SPECSIA: RELIABLE SOURCE INFERENCE IN FEDERATED LEARNING VIA ADAPTIVE SPECTRAL FILTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables collaborative model training without directly sharing raw data, yet client-specific updates can still reveal which client contributed a target sample, exposing FL to Source Inference Attack (SIA). Existing passive attacks typically attribute the target to the client model with the lowest prediction loss, which faces two limitations: first, global aggregation propagates target-relevant knowledge across clients, obscuring each client's original contribution. Second, source attribution can be unstable across diverse federated configurations. Therefore, we propose Spectral Contribution-based Source Inference Attack (SpecSIA), which estimates target-relevant client contributions from current-round updates rather than relying on endpoint prediction loss. By combining normalized contribution attribution with adaptive spectral filtering, SpecSIA improves stability across communication rounds and federated settings. The attack operates using only information available in the current communication round and requires neither local optimization trajectories nor historical client updates. We evaluate source inference from four complementary perspectives: attack effectiveness, temporal reliability, statistical reliability across repeated runs, and cross-configuration consistency. Compared with the conventional loss-based attack, SpecSIA improves mean attack success rate (ASR) by up to 51.88% and reduces temporal variation by up to 54.34%, while maintaining competitive performance across heterogeneous federated configurations.

1 INTRODUCTION

Federated Learning (FL) enables multiple clients to collaboratively train a shared model without centrally collecting their raw data (McMahan et al., 2017; Yang et al., 2019). In each communication round, a server broadcasts the current global model, participating clients optimize it on their private datasets, and the server aggregates their uploaded model updates to produce the next global model (Li et al., 2020; Reddi et al., 2021). However, keeping raw data local does not provide a complete privacy guarantee (Truong et al., 2021; Truex et al., 2019). During the training process, the shared model updates may encode information about the local data distributions and individual training samples (Zhu et al., 2019; Geiping et al., 2020). A curious server that follows the FL protocol may exploit client-specific updates to perform Source Inference Attack (SIA) (Hu et al., 2021; 2024; Song et al., 2026).

A Source Inference Attack in FL goes beyond deciding whether a target sample participated in the training process, it seeks to identify the client that contributed the target (Hu et al., 2021; 2024; Li et al., 2025; Song et al., 2026; Chen et al., 2025). Such client-level provenance reveals the otherwise private association between individual training samples and their source clients, exposing information about data ownership and local data characteristics beyond membership alone (Shokri et al., 2017; He et al., 2024; Song et al., 2026; Nguyen et al., 2023; Nasr et al., 2019).

However, prediction-loss-based attribution can be unstable along two complementary dimensions. Across communication rounds, repeated aggregation propagates target-relevant knowledge across clients and can weaken source distinguishability, as reflected by the declining ASR in Figure 1. Across federated configurations, changes in heterogeneity, local training intensity, and data partitions can also lead to inconsistent attack performance, as shown in Figure 2.

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.

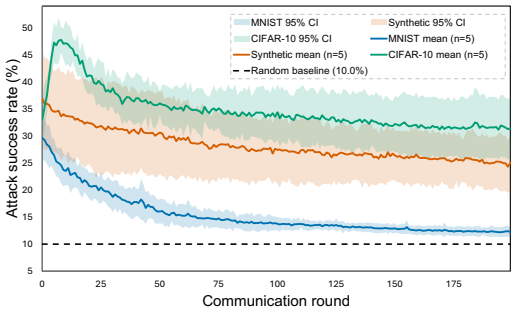


Figure 1: Loss-based ASR declines as communication rounds increase.

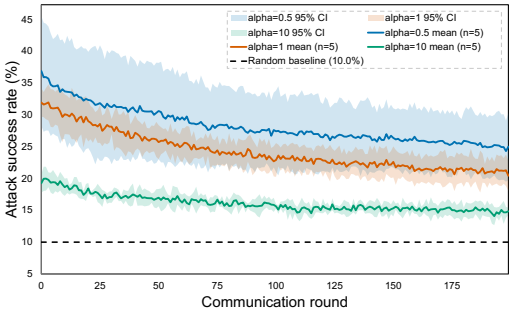


Figure 2: Heterogeneity improves performance but remains unstable.

