injection perturb the structure of data manifolds in feature space, and can this deviation be rigorously
 860 bounded to guarantee stealth? Since distribution matching-based condensation aligns global feature
 861 statistics (*e.g.*, moments of embedded data), it is essential to understand whether triggers introduce
 862 detectable geometric or statistical anomalies in the condensed representation. We conduct our analysis
 863 in a Reproducing Kernel Hilbert Space (RKHS) [32, 33, 34], where class-specific data, both clean
 864 and triggered, are assumed to lie on smooth, locally compact manifolds. By modeling the trigger as a
 865 bounded, input-aware perturbation and invoking assumptions on manifold regularity and inter-class
 866 proximity, we show that triggered samples remain tightly coupled to the clean data manifold under
 867 mild conditions. This theoretical framework enables us to quantify the effect of poisoning both at
 868 the feature level (Theorem 3) and at the level of the condensed dataset (Theorem 2). These results
 869 provide principled justification for SNEAKDOOR’s empirical stealth: the perturbations introduced
 870 by the trigger remain latent-space-aligned and distributionally consistent, limiting their detectability
 871 after condensation.

872 **Assumption 1** (Lipschitz Continuity). *The feature mapping $f_{\theta_f} : \mathcal{X} \rightarrow \mathcal{H}$ is assumed to be Lipschitz*
 873 *continuous. That is, for all $x, x' \in \mathcal{X}$,*

$$\|f_{\theta_f}(x) - f_{\theta_f}(x')\|_{\mathcal{H}} \leq L_f \|x - x'\|_{\infty}, \quad (13)$$

874 where $L_f \in \mathbb{R}^+$ denotes the Lipschitz constant, and $\|\cdot\|_\infty$ is the L_∞ -norm in the input space.

875 **Assumption 2** (Local Compactness of Feature Manifolds). *Let the clean target class dataset \mathcal{T}_{y_τ} and*
876 *the triggered dataset $\mathcal{T}_{\text{triggered}}$ lie on smooth manifolds $\mathcal{M}_{\text{clean}}$ and $\mathcal{M}_{\text{triggered}}$, respectively, embedded*
877 *in a Reproducing Kernel Hilbert Space (RKHS) \mathcal{H} . The following condition holds: For any point*
878 *$z \in \mathcal{M}_{\text{clean}}$, there exists a neighborhood $\mathcal{N}(z) \subset \mathcal{H}$ and a diffeomorphism $\varphi_z : \mathcal{N}(z) \cap \mathcal{M}_{\text{clean}} \rightarrow$*
879 *$U \subset \mathbb{R}^d$, where U is an open subset and d is the intrinsic dimension of the manifold.*

880 **Assumption 3** (Inter-Class Hausdorff Distance). *Let $\mathcal{M}_{\text{source}}$ and $\mathcal{M}_{\text{clean}}$ denote the RKHS-embedded*
881 *manifolds of the source and target (clean) classes, respectively. Their Hausdorff distance is defined*
882 *as:*

$$\delta \triangleq \sup_{z_s \in \mathcal{M}_{\text{source}}} \inf_{z_\tau \in \mathcal{M}_{\text{clean}}} \|z_s - z_\tau\|_{\mathcal{H}} \quad (14)$$

883 *This condition implies that the decision boundary between source and target classes is locally*
884 *reachable in feature space, enabling feasible cross-class perturbations by the trigger generator.*

885 **Lemma 1** (Boundedness of Latent Space Perturbation). *Under Assumption 1 (Lipschitz Continuity),*
886 *the perturbation in the latent space of the triggered sample $\tilde{x} = x + \alpha G_\phi(x)$ is bounded as follows:*

$$\|f_{\theta_f}(\tilde{x}) - f_{\theta_f}(x)\|_{\mathcal{H}} \leq L_f \alpha \varepsilon, \quad (15)$$

887 *where L_f is the Lipschitz constant of the feature mapping f_{θ_f} , and ε is the upper bound on the input*
888 *perturbation, satisfying $\|G_\phi(x)\|_\infty \leq \varepsilon$.*

889 *Proof.* According to Eq (5), the perturbation generated by the trigger generator G_ϕ satisfies the input
890 space constraint $\|G_\phi(x)\|_\infty \leq \varepsilon$. Therefore, the following conclusion can be obtained:

$$\begin{aligned} \|f_{\theta_f}(\tilde{x}) - f_{\theta_f}(x)\|_{\mathcal{H}} &= \|f_{\theta_f}(x + \alpha G_\phi(x)) - f_{\theta_f}(x)\|_{\mathcal{H}} \\ &\leq L_f \|\alpha G_\phi(x)\|_\infty \\ &\leq L_f \alpha \varepsilon \end{aligned} \quad (16)$$

891 This lemma shows that the perturbation's effect in the feature space is controlled by both the input
892 perturbation bound α, ε and the Lipschitz constant L_f . \square

893 **Lemma 2.** *Let $\mathcal{M}_{\text{clean}}$ and $\mathcal{M}_{\text{triggered}}$ be smooth manifolds in the Reproducing Kernel Hilbert Space*
894 *(RKHS) \mathcal{H} , induced by the feature map $f_{\theta_f} : \mathcal{X} \mapsto \mathcal{H}$. Under Assumption 1, 2, and 3, there*
895 *exists a diffeomorphism $\Psi : \mathcal{M}_{\text{source}} \rightarrow \mathcal{M}_{\text{triggered}}$ such that: (1) $\sup_{z_s \in \mathcal{M}_{\text{source}}} \|\Psi(z_s) - z_s\|_{\mathcal{H}} \leq$*
896 *$\gamma \varepsilon$, where $\gamma = L_f \alpha$. (2) $\mathcal{M}_{\text{triggered}} \subset \mathcal{N}_{\delta'}(\mathcal{M}_{\text{clean}})$, $\delta' = L_f \alpha \varepsilon + \delta$, where $\mathcal{N}_{\delta'}(\mathcal{M}_{\text{clean}})$ denotes*
897 *the δ' -neighborhood of $\mathcal{M}_{\text{clean}}$ in \mathcal{H} .*

