ANCHOR IN EVIDENCE: DISTRIBUTION-PRESERVING BACKDOOR ATTACKS IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (FL) enables collaborative model training without sharing local data, but its decentralized nature also makes it vulnerable to backdoor attacks. Among these attacks, in-distribution (ID) backdoor attacks are particularly attractive because they preserve distribution consistency and are difficult to detect. However, we observe that existing ID backdoor attacks become highly unstable under heterogeneous client distributions, where attack success can abruptly deteriorate during federated training. We identify this previously overlooked phenomenon as *Attack Collapse* and show that it arises from the continuously evolving feature space in FL, which progressively misaligns injected backdoor representations with the target-class distribution. Motivated by this observation, we propose IDEA, a distribution-preserving backdoor attack that maintains stable target alignment throughout federated training. IDEA first leverages evidential deep learning to identify reliable target-class samples and construct robust target regions. It then jointly performs region-aware alignment and evidential geometric refinement to preserve both target affinity and local semantic structure, enabling poisoned samples to remain compatible with the evolving feature space. Extensive experiments on multiple benchmark datasets demonstrate that IDEA consistently mitigates Attack Collapse and significantly improves attack efficiency and stability over state-of-the-art backdoor attacks, while maintaining competitive clean-task performance.

1 INTRODUCTION

Federated Learning (FL) (Konečnỳ et al., 2016; McMahan et al., 2017) enables collaborative model training across multiple clients without sharing local data, making it a widely adopted paradigm for privacy-preserving machine learning (Hanser et al., 2025; Ye et al., 2024; Li et al., 2024). However, its decentralized optimization process is inherently vulnerable to malicious or compromised clients that can poison local updates and manipulate the global model (Zhang et al., 2023; Xie et al., 2026). Among various poisoning threats, backdoor attacks are particularly attractive because they can implant attacker-specified behaviors into the global model while maintaining high performance on benign inputs, making them difficult to detect (Li et al., 2025).

Despite recent advances in backdoor attacks, existing methods often generate poisoned samples that exhibit out-of-distribution (OOD) property relative to clean data (Jia et al., 2022; Liu & Zhang, 2025). Such distribution mismatch makes poisoned representations more distinguishable from benign ones, increasing the risk of backdoor detection (Xun et al., 2025; Zhang et al., 2025). To improve attack stealthiness, recent studies have shifted toward in-distribution (ID) backdoor attacks, which constrain poisoned samples to remain consistent with the clean data distribution while maintaining attack effectiveness (Tao et al., 2024; Zhang et al., 2024). DRUPE (Tao et al., 2024) first demonstrated the feasibility of this paradigm, and Mirage (Li et al., 2025) further extended it to FL, highlighting the potential of ID backdoor attacks in heterogeneous federated settings.

However, current ID backdoor attacks in FL (Li et al., 2025) have largely overlooked the stability challenges introduced by heterogeneous client data. Specifically, under heterogeneous client distributions, we observe a previously overlooked phenomenon, termed *Attack Collapse*, where attack effectiveness abruptly deteriorates during training, manifested as a sharp drop in attack success rate (ASR). As shown in Fig. 1, Attack Collapse becomes increasingly frequent and severe as client het-

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.

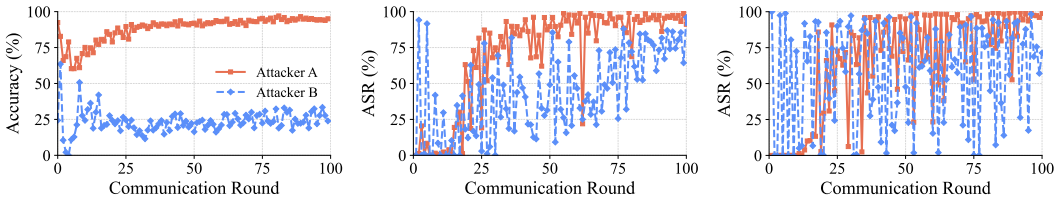


Figure 1: ID backdoor instability over communication rounds under heterogeneous FL with different Dirichlet concentration parameters κ. From left to right: task accuracy with $\kappa = 1$, backdoor ASR with $\kappa = 1$, and backdoor ASR with $\kappa = 0.5$. Lower κ indicates stronger data heterogeneity.

erogeneity increases, suggesting that ID backdoors are particularly vulnerable to the continuously evolving feature space in FL. Interestingly, we further find that stable attack performance consistently accompanies stable target-class semantics, indicating that maintaining stable alignment with the target distribution is crucial for sustaining effective ID backdoors throughout federated training.

To understand the root cause of Attack Collapse, we investigate how ID backdoors interact with the evolving feature space in FL. An effective ID backdoor relies on maintaining alignment between injected backdoor representations and the target-class feature distribution. However, unlike centralized training, heterogeneous local updates in FL continuously reshape the global feature space across communication rounds, causing the target distribution itself to evolve over time. As a result, backdoor representations that are well aligned in earlier rounds can gradually become misaligned with the evolving target distribution, weakening backdoor propagation and ultimately leading to Attack Collapse.

These observations and analyses lead to three key requirements for an effective ID backdoor attack in FL: (1) the injected backdoor should induce the desired target behavior on triggered inputs; (2) the backdoored model should preserve clean-model behavior on benign inputs; and (3) the injected backdoor should remain aligned with the evolving feature space to ensure stable propagation across communication rounds. Collectively, they point to a central challenge: *how to maintain effective and stable ID backdoor behavior under heterogeneous client distributions despite the continuously evolving feature space of FL.*

To address this challenge, we propose **IDEA** (**I**n-**D**istribution **E**vidential Backdoor **A**ttack), a stable distribution-preserving backdoor attack for FL. First, IDEA identifies reliable target-class samples using Evidential Deep Learning (EDL), based on prediction concentration and opinion conflict. These samples are then used as anchors to construct structured target regions that provide robust references for attack optimization. Second, it jointly leverages region-aware alignment to attract poisoned samples toward compatible region and evidential geometric refinement to preserve target affinity and separation from nearby non-target semantics. Therefore, by combining these complementary constraints, IDEA enables trigger optimization to produce stable distribution-preserving backdoors throughout heterogeneous federated training.

- We identify and systematically characterize Attack Collapse, a previously overlooked phenomenon in ID backdoor attacks under heterogeneous FL. To the best of our knowledge, this work represents the first attempt to explicitly investigate and mitigate this instability.
- We propose **IDEA**, a stable distribution-preserving backdoor attack for FL, which combines target-region attraction with evidential geometric refinement to guide poisoned samples toward reliable target regions while preserving their relationships with neighboring samples, thereby enabling stable poisoning under heterogeneous federated training.
- We conduct extensive experiments across diverse datasets and heterogeneous federated settings. The results demonstrate that IDEA consistently improves attack effectiveness and backdoor stability while preserving competitive clean-task performance.

2 Threat Model

Problem Statement. We consider a FL system with K clients $\mathcal{C} = \{C_k\}_{k=1}^{K}$, where each client C_k privately holds a local dataset $\mathcal{D}_k \sim P_k$. The local data distributions $\{P_k\}_{k=1}^{K}$ differ due to data heterogeneity. Among these clients, $\mathcal{C}_{\text{att}} \subseteq \mathcal{C}$ denotes a set of N non-colluding malicious clients.

Each malicious client $C_n \in \mathcal{C}_{\text{att}}$ independently specifies a target class c_n and selects a source sample set $\mathcal{X}_n \subseteq \mathcal{D}_n$, from which a trigger δ_n is learned. The trigger is injected into each source sample $x_i \in \mathcal{X}_n$ to construct its triggered counterpart $\tilde{x}_i = x_i \oplus \delta_n$, where $\oplus$ denotes the trigger injection operation. The trigger is optimized to achieve the attacker's predefined objectives.

Attack Objective. The attacker aims to implant an ID backdoor into the global model through federated training process while preserving benign functionality and maintaining attack stealth and persistence. Specifically, the attack seeks to: (1) achieve a high ASR on the target class, (2) preserve benign model performance, (3) maintain the in-distribution characteristics of triggered samples, and (4) sustain a high ASR throughout federated training under heterogeneous client distributions.

Attack Knowledge and Capabilities. We assume that each malicious client is an FL participant, allowing the attacker to control its local data and training process. Specifically, the attacker has access to its local dataset and the local training procedure, and can manipulate selected samples and their labels during local optimization. Beyond its own local information, the attacker has no knowledge of the private data or training details of benign clients or other malicious clients, including their data distributions, source samples, or attack objectives.

3 PRELIMINARIES

3.1 ID BACKDOOR ATTACK

To construct triggered counterparts that induce the desired target behavior while preserving their ID characteristics, ID backdoor mapping is formulated from two complementary perspectives: Target Mapping and Distribution Mapping.

Target Mapping. To map source samples to the attacker-specified target while facilitating the transition from source representations toward the target region, malicious client C_n optimizes the following target-oriented objective over its selected source sample set $\mathcal{X}_n$:

$$\mathcal{L}_{\text{target}} = \frac{1}{|\mathcal{X}_n|} \sum_{x_i \in \mathcal{X}_n} \left[\text{CE}\left(g_{\theta_g}(\tilde{x}_i), \tilde{y}_i\right) + \cos\left(f_{\theta_g}(\tilde{x}_i), f_{\theta_g}(x_i)\right) \right], \tag{1}$$

where $\tilde{y}_i$ denotes the target label corresponding to the attacker-specified target class c_n for x_i, while $g_{\theta_g}(\cdot)$ and $f_{\theta_g}(\cdot)$ denote the logits output and feature representation of the global model, respectively.

Distribution Mapping. To align triggered samples with the target distribution, a binary discriminator D_{ψ_n} is introduced to distinguish triggered samples from clean samples. The feature extractor of the global model $f_{\theta_g}(\cdot)$ is frozen, with only the binary classification head being optimized:

$$\mathcal{L}_{\text{disc}} = \frac{1}{|\mathcal{D}_n^{\text{disc}}|} \sum_{(x_j, y_j) \in \mathcal{D}_n^{\text{disc}}} \text{CE}\left(D_{\psi_n}(x_j), y_j\right), \tag{2}$$

where $\mathcal{D}_n^{\text{disc}} = \{(x \oplus \delta_n, 1) \mid x \in \mathcal{X}_n\} \cup \{(x, 0) \mid x \in \mathcal{D}_n^{c_n}\}$, with $\mathcal{D}_n^{c_n}$ denoting clean samples from the target class c_n.

After optimizing the discriminator, its parameters ψ_n^* are fixed, and the trigger is optimized to reduce the distinguishability between triggered and clean target-class samples:

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\mathcal{X}_n|} \sum_{x_i \in \mathcal{X}_n} \text{CE}\left(D_{\psi_n^*}(\tilde{x}_i), 0\right). \tag{3}$$

The overall ID backdoor attack objective is

$$\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{target}} + \mathcal{L}_{\text{dist}}. \tag{4}$$

3.2 DIRICHLET-BASED EVIDENTIAL MODELING

Dirichlet-based EDL (Sensoy et al., 2018) models predictive uncertainty through class-specific evidence. For M classes, a deep neural network maps an input to non-negative evidence $\mathbf{e} = (e_1, \ldots, e_M)^\top$, which parameterizes a Dirichlet distribution over the class probability simplex.

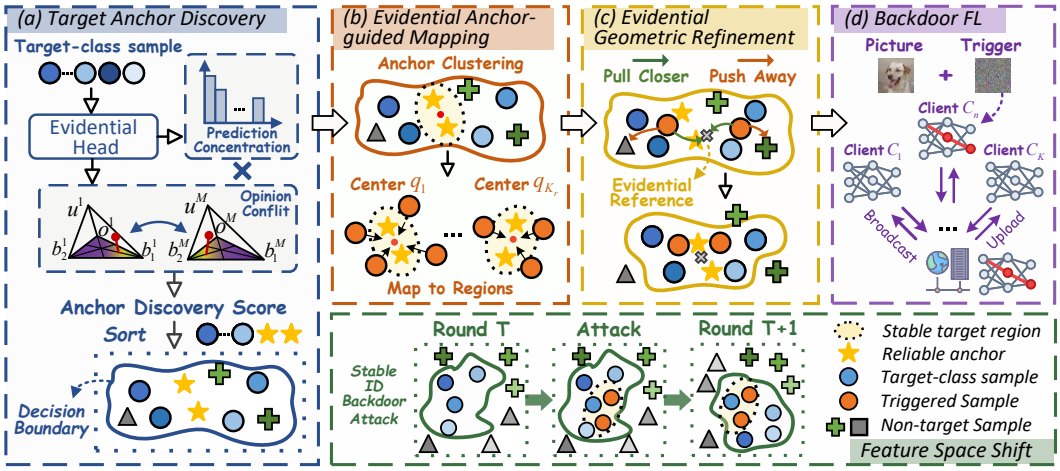


Figure 2: Illustration of IDEA. The attacker first evaluates target-class samples to select reliable anchors based on their ADS. These anchors are clustered into structured target regions, with representative centers (e.g., $\mathbf{q}_1$) guiding each triggered sample toward its most compatible center. The triggered samples are then refined through target alignment and non-target separation. The optimized poisoned samples are used by malicious clients (e.g., $\mathcal{C}_n$) to train a backdoored model, which is subsequently used to poison the global model. Through this process, IDEA maintains a stable ID backdoor attack across federated rounds despite the evolving feature space.