Together, these observations suggest that reliable source inference should remain robust to both temporal variation during training and configuration variation across federated settings. This motivates two complementary research questions:

RQ1: *How can source inference maintain reliable attribution across communication rounds as federated training progresses?*

RQ2: *How can source inference maintain reliable performance across diverse federated configurations?*

To address these questions, we develop Spectral Contribution-based Source Inference Attack(SpecSIA), a passive and reference-free source inference method that attributes a target sample based on the target-relevant contribution of each client's current-round update. Round-Anchored Contribution Attribution (RCA) linearizes the target-loss reduction around the common pre-round model and normalizes the resulting first-order contribution, yielding a scale-invariant measure of target-update alignment. Adaptive Spectral Filtering(ASF) further suppresses dominant directions shared across clients while preserving more discriminative client-specific components. Together, they improve temporal stability across communication rounds and relative performance stability across diverse federated configurations. Our main contributions are:

- We characterize reliable source inference beyond average attack success by considering temporal reliability and cross-configuration consistency.
- We propose SpecSIA, which combines round-anchored normalized first-order attribution with adaptive spectral filtering to reduce update-scale sensitivity and dominant cross-client spectral structure.
- Across three datasets and 27 FedAvg configurations, SpecSIA improves mean ASR by up to 51.88% and reduces Temporal Std by up to 54.34% over loss-based SIA, while remaining competitive with LTMIA.

2 Related Work

2.1 Federated Learning

Federated Learning (FL) enables multiple clients to collaboratively train a shared model while keeping their raw data local (McMahan et al., 2017; Yang et al., 2019; Truong et al., 2021; Lai et al., 2022; Kairouz et al., 2021). In each communication round, the server distributes the current global model, clients perform local optimization on their private datasets, and the resulting model updates are aggregated to obtain the next global model (Li et al., 2020; Reddi et al., 2021). A major challenge in FL is statistical heterogeneity, where clients follow different data distributions, leading to divergent local updates and substantially affecting both local and global model behavior (Karimireddy et al., 2020; Wang et al., 2020). Although raw data are not directly exchanged, FL does not eliminate privacy leakage, as the updates may still encode information about local training data. Such information can be exploited to infer sensitive properties of client datasets (Truong et al., 2021;

Truex et al., 2019; Melis et al., 2019). Prior work has extensively studied membership inference attacks(MIA), which determine whether a target sample participated in model training (Shokri et al., 2017; Nasr et al., 2019; He et al., 2024; Song et al., 2026). SIA further extends this privacy threat by identifying which client contributed a known training sample in FL scenario.

2.2 Source Inference Attack

SIAs aim to identify the client that contributed a target training sample and can be broadly grouped into passive, active, and subject-level variants. Active SIAs strengthen source attribution by intervening in the federated training process. For example, Enhanced SIA actively amplifies client-specific differences to improve attribution performance (Chen et al., 2025). Subject-level SIAs extend the inference target from individual samples to subjects represented by multiple records. For example, SLSIA uses auxiliary target-subject data and an additional attack classifier to identify clients containing data from a target subject (Li et al., 2025).

Although active and subject-level SIAs can provide stronger or broader inference capabilities, they rely on additional intervention or auxiliary information. Passive SIAs avoid these stronger assumptions by performing source attribution directly from information naturally exposed during federated training. For example, loss-based SIAs evaluate a target sample on each local client model and attribute it to the client with the lowest prediction loss (Hu et al., 2021; 2024). However, this prediction-loss-based attribution still faces two key limitations. First, prediction loss may reflect target-relevant knowledge already inherited from the shared global model rather than only the contribution introduced by the client's current local update. Second, its attribution effectiveness can vary across federated configurations, making it difficult to maintain consistent relative performance under different heterogeneity levels, local-training settings, and FL protocols.

3 Problem Formulation

Federated Learning. We consider a FL system consisting of a central server and K clients, indexed by $\mathcal{K} = \{1, \ldots, K\}$. Each client i owns a private local dataset $\mathcal{D}_i$, which is never directly revealed to the server. At round t, the server broadcasts the pre-round global model θ_t to the participating clients. Each client initializes its local training from θ_t, performs local optimization on its private dataset, and obtains a locally trained model $\theta_{i,t}$. We represent the resulting client contribution by

$$\Delta\theta_{i,t} = \theta_t - \theta_{i,t}, \tag{1}$$

The server receives the current-round client updates and uses them for the prescribed federated aggregation procedure. Importantly, the individual update $\Delta\theta_{i,t}$ reflects the change induced by client i from the same pre-round model θ_t, providing a natural reference for comparing client-specific contributions within the current round.