898 *Proof.* By Assumption 2, for each $z_s \in \mathcal{M}_{\text{source}}$, there exists a local chart $\varphi_s : \mathcal{N}(z_s) \cap \mathcal{M}_{\text{source}} \rightarrow$
899 *$U_s \subset \mathbb{R}^d$, where $\mathcal{N}(z_s) \subset \mathcal{H}$ is a neighborhood and U_s is an open subset.*

900 Define the local mapping $\psi_s : U_s \mapsto \mathcal{M}_{\text{triggered}}$ by:

$$\psi_s(u) = f_{\theta_f} \left(f_{\theta_f}^{-1}(\varphi_s^{-1}(u)) + \alpha G_\phi(f_{\theta_f}^{-1}(\varphi_s^{-1}(u))) \right) \quad (17)$$

901 The smoothness of ψ_s follows from the differentiability of G_ϕ and f_{θ_f} . Then, by Lemma 1, we can
902 obtain: $\|\psi_s(u) - \varphi_s^{-1}(u)\|_{\mathcal{H}} \leq L_f \alpha \varepsilon = \gamma \varepsilon$.

903 To construct a global diffeomorphism, take a finite open cover $\{\mathcal{N}(z_{s_i})\}_{i=1}^k$ of $\mathcal{M}_{\text{source}}$, with corre-
904 sponding charts φ_{s_i} and a smooth partition of unity $\{\rho_i\}$:

$$\Psi(z_s) = \sum_{i=1}^k \rho_i(z_s) \cdot \psi_{s_i}(\varphi_{s_i}(z_s)). \quad (18)$$

905 We now bound the total perturbation:

$$\begin{aligned}
\|\Psi(z_s) - z_s\|_{\mathcal{H}} &\leq \sum_{i=1}^k \rho_i(z_s) \|\psi_{s_i}(\varphi_{s_i}(z_s)) - z_s\|_{\mathcal{H}} \\
&\leq \sum_{i=1}^k \rho_i(z_s) L_f \alpha \varepsilon \\
&= L_f \alpha \varepsilon \\
&= \gamma \varepsilon
\end{aligned} \tag{19}$$

906 For any $z_t \in \mathcal{M}_{\text{triggered}}$, there exists $z_s \in \mathcal{M}_{\text{source}}$ such that $z_t = \Psi(z_s)$. By Assumption 3, there
907 exists $z_\tau \in \mathcal{M}_{\text{clean}}$ with $\|z_s - z_\tau\|_{\mathcal{H}} \leq \delta$. Then by the triangle inequality:

$$\begin{aligned}
\|z_t - z_\tau\|_{\mathcal{H}} &\leq \|z_t - z_s\|_{\mathcal{H}} + \|z_s - z_\tau\|_{\mathcal{H}} \\
&\leq L_f \alpha \varepsilon + \delta = \delta'
\end{aligned} \tag{20}$$

908 Hence, $\mathcal{M}_{\text{triggered}} \subset \mathcal{N}_{\delta'}(\mathcal{M}_{\text{clean}})$.

909 To verify Ψ is a diffeomorphism:

- 910 • Injectivity: Follows from local injectivity of each ψ_{s_i} and the partition of unity.
- 911 • Surjectivity: For any $z_t \in \mathcal{M}_{\text{triggered}}$, there exists $x \in \mathcal{T}_{y_s}$ such that $z_t = f_{\theta_f}(x + \alpha G_\phi(x)) =$
912 $\Psi(f_{\theta_f}(x))$.
- 913 • Smooth Inverse: Local inverses $\psi_{s_i}^{-1}$ exist by the inverse function theorem and can be
914 smoothly blended via $\{\rho_i\}$.

915 □

916 **Theorem 3** (Upper Bound on Feature-Manifold Deviation under Poisoning). *Let \mathcal{T}_{y_τ} denote the clean*
917 *target-class dataset and $\mathcal{T}_{\text{triggered}}$ the triggered (poisoned) dataset, with corresponding feature-space*
918 *distributions $P_{\mathcal{M}_{\text{clean}}}$ and $P_{\mathcal{M}_{\text{triggered}}}$, respectively. Define the mixed distribution as:*

$$P_{\mathcal{M}_{\text{mixed}}} = (1 - \rho)P_{\mathcal{M}_{\text{clean}}} + \rho P_{\mathcal{M}_{\text{triggered}}},$$

919 *where $\rho \in [0, 1]$ denotes the poisoning ratio. Under Assumptions 1, 2, and 3, the expected deviation*
920 *of samples from the mixed distribution to the target feature manifold satisfies:*

$$\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{mixed}}}} \left[\inf_{z_\tau \in \mathcal{M}_{\text{clean}}} \|z - z_\tau\|_{\mathcal{H}} \right] \leq \rho(\gamma \varepsilon + \delta), \tag{21}$$

921 *where \mathcal{H} is the RKHS associated with the feature encoder.*

922 *Proof.* By the linearity of expectation and the definition of $P_{\mathcal{M}_{\text{mixed}}}$, we have:

$$\begin{aligned}
&\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{mixed}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right] \\
&= (1 - \rho) \underbrace{\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{clean}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right]}_{=0} \\
&\quad + \rho \mathbb{E}_{z \sim P_{\mathcal{M}_{\text{triggered}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right].
\end{aligned} \tag{22}$$