Definition 1 (Dirichlet Distribution)**.** *For M classes, the Dirichlet distribution is defined over the probability simplex $\Delta_p^{M-1} = \left\{\mathbf{p} \in \mathbb{R}^M \;\middle|\; \sum_{m=1}^M p_m = 1,\ p_m \geq 0\right\}$, and parameterized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^\top$ with $\alpha_m > 0$. Its probability density function is*

$$\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})} \prod_{m=1}^M p_m^{\alpha_m - 1}, & \mathbf{p} \in \Delta_p^{M-1}, \\ 0, & \textit{otherwise}, \end{cases} \tag{5}$$

where $B(\boldsymbol{\alpha})$ denotes the multivariate beta function.

In EDL, the class-specific evidence parameterizes the Dirichlet distribution as $\boldsymbol{\alpha} = \mathbf{e} + \mathbf{1}$, with the corresponding Dirichlet strength $S = \sum_{m=1}^M \alpha_m = \|\mathbf{e}\|_1 + M$. Following the evidence-based interpretation of Dempster–Shafer Theory (Dempster, 1968; Shafer, 1976) and Subjective Logic (Jsang, 2018), the prediction is modeled as a multinomial opinion.

Definition 2 (Multinomial Opinion)**.** *A multinomial opinion $\mathbf{w} = (\mathbf{b}, u, \mathbf{a})$ consists of a belief-mass vector $\mathbf{b} = (b_1, \ldots, b_M)^\top$, an uncertainty mass u, and a base-rate distribution $\mathbf{a} = (a_1, \ldots, a_M)^\top$, subject to*

$$\sum_{m=1}^M b_m + u = 1, \qquad b_m, u \geq 0, \qquad \sum_{m=1}^M a_m = 1. \tag{6}$$

Under Dirichlet-based EDL, the belief mass, uncertainty mass, and predictive probability are given by

$$b_m = \frac{e_m}{S}, \qquad u = \frac{M}{S}, \qquad p_m = b_m + a_m u. \tag{7}$$

With a uniform base rate $a_m = 1/M$, the predictive probability reduces to $p_m = \alpha_m / S$.

4 Methodology

4.1 Motivation of IDEA

The evolving feature space poses a fundamental challenge to maintaining consistent ID alignment during federated training, as feature representations can vary substantially across communication

rounds and clients. During poisoning, however, the attacker cannot observe the feature space in the subsequent participating round, making alignment with instantaneous target representations potentially unstable. This motivates a key intuition: stable target references and the underlying geometry of the feature space can provide a robust basis for guiding poisoned samples toward the target regions despite representation shifts, as illustrated by the attack effectiveness in Fig. 3.

RQ1: How can reliable target regions be identified? Unlike existing ID poisoning methods that derive target representations from global feature spaces observed in previous rounds, IDEA instead identifies target-specific structures that are more robust to representation shifts. Specifically, target samples exhibiting strong predictive and evidential consistency in their local neighborhoods are selected as anchors to characterize the target regions, providing stable references for subsequent target-oriented poisoning.

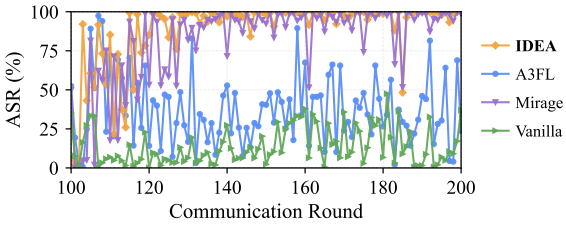


Figure 3: ASR evolution of different attack methods.

RQ2: How can poisoned samples be mapped toward the identified target regions? Given the anchor-based references, IDEA guides poisoned samples toward the identified target regions while preserving their evidential semantics. To reduce ambiguity with competing non-target structures, IDEA jointly exploits evidential cues to promote target alignment and non-target separation, enabling fine-grained target-oriented poisoning.

4.2 Target Anchor Discovery

To assess the suitability of target-class samples as anchors, we characterize their local neighborhoods in terms of predictive and evidential consistency. Specifically, for each target-class sample x_i, we define $\mathcal{N}_i$ as its set of k nearest neighbors and evaluate it from two complementary perspectives: prediction concentration and opinion conflict.

Prediction concentration. To quantify the concentration of evidential predictions within the local neighborhood $\mathcal{N}_i$, the prediction concentration metric measures the extent to which neighboring predictions favor a dominant class:

$$\mathrm{Conc}(x_i) = \frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \log\left(\frac{S_j}{\max_m \alpha_{jm}} - 1\right), \tag{8}$$

where α_{jm} denotes the Dirichlet parameter of sample x_j for class m, and $S_j = \sum_{m=1}^{M} \alpha_{jm}$ denotes its total evidential strength. A smaller value indicates stronger concentration toward a dominant class, suggesting stronger local agreement with the target semantics.

Opinion conflict. To quantify evidential inconsistency within the local neighborhood, the opinion conflict metric measures the discrepancy between the opinions of x_i and its neighbors:

$$\mathrm{Conf}(x_i) = \frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \left\| \frac{\boldsymbol{\alpha}_i}{S_i} - \frac{\boldsymbol{\alpha}_j}{S_j} \right\|_1 (1-u_i)(1-u_j), \tag{9}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iM})^\top$ and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jM})^\top$ denote the Dirichlet parameter vectors of x_i and x_j, respectively, and u_i and u_j denote their corresponding uncertainty masses. A smaller value therefore indicates stronger local agreement among confident predictions.

Anchor Discovery Score. Prediction concentration captures the clarity of local predictions, while opinion conflict reflects their evidential consistency. Together, they provide complementary cues for identifying target regions that are less susceptible to representation shifts. Accordingly, the ADS is defined as

$$\mathrm{ADS}(x_i) = \mathrm{Conc}(x_i) \cdot \mathrm{Conf}(x_i). \tag{10}$$

4.3 EVIDENTIAL ANCHOR-GUIDED MAPPING

Stable Region Construction. To construct stable target regions for poisoning, IDEA selects reliable target samples as anchors based on their ADS scores. Unlike an individual target representation that may vary across federated rounds, these anchors provide region-level references for target alignment. Specifically, sorting target-class samples in ascending order of their ADS values, the reliable anchor set is defined as

$$\mathcal{A} = \left\{ x_{\pi(d)} \;\middle|\; 1 \le d \le K_d \right\}, \qquad K_d = \lfloor \rho |\mathcal{D}_n| \rfloor, \tag{11}$$

where π denotes the corresponding permutation and ρ controls the fraction of target-class samples retained as reliable anchors.

Since IDEA aims to map poisoned samples toward target regions for precise region-level alignment rather than toward a single target representation, the selected anchors serve as a reliable basis for representing the target regions. Specifically, K-means is applied to the representations of $\mathcal{A}$, yielding K_r representative centers $\mathcal{R} = \{\mathbf{q}_1, \ldots, \mathbf{q}_{K_r}\}$.

Target-region Attraction. With the target regions established by the representative centers $\mathcal{R}$, IDEA guides each triggered sample $\tilde{x}_i$ toward its most compatible target-region center, rather than a fixed target point. This region-aware attraction allows the alignment to adapt to the spatial structure of the target regions:

$$\mathcal{L}_{\text{anchor}} = -\frac{1}{|\mathcal{X}_n|} \sum_{x_i \in \mathcal{X}_n} \max_{1 \le r \le K_r} \cos\left(f_{\theta_g}(\tilde{x}_i), \mathbf{q}_r \right), \tag{12}$$

Evidential Geometric Refinement. While anchor attraction guides poisoned samples toward stable target regions, a successful ID backdoor also requires preserving the relationships and interactions among poisoned, target-class, and non-target samples. Our feature-space analysis of Attack Collapse in Appendix B.4 shows that collapse occurs when poisoned representations drift away from the target regions and become entangled with nearby non-target samples. This observation suggests a three-party relationship: poisoned samples should maintain a compact and coherent local structure, preserve sufficient target affinity with benign target-class samples, and stay sufficiently separated from competing non-target samples.

Motivated by this observation, evidential geometric refinement adopts an asymmetric metric design, using the ℓ_2 distance for target alignment to capture overall evidential deviation and the ℓ_1 distance for non-target separation to emphasize coordinate-wise differences in class-wise evidence:

$$\mathcal{L}_{\text{evid}} = \frac{1}{|\mathcal{X}_n|} \sum_{x_i \in \mathcal{X}_n} \left[\left\| \tilde{\boldsymbol{\alpha}}_i - \boldsymbol{\mu}_n^{\alpha} \right\|_2 - \min_{x_j \in \mathcal{D}_n^{\neg c_n}} \left\| \tilde{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_j \right\|_1 \right], \tag{13}$$

where $\boldsymbol{\mu}_n^{\alpha} = \frac{1}{|\mathcal{D}_n^{c_n}|} \sum_{x_i \in \mathcal{D}_n^{c_n}} \boldsymbol{\alpha}_i$ denotes the target evidential reference, computed as the mean Dirichlet parameter vector over benign samples from the target class, $\tilde{\boldsymbol{\alpha}}_i$ denotes the Dirichlet parameter vector of the triggered sample $\tilde{x}_i$, and $\mathcal{D}_n^{\neg c_n}$ denotes the set of non-target samples.

Trigger Optimization. The overall trigger objective jointly combines the ID poisoning loss, target-region anchor attraction, and evidential geometric refinement:

$$\mathcal{L}_{\text{IDEA}} = \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{anchor}} + \lambda_{\text{evid}} \mathcal{L}_{\text{evid}}. \tag{14}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Data. We evaluate IDEA on four widely used benchmark datasets: F-MNIST (Xiao et al., 2017), CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and GTSRB (Houben et al., 2013). Detailed dataset descriptions are provided in Appendix B.2. To simulate heterogeneous client distributions, we adopt a Dirichlet-based partitioning scheme, where the concentration parameter κ controls the degree of non-IID skew across clients. Further details are provided in Appendix B.3.

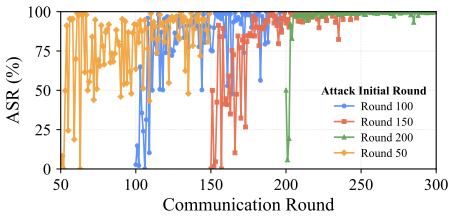


Figure 4: Impact of attack initiation on attack performance on CIFAR-10 dataset.

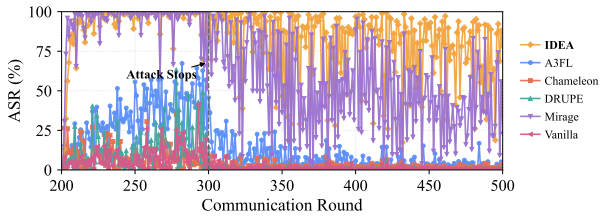


Figure 5: Backdoor persistence under continued training on CIFAR-10 dataset.

Baselines. We compare IDEA with representative state-of-the-art backdoor attacks, including static-trigger methods such as Vanilla (Bagdasaryan et al., 2020) and Chameleon (Dai & Li, 2023), as well as trigger-optimization methods including A3FL (Zhang et al., 2023), DRUPE (Tao et al., 2024), and Mirage (Li et al., 2025).

We further evaluate these attacks under six state-of-the-art defenses: Multi-Krum (Blanchard et al., 2017), Bulyan (Guerraoui et al., 2018), Foolsgold (Fung et al., 2020), FLAME (Nguyen et al., 2022), BackdoorIndicator (Li & Dai, 2024), and Mars (Wan et al., 2026). Unless otherwise specified, all baseline attacks and defenses are implemented following their original implementations and settings.