Source Inference Attack. We study source inference for a target sample $z = (x, y)$ that is known to be a training member of one of the clients. The objective of Source Inference Attack is to identify the client whose private dataset contains the target sample. Given z, the attacker outputs a predicted source client

$$\widehat{i}_t(z) \in \mathcal{K}, \tag{2}$$

and the attribution is correct when $z \in \mathcal{D}_{\widehat{i}_t(z)}$.

In federated learning, the server naturally observes the global model and the client updates communicated during training. A *passive* SIA performs source attribution solely from information available to the server under the prescribed federated learning protocol. The server follows the training procedure normally and does not modify client selection, local optimization, model aggregation, or communicated updates for the purpose of the attack.

Under this setting, the attacker is given the target sample z, the pre-round global model θ_t, and the current-round client updates $\{\Delta\theta_{i,t}\}_{i\in\mathcal{K}}$. The goal is to infer which client contributed z to federated training without accessing the clients' private datasets.

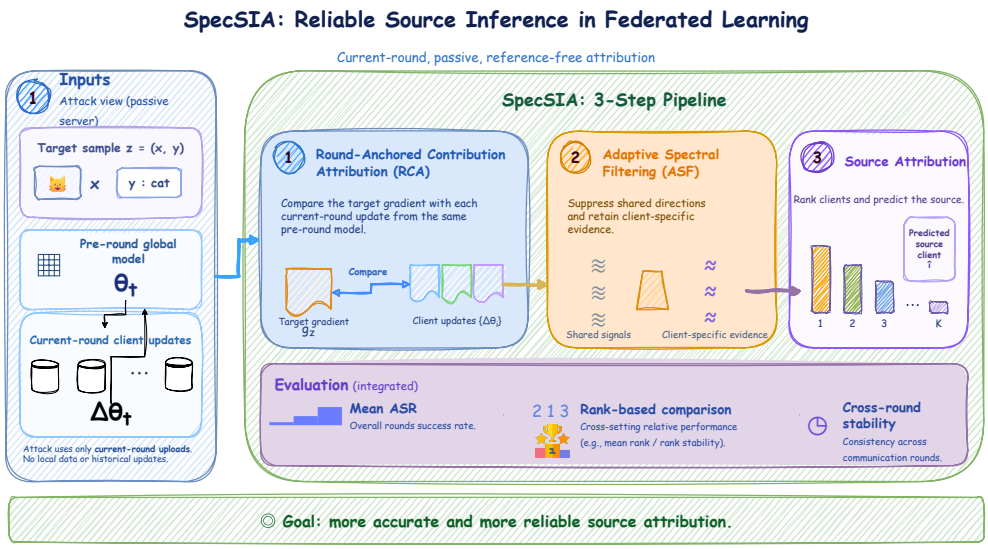


Figure 3: Overview of SpecSIA. RCA measures normalized target–update alignment from the pre-round model, while ASF attenuates dominant spectral directions with strong aggregate support across client updates. The resulting scores are used to infer the source client.

4 METHODOLOGY

4.1 FRAMEWORK OVERVIEW

As illustrated in fig. 3, our framework consists of two complementary modules: Round-Anchored Contribution Attribution (RCA) for target-conditioned contribution attribution and Adaptive Spectral Filtering (ASF) for adaptive spectral filtering.

Given a target sample $z = (x, y)$, the pre-round global model θ_t, and the current-round client updates $\{\Delta\theta_{i,t}\}$, RCA measures the normalized alignment between the target gradient and each client update from the same round anchor. ASF then suppresses dominant update directions shared across clients while preserving more client-specific evidence. Finally, SpecSIA ranks the resulting attribution scores and predicts the highest-ranked client as the source. The complete attack procedure is summarized in Algorithm 1 in Appendix A.

4.2 ROUND-ANCHORED CONTRIBUTION ATTRIBUTION (RCA)

Existing loss-based SIAs attribute a target sample according to the prediction losses of locally trained client models. ***However, endpoint prediction loss does not explicitly isolate the contribution introduced by the current client update.*** We therefore measure target-loss reduction from the common pre-round model θ_t and linearize it to expose its directional contribution.

Round-Anchored Contribution. Let the participating client uploads: $\Delta\theta_{i,t} = \theta_t - \theta_{i,t}$, and let $g_{z,t} = \nabla_\theta \ell(\theta_t; z)$ be the target gradient. The target-loss reduction induced by client i is $\ell(\theta_t; z) - \ell(\theta_{i,t}; z)$. Using a first-order expansion around θ_t:

$$\ell(\theta_t; z) - \ell(\theta_{i,t}; z) \approx g_{z,t}^\top \Delta\theta_{i,t}. \tag{3}$$

This term provides a target-specific approximation of the current client update's marginal contribution to target-loss reduction. A detailed error characterization is provided in Appendix C.1.