923 Since clean samples $z \sim P_{\mathcal{M}_{\text{clean}}}$ lie on the target manifold, their distance minimum distance to the
924 target manifold is zero. Therefore:

$$\begin{aligned}
&\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{mixed}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right] \\
&= \rho \mathbb{E}_{z \sim P_{\mathcal{M}_{\text{triggered}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right].
\end{aligned} \tag{23}$$

925 By Lemma 2, for any $z_t \in \mathcal{M}_{\text{triggered}}$, there exists $z_\tau \in \mathcal{M}_{\text{clean}}$ such that:

$$\|z_t - z_\tau\|_{\mathcal{H}} \leq \delta' = \gamma\varepsilon + \delta. \quad (24)$$

926 Hence,

$$\inf_{z_\tau \in \mathcal{M}_{\text{clean}}} \|z_t - z_\tau\|_{\mathcal{H}} \leq \delta'. \quad (25)$$

927 Taking the expectation over $P_{\mathcal{M}_{\text{triggered}}}$, we obtain:

$$\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{triggered}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right] \leq \delta'. \quad (26)$$

928 Substituting into Eq.(22) yields:

$$\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{mixed}}}} \left[\inf_{z_\tau} \|z - z_\tau\|_{\mathcal{H}} \right] \leq \rho(\gamma\varepsilon + \delta). \quad (27)$$

929 □

930 **Theorem 4** (Upper Bound on the Discrepancy Between Poisoned and Clean Condensation Datasets).
 931 Let \mathcal{T}_{y_τ} denote the clean target-class dataset and $\mathcal{T}_{\text{mixed}} = \mathcal{T}_{y_\tau} \cup \mathcal{T}_{\text{triggered}}$, where $\mathcal{T}_{\text{triggered}}$ consists
 932 of source-class samples $x \in \mathcal{T}_{y_s}$ perturbed by a trigger generator G_ϕ and relabeled as the target
 933 class.

934 Let $\mathcal{S}_{\text{clean}}$ and $\mathcal{S}_{\text{poison}}$ denote the condensation datasets distilled from \mathcal{T}_{y_τ} and $\mathcal{T}_{\text{mixed}}$, respectively,
 935 by minimizing:

$$\mathcal{S}^* = \arg \min_{\mathcal{S}} \text{MMD}(\mathcal{T}, \mathcal{S}) + \lambda \mathcal{R}(\mathcal{S}), \quad (28)$$

936 where $\mathcal{T} \in \{\mathcal{T}_{y_\tau}, \mathcal{T}_{\text{mixed}}\}$, $\lambda > 0$, and \mathcal{R} is a strongly convex regularizer.

937 Under Assumptions 1, 2, and 3, the MMD between $\mathcal{S}_{\text{clean}}$ and $\mathcal{S}_{\text{poison}}$ satisfies:

$$\text{MMD}(\mathcal{S}_{\text{clean}}, \mathcal{S}_{\text{poison}}) \leq \frac{L_f^2 \rho(\gamma\varepsilon + \delta)}{\lambda \mu_R}$$

938 where $\gamma = L_f \alpha$, $\delta = \sup_{z_s \in \mathcal{M}_{\text{source}}} \inf_{z_\tau \in \mathcal{M}_{\text{clean}}} \|z_s - z_\tau\|_{\mathcal{H}}$, ρ is the poisoning rate, and ε bounds
 939 the input perturbation.

940 *Proof.* By Theorem 3:

$$\mathbb{E}_{z \sim P_{\mathcal{M}_{\text{mixed}}}} \left[\inf_{z_\tau \in \mathcal{M}_{\text{clean}}} \|z - z_\tau\|_{\mathcal{H}} \right] \leq \rho(\gamma\varepsilon + \delta). \quad (29)$$

941 This inequality constrains the average deviation of the mixed distribution from the clean target
 942 manifold by $\rho(\gamma\varepsilon + \delta)$.

943 In RKHS, MMD can be expressed via the norm of mean embeddings:

$$\text{MMD}(\mathcal{T}_{y_\tau}, \mathcal{T}_{\text{mixed}}) = \|\mu_{\text{clean}} - \mu_{\text{mixed}}\|_{\mathcal{H}}. \quad (30)$$

944 where

$$\begin{aligned} \mu_{\text{clean}} &= \mathbb{E}_{x \sim P_{\mathcal{T}_{y_\tau}}} [f_{\theta_f}(x)] \\ \mu_{\text{mixed}} &= \mathbb{E}_{x \sim P_{\mathcal{T}_{y_{\text{mixed}}}}} [f_{\theta_f}(x)] \end{aligned}$$

946 Using the decomposition, the mean embedding of the mixed distribution can be written as::

$$\mu_{\text{mixed}} = (1 - \rho)\mu_{\text{clean}} + \rho\mu_{\text{triggered}} \quad (31)$$

947 we get:

$$\mu_{\text{clean}} - \mu_{\text{mixed}} = \rho(\mu_{\text{clean}} - \mu_{\text{triggered}}) \quad (32)$$

948 Hence:

$$\begin{aligned} \text{MMD}(\mathcal{T}_{y_\tau}, \mathcal{T}_{\text{mixed}}) &= \rho \|\mu_{\text{clean}} - \mu_{\text{triggered}}\|_{\mathcal{H}} \\ &\leq \rho(\gamma\varepsilon + \delta) \end{aligned} \quad (33)$$