Federated Learning and Attack Setup. We consider an FL system with 100 clients, 20% of which are designated as malicious, with the server randomly selecting 10 clients in each communication round. ResNet-18 (He et al., 2016) is adopted as the global model. Following prior work (Li et al., 2025), we partition the training data across clients using a Dirichlet distribution with $\kappa = 0.5$ to induce label heterogeneity. Each selected client performs local training using SGD with a learning rate of 0.1, momentum of 0.9, weight decay of 0.0005, and a batch size of 64. Blend (Chen et al., 2017) is adopted as the backdoor trigger across all attacks, while BackdoorIndicator is used as the default defense method. The training runs for 200 communication rounds, with the defense enabled throughout training and malicious clients launching backdoor attacks from round 100 onward. All experiments are repeated three times.

Evaluation Metrics. To evaluate IDEA in terms of attack effectiveness and stability, we consider three metrics: (1) Test Accuracy (Acc), which measures predictive performance on clean test samples; (2) ASR, which measures backdoor effectiveness by quantifying the proportion of triggered samples classified as the target class; and (3) Attack Quality (AQ), which comprehensively evaluates the strength and stability of an attack across communication rounds. Specifically, $\mathrm{AQ} = \overline{\mathrm{ASR}} \left(1 - \frac{1}{T-1} \sum_{t=2}^{T} |\mathrm{ASR}_t - \mathrm{ASR}_{t-1}|\right)$, where $\overline{\mathrm{ASR}}$ denotes the average ASR.

5.2 Attack Initiation and Lifespan

Our first experiment evaluates the effectiveness and durability of IDEA in an evolving FL environment. We vary the attack initiation round while fixing the attack duration at 100 communication rounds. The results are shown in Fig. 4 and Fig. 5.

Existing attack settings typically assume that the global model has already converged and inject backdoors only after thousands of communication rounds (Zhang et al., 2023; Li et al., 2025). However, practical FL involves heterogeneous client data, causing the global representation space to evolve throughout training. We therefore vary the attack initiation round to evaluate IDEA at different stages of training. Our experiments show that attack collapse becomes less frequent and less pronounced as training progresses and the representation space stabilizes, highlighting the challenge of maintaining a persistent backdoor in an evolving FL environment. Despite this challenge, IDEA rapidly achieves a high ASR regardless of the attack initiation round. Notably, under more unstable training conditions, earlier attack initiation yields more consistent backdoor performance at the same subsequent round, suggesting that IDEA benefits from a longer adaptation period.

We further evaluate backdoor persistence by terminating the attack and continuing benign federated training. The ASR of all methods gradually declines as subsequent model updates alter the learned representation space. For most baselines, the backdoor effect deteriorates rapidly and nearly disappears shortly after attack termination, with some exhibiting pronounced Attack Collapse during

Table 1: Comparison of different backdoor attacks under SOTA defense methods. Acc, ASR, and AQ denote clean accuracy, attack success rate, and attack quality, respectively, with higher values indicating better performance. The best-performing results are highlighted in **bold**, while the second-best results are underlined. ASR values above 90% are marked with " $*$ ".

Dataset	Defense	Vanilla			Chameleon			A3FL			DRUPE			Mirage			IDEA		
		Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑
CIFAR-10	*No Defense*	69.50	15.49	7.43	68.97	21.38	11.29	69.63	41.05	28.50	67.50	33.53	20.61	70.81	96.71*	71.14	**71.19**	**99.37***	**77.06**
	Multi-Krum	65.47	1.76	1.72	**66.01**	6.25	3.83	65.67	19.22	16.22	63.17	13.37	8.30	65.32	73.88	59.22	65.73	**75.05**	**61.34**
	Bulyan	**52.47**	3.48	1.51	46.59	6.82	3.04	45.21	18.39	11.44	51.46	6.56	3.43	46.74	93.09*	58.15	48.71	**95.63***	**64.63**
	FoolsGold	74.14	1.32	1.28	71.22	2.01	2.44	73.95	21.27	17.99	**74.27**	8.40	7.50	71.66	96.74*	87.13	73.02	**99.13***	**90.40**
	FLAME	66.20	2.51	2.02	67.06	6.79	6.38	66.39	34.05	25.27	64.87	22.25	12.20	68.88	92.90*	69.25	**68.97**	**94.34***	**73.87**
	Indicator	68.08	6.07	4.39	62.83	10.85	10.77	66.03	32.18	27.23	68.71	11.69	10.73	68.23	96.65*	67.64	**69.30**	**98.71***	**74.46**
	Mars	69.21	5.17	3.88	67.80	5.92	3.79	69.16	32.01	26.67	68.90	9.54	6.01	70.84	96.59*	74.41	**71.25**	**97.22***	**77.60**
CIFAR-100	*No Defense*	51.20	28.64	13.81	46.69	16.25	7.37	51.18	42.29	29.36	50.44	30.27	21.18	51.11	94.02*	82.59	**51.33**	**97.96***	**87.25**
	Multi-Krum	46.00	0.14	0.45	46.03	0.17	0.53	45.32	7.25	5.70	**50.17**	7.37	6.30	45.44	45.39	49.65	46.13	**72.95**	**63.91**
	Bulyan	21.93	1.03	0.52	46.59	6.82	3.04	21.24	1.91	3.81	**51.46**	10.56	4.43	23.46	36.34	23.42	33.57	**41.40**	**32.55**
	FoolsGold	51.09	0.33	0.51	**51.12**	1.02	1.84	50.63	4.59	5.60	50.27	13.42	13.58	49.87	84.12	62.40	50.70	**85.06**	**73.07**
	FLAME	**47.65**	4.57	2.89	47.39	0.37	0.56	46.31	9.36	9.16	47.33	15.73	10.20	47.01	90.19*	76.48	47.31	**90.13***	**81.50**
	Indicator	**48.05**	1.83	1.42	44.65	2.17	4.52	45.65	10.26	8.96	47.77	10.36	10.67	47.65	80.89	33.34	47.63	**83.29**	**36.72**
	Mars	51.05	6.43	3.97	49.64	0.46	0.55	51.08	12.60	9.06	**51.64**	13.18	12.01	50.88	91.34*	77.06	51.33	**91.66***	**81.25**
GTSRB	*No Defense*	93.41	83.61	73.55	90.32	27.86	21.94	93.53	91.31*	72.87	93.01	63.55	53.62	93.09	97.04*	92.23	**93.73**	**99.91***	**98.96**
	Multi-Krum	90.81	1.41	0.98	90.36	2.42	1.28	**92.97**	46.46	26.94	81.45	22.00	11.02	86.16	95.27*	88.60	92.18	**99.81***	**95.16**
	Bulyan	21.50	0.25	0.50	83.43	4.96	1.48	83.95	62.93	39.12	84.20	42.37	32.76	81.84	96.32*	87.27	**88.21**	**98.28***	**95.01**
	FoolsGold	92.68	71.22	46.81	91.51	21.12	16.57	92.82	71.80	50.90	92.91	41.49	30.93	**92.96**	97.93*	90.13	92.92	**99.81***	**92.58**
	FLAME	88.34	79.00	69.06	87.74	3.55	1.68	85.58	86.18	70.38	86.93	56.44	49.40	**89.71**	97.24*	92.78	89.04	**99.61***	**98.58**
	Indicator	**88.53**	73.37	45.85	83.32	18.69	8.54	86.84	83.73	70.00	87.96	50.95	41.80	84.16	97.67*	79.79	84.69	**99.79***	**88.56**
	Mars	85.68	11.29	61.28	90.22	20.72	6.85	90.12	75.49	65.58	91.42	51.49	41.77	92.39	96.87*	91.51	**93.32**	**99.87***	**98.92**

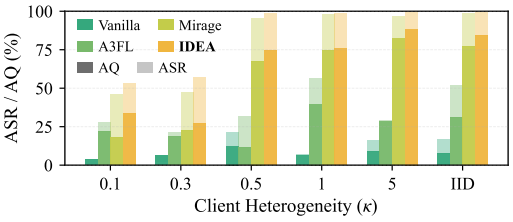


Figure 6: Impact of client heterogeneity on attack performance on CIFAR-10.

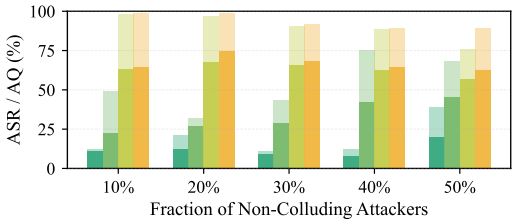


Figure 7: Impact of the number of attackers on attack performance on CIFAR-10.

the decay process. In contrast, IDEA maintains a smoother ASR decline and preserves effective backdoor behavior for a longer period, indicating stronger resilience to continued model evolution and milder attack collapse.

5.3 Attack Performance

We evaluate the robustness of backdoor attacks against state-of-the-art (SOTA) defenses on CIFAR-10, CIFAR-100, and GTSRB, with FMNIST results deferred to Section F.2. The results are summarized in Table 1.

Across all datasets and defense settings, IDEA consistently achieves the highest ASR and AQ, indicating strong and stable backdoor performance. In contrast, some existing attacks achieve relatively high AQ despite much lower ASR, suggesting that their performance is stable but consistently weak. IDEA remains effective under diverse defenses, maintaining ASRs above 90% in most settings. This advantage is particularly evident on CIFAR-100, demonstrating its robustness on more challenging classification tasks.

Importantly, IDEA achieves strong attack performance without sacrificing clean accuracy. ID backdoor attacks such as DRUPE, Mirage, and IDEA naturally better preserve benign model utility because their poisoned samples remain within the target data distribution. This property makes ID poisoning particularly advantageous in FL. Meanwhile, defense methods such as Multi-Krum and Bulyan can substantially reduce clean accuracy under heterogeneous settings, regardless of the attack method, likely because they filter client updates with distinctive yet legitimate local distributions. Despite this effect, IDEA generally maintains competitive accuracy while achieving higher ASR and AQ, demonstrating a favorable balance between attack effectiveness and model utility.

5.4 Robustness under Varying FL Conditions

Data Distribution Variation. We investigate the impact of the client heterogeneity, with the results are presented in Fig. 6. As client heterogeneity increases, existing attacks can experience sub-

stantial performance degradation beyond a certain level, indicating that severe distributional heterogeneity poses a significant challenge to backdoor propagation in FL. This degradation is particularly pronounced for the other three methods, whose ASR drops considerably under strong non-IID settings. In contrast, IDEA is substantially less sensitive to client heterogeneity, maintaining effective and stable attack performance even under extreme heterogeneity. Moreover, its relative advantage over the baselines becomes increasingly pronounced as heterogeneity intensifies, with IDEA consistently achieving higher ASR and AQ across all settings.

Attacker Variation. We further investigate the impact of the fraction of non-colluding attackers, with the results presented in Fig. 7. As the fraction of attackers increases, non-ID attacks generally achieve higher ASR, suggesting that a larger number of malicious participants facilitates the accumulation and propagation of OOD backdoor patterns even without collaboration. In contrast, ID attacks maintain consistently high ASR across different attacker fractions, indicating substantially lower sensitivity to changes in the attacker fraction. Among them, IDEA demonstrates the strongest robustness, maintaining high attack effectiveness across all settings. Moreover, its stability advantage becomes increasingly pronounced as the attacker fraction grows, highlighting its ability to sustain reliable backdoor behavior under varying attacker settings.

Model Architecture Variation. We further compare IDEA with the strongest baseline, Mirage, across different model architectures. The results are reported in Table 2. IDEA maintains effective attack performance across architectures, without a clear dependence on model scale. In contrast, Mirage suffers a pronounced performance degradation on ResNet34, suggesting greater sensitivity to architectural variations. Notably, IDEA consistently achieves higher attack quality than Mirage across all four architectures, demonstrating that its advantage extends beyond attack effectiveness to the stability of backdoor behavior.

Table 2: Performance comparison across model architectures.

Method	ResNet18			ResNet34			VGG11			VGG19		
	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑
Mirage	68.23	96.65	67.64	34.86	66.15	33.03	**57.81**	75.00	48.21	45.81	75.68	50.16
IDEA	**69.30**	**98.71**	**74.46**	**42.73**	**85.11**	**59.74**	56.54	**83.61**	**60.83**	**47.58**	**93.91**	**65.73**

5.5 Ablation Study

To investigate the contribution of each component, we conduct a comprehensive ablation study under heterogeneous settings. The results are reported in Table 3.