Scale Sensitivity. The first-order contribution can be decomposed as

$$g_{z,t}^\top \Delta\theta_{i,t} = \underbrace{\|g_{z,t}\|_2 \|\Delta\theta_{i,t}\|_2}_{\text{Magnitude}} \underbrace{\cos\angle(g_{z,t}, \Delta\theta_{i,t})}_{\text{Directional target relevance}}, \tag{4}$$

which mixes directional target relevance with client-update magnitude. While the directional term reflects how well a client update aligns with reducing the target loss, its magnitude can also be affected by local optimization conditions such as the number of local steps and data heterogeneity. Consequently, two clients with similar target relevance may receive different raw contribution scores simply because their update scales differ. Directly using this quantity for source attribution may therefore introduce undesirable scale sensitivity across clients and federated configurations.

Normalized Attribution. To isolate directional target relevance, we then normalize both the target gradient and client update:

$$\widehat{g}_{z,t} = \frac{g_{z,t}}{\|g_{z,t}\|_2}, \qquad b_{i,t} = \frac{\Delta\theta_{i,t}}{\|\Delta\theta_{i,t}\|_2}. \tag{5}$$

This normalization removes their magnitude terms and retains only the directional relation between the target gradient and each client update. The resulting attribution score is

$$s_{i,t}^{\mathrm{RCA}}(z) = \widehat{g}_{z,t}^{\top} b_{i,t}. \tag{6}$$

A larger score indicates stronger alignment between the client's update direction and the direction associated with reducing the target loss. Thus, RCA transforms a scale-sensitive first-order loss reduction into a round-anchored, scale-invariant measure of target-update alignment.

4.3 Adaptive Spectral Filtering (ASF)

Although RCA removes raw update-scale effects, ***normalized client updates may still contain dominant spectral directions with strong aggregate support across clients.*** These shared directions capture common update structure rather than source-specific variation, and therefore provide limited discrimination for source attribution. We introduce ASF to adaptively suppress such dominant cross-client structure while preserving relatively more client-specific evidence.

Spectral Representation. We collect the normalized client updates as

$$B_t = [b_{1,t}, \ldots, b_{K,t}] \in \mathbb{R}^{d\times K}, \tag{7}$$

where $B_t^{\top} B_t$ captures the cross-client geometry by pairwise alignment among normalized updates. We decompose $B_t = U_t \Sigma_t V_t^{\top}$, and define the normalized spectral energy of the j-th direction:

$$\mu_{j,t} = \frac{\sigma_{j,t}^2}{K}, \tag{8}$$

where $\sigma_{j,t}$ is the corresponding singular value and K is the number of client updates in the current round. A larger $\mu_{j,t}$ indicates stronger aggregate projection energy across the client updates.

Adaptive Spectral Retention. To determine how much of each spectral direction should be preserved, we calibrate the filtering strength according to the current-round spectrum:

$$\widehat{\lambda}_t = \mathrm{median}\left\{\mu_{j,t} : \mu_{j,t} > 0\right\}. \tag{9}$$

The retention factor for the j-th spectral direction is then

$$\rho_{j,t} = \frac{\widehat{\lambda}_t}{\widehat{\lambda}_t + \mu_{j,t}} = \frac{1}{1 + \mu_{j,t}/\widehat{\lambda}_t}. \tag{10}$$

Since $\rho_{j,t}$ decreases monotonically with $\mu_{j,t}$, directions with stronger aggregate support across clients are attenuated more heavily, while lower-energy directions preserve relatively more variation. This suppresses broadly shared spectral components that are less discriminative for source attribution, while retaining directions more likely to capture source-specific differences.

Filtered Representation. Using these retention factors, we rescale the spectral components of B_t and reconstruct the filtered update matrix as

$$\widetilde{B}_t = U_t \Sigma_t \,\mathrm{diag}\left(\rho_{1,t}, \ldots, \rho_{K,t}\right) V_t^{\top}. \tag{11}$$

Thus, $\rho_{j,t}$ explicitly controls how much of the j-th spectral direction is retained in the filtered client representations. Equivalently, this transformation can be written in client space as

$$\widetilde{B}_t = B_t \left[K\widehat{\lambda}_t \left(B_t^{\top} B_t + K\widehat{\lambda}_t I \right)^{-1} \right]. \tag{12}$$

Let $\widetilde{b}_{i,t}$ denote the i-th column of $\widetilde{B}_t$. Compared with the original normalized update $b_{i,t}$, $\widetilde{b}_{i,t}$ suppresses shared spectral structure while retaining relatively more client-specific variation. We use $\widetilde{b}_{i,t}$ in place of $b_{i,t}$ for the final target–update attribution.