949 Let the clean and poisoned synthetic datasets, $\mathcal{S}_{\text{clean}}$ and $\mathcal{S}_{\text{poison}}$, be obtained by solving the following
950 optimization problems:

$$\begin{aligned} \mathcal{S}_{\text{clean}} &= \arg \min_{\mathcal{S}} \text{MMD}(\mathcal{T}_{y_\tau}, \mathcal{S}) + \lambda \mathcal{R}(\mathcal{S}), \\ \mathcal{S}_{\text{poison}} &= \arg \min_{\mathcal{S}} \text{MMD}(\mathcal{T}_{\text{mixed}}, \mathcal{S}) + \lambda \mathcal{R}(\mathcal{S}) \end{aligned} \quad (34)$$

951 According to the first-order optimality condition, the solutions $\mathcal{S}_{\text{clean}}$ and $\mathcal{S}_{\text{poison}}$ satisfy:

$$\begin{aligned} \nabla \text{MMD}_{\mathcal{S}}(\mathcal{T}_{y_\tau}, \mathcal{S}_{\text{clean}}) + \lambda \nabla \mathcal{R}(\mathcal{S}_{\text{clean}}) &= 0 \\ \nabla \text{MMD}_{\mathcal{S}}(\mathcal{T}_{\text{mixed}}, \mathcal{S}_{\text{poison}}) + \lambda \nabla \mathcal{R}(\mathcal{S}_{\text{poison}}) &= 0 \end{aligned} \quad (35)$$

952 Subtracting the optimality conditions:

$$\begin{aligned} \lambda(\nabla \mathcal{R}(\mathcal{S}_{\text{clean}}) - \nabla \mathcal{R}(\mathcal{S}_{\text{poison}})) &= \nabla \text{MMD}_{\mathcal{S}}(\mathcal{T}_{\text{mixed}}, \mathcal{S}_{\text{poison}}) \\ &\quad - \nabla \text{MMD}_{\mathcal{S}}(\mathcal{T}_{y_\tau}, \mathcal{S}_{\text{clean}}) \end{aligned} \quad (36)$$

953 Since \mathcal{R} is $\mu_{\mathcal{R}}$ -strongly convex, we obtain:

$$\begin{aligned} \langle \nabla \mathcal{R}(\mathcal{S}_{\text{clean}}) - \nabla \mathcal{R}(\mathcal{S}_{\text{poison}}), \mathcal{S}_{\text{clean}} - \mathcal{S}_{\text{poison}} \rangle \\ \geq \mu_{\mathcal{R}} \|\mathcal{S}_{\text{clean}} - \mathcal{S}_{\text{poison}}\|^2 \end{aligned} \quad (37)$$

954 Then, we can obtain:

$$\begin{aligned} &\|\mathcal{S}_{\text{clean}} - \mathcal{S}_{\text{poison}}\| \\ &\leq \frac{\|\nabla_{\mathcal{S}} \text{MMD}(\mathcal{T}_{y_\tau}, \mathcal{S}_{\text{clean}}) - \nabla_{\mathcal{S}} \text{MMD}(\mathcal{T}_{\text{mixed}}, \mathcal{S}_{\text{poison}})\|}{\lambda \mu_{\mathcal{R}}} \\ &\leq \frac{L_f \text{MMD}(\mathcal{T}_{y_\tau}, \mathcal{T}_{\text{mixed}})}{\lambda \mu_{\mathcal{R}}} \\ &\leq \frac{L_f \rho(\gamma\varepsilon + \delta)}{\lambda \mu_{\mathcal{R}}} \end{aligned} \quad (38)$$

955 According to Assumption 1:

$$\begin{aligned} \text{MMD}(\mathcal{S}_{\text{clean}}, \mathcal{S}_{\text{poison}}) &\leq L_f \|\mathcal{S}_{\text{clean}} - \mathcal{S}_{\text{poison}}\| \\ &\leq \frac{L_f^2 \rho(\gamma\varepsilon + \delta)}{\lambda \mu_{\mathcal{R}}}. \end{aligned} \quad (39)$$

956

□

957 C Additional Experiments

958 In dataset distillation, simple architectures such as ConvNet or AlexNetBN are typically employed
959 as distillation networks, rather than more complex models. This design choice is motivated by
960 several factors. First, computational efficiency and stability: simpler networks are faster and less
961 resource-intensive to train, which is essential given the iterative optimization cycles required in dataset
962 distillation. In contrast, deeper architectures substantially increase computational cost and introduce
963 greater instability during optimization. Second, optimization tractability: simple models possess
964 smoother and more navigable loss landscapes, facilitating the extraction of effective gradients from
965 synthetic data. Complex architectures, with highly non-convex objectives, complicate this process
966 and hinder optimization. Third, fairness and generality: the distilled data is intended to generalize
967 across a range of architectures. Relying on a highly specialized, deep network risks overfitting the
968 synthetic data to its unique characteristics. Employing a lightweight, generic model encourages the
969 generation of broadly transferable synthetic datasets.

To further substantiate the choice of AlexNetBN as the distillation network, we report additional experimental results in the appendix. While ConvNet is widely adopted in dataset distillation for its simplicity, AlexNetBN introduces greater depth and batch normalization, offering a complementary evaluation of the distilled data’s robustness and generalizability. These experiments assess whether the performance patterns observed with ConvNet persist under a moderately more complex architecture, thereby strengthening the evidence for the reliability of the distilled datasets.

C.1 Effectiveness on Different Datasets and Settings

Firstly, for completeness, we report the results of the Naive attack in Table 6.