Removing any component degrades overall performance, confirming their distinct and complementary roles in IDEA. Among them, removing prediction concentration (Conc) causes the largest decline in AQ, indicating its key role in preserving target-consistent semantics and supporting attack stability. Removing ADS leads to the most pronounced drop in ASR, highlighting the importance of reliable target selection for constructing stable target regions. For L_{anchor}, removing the explicit target-region attraction reduces both ASR and AQ, showing its contribution to aligning poisoned samples with the target class. In contrast, removing L_{evid} causes the largest degradation in clean accuracy, demonstrating that evidential geometric refinement helps preserve clean performance while maintaining the desired geometric relationships among samples. Opinion conflict (Conf) produces a more balanced impact across the three metrics, indicating its role in regulating the poisoning process.

Table 3: Ablation study of IDEA.

Component	Acc↑	ASR↑	AQ↑
IDEA	**69.30**$\pm$**1.63**	**98.52**$\pm$**0.17**	**74.46**$\pm$**0.14**
w/o Conc	66.87$\pm$1.31	94.56$\pm$0.94	65.92$\pm$2.01
w/o Conf	67.36$\pm$3.22	97.08$\pm$0.42	73.41$\pm$0.70
w/o ADS	68.74$\pm$2.39	93.48$\pm$0.84	71.19$\pm$1.32
w/o L_{anchor}	68.23$\pm$0.82	95.87$\pm$0.82	73.63$\pm$0.78
w/o L_{evid}	65.22$\pm$2.16	95.95$\pm$1.23	68.60$\pm$1.69

6 Conclusion

In this work, we study in-distribution backdoor attacks under heterogeneous FL and characterize Attack Collapse as a key phenomenon affecting their effectiveness. We propose IDEA, a stable distribution-preserving backdoor attack that constructs reliable target regions for stable target-oriented guidance and employs evidential geometric refinement to preserve the local relationships among neighboring samples. Extensive experiments across diverse settings demonstrate the superior attack effectiveness and stability of IDEA while preserving competitive clean-task accuracy.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing and improving the readability of the manuscript, as well as with creating and modifying code and scientific figures. We did not use generative AI tools to develop the methodology, design the experiments, formulate mathematical claims, interpret experimental results, or generate synthetic data. All AI-assisted content was carefully reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and scientific artifacts.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or the collection of new sensitive data. The datasets used in our experiments are publicly available and are used solely for research purposes. Although our work studies backdoor attacks in federated learning, the proposed methodology is developed and evaluated strictly for research purposes, with the goal of understanding vulnerabilities and improving the robustness and security of federated learning systems. We are not aware of any ethical concerns or conflicts of interest arising from this work.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our work, we provide the complete source code and implementation details in the Supplementary Material. The methodology, experimental settings, and implementation details are also described in detail throughout the main paper and appendix. The released code is organized to enable the experiments to be directly executed under the configurations described in these materials. We additionally provide the corresponding experimental results in the Supplementary Material, allowing the reported results to be readily verified. Together, these materials are intended to support faithful reproduction and verification of our results.

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11439–11449, 2024.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Yanbo Dai and Songze Li. Chameleon: Adapting to peer images for planting durable backdoors in federated learning. In *International conference on machine learning*, pp. 6712–6725. PMLR, 2023.

Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Jun Feng, Yuzhe Lai, Hong Sun, and Bocheng Ren. Sadba: Self-adaptive distributed backdoor attack against federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16568–16576, 2025.

Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International symposium on research in attacks, intrusions and defenses (RAID 2020)*, pp. 301–316, 2020.

Junyu Gao, Mengyuan Chen, Liangyu Xiang, and Changsheng Xu. A comprehensive survey on evidential deep learning and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, pp. 3521–3530. PMLR, 2018.

Thierry Hanser, Ernst Ahlberg, Alexander Amberg, Lennart T Anger, Chris Barber, Richard J Brennan, Alessandro Brigo, Annie Delaunois, Susanne Glowienke, Nigel Greene, et al. Data-driven federated learning in drug discovery with knowledge distillation. *Nature Machine Intelligence*, 7 (3):423–436, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, pp. 1–8. Ieee, 2013.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059. IEEE, 2022.

Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Songze Li and Yanbo Dai. {BackdoorIndicator}: Leveraging {OOD} data for proactive backdoor detection in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4193–4210, 2024.

Ye Li, Yanchao Zhao, Chengcheng Zhu, and Jiale Zhang. Infighting in the dark: Multi-label backdoor attack in federated learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25770–25779. IEEE, 2025.

Yehui Li, Dalin Qin, H Vincent Poor, and Yi Wang. Introducing edge intelligence to smart meters via federated split learning. *Nature communications*, 15(1):9044, 2024.

Haiyan Lin, Qi Shen, Fei Zhu, Zixuan Qin, Liu Yang, and Yuping Duan. Feddna: Dna sequence reconstruction via deep evidential learning and personalized federated aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23523–23531, 2026.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

Zhaoyi Liu and Huan Zhang. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25060–25070. IEEE, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al. {FLAME}: Taming backdoors in federated learning. In *31st USENIX security symposium (USENIX Security 22)*, pp. 1415–1432, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2029–2047. IEEE, 2024.

Wei Wan, Ning Yuxuan, Zhicong Huang, Cheng Hong, Shengshan Hu, Ziqi Zhou, Yechao Zhang, Tianqing Zhu, Wanlei Zhou, and Leo Yu Zhang. Mars: A malignity-aware backdoor defense in federated learning. *Advances in Neural Information Processing Systems*, 38:154985–155017, 2026.

Dawood Wasif, Terrence Moore, and Jin-Hee Cho. Resfl: An uncertainty-aware framework for responsible federated learning by balancing privacy, fairness and utility. In *International Conference on Learning Representations*, volume 2026, pp. 142238–142275, 2026.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. *ICLR*, 2020:1–19, 2020.

Yuhan Xie, Chen Lyu, and Jingrong Huang. Besplit: Bias-compensated split federated learning with evidential aggregation. In *Forty-third International Conference on Machine Learning*, 2026.

Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, Jun Chen, and Xiaochun Cao. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. *IEEE Transactions on Information Forensics and Security*, 20:12404–12416, 2025.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6137–6147, 2024.

Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in neural information processing systems*, 36: 61213–61233, 2023.

Yechao Zhang, Yuxuan Zhou, Tianyu Li, Minghui Li, Shengshan Hu, Wei Luo, and Leo Yu Zhang. Secure transfer learning: Training clean model against backdoor in pre-trained encoder and downstream dataset. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 1–19. IEEE, 2025.

Zhiwei Zhang, Minhua Lin, Enyan Dai, and Suhang Wang. Rethinking graph backdoor attacks: A distribution-preserving perspective. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4386–4397, 2024.

A Related Work

A.1 Backdoor Attacks in Federated Learning

Backdoor attacks in FL aim to implant malicious behaviors into the global model through poisoned client updates, causing triggered inputs to produce adversary-specified outputs while preserving normal performance on clean inputs (Tao et al., 2024). Existing attacks can be broadly classified

into static-trigger attacks (Bagdasaryan et al., 2020; Xie et al., 2020; Dai & Li, 2023) and trigger-optimization attacks (Zhang et al., 2023; Feng et al., 2025; Li et al., 2025) based on their trigger construction strategies.

Static-trigger attacks employ pre-defined triggers to generate poisoned samples and typically enhance attack effectiveness through additional update manipulation. DBA (Xie et al., 2020) decomposes a global trigger into multiple sub-triggers distributed across compromised clients to improve stealthiness. Chameleon (Dai & Li, 2023) enhances backdoor durability via contrastive representation learning by separating poisoned samples from same-label images while aligning them with target-label images. In contrast, trigger-optimization attacks explicitly optimize triggers to improve attack effectiveness and persistence. A3FL (Zhang et al., 2023) adversarially optimizes triggers against both current and backdoor-unlearned models to improve persistence. SADBA (Feng et al., 2025) jointly optimizes data and model poisoning to enhance persistence while reducing the required fraction of malicious clients. Mirage (Li et al., 2025) combines adversarial trigger optimization with constrained distribution optimization to construct persistent in-distribution backdoors.

Despite these advances, existing FL backdoor attacks largely focus on improving attack effectiveness, stealthiness, or persistence, while overlooking the instability introduced by heterogeneous client data distributions. In particular, the non-IID nature of FL can cause substantial fluctuations or even sudden degradation in attack effectiveness during training, making backdoor performance difficult to maintain consistently across communication rounds. This observation highlights the need for backdoor attacks that remain robust under heterogeneous federated environments.

A.2 Evidential Deep Learning for Federated Learning

Evidential Deep Learning (EDL) (Sensoy et al., 2018) provides a probabilistic framework for predictive uncertainty estimation by learning class-specific evidence rather than directly producing point-valued class probabilities (Liu et al., 2025). Through Dirichlet-based evidential modeling, EDL captures both predictive belief and uncertainty, providing a principled basis for assessing prediction confidence and model reliability (Gao et al., 2025).

EDL has been increasingly adopted in FL to facilitate uncertainty-aware learning across diverse settings, including domain shifts, fairness and privacy, client heterogeneity, and personalization (Gao et al., 2025; Xie et al., 2026). For example, FEAL (Chen et al., 2024) integrates Dirichlet-based EDL into federated active learning to jointly capture aleatoric and epistemic uncertainty for informative sample selection under domain shifts. FedDNA (Lin et al., 2026) further leverages the decomposition of aleatoric and epistemic uncertainty to adaptively weight client updates for personalized FL under data heterogeneity. RESFL (Wasif et al., 2026) leverages evidential uncertainty to guide client aggregation for improved fairness and privacy. More recently, BESplit Xie et al. (2026) exploits EDL-derived evidence and uncertainty to characterize client heterogeneity and adaptively reweight updates for robust aggregation in Split Federated Learning.

Despite these advances, the application of EDL to federated backdoor attacks remains largely unexplored. In particular, the potential of evidential information to guide the construction of effective backdoor samples has yet to be investigated. This gap motivates the exploration of evidential signals for backdoor sample construction in FL.

B Implementation Details

B.1 Baseline Details

To provide a comprehensive comparison, we evaluate IDEA against representative state-of-the-art backdoor attacks and defenses.

Static-trigger Backdoor Attacks:

- **Vanilla Attacks** (Bagdasaryan et al., 2020) directly train malicious clients on a mixture of benign and backdoor samples and scale the resulting model update to facilitate effective backdoor implantation.

- **Chameleon** (Dai & Li, 2023) employs supervised contrastive learning to align poisoned samples with facilitator images and separate them from interferers, followed by classifier training to improve backdoor durability.

Trigger-Optimization Backdoor Attacks:

- **A3FL** (Zhang et al., 2023) adversarially optimizes the trigger against both the current global model and an adversarially adapted model that simulates backdoor unlearning, thereby improving backdoor persistence under evolving global dynamics.
- **DRUPE** (Tao et al., 2024) was originally proposed for self-supervised learning. Following its implementation, we adapt the attack to the federated learning setting by transforming poisoned samples toward the clean data distribution while reducing their feature concentration, making the backdoor less distinguishable in the feature space.
- **Mirage** (Li et al., 2025) considers non-colluding attackers with different target labels and constructs persistent in-distribution backdoor mappings to mitigate conflicts among independently injected backdoors.

We further evaluate these attacks against representative defenses based on robust aggregation, update similarity, and anomaly detection, thereby assessing the robustness of IDEA against diverse defense mechanisms.

Backdoor Defenses:

- **Multi-Krum** (Blanchard et al., 2017) selects the m client updates with the smallest Krum scores, where each score is computed from the distances to the $N-f-2$ nearest neighbors. Here, N denotes the number of participating clients and f the assumed upper bound on the number of Byzantine clients. The selected updates are then averaged to mitigate potentially malicious contributions.
- **Bulyan** (Guerraoui et al., 2018) combines a Byzantine-resilient aggregation rule with coordinate-wise median-based filtering to constrain the influence of malicious updates. In our implementation, Krum is used as the underlying aggregation rule.
- **FoolsGold** (Fung et al., 2020) detects Sybil-based poisoning by measuring cosine similarity among clients' historical updates and downweighting clients with highly similar update patterns, thereby limiting coordinated malicious contributions.
- **FLAME** (Nguyen et al., 2022) combines dynamic HDBSCAN-based outlier filtering, adaptive norm clipping, and calibrated Gaussian noising to remove high-impact malicious updates while limiting utility degradation.
- **BackdoorIndicator** (Li & Dai, 2024) injects an OOD indicator task into the global model and identifies backdoored client models through anomalously high indicator-task accuracy, while restoring batch normalization statistics to minimize interference with the main task.
- **MARS** (Wan et al., 2026) extracts concentrated backdoor energy from local models and employs Wasserstein distance-based clustering to identify malicious updates, selecting the cluster with the smaller center norm as benign while retaining both clusters when their distributions are sufficiently similar.

All baseline attacks and defenses are implemented or configured following their original implementations and recommended settings to ensure fair and reproducible comparisons.