4.4 Overall Source Attribution

Final Attribution. RCA provides a scale-invariant target gradient, while ASF produces client representations with shared spectral structure adaptively suppressed. Given the filtered representation $\widetilde{b}_{i,t}$, the final SpecSIA score measures its alignment with the normalized target gradient:

$$s_{i,t}^{\mathrm{SpecSIA}}(z) = \widehat{g}_{z,t}^{\top} \widetilde{b}_{i,t}. \tag{13}$$

A larger score indicates stronger target-specific alignment of the filtered client contribution. The source client is therefore inferred as $\widehat{i}_t(z) = \arg\max_{i \in \mathcal{K}} s_{i,t}^{\mathrm{SpecSIA}}(z)$. The resulting score combines the target-specific directional evidence from RCA with the client-specific structure retained by ASF.

Computational Cost. ASF operates on the $K \times K$ Gram matrix $B_t^{\top} B_t$, rather than directly in the d-dimensional parameter space. Its main additional cost is constructing the Gram matrix and solving a $K \times K$ linear system, which remains practical when $K \ll d$.

5 Experiments

5.1 Experimental Setup and Metrics

Datasets. Under the standard FedAvg protocol (McMahan et al., 2017), we evaluate SpecSIA on three datasets spanning image classification and controlled synthetic settings.

- **CIFAR-10.** (Krizhevsky et al., 2009) is an image classification benchmark containing 60,000 32×32 color images from 10 classes, with 50,000 training and 10,000 test samples.
- **MNIST.** (Lecun et al., 1998) consists of 70,000 grayscale handwritten-digit images from 10 classes, including 60,000 training and 10,000 test samples.
- **Synthetic.** We adopt the synthetic dataset and generation protocol released with the loss-based SIA (Hu et al., 2021), which provides a controlled setting for studying source inference under different levels of federated data heterogeneity.

Heterogeneity Setting. Client data are partitioned using a Dirichlet distribution (Hsu et al., 2019) with $\alpha \in \{0.5, 1, 10\}$, where smaller α induces stronger statistical heterogeneity across clients.

Training Details. We use $K = 10$ clients, with all clients participating in each communication round. Local training uses SGD with a learning rate of 0.01, momentum 0.9, and a batch size of 64. We vary the number of local epochs as $E \in \{1, 5, 10\}$ and train each federated configuration for 200 communication rounds. For each configuration, we conduct five independent runs using different random seeds, and report results averaged over these runs.

Counterparts. We compare SpecSIA with three representative passive SIAs:

- **Loss-based** [TDSC'24] (Hu et al., 2024). It performs source attribution based on the prediction loss of the client on the target sample, identifying the client with the lowest loss.
- **LTMIA** [JISA'26] (Song et al., 2026). It extends source inference beyond a single model snapshot by exploiting information accumulated over multiple communication rounds. We adapt LTMIA to our round-wise evaluation using a sliding loss window of length 10.
- **Residual**. A leave-one-client-out subspace baseline that projects the target gradient onto the orthogonal complement of the subspace spanned by the other clients' updates.

Metrics. We evaluate source inference from five complementary perspectives. Mean ASR measures overall attack effectiveness across communication rounds, while Temporal Std characterizes round-to-round variation and thus temporal stability. The lower endpoint of a normal-approximation 95% confidence interval (95% LCB) captures cross-run statistical reliability by considering both the

Table 1: Absolute SIA performance under different data heterogeneity levels α and local epochs E. ASR measures attack success rate, Std measures variation across communication rounds, and 95% LCB denotes the lower confidence bound across repeated runs. **Bold** denotes the best result.