Table 6: Effectiveness on Different Datasets

Dataset	Method	SNEAKDOOR		NAIVE	
		CTA	ASR	CTA	ASR
CIFAR10	DM	0.626±0.001	0.989±0.000	0.632±0.001	0.113±0.012
	DC	0.537±0.000	0.996±0.000	0.552±0.001	0.102±0.007
	IDM	0.643±0.002	0.975±0.001	0.652±0.001	0.103±0.006
	DAM	0.591±0.001	0.979±0.001	0.582±0.001	0.086±0.003
STL10	DM	0.598±0.001	0.973±0.000	0.621±0.001	0.103±0.006
	DC	0.565±0.001	0.998±0.001	0.583±0.001	0.090±0.007
	IDM	0.658±0.001	0.979±0.001	0.667±0.001	0.102±0.007
	DAM	0.532±0.001	0.992±0.001	0.549±0.001	0.088±0.009
FMNIST	DM	0.876±0.001	0.998±0.000	0.887±0.001	0.090±0.008
	DC	0.851±0.001	0.998±0.000	0.857±0.001	0.086±0.002
	IDM	0.877±0.001	1.000±0.000	0.887±0.001	0.093±0.007
	DAM	0.877±0.000	0.996±0.000	0.881±0.001	0.098±0.005
SVHN	DM	0.800±0.000	1.000±0.000	0.799±0.000	0.111±0.006
	DC	0.687±0.000	1.000±0.000	0.699±0.001	0.115±0.011
	IDM	0.831±0.001	0.986±0.001	0.840±0.000	0.122±0.010
	DAM	0.782±0.001	1.000±0.000	0.770±0.000	0.112±0.006
TINY IMAGENET	DM	0.503±0.001	1.000±0.000	0.497±0.002	0.070±0.002
	DC	0.432±0.002	1.000±0.000	0.421±0.002	0.019±0.001
	IDM	0.517±0.004	1.000±0.000	0.501±0.008	0.042±0.004
	DAM	0.482±0.003	1.000±0.000	0.462±0.003	0.042±0.002

Table 7 and 8 reports the ASR and CTA of different dataset distillation methods using AlexNetBN as the distillation network across multiple datasets. The results reveal how distilled data behaves under both clean and backdoor settings when applied to AlexNetBN. This provides a comprehensive view of each attack’s robustness and generalization in adversarial contexts.

Table 7: Effectiveness on Different Datasets condensed with AlexNetBN

Dataset	Method	SNEAKDOOR		NAIVE		DOORPING	
		CTA	ASR	CTA	ASR	CTA	ASR
CIFAR10	DM	0.595±0.001	0.947±0.004	0.608±0.002	0.093±0.011	0.505±0.001	1.000±0.000
	DC	0.222±0.001	0.003±0.001	0.140±0.001	0.000±0.000	0.319±0.007	0.000±0.000
	IDM	0.700±0.002	0.946±0.003	0.739±0.002	0.104±0.009	0.639±0.003	1.000±0.000
	DAM	0.606±0.001	0.721±0.013	0.609±0.001	0.096±0.010	0.565±0.001	1.000±0.000
STL10	DM	0.562±0.001	0.993±0.000	0.573±0.004	0.104±0.010	0.557±0.004	1.000±0.000
	DC	0.155±0.006	0.003±0.002	0.178±0.001	0.000±0.000	0.278±0.003	1.000±0.000
	IDM	0.723±0.002	0.986±0.002	0.729±0.003	0.100±0.007	0.646±0.003	1.000±0.000
	DAM	0.584±0.001	0.962±0.003	0.603±0.004	0.101±0.010	0.565±0.000	1.000±0.000
FMNIST	DM	0.822±0.000	1.000±0.000	0.844±0.001	0.090±0.010	0.636±0.005	1.000±0.000
	DC	0.287±0.000	0.000±0.000	0.172±0.003	0.320±0.018	0.516±0.010	1.000±0.000
	IDM	0.844±0.001	0.978±0.002	0.858±0.001	0.113±0.003	0.736±0.001	1.000±0.000
	DAM	0.831±0.003	1.000±0.000	0.821±0.002	0.100±0.003	0.758±0.003	1.000±0.000
SVHN	DM	0.622±0.020	1.000±0.000	0.697±0.007	0.124±0.006	0.774±0.001	1.000±0.000
	DC	0.108±0.001	0.984±0.001	0.095±0.001	0.069±0.010	0.379±0.006	1.000±0.000
	IDM	0.880±0.001	0.966±0.001	0.886±0.001	0.116±0.010	0.781±0.002	1.000±0.000
	DAM	0.672±0.006	0.999±0.000	0.701±0.002	0.112±0.008	0.593±0.003	1.000±0.000
TINY IMAGENET	DM	0.463±0.002	0.920±0.013	0.457±0.003	0.011±0.002	0.485±0.002	1.000±0.000
	DC	0.247±0.003	1.000±0.000	0.269±0.005	0.013±0.003	0.260±0.004	0.000±0.000
	IDM	0.260±0.005	0.860±0.013	0.284±0.007	0.000±0.000	0.293±0.006	1.000±0.000
	DAM	0.442±0.006	0.972±0.010	0.430±0.013	0.010±0.001	0.419±0.010	1.000±0.000