B.2 Dataset Details

We evaluate IDEA on four widely used image classification benchmarks that cover diverse image modalities, visual complexity, and numbers of classes. These datasets are commonly adopted in studies of image classification and backdoor attacks, providing a diverse testbed for our evaluation.

F-MNIST (Xiao et al., 2017): A dataset of 70,000 grayscale images from 10 fashion categories, with a resolution of 28×28 pixels. It contains 60,000 training images and 10,000 test images and provides a relatively simple visual setting.

CIFAR-10 (Krizhevsky et al., 2009): A dataset of 60,000 color images from 10 object classes at a resolution of 32×32 pixels. It includes 50,000 training images and 10,000 test images and serves as a standard benchmark for natural image classification.

CIFAR-100 (Krizhevsky et al., 2009): A dataset of 60,000 color images of size 32×32 from 100 classes, including 50,000 training images and 10,000 test images. The classes are organized into 20 superclasses, providing a more fine-grained classification setting than CIFAR-10.

GTSRB (Houben et al., 2013): A traffic sign recognition dataset containing 43 classes, with 39,209 training images and 12,630 test images. It features substantial variations in image appearance, scale, illumination, and viewpoint, providing a challenging real-world classification setting.

B.3 Heterogeneous Data Partitioning

IID Allocation. To provide a homogeneous baseline, the training data are randomly shuffled and evenly distributed across the K clients, resulting in comparable local data distributions.

Non-IID Allocation. To simulate heterogeneous federated settings, we adopt a class-wise Dirichlet partitioning strategy to induce label heterogeneity. For each class $m \in \{1, \ldots, M\}$, its samples are first randomly shuffled, and the client-wise allocation proportions are sampled as

$$(p_{m,1}, \ldots, p_{m,K}) \sim \mathrm{Dirichlet}(\kappa, \ldots, \kappa).$$

The sampled proportions are scaled by the total number of samples in class m and used to allocate its samples across clients without duplication. The parameter κ controls the degree of label heterogeneity, where smaller values produce more skewed client distributions. Unless otherwise specified, we set $\kappa = 0.5$ to construct a strongly heterogeneous non-IID setting.

Impact on ID Backdoor Attacks. Non-IID data introduce a fundamental challenge for ID backdoor construction. Since each malicious client can only access its local data, the samples available for constructing the backdoor may substantially deviate from the global data distribution under strong heterogeneity. In particular, local target-class samples provide a biased and incomplete reference for the global target distribution, making it difficult to construct poisoned samples that remain aligned with the target ID regions. As the heterogeneity becomes stronger, this local-global distribution mismatch becomes more pronounced, which can impair both the effectiveness and persistence of ID backdoors during federated training.

B.4 Feature-space Analysis of Attack Collapse

ID backdoor attacks can abruptly deteriorate under heterogeneous federated training, a phenomenon we term *Attack Collapse*. To understand why, we examine the evolution of poisoned representations in the feature space. Fig. 8 visualizes two sequences of three consecutive training rounds, with the middle round in each sequence corresponding to the occurrence of Attack Collapse.

Our Finding. We observe that poisoned representations undergo distinct geometric changes across FL training rounds as the feature space continuously evolves under heterogeneous training. When the attack remains effective, poisoned samples stay within or near their corresponding target-class regions while maintaining a compact local structure. During Attack Collapse, however, poisoned representations drift away from the target region and become increasingly entangled with nearby non-target regions. This disruption of their spatial relationships with surrounding samples coincides with a sharp degradation in attack effectiveness.

Our Motivation. These observations indicate that the success of an ID backdoor depends not only on the proximity of poisoned samples to the target class, but also on how they interact with surrounding samples. Therefore, stable backdoor behavior requires preserving the relative relationships among poisoned samples, benign target-class samples, and nearby non-target samples as the feature space evolves.

In particular, poisoned samples should satisfy three complementary geometric requirements. First, poisoned samples should maintain intra-backdoor cohesion, forming a compact and coherent local structure rather than dispersing arbitrarily in the feature space. Notably, ID backdoor samples need not faithfully imitate benign target-class samples; they may occupy a distinct local region while

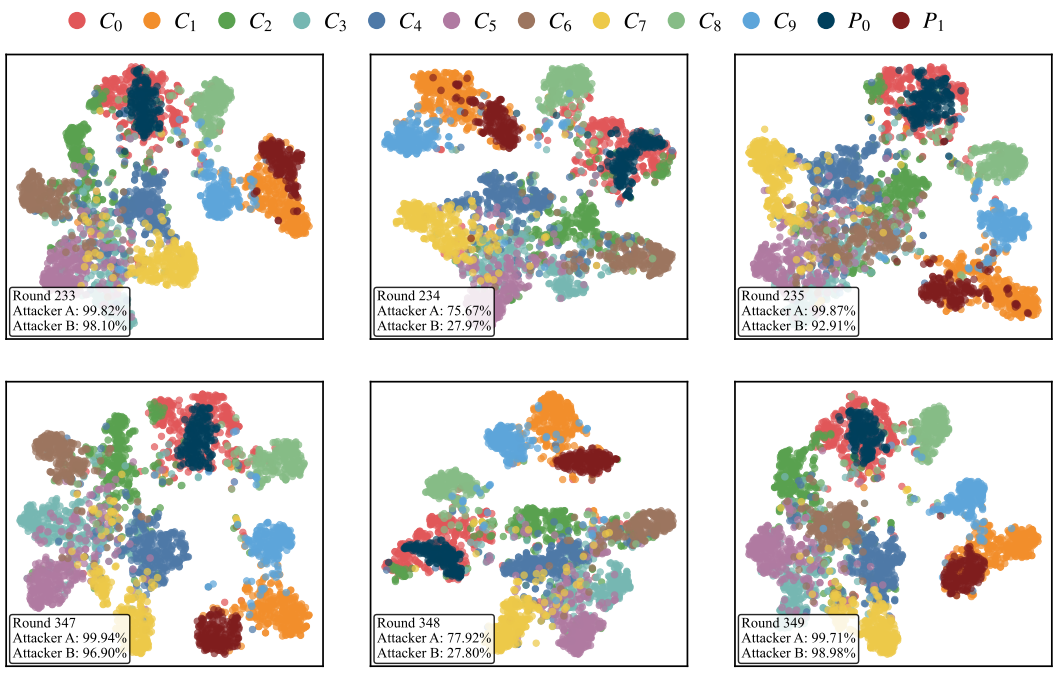


Figure 8: Feature-space evolution of ID backdoor representations before, during, and after Attack Collapse. C_0–C_9 denote benign samples from the corresponding classes, while P_0 and P_1 denote poisoned samples targeting classes 0 and 1, respectively, injected by attackers A and B.

remaining target-consistent. Second, poisoned samples should preserve sufficient target affinity with benign samples from the target class, which anchors them to the desired semantic region. Third, they should maintain sufficient non-target separation from nearby competing classes to avoid being absorbed into non-target regions. Attack Collapse is therefore associated with the disruption of these relationships, particularly the loss of target affinity accompanied by increasing proximity to competing non-target regions.

Implication. This finding motivates IDEA to explicitly preserve these complementary relationships during trigger optimization. Specifically, the preceding anchor attraction establishes a coherent target-oriented structure, while the evidential geometric constraints further enforce target affinity and non-target separation as the feature space evolves.

C Trigger Construction and Attack Details

We provide additional implementation details of trigger construction and poisoning for IDEA. Unless otherwise specified, the following settings are used for the ResNet18 model on CIFAR-10. We consider a total of 100 clients, 20% of which are malicious. In each round, 10% of the clients participate, among which two malicious clients are included, with their target labels set to 0 and 1, respectively. The trigger is initialized with a constant value of 0.5 and injected using the blend mode with a blending coefficient of 0.3. For trigger optimization, we first train a binary discriminator using SGD with a learning rate of 0.01 for 30 iterations. The trigger is then optimized using Adam with a learning rate of 0.001 for 50 iterations. After each optimization step, the trigger values are clipped to $[-2, 2]$ to stabilize the optimization.

During poisoned local training, each batch contains 64 samples, among which 8 non-target samples are selected for poisoning, corresponding to a poisoning ratio of 12.5%. The optimized trigger is injected with a blending coefficient of 0.3, and the labels of the selected samples are replaced with the corresponding target labels. Malicious clients are trained for 5 epochs using a fixed learning rate of 0.01, whereas benign clients are trained for 2 epochs with their learning rate linearly decayed from 0.1 to 0.01.

These settings are consistently applied in the corresponding experiments to ensure a fair and reproducible evaluation of IDEA.

D Theoretical Analysis

This section provides a theoretical analysis of IDEA in three stages. We first establish the local and inter-round stability of reliable target anchors. We then show that the resulting representative-center regions provide robust target alignment under bounded inter-round drift and admit a bounded approximation to the underlying target regions. Finally, we analyze the projected sign-gradient trigger optimization and derive a finite-step stationarity bound for the IDEA objective.

D.1 Anchor Stability

IDEA selects target-class anchors using the ADS, which jointly captures prediction concentration and opinion conflict. We first characterize the local predictive consistency of reliable anchors and then relate it to their feature-space compactness and inter-round stability.

Lemma 1 (Monotonicity of Prediction Concentration)**.** *Let* $p_{j,\max} = \frac{\max_m \alpha_{jm}}{S_j}$*. Then Eq. 8 can be rewritten as*

$$\mathrm{Conc}(x_i) = \frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \log \frac{1 - p_{j,\max}}{p_{j,\max}}. \tag{15}$$

Moreover, $\mathrm{Conc}(x_i)$ *decreases monotonically as the maximum predictive probability of neighboring samples increases.*

Proof. For $p \in (0, 1)$,

$$\frac{\mathrm{d}}{\mathrm{d}p} \log \frac{1-p}{p} = -\frac{1}{p(1-p)} < 0.$$

Thus, larger $p_{j,\max}$ yields smaller $\mathrm{Conc}(x_i)$, corresponding to more concentrated neighboring predictions. □

Lemma 2 (Local Predictive Consistency)**.** *Let* x_i *be a reliable target-class anchor satisfying*

$$\mathrm{Conc}(x_i) \le \tau_{\mathrm{conc}}, \qquad \mathrm{Conf}(x_i) \le \tau_{\mathrm{conf}},$$

where τ_{conc} *and* τ_{conf} *are the respective concentration and conflict thresholds. Assume that*

$$u_i \le \delta_u < 1, \qquad u_j \le \delta_u < 1, \quad \forall x_j \in \mathcal{N}_i.$$

Then the average predictive discrepancy between x_i *and its neighbors satisfies*

$$\frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \|\mathbf{p}_i - \mathbf{p}_j\|_1 \le \frac{\tau_{\mathrm{conf}}}{(1-\delta_u)^2}. \tag{16}$$

Proof. For every $x_j \in \mathcal{N}_i$,

$$(1-u_i)(1-u_j) \ge (1-\delta_u)^2.$$

Using the definition of $\mathrm{Conf}(x_i)$ in Eq. 9,

$$\begin{aligned} \mathrm{Conf}(x_i) &= \frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \|\mathbf{p}_i - \mathbf{p}_j\|_1 (1-u_i)(1-u_j) \\ &\ge (1-\delta_u)^2 \frac{1}{|\mathcal{N}_i|} \sum_{x_j \in \mathcal{N}_i} \|\mathbf{p}_i - \mathbf{p}_j\|_1. \end{aligned} \tag{17}$$

Combining this inequality with $\mathrm{Conf}(x_i) \le \tau_{\mathrm{conf}}$ yields Eq. 16. □

To relate predictive consistency to feature compactness, we make the following local evidential–feature consistency assumption.

Assumption 1 (Local Evidential–Feature Consistency)**.** *There exists a positive, non-increasing function $\mu_\alpha : \mathbb{R} \to \mathbb{R}_{>0}$ such that, for a target-class anchor x_i and every $x_j \in \mathcal{N}_i$,*

$$\mu_\alpha(\mathrm{Conc}(x_i))\,\|\mathbf{z}_i - \mathbf{z}_j\|_2 \le \|\mathbf{p}_i - \mathbf{p}_j\|_1, \tag{18}$$

where $\mathbf{z}_i = f_\theta(x_i)$ denotes the feature representation of x_i.