Dataset	α	E	Loss-based			LTMIA			Ours		
			ASR(↑)	Std(↓)	95% LCB(↑)	ASR(↑)	Std(↓)	95% LCB(↑)	ASR(↑)	Std(↓)	95% LCB(↑)
CIFAR-10	0.5	1	33.57	**1.58**	30.97	**35.13**	1.74	**32.49**	32.35	2.15	29.54
		5	34.57	4.17	31.35	**36.77**	4.79	**33.65**	34.17	**2.37**	31.19
		10	34.47	4.77	31.54	**37.15**	5.68	**34.03**	34.61	**2.26**	31.84
	1	1	27.14	**1.75**	25.52	**28.81**	2.84	**27.06**	26.89	2.15	25.12
		5	27.74	5.79	26.44	**29.70**	7.16	**28.01**	29.37	**3.53**	27.19
		10	27.60	6.69	26.48	**30.29**	8.71	**28.57**	29.83	**3.45**	27.59
	10	1	17.03	**2.60**	16.30	**18.77**	4.90	**18.11**	18.24	2.89	17.70
		5	18.23	7.89	17.25	**20.11**	10.51	**19.26**	19.54	**6.41**	18.87
		10	18.49	9.13	17.71	**20.55**	12.52	**19.88**	19.91	**6.68**	19.03
	Overall		26.54	4.93	24.84	**28.59**	6.54	**26.78**	27.21	**3.54**	25.34
MNIST	0.5	1	20.43	4.43	18.94	22.48	4.75	20.95	**28.43**	**2.68**	**26.50**
		5	15.37	3.93	14.51	16.85	5.14	15.68	**28.72**	**2.01**	**26.18**
		10	14.09	3.46	13.33	15.56	4.95	14.63	**27.88**	**1.86**	**25.70**
	1	1	16.11	3.02	14.99	17.65	3.88	16.24	**22.85**	**2.45**	**20.59**
		5	13.33	2.60	12.39	14.18	3.51	13.14	**23.20**	**1.66**	**20.58**
		10	12.76	2.21	12.01	13.65	3.29	12.78	**22.22**	**1.75**	**19.68**
	10	1	11.25	**0.85**	11.06	11.47	0.97	11.22	**12.17**	1.06	**11.54**
		5	11.01	0.74	10.93	11.14	1.03	11.08	**12.36**	**0.73**	**11.27**
		10	10.87	0.79	10.77	11.10	1.12	10.98	**12.36**	**0.66**	**11.34**
	Overall		13.91	2.45	13.21	14.90	3.18	14.08	**21.13**	**1.65**	**19.26**
Synthetic	0.5	1	34.63	1.49	28.08	**34.93**	1.52	**28.10**	**34.93**	**1.18**	27.42
		5	28.27	2.84	23.94	28.89	3.01	24.35	**32.63**	**0.84**	**25.65**
		10	25.68	2.79	22.02	26.32	3.14	22.56	**31.85**	**0.84**	**25.25**
	1	1	29.97	1.81	28.01	**30.54**	1.58	**28.58**	30.16	**1.10**	28.15
		5	24.27	2.99	22.83	25.01	3.13	23.55	**29.10**	**0.94**	**26.46**
		10	21.55	3.07	19.96	22.23	3.41	20.65	**27.89**	**1.10**	**25.23**
	10	1	18.06	1.15	17.41	**18.71**	**0.85**	**18.09**	17.16	0.98	16.00
		5	16.14	1.44	15.73	16.65	1.40	16.14	**17.62**	**0.85**	**16.44**
		10	15.31	1.32	14.84	15.76	1.54	15.21	**17.37**	**0.80**	**16.33**
	Overall		23.76	2.10	21.42	24.34	2.18	21.91	**26.52**	**0.96**	**22.99**

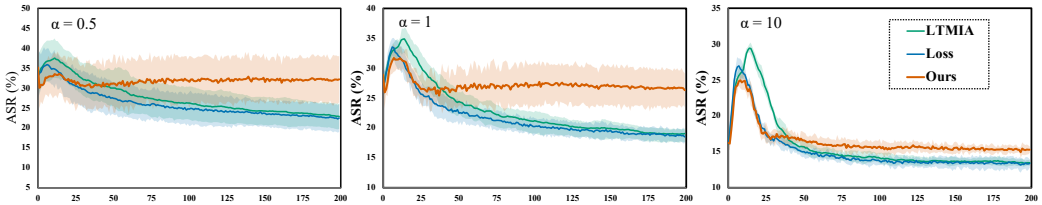


Figure 4: Mean ASR over communication rounds under different data heterogeneity levels. Each curve reports the ASR averaged across CIFAR-10, MNIST, and Synthetic, while the shaded region denotes the 95% confidence interval across five independent runs. More analysis see section 5.

mean performance and its variation across independent runs. To further assess performance across different federated configurations, we report Mean Rank, which summarizes the average relative ranking of each method, and Rank IQR, which measures the dispersion of these ranks across configurations. Additional implementation details and metric definitions are provided in Appendices A