Table 8: Effectiveness on Different Datasets condensed with AlexNetBN

Dataset	Method	SNEAKDOOR		SIMPLE		RELAX	
		CTA	ASR	CTA	ASR	CTA	ASR
CIFAR10	DM	0.595±0.001	0.947±0.004	0.581±0.001	0.183±0.013	0.603±0.001	0.704±0.022
	DC	0.222±0.001	0.003±0.001	0.169±0.002	0.000±0.000	0.152±0.001	0.047±0.018
	IDM	0.700±0.002	0.946±0.003	0.727±0.001	0.146±0.009	0.252±0.002	0.636±0.024
	DAM	0.606±0.001	0.721±0.013	0.584±0.001	0.204±0.024	0.591±0.002	0.978±0.004
STL10	DM	0.562±0.001	0.993±0.000	0.544±0.002	0.092±0.007	0.550±0.003	0.706±0.010
	DC	0.155±0.006	0.003±0.002	0.121±0.008	0.117±0.013	0.144±0.003	0.574±0.036
	IDM	0.723±0.002	0.986±0.002	0.724±0.003	0.102±0.013	0.719±0.002	0.668±0.029
	DAM	0.584±0.001	0.962±0.003	0.568±0.003	0.098±0.010	0.566±0.005	0.872±0.022
FMNIST	DM	0.822±0.000	1.000±0.000	0.812±0.006	0.952±0.009	0.816±0.003	1.000±0.000
	DC	0.287±0.000	0.000±0.000	0.161±0.001	0.895±0.018	0.171±0.001	0.646±0.033
	IDM	0.844±0.001	0.978±0.002	0.849±0.001	0.231±0.028	0.856±0.001	0.719±0.015
	DAM	0.831±0.003	1.000±0.000	0.806±0.002	0.482±0.128	0.811±0.002	1.000±0.000
SVHN	DM	0.622±0.020	1.000±0.000	0.484±0.010	0.071±0.005	0.672±0.009	0.978±0.007
	DC	0.108±0.001	0.984±0.001	0.157±0.006	0.060±0.006	0.137±0.004	0.119±0.027
	IDM	0.880±0.001	0.966±0.001	0.880±0.001	0.118±0.008	0.874±0.001	1.000±0.001
	DAM	0.672±0.006	0.999±0.000	0.693±0.006	0.092±0.007	0.692±0.003	0.996±0.003
TINY IMAGENET	DM	0.463±0.002	0.920±0.013	0.457±0.003	0.011±0.002	0.449±0.003	0.835±0.017
	DC	0.247±0.003	1.000±0.000	0.200±0.008	0.000±0.000	0.259±0.002	0.471±0.023
	IDM	0.260±0.005	0.860±0.013	0.337±0.006	0.053±0.008	0.313±0.007	0.759±0.058
	DAM	0.442±0.006	0.972±0.010	0.443±0.007	0.013±0.002	0.441±0.004	0.787±0.027

Moreover, we have expanded our evaluation in two key directions: (1) *incorporating a larger, higher-resolution dataset*, ImageNette (resolution $3 \times 224 \times 224$), as shown in Table 9, and (2) *evaluating key parameters* on STL10 (resolution $3 \times 96 \times 96$), including *ipc* (the number of synthetic samples per class), *perturbation bound* ε , and *poisoning ratio*, as shown in Table 10, 11, and 12.

Table 9 reports SNEAKDOOR’s attack performance under DM and DAM on the ImageNette dataset, demonstrating that **SNEAKDOOR remains effective on higher-resolution, larger-scale data**. Due to computational resources constraints, we could not include results for DC and IDM, as a single run with DC or IDM takes about three to four days, making full tuning impractical. We plan to include these results in a future version to provide a more complete picture of performance across algorithms and settings.

Table 9: Attack Performance of SNEAKDOOR on the ImageNette Dataset.

Method	ASR	CTA	PNSR	SSIM	IS
DM	0.9809±0.0000	0.5625±0.0007	68.62	0.6673	2.25e-4
DAM	0.9429±0.0008	0.4598±0.0003	72.16	0.6814	2.08e-4

Table 10: Impact of IPC on Attack Performance

Method	ipc	ASR	CTA	PNSR	SSIM	IS
DM	10	0.8735±0.0009	0.4347±0.0003	73.0381	0.8211	9.05e-5
DM	20	0.9872±0.0005	0.4882±0.0008	73.5021	0.7950	1.32e-4
DM	50	0.9725±0.0000	0.5979±0.0006	70.1216	0.8066	1.41e-4
IDM	10	0.9778±0.0015	0.5965±0.0004	74.1393	0.8199	1.05e-4
IDM	20	0.9573±0.0009	0.6217±0.0006	73.9608	0.8049	2.39e-4
IDM	50	0.9790±0.0009	0.6582±0.0005	70.1548	0.7554	1.40e-4
DAM	10	0.8910±0.0015	0.3678±0.0006	73.6366	0.8106	9.21e-5
DAM	20	0.8902±0.0025	0.4522±0.0004	73.8535	0.8146	9.22e-5
DAM	50	0.9918±0.0006	0.5324±0.0007	73.7877	0.8245	9.14e-5
DC	10	0.9258±0.0035	0.4675±0.0006	73.1598	0.8072	9.54e-5
DC	20	0.9243±0.0035	0.5282±0.0002	73.0987	0.8018	9.05e-5
DC	50	0.9975±0.0008	0.5653±0.0011	71.2365	0.7550	7.26e-5

As shown in Table 10, varying ipc notably affects CTA, while ASR and STE metrics (PSNR, SSIM, IS) remain relatively stable. This is expected, as fewer samples per class reduce the fidelity of clean distribution modeling, impacting generalization. In contrast, ASR stays high across ipc values, indicating that once embedded, the backdoor remains effective even with limited data. STE metrics also show minimal change, suggesting the perturbations remain visually subtle and robust.

As shown in Table 11, increasing the perturbation bound ε improves ASR but reduces STE, as reflected in lower PSNR, SSIM, and IS. This is expected, since a larger ε allows stronger and more

noticeable triggers, enhancing attack success at the expense of stealth. Notably, CTA remains stable across ε values, indicating that stronger triggers do not significantly harm generalization on clean data. These results highlight a trade-off between ASR and STE controlled by ε .