Theorem 1 (Local Representation Compactness)**.** *Under the conditions of Lemma 2 and Assumption 1, the average local feature variation of a reliable target-class anchor satisfies*

$$\frac{1}{|\mathcal{N}_i|}\sum_{x_j\in\mathcal{N}_i}\|\mathbf{z}_i - \mathbf{z}_j\|_2 \le \frac{\tau_{\mathrm{conf}}}{\mu_\alpha(\tau_{\mathrm{conc}})(1-\delta_u)^2}. \tag{19}$$

Proof. By Assumption 1,

$$\|\mathbf{z}_i - \mathbf{z}_j\|_2 \le \frac{\|\mathbf{p}_i - \mathbf{p}_j\|_1}{\mu_\alpha(\mathrm{Conc}(x_i))}.$$

Averaging over $x_j \in \mathcal{N}_i$ gives

$$\frac{1}{|\mathcal{N}_i|}\sum_{x_j\in\mathcal{N}_i}\|\mathbf{z}_i - \mathbf{z}_j\|_2 \le \frac{1}{\mu_\alpha(\mathrm{Conc}(x_i))}\frac{1}{|\mathcal{N}_i|}\sum_{x_j\in\mathcal{N}_i}\|\mathbf{p}_i - \mathbf{p}_j\|_1. \tag{20}$$

Since $\mu_\alpha(\cdot)$ is non-increasing and $\mathrm{Conc}(x_i) \le \tau_{\mathrm{conc}}$,

$$\mu_\alpha(\mathrm{Conc}(x_i)) \ge \mu_\alpha(\tau_{\mathrm{conc}}).$$

Combining this inequality with Eq. 16 yields Eq. 19. □

Assumption 2 (Lipschitz Feature Evolution)**.** *Let $\mathbf{z}_i^{(t)} = f_{\theta_t}(x_i)$ denote the representation of x_i at round t. Assume that, along the parameter trajectory, the feature extractor satisfies*

$$\left\|f_{\theta_{t+1}}(x) - f_{\theta_t}(x)\right\|_2 \le L_\theta\left\|\theta_{t+1} - \theta_t\right\|_2, \tag{21}$$

and that the inter-round parameter update is bounded by

$$\left\|\theta_{t+1} - \theta_t\right\|_2 \le \delta_\theta. \tag{22}$$

Proposition 1 (Inter-Round Representation Stability)**.** *Under Assumption 2, the representation drift between consecutive rounds satisfies*

$$\left\|\mathbf{z}_i^{(t+1)} - \mathbf{z}_i^{(t)}\right\|_2 \le L_\theta\delta_\theta. \tag{23}$$

Moreover, over T consecutive rounds,

$$\left\|\mathbf{z}_i^{(T)} - \mathbf{z}_i^{(0)}\right\|_2 \le TL_\theta\delta_\theta. \tag{24}$$

Proof. The first inequality follows directly from Eq. 21 and Eq. 22. By the triangle inequality,

$$\begin{aligned}\left\|\mathbf{z}_i^{(T)} - \mathbf{z}_i^{(0)}\right\|_2 &\le \sum_{t=0}^{T-1}\left\|\mathbf{z}_i^{(t+1)} - \mathbf{z}_i^{(t)}\right\|_2 \\ &\le TL_\theta\delta_\theta.\end{aligned} \tag{25}$$

□

Theorem 1 and Proposition 1 establish complementary spatial and temporal stability of reliable anchors. The former bounds their local feature variation within a round, while the latter bounds their representation drift across rounds. Together, these properties motivate constructing structured target regions from reliable anchors rather than relying on a single round-specific target point.

D.2 Region-level Mapping Robustness

The representative centers provide region-level references for target mapping. We characterize the stability of these references under inter-round representation evolution and their approximation to the underlying target regions.

Let $\mathcal{R}^{(t)} = \{\mathbf{q}_1^{(t)}, \dots, \mathbf{q}_{K_r}^{(t)}\}$ denote the representative centers at communication round t. For a representation $\mathbf{z}$, define its distance to the target region as

$$d\Big(\mathbf{z}, \mathcal{R}^{(t)}\Big) = \min_{\mathbf{q}\in\mathcal{R}^{(t)}} d(\mathbf{z}, \mathbf{q}), \qquad d(\mathbf{z}, \mathbf{q}) = \|\mathbf{z} - \mathbf{q}\|_2. \tag{26}$$

This distance characterizes the geometric proximity of a representation to the representative-center region. The anchor attraction in Eq. 12 instead selects the most compatible center according to cosine similarity, while Eq. 26 is used to characterize the geometric stability of the representative-center region.

Assumption 3 (Bounded Target-region Drift)**.** *Let $\mathcal{R}^{(t)}$ and $\mathcal{R}^{(t+1)}$ denote the representative-center sets at two consecutive communication rounds. There exists a constant $\delta_R \geq 0$ such that*

$$d_H\left(\mathcal{R}^{(t)}, \mathcal{R}^{(t+1)}\right) \leq \delta_R, \tag{27}$$

where δ_R bounds the inter-round displacement of the target region.

Theorem 2 (Region-level Alignment Robustness)**.** *Under Assumption 3, for any representation $\mathbf{z}$, the distance to the target region satisfies*

$$\left|d\Big(\mathbf{z}, \mathcal{R}^{(t+1)}\Big) - d\Big(\mathbf{z}, \mathcal{R}^{(t)}\Big)\right| \leq \delta_R. \tag{28}$$

Proof. Let $\mathbf{q}^* \in \mathcal{R}^{(t)}$ attain $d(\mathbf{z}, \mathcal{R}^{(t)})$. By the definition of the Hausdorff distance, there exists $\mathbf{q}' \in \mathcal{R}^{(t+1)}$ such that

$$d(\mathbf{q}^*, \mathbf{q}') \leq d_H(\mathcal{R}^{(t)}, \mathcal{R}^{(t+1)}) \leq \delta_R.$$

By the triangle inequality,

$$d(\mathbf{z}, \mathcal{R}^{(t+1)}) \leq d(\mathbf{z}, \mathbf{q}') \leq d(\mathbf{z}, \mathbf{q}^*) + d(\mathbf{q}^*, \mathbf{q}') \leq d(\mathbf{z}, \mathcal{R}^{(t)}) + \delta_R.$$

Exchanging $\mathcal{R}^{(t)}$ and $\mathcal{R}^{(t+1)}$ gives

$$d(\mathbf{z}, \mathcal{R}^{(t)}) \leq d(\mathbf{z}, \mathcal{R}^{(t+1)}) + \delta_R,$$

which proves Eq. 28. □

Consequently, if a poisoned representation $\tilde{\mathbf{z}}$ satisfies $d(\tilde{\mathbf{z}}, \mathcal{R}^{(t)}) \leq \varepsilon$ at round t, then $d(\tilde{\mathbf{z}}, \mathcal{R}^{(t+1)}) \leq \varepsilon + \delta_R$. Thus, bounded inter-round region drift limits the change in the region-level alignment distance.

We next characterize how well the representative centers approximate the underlying target representation region.

Assumption 4 (Target-region Approximation)**.** *Let $\mathcal{T}^{(t)}$ denote the underlying target representation region at communication round t. Assume that the representative centers approximate $\mathcal{T}^{(t)}$ with bounded Hausdorff error:*

$$d_H\left(\mathcal{T}^{(t)}, \mathcal{R}^{(t)}\right) \leq \epsilon_R, \tag{29}$$

where ϵ_R characterizes the approximation error of the representative-center construction.

Theorem 3 (Region-level Approximation Bound)**.** *Under Assumption 4, every target representation $\mathbf{z} \in \mathcal{T}^{(t)}$ satisfies*

$$d\Big(\mathbf{z}, \mathcal{R}^{(t)}\Big) \leq \epsilon_R. \tag{30}$$

Proof. For any $\mathbf{z} \in \mathcal{T}^{(t)}$, Eq. 29 implies the existence of $\mathbf{q} \in \mathcal{R}^{(t)}$ such that $d(\mathbf{z}, \mathbf{q}) \leq \epsilon_R$. Therefore,

$$d\Big(\mathbf{z}, \mathcal{R}^{(t)}\Big) = \min_{\mathbf{q}\in\mathcal{R}^{(t)}} d(\mathbf{z}, \mathbf{q}) \leq d(\mathbf{z}, \mathbf{q}) \leq \epsilon_R,$$

which proves the claim. □

To compare region-level mapping with a single-point reference, consider an arbitrary target representation $\mathbf{q}^{(t)} \in \mathcal{T}^{(t)}$. For any $\mathbf{z} \in \mathcal{T}^{(t)}$,

$$d(\mathbf{z}, \mathbf{q}^{(t)}) \leq \operatorname{diam}(\mathcal{T}^{(t)}),$$

where

$$\operatorname{diam}(\mathcal{T}^{(t)}) = \sup_{\mathbf{z},\mathbf{z}' \in \mathcal{T}^{(t)}} d(\mathbf{z}, \mathbf{z}').$$

In contrast, Theorem 3 gives the uniform bound $d(\mathbf{z}, \mathcal{R}^{(t)}) \leq \epsilon_R$ over the entire target region. Hence, when $\epsilon_R < \operatorname{diam}(\mathcal{T}^{(t)})$, the representative-center mapping admits a tighter worst-case upper bound than the generic diameter bound associated with an arbitrary single-point reference.

D.3 Trigger Optimization Analysis

IDEA optimizes the trigger parameters using automatic differentiation and projected sign-gradient updates. We establish a finite-step stationarity bound for this optimization procedure.

Trigger Update. Let $\mathbf{p} \in \mathbb{R}^d$ denote the trigger parameters. For a triggered sample $\mathcal{T}_{\mathbf{p}}(x_i)$, let $\tilde{\mathbf{z}}_i(\mathbf{p})$ and $\tilde{\boldsymbol{\alpha}}_i(\mathbf{p})$ denote its feature representation and Dirichlet parameter vector, respectively. We denote the IDEA trigger objective in Eq. 14 by $F(\mathbf{p}) = \mathcal{L}_{\mathrm{IDEA}}(\mathbf{p})$. The objective is generally non-smooth due to the selection operations in the anchor and evidential losses: the anchor loss selects the center with the largest cosine similarity, whereas the evidential loss selects the nearest non-target sample under the ℓ_1 distance. These selections may change with $\mathbf{p}$, and the ℓ_1 distance is non-differentiable at zero coordinate-wise differences. The analysis below is stated for differentiable iterates of F.

We constrain $\mathbf{p}$ to $\mathcal{P} = [-2, 2]^d$ and update it as

$$\mathbf{p}_{k+1} = \Pi_{\mathcal{P}} \left(\mathbf{p}_k - \eta \, \operatorname{sign}(\mathbf{g}_k) \right), \tag{31}$$

where $\mathbf{g}_k = \nabla F(\mathbf{p}_k)$, $\eta > 0$ is the trigger learning rate, and $\Pi_{\mathcal{P}}$ denotes projection onto $\mathcal{P}$.

Projected Stationarity. To characterize the effective update after projection, let $\mathbf{h}_k = (\mathbf{p}_k - \mathbf{p}_{k+1})/\eta$. Under the box projection, each coordinate satisfies $|h_{k,l}| \leq 1$ and $g_{k,l} h_{k,l} \geq 0$. We therefore define the projected sign-gradient stationarity measure as

$$\mathcal{S}_\eta(\mathbf{p}_k) = \langle \mathbf{g}_k, \mathbf{h}_k \rangle . \tag{32}$$

This measure quantifies the first-order descent component associated with the actual projected update.

Assumption 5 (Pathwise Quadratic Upper Bound)**.** *There exists a constant $L > 0$ such that, for every consecutive pair $(\mathbf{p}_k, \mathbf{p}_{k+1})$ along the optimization path,*

$$F(\mathbf{p}_{k+1}) \leq F(\mathbf{p}_k) + \langle \mathbf{g}_k, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle + \frac{L}{2} \|\mathbf{p}_{k+1} - \mathbf{p}_k\|_2^2 . \tag{33}$$

Moreover, F is lower bounded on $\mathcal{P}$ by a finite constant $F_{\inf}$.

Lemma 3 (One-Step Descent)**.** *Under Assumption 5, the update in Eq. 31 satisfies*

$$F(\mathbf{p}_{k+1}) \leq F(\mathbf{p}_k) - \eta \mathcal{S}_\eta(\mathbf{p}_k) + \frac{L \eta^2 d}{2}. \tag{34}$$

Proof. By the definition of $\mathbf{h}_k$, $\mathbf{p}_{k+1} - \mathbf{p}_k = -\eta \mathbf{h}_k$, and therefore

$$\langle \mathbf{g}_k, \mathbf{p}_{k+1} - \mathbf{p}_k \rangle = -\eta \mathcal{S}_\eta(\mathbf{p}_k).$$

Moreover, $|h_{k,l}| \leq 1$ implies $\|\mathbf{h}_k\|_2^2 \leq d$ and hence

$$\|\mathbf{p}_{k+1} - \mathbf{p}_k\|_2^2 = \eta^2 \|\mathbf{h}_k\|_2^2 \leq \eta^2 d.$$

Substituting these relations into Eq. 33 yields the claimed bound. □

Algorithm 1 IDEA at the n-th attacker

Input: Global model θ_g, local model θ_n, local dataset $\mathcal{D}_n$, target class c_n, source sample set $\mathcal{X}_n$, trigger δ_n, learning rate η, anchor ratio ρ, number of region centers K_r, λ_{evid}, and trigger optimization epochs E.