Table 11: Impact of Perturbation Bound ε on Attack Performance

Method	ε	ASR	CTA	PSNR	SSIM	IS
DM	0.1	0.7755±0.0049	0.6045±0.0009	82.1241	0.9548	2.97e-5
DM	0.2	0.9332±0.0006	0.5824±0.0008	76.9565	0.8769	5.46e-5
DM	0.3	0.9732±0.000	0.5981±0.0010	74.0076	0.7963	6.32e-5
IDM	0.1	0.5400±0.0076	0.6627±0.0010	78.7475	0.7914	1.14e-4
IDM	0.2	0.7905±0.0073	0.6624±0.0013	76.4274	0.7931	1.30e-4
IDM	0.3	0.9790±0.0009	0.6582±0.0005	70.1548	0.8054	1.40e-4
DAM	0.1	0.6785±0.0022	0.5278±0.0012	82.0221	0.9594	3.06e-5
DAM	0.2	0.8715±0.0015	0.5389±0.0007	76.8882	0.8916	5.51e-5
DAM	0.3	0.9918±0.0006	0.5324±0.0007	73.7877	0.8245	9.14e-5
DC	0.1	0.6128±0.004	0.5743±0.0002	78.8841	0.7633	7.54e-5
DC	0.2	0.7828±0.0056	0.58±0.0011	73.3082	0.5337	1.06e-4
DC	0.3	0.9980 ± 0.0010	0.5650±0.0010	71.2365	0.5551	7.25e-5

Table 12: Impact of Poisoning Ratio on Attack Performance

Method	poison ratio	ASR	CTA	PSNR	SSIM	IS
DM	0.10	0.8810±0.0020	0.5986±0.001	74.0086	0.8285	8.82e-5
DM	0.25	0.8970±0.0019	0.6009±0.0009	73.7735	0.7942	9.55e-5
DM	0.5	0.9725±0.0000	0.5979±0.0006	73.0076	0.7963	1.14e-4
IDM	0.10	0.8205±0.0026	0.6645±0.0015	74.0362	0.7803	2.61e-4
IDM	0.25	0.8615±0.0044	0.6592±0.0007	70.2375	0.7788	1.33e-4
IDM	0.5	0.9790±0.0009	0.6582±0.0005	70.1548	0.7554	1.40e-4
DAM	0.10	0.5073±0.0035	0.5526±0.0003	74.2949	0.8200	8.10e-5
DAM	0.25	0.7820±0.0017	0.5488±0.0006	73.5737	0.8429	1.11e-4
DAM	0.5	0.9918±0.0006	0.5324±0.0007	73.7877	0.8245	9.14e-5
DC	0.10	0.7912±0.0041	0.5745±0.0007	69.7258	0.5573	1.32e-4
DC	0.25	0.8627±0.0031	0.5851±0.0005	70.4030	0.5113	1.49e-4
DC	0.5	0.9980±0.0010	0.5650±0.0010	71.2365	0.5551	7.25e-5

As shown in Table 12, increasing the poisoning ratio improves the ASR, which aligns with the intuition that more poisoned samples enhance the trigger’s influence in the condensed dataset. However, this improvement comes with a slight degradation in CTA. Interestingly, the decline in CTA is relatively limited even at higher poisoning ratios (e.g., 0.5), suggesting that the trigger’s interference with the clean distribution remains modest. Nevertheless, the reliance on a relatively high poisoning ratio to achieve optimal attack effectiveness highlights a limitation of the current approach.

C.2 Stealthiness on CIFAR10, SVHN, and FMNIST

We have included stealthiness for the remaining datasets, *i.e.*, CIFAR10, SVHN, and FMNIST. These additional results offer a comprehensive assessment of SNEAKDOOR’s visual imperceptibility across diverse datasets. Notably, we omit the Inception Score (IS) evaluation for FMNIST because it is a single-channel (grayscale) dataset, which is incompatible with the standard IS computation that relies on a pre-trained Inception network trained on RGB images. Applying IS directly to grayscale data would yield unreliable and uninformative results.

C.3 Effectiveness on Cross Architectures

We further include cross-architecture evaluations with AlexNetBN. This setting tests the transferability of the backdoor attack to a moderately different network from the distillation model. The results offer additional evidence of the generalization and robustness of SNEAKDOOR across architectures. This property is critical for practical deployment in real-world scenarios.

C.4 Visual Analysis of Trigger Stealthiness

We provide visualizations of original images after injecting the trigger during inference. Figure 5 illustrates the effect following trigger injection. The images demonstrate the trigger’s subtlety and