$\mathcal{D}_n^{c_n} \leftarrow \{x \in \mathcal{D}_n \mid y = c_n\}$, $\mathcal{D}_n^{\neg c_n} \leftarrow \{x \in \mathcal{D}_n \mid y \neq c_n\}$;
$\theta_n^t \leftarrow \theta_g$;
/ Stable target region construction */*
for $x_i \in \mathcal{D}_n^{c_n}$ **do**
  Compute the multinomial opinion $\mathbf{w}_i = (\mathbf{b}_i, u_i, \mathbf{a}_i)$;
  Compute $\text{ADS}(x_i)$ from $\mathbf{w}_i$ by Eq. (10);
end for
Select $K_d = \lfloor \rho |\mathcal{D}_n^{c_n}| \rfloor$ reliable anchors $\mathcal{A}$ with the lowest ADS scores;
$\mathcal{R} = \{\mathbf{q}_1, \ldots, \mathbf{q}_{K_r}\} \leftarrow \text{KMeans}(\{f_{\theta_g}(x) \mid x \in \mathcal{A}\}, K_r)$;
Compute the target evidential center $\boldsymbol{\mu}_n^{\alpha} \leftarrow \dfrac{1}{|\mathcal{D}_n^{c_n}|} \displaystyle\sum_{x_i \in \mathcal{D}_n^{c_n}} \boldsymbol{\alpha}_i$;
/ Trigger optimization */*
for $e = 1, \ldots, E$ **do**
  for $x_i \in \mathcal{X}_n$ **do**
    Construct the triggered sample $\tilde{x}_i = x_i \oplus \delta_n$;
    Compute the Dirichlet parameter $\tilde{\boldsymbol{\alpha}}_i$ for $\tilde{x}_i$;
    Identify the nearest non-target sample x_j;
  end for
  Compute $\mathcal{L}_{\text{ID}}$ by Eq. (4);
  Compute $\mathcal{L}_{\text{anchor}}$ by Eq. (12) using $\mathcal{R}$;
  Compute $\mathcal{L}_{\text{evi}}$ by Eq. (13) using $\tilde{\boldsymbol{\alpha}}_i$, x_j, and $\boldsymbol{\mu}_n^{\alpha}$;
  $\mathcal{L}_{\text{IDEA}} \leftarrow \mathcal{L}_{\text{ID}} + \mathcal{L}_{\text{anchor}} + \lambda_{\text{evid}} \mathcal{L}_{\text{evid}}$;
  $\delta_n \leftarrow \delta_n - \eta \nabla_{\delta_n} \mathcal{L}_{\text{IDEA}}$;
end for
/ Local training */*
Construct the backdoor dataset $\widetilde{\mathcal{D}}_n$ using the optimized δ_n;
Locally optimize θ_n^{t+1} on $\widetilde{\mathcal{D}}_n$;
$\Delta_n^{t+1} \leftarrow \theta_n^{t+1} - \theta_n^t$;
Upload Δ_n^{t+1} to the server;

Theorem 4 (Finite-Step Stationarity Bound)**.** *Under Assumption 5, the iterates generated by Eq. 31 satisfy*

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathcal{S}_{\eta}(\mathbf{p}_k) \leq \frac{F(\mathbf{p}_0) - F_{\text{inf}}}{\eta T} + \frac{L \eta d}{2}. \tag{35}$$

In particular, there exists an iterate $\mathbf{p}_k$, $0 \leq k < T$, *such that*

$$\min_{0 \leq k < T} \mathcal{S}_{\eta}(\mathbf{p}_k) \leq \frac{F(\mathbf{p}_0) - F_{\text{inf}}}{\eta T} + \frac{L \eta d}{2}. \tag{36}$$

Proof. Rearranging Lemma 3 and summing over $k = 0, \ldots, T-1$ gives

$$\eta \sum_{k=0}^{T-1} \mathcal{S}_{\eta}(\mathbf{p}_k) \leq F(\mathbf{p}_0) - F(\mathbf{p}_T) + \frac{L \eta^2 d T}{2}.$$

Since $\mathbf{p}_T \in \mathcal{P}$ and $F(\mathbf{p}_T) \geq F_{\text{inf}}$, division by ηT yields Eq. 35. The second result follows because the minimum of a finite set does not exceed its average. □

For a fixed learning rate η, the transient term $(F(\mathbf{p}_0) - F_{\text{inf}})/(\eta T)$ vanishes as T increases, leaving the residual bound $L\eta d/2$. Thus, increasing the number of trigger optimization steps reduces the transient component, while a smaller learning rate tightens the residual bound. The analysis applies directly to the original IDEA objective and the projected sign-gradient update used by IDEA, without introducing a surrogate objective.

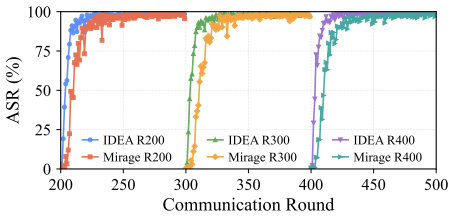


Figure 9: Attack performance under different attack initiation rounds on CIFAR-100 dataset.

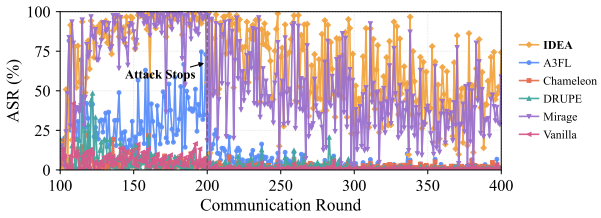


Figure 10: Backdoor persistence under continued benign training on CIFAR-10 dataset.

E Algorithm Overview

The overall optimization procedure of IDEA at the n-th attacker is presented in Algorithm 1, which follows a three-stage pipeline: stable target region construction, trigger optimization, and local model training. Given the global model θ_g, local dataset $\mathcal{D}_n$, attacker-specified target class c_n, source sample set $\mathcal{X}_n$, and initial trigger δ_n, IDEA first initializes the local model θ_n^t with θ_g and partitions $\mathcal{D}_n$ into target-class samples $\mathcal{D}_n^{c_n}$ and non-target samples $\mathcal{D}_n^{\neg c_n}$ (Lines 1–2).

In the first stage, IDEA constructs stable target regions from the target-class samples $\mathcal{D}_n^{c_n}$ (Lines 3–9). For each target-class sample x_i, its multinomial opinion $\mathbf{w}_i$ is obtained and used to compute the corresponding ADS (Lines 3–6). The $\rho\%$ of target-class samples with the lowest ADS scores are then selected as reliable anchors (Line 7), whose feature representations are clustered via K-means to obtain the representative region centers $\mathcal{R}$ (Line 8). Additionally, the Dirichlet parameters of the benign target-class samples are averaged to construct the target evidential center $\boldsymbol{\mu}_n^{\alpha}$ (Line 9), providing an evidential reference for subsequent trigger optimization.

In the second stage, IDEA optimizes the trigger δ_n over E epochs (Lines 10–21). For each source sample $x_i \in \mathcal{X}_n$, a triggered sample $\tilde{x}_i$ is first constructed, followed by the computation of its Dirichlet parameter $\tilde{\boldsymbol{\alpha}}_i$ (Lines 11–13). The nearest non-target sample $x_j \in \mathcal{D}_n^{\neg c_n}$ is then identified to provide a competing evidential reference (Line 14). Based on the resulting representations and evidential statistics, IDEA computes the target-mapping loss $\mathcal{L}_{\text{ID}}$, the target-region anchor attraction loss $\mathcal{L}_{\text{anchor}}$, and the evidential geometric refinement loss $\mathcal{L}_{\text{evid}}$ (Lines 16–18). These objectives are jointly formulated as $\mathcal{L}_{\text{IDEA}}$, encouraging triggered samples to exhibit the target behavior, align with the stable target regions, and maintain target-consistent evidential geometry (Line 19). The trigger is subsequently updated via gradient descent with respect to δ_n (Line 20).

In the final stage, IDEA uses the optimized trigger to construct the backdoor dataset $\widetilde{\mathcal{D}}_n$ (Line 22). The attacker then performs local training on $\widetilde{\mathcal{D}}_n$ to obtain the updated model θ_n^{t+1} and computes the corresponding model update $\Delta_n^{t+1} = \theta_n^{t+1} - \theta_n^t$ (Lines 23–24). The resulting update is finally uploaded to the server for federated aggregation (Line 25).

F Additional Experiments

F.1 Further Analysis of Attack Initiation and Lifespan

We further analyze attack initiation and backdoor lifespan to complement the results in Section 5.2. For attack initiation, the starting round is varied while fixing the attack duration at 100 communication rounds, as shown in Fig. 9. For backdoor lifespan, the attack is terminated after 100 rounds, followed by continued benign federated training to evaluate backdoor persistence, as shown in Fig. 10. All experiments use BackdoorIndicator as the defense.

We first compare IDEA with the strongest baseline, Mirage. As shown in Fig. 9, attacks are initiated at rounds 200, 300, and 400. Compared with the earlier setting in Fig. 4, these attacks are launched at later stages of training, when the global model is closer to convergence. The global representation space is thus more stable, resulting in similar overall attack behavior across different initiation points. Nevertheless, IDEA consistently reaches a high ASR earlier, indicating that it can establish

Table 4: Comparison of different backdoor attacks under SOTA defense methods on the FMNIST dataset. Acc, ASR, and AQ denote clean accuracy, attack success rate, and attack quality, respectively, with higher values indicating better performance. The best-performing results are highlighted in **bold**, while the second-best results are underlined. ASR values above 90% are marked with " $*$ ".

Defense	Vanilla			Chameleon			A3FL			DRUPE			Mirage			IDEA		
	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑	Acc↑	ASR↑	AQ↑
No Defense	90.21	84.87	47.65	90.37	50.74	47.62	**90.42**	92.69*	64.20	90.17	90.53	73.72	90.20	98.76*	91.34	90.31	**100.00***	**92.53**
Multi-Krum	88.45	10.14	5.08	88.31	48.74	39.52	**88.89**	77.75	39.83	88.73	88.45	72.83	88.06	87.07	75.99	88.05	**99.81***	**80.05**
Bulyan	**86.94**	48.27	22.48	76.56	45.77	42.03	79.53	89.76	57.79	80.59	82.79	52.49	80.51	98.75*	83.89	81.39	**99.80***	**84.36**
FoolsGold	88.89	5.26	1.55	88.72	49.16	43.65	89.10	65.67	41.75	**89.33**	60.79	31.94	89.19	99.66*	92.08	88.20	**99.96***	**92.23**
FLAME	88.33	39.98	12.30	87.36	49.75	44.89	87.84	90.44*	60.10	87.56	78.67	53.19	88.08	99.79*	94.59	88.18	**100.00***	**95.16**
Indicator	82.68	20.76	20.29	83.28	40.30	22.08	81.17	43.09	31.97	82.71	42.77	33.96	82.50	95.81*	66.57	**84.35**	**98.82***	**69.39**
Mars	89.47	79.33	39.19	**89.70**	49.01	37.59	89.49	91.93*	55.55	89.37	81.27	42.49	89.61	98.73*	75.36	89.00	**99.91***	**77.02**

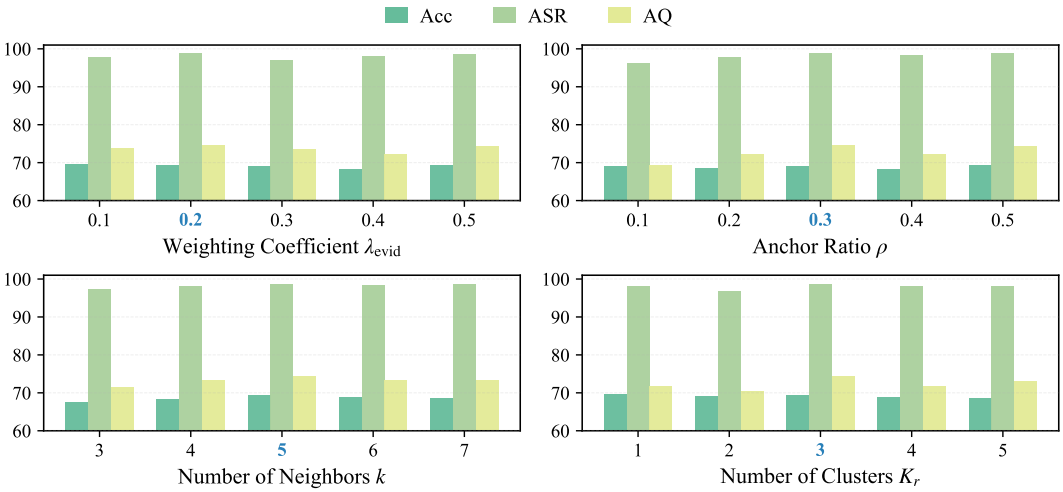


Figure 11: Sensitivity analysis of four IDEA hyperparameters: weighting coefficient λ_{evid}, anchor ratio ρ, number of neighbors k, and number of clusters K_r.

the backdoor with fewer poisoning rounds. Moreover, IDEA exhibits limited Attack Collapse and maintains a more stable ASR trajectory across different attack initiation points.

For backdoor lifespan, as shown in Fig. 10, most baseline methods experience a rapid decline in ASR after the attack ends, as subsequent benign updates progressively alter the learned representation space, leading to Attack Collapse to varying degrees. In contrast, IDEA exhibits a substantially smoother ASR decline and milder Attack Collapse, allowing the backdoor to remain effective for a longer period under continued benign training. These results demonstrate that IDEA achieves both faster backdoor establishment and stronger persistence against subsequent benign updates.

F.2 Accuracy, Attack Success, and Stability

In Section 5.3, we investigate the overall performance of different backdoor attacks across multiple datasets and defense mechanisms. To further assess IDEA in terms of clean accuracy, attack success, and attack stability, we conduct an additional evaluation on FMNIST under a strong non-IID setting with $\kappa = 0.3$. The results are reported in Table 4.

The results further reveal that final clean accuracy and ASR are not directly coupled. Although a relatively low clean accuracy does not necessarily indicate weak backdoor performance, pronounced degradation in clean accuracy is often accompanied by greater instability of the global model during training. This instability can perturb the evolving representation space, making the learned backdoor representation more susceptible to Attack Collapse and causing fluctuations in attack performance. AQ captures this temporal behavior by measuring the persistence and consistency of the attack beyond its final ASR. Therefore, evaluating IDEA jointly in terms of clean accuracy, ASR, and AQ provides a more complete view of its performance, particularly under heterogeneous federated training. From this perspective, IDEA achieves high and stable attack success while maintaining com-

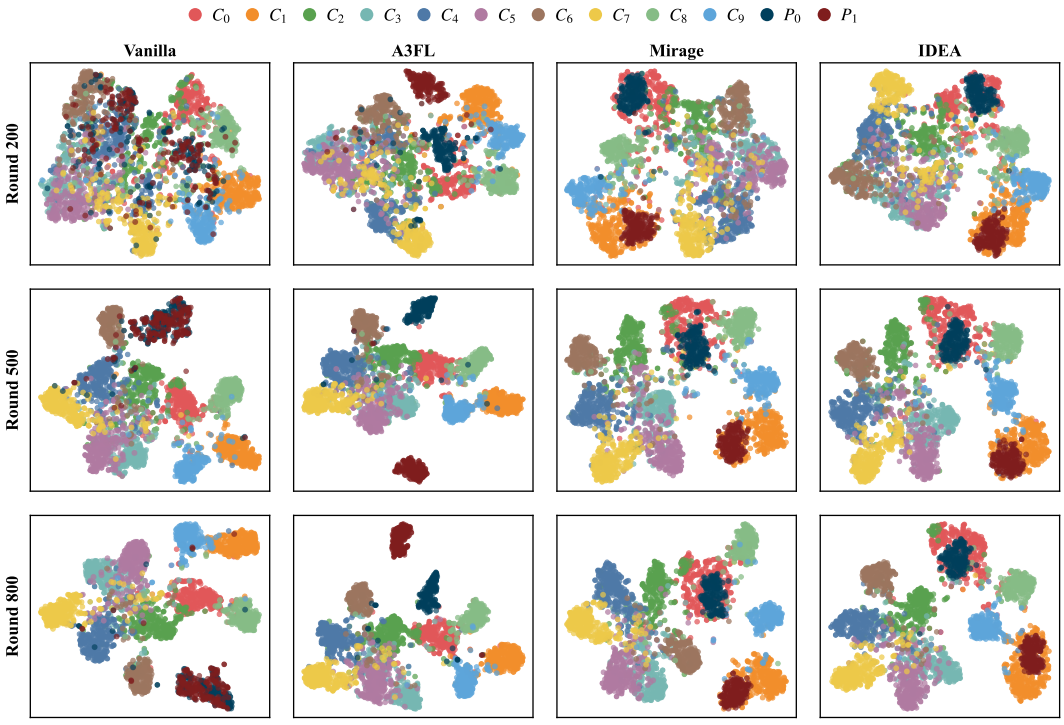


Figure 12: t-SNE visualization of feature-space evolution across training rounds and attack methods.

petitive clean accuracy, demonstrating that its backdoor performance does not come at the expense of benign model utility.

F.3 Hyperparameter Sensitivity

We further investigate the sensitivity of IDEA to four key hyperparameters, including the evidential weighting coefficient λ_{evid}, anchor ratio ρ, number of neighbors k, and number of clusters K_r. The results are shown in Fig. 11.

IDEA exhibits relatively stable performance across a broad range of configurations, with only moderate variations in clean accuracy, ASR, and AQ. The evidential weighting coefficient λ_{evid} has a limited effect on all three metrics and does not exhibit a consistent monotonic trend, suggesting that IDEA is relatively insensitive to its specific value. For the anchor ratio ρ, a small number of anchors provides insufficient evidential support for reliably characterizing the target regions, whereas an excessively large anchor set may incorporate less reliable information. An intermediate ratio therefore provides a more appropriate trade-off, yielding higher ASR and AQ. Similarly, increasing the number of neighbors k improves attack effectiveness and stability up to $k = 5$, after which the performance largely saturates. Varying the number of clusters K_r results in only modest performance fluctuations, with $K_r = 3$ achieving the strongest overall results.

Based on these observations, the corresponding values are adopted as the default hyperparameter configuration throughout our experiments.

F.4 Feature-Space Evolution under Different Attacks

We further investigate the feature-space behavior of different attacks. Specifically, BackdoorIndicator is enabled throughout training, and the feature representations of Vanilla, A3FL, Mirage, and IDEA are visualized at rounds 200, 500, and 800 in Fig. 12.

The results reveal a clear distinction between OOD- and ID-based backdoor attacks. As OOD-based attacks, Vanilla and A3FL produce poisoned samples that are noticeably separated from the benign feature distribution, making their backdoor representations more readily identifiable under the

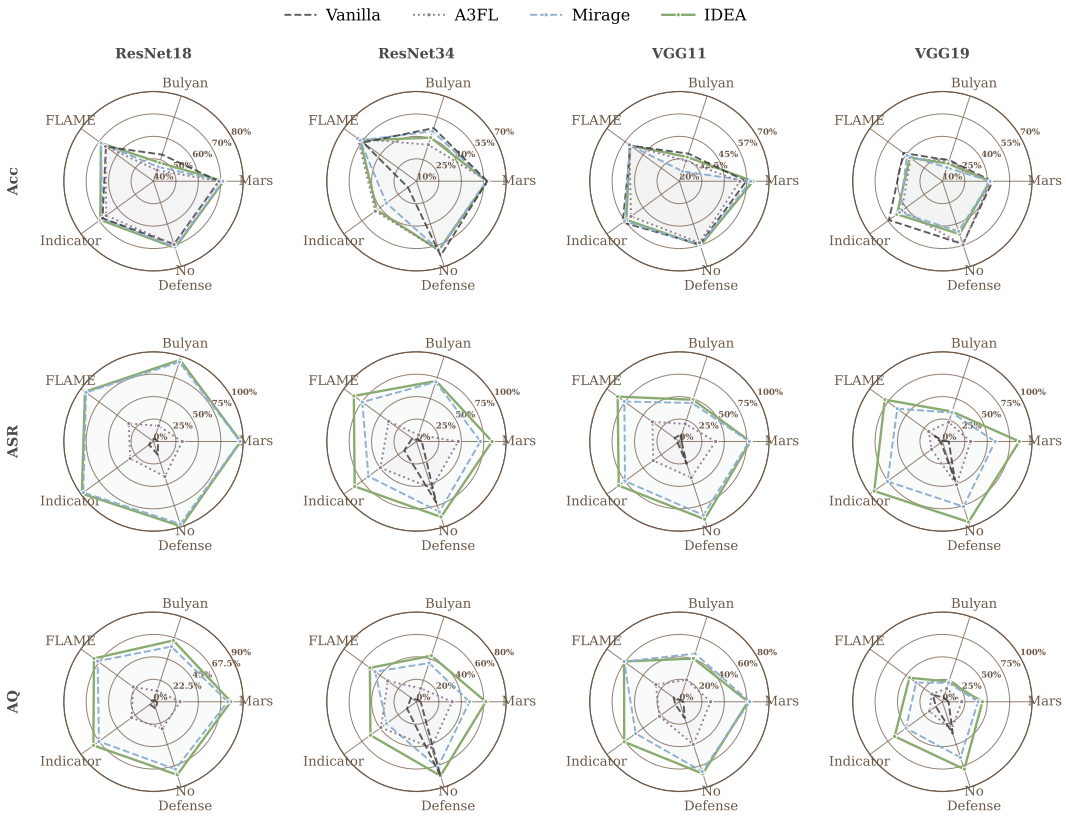


Figure 13: Comparison of different backdoor attacks across model architectures and defense methods in terms of clean accuracy (Acc), attack success rate (ASR), and attack quality (AQ).

BackdoorIndicator defense. Meanwhile, their benign non-target representations exhibit substantial overlap, with different classes becoming less well separated in the feature space. Such entanglement compromises the discriminability of clean samples and is consistent with their lower clean accuracy. In contrast, Mirage and IDEA embed poisoned samples within the ID feature space while better preserving the overall class structure of benign representations.

Compared with Mirage, IDEA tends to maintain more consistent alignment between poisoned samples and the target-class feature regions across training rounds, while preserving clearer class-wise organization among benign non-target samples. This suggests that IDEA introduces less interference with benign representations, which is consistent with its higher clean accuracy. Overall, the visualization provides qualitative evidence that IDEA can achieve effective ID poisoning while better preserving the underlying discriminative structure of the feature space.

F.5 Comprehensive Comparison across Architectures and Defenses

To further evaluate the generality and overall effectiveness of IDEA, we compare representative backdoor attacks under five defense settings, including Mars, Bulyan, FLAME, Indicator, and No Defence, across four model architectures: ResNet18, ResNet34, VGG11, and VGG19. The corresponding results are reported in Fig. 13.

IDEA achieves the most favorable overall performance profile across the evaluated architectures and defense mechanisms, with a consistent advantage in both ASR and AQ while maintaining competitive accuracy. Vanilla and A3FL exhibit substantially weaker attack effectiveness, whereas Mirage achieves strong ASR but shows greater variation in AQ across different settings. In contrast, IDEA maintains high ASR across all architecture–defense combinations, indicating that its attack effectiveness is less dependent on a specific model architecture or defense mechanism.

The advantage of IDEA becomes particularly evident when considering its robustness to different defenses and architectures. On ResNet34, IDEA consistently outperforms Mirage in both ASR and AQ under all five defense settings, demonstrating a clear advantage in sustaining effective poisoning against defensive intervention. A similar trend is observed on VGG19, where IDEA achieves higher ASR and AQ than Mirage across all defenses, with particularly pronounced gains under Mars, FLAME, IndicatorBackdoor, and No Defence. On VGG11, IDEA also maintains higher ASR than Mirage under all defenses, although Mirage remains competitive in AQ under a small number of defenses. Together, these results demonstrate that IDEA can preserve strong attack effectiveness as both the underlying architecture and defense mechanism vary.

IDEA further achieves a favorable balance between attack effectiveness and clean model utility. Its consistently higher ASR and generally stronger AQ are not accompanied by substantial degradation in Acc, which remains competitive with the baselines across most settings. This suggests that IDEA does not rely on sacrificing global model utility to obtain stronger poisoning effectiveness. Instead, its target-oriented poisoning strategy enables effective and stable attacks while maintaining comparable clean performance across heterogeneous architectures and defense mechanisms.