Table 13: PSNR, SSIM, and IS on CIFAR10, SVHN, and FMNIST

Method	Backdoor	CIFAR-10			SVHN			FMNIST		
		PSNR	SSIM	IS	PSNR	SSIM	IS	PSNR	SSIM	IS
DM	SNEAKDOOR	73.94	0.61	5.80e-05	74.68	0.77	3.90e-05	58.41	0.39	–
	Doorping	59.85	0.08	2.30e-04	60.27	0.08	2.08e-04	55.68	0.12	–
	Relax	60.97	-0.01	2.48e-04	61.47	-0.14	2.45e-04	51.88	-0.07	–
	naive	63.67	0.15	3.56e-04	62.27	0.10	4.60e-04	54.15	0.10	–
	Simple	60.98	0.69	8.10e-05	61.59	0.74	7.95e-05	54.01	0.00	–
DC	SNEAKDOOR	70.48	0.46	7.10e-05	73.15	0.42	8.10e-05	57.39	0.24	–
	Doorping	59.22	0.05	2.43e-04	61.25	0.06	2.00e-04	60.11	0.52	–
	Relax	61.37	0.04	2.38e-04	62.17	-0.04	2.43e-04	52.15	-0.11	–
	naive	64.46	0.18	3.62e-04	60.45	0.04	4.92e-04	54.21	0.06	–
	Simple	60.74	0.66	8.70e-05	61.44	0.72	8.08e-05	53.99	0.00	–
IDM	SNEAKDOOR	74.88	0.77	4.40e-05	72.19	0.68	6.30e-05	57.16	0.10	–
	Doorping	59.23	0.10	2.23e-04	59.66	0.06	2.17e-04	57.26	0.06	–
	Relax	61.18	0.02	2.46e-04	61.17	-0.20	2.70e-04	52.04	-0.08	–
	naive	64.23	0.14	3.44e-04	62.05	0.07	5.02e-04	54.15	0.05	–
	Simple	61.05	0.69	8.60e-05	61.21	0.70	8.00e-05	54.23	0.00	–
DAM	SNEAKDOOR	74.40	0.74	4.50e-05	78.91	0.74	4.30e-05	57.39	0.24	–
	Doorping	59.52	0.08	1.62e-04	59.67	0.08	1.05e-04	57.16	0.10	–
	Relax	61.19	0.02	2.31e-04	62.36	-0.24	2.04e-04	51.83	-0.10	–
	naive	62.99	0.13	4.53e-04	60.43	0.04	5.39e-04	55.07	0.12	–
	Simple	60.85	0.64	8.70e-05	61.78	0.75	7.95e-05	54.07	0.00	–

Table 14: Cross-architecture CTA and ASR condensed with AlexNetBN

Dataset	Network	DM		DC		IDM		DAM	
		CTA	ASR	CTA	ASR	CTA	ASR	CTA	ASR
CIFAR10	VGG11	0.544±0.000	0.961±0.000	0.209±0.000	0.009±0.000	0.673±0.000	0.945±0.001	0.542±0.000	0.733±0.001
	ResNet	0.495±0.001	0.915±0.002	0.186±0.000	0.009±0.000	0.671±0.001	0.926±0.001	0.500±0.001	0.491±0.001
	ConvNet	0.585±0.001	0.807±0.002	0.216±0.001	0.004±0.001	0.638±0.001	0.951±0.002	0.582±0.001	0.457±0.005
STL10	VGG11	0.527±0.001	0.921±0.000	0.195±0.001	0.012±0.001	0.694±0.000	0.947±0.002	0.547±0.001	0.924±0.002
	ResNet	0.413±0.001	0.999±0.000	0.160±0.001	0.011±0.001	0.644±0.001	0.991±0.001	0.445±0.002	0.995±0.000
	ConvNet	0.532±0.000	0.841±0.002	0.180±0.000	0.152±0.005	0.693±0.001	0.828±0.011	0.555±0.001	0.997±0.001
TINY IMAGENET	VGG11	0.427±0.001	0.920±0.000	0.174±0.002	0.860±0.000	0.435±0.003	0.588±0.024	0.437±0.002	0.960±0.000
	ResNet	0.361±0.002	0.800±0.000	0.227±0.002	0.716±0.008	0.228±0.004	0.360±0.036	0.391±0.002	1.000±0.000
	ConvNet	0.443±0.003	0.604±0.008	0.217±0.003	0.932±0.010	0.335±0.009	0.604±0.015	0.430±0.004	0.884±0.015

stealthiness. Changes to the original images are minimal and barely perceptible. Despite this, the trigger effectively activates the backdoor in the model. These visual results emphasize the challenge of detecting such backdoors through simple inspection. They also underscore the importance of robust defenses against stealthy triggers.

C.5 Hyper-parameter Settings

We have provided the full set of optimization hyperparameters used for SNEAKDOOR on the STL10 dataset across four condensation baselines: DM, DC, IDM, and DAM, including learning rates, number of epochs, batch sizes, etc. These details are listed in Tab.5 - Tab.8, allowing replication of our experiments. In addition, we will release the full source code in a future version of the paper. This will include the complete training pipeline for both the trigger generator and dataset condensation procedures. Our goal is to ensure that the community can easily reproduce and extend our work.

The overall method is divided into four stages:

1. Training the Surrogate Model. The surrogate model serves two key purposes: (i) estimating inter-class boundary vulnerability (ICBV), and (ii) guiding the training of the trigger generator.
2. Training the Trigger Generator G_ϕ . The generator learns to produce input-aware perturbations that cause misclassification.
3. Malicious Condensation. This phase incorporates the trigger signal into the synthetic dataset via a standard condensation framework.
4. Downstream Model Training. Standard training on the poisoned condensed dataset using typical optimization settings.



Figure 5: STL10 Stealthiness Illustration

Table 15: Hyperparameters for Surrogate Model Training

Hyperparameter	Value
Optimizer	SGD
Batch size	256
Learning rate	0.01
Momentum	0.9
Weight decay	0.0005
Epochs	50

Table 16: Hyperparameters for Trigger Generator Training

Hyperparameter	Value
Learning rate	5e-5
Perturbation scaling factor α	0.25
Maximum perturbation bound ε	0.5

Table 17: Hyperparameters for Malicious Dataset Condensation

Hyperparameter	Value
Images per class (IPC)	50
Condensation epochs	20000
Synthesis learning rate	1.0
Batch size	256
Optimizer	Adam

Table 18: Hyperparameters for Downstream Model Training

Hyperparameter	Value
Optimizer	SGD
Batch size	256
Learning rate	0.01
Momentum	0.9
Weight decay	0.0005
Epochs	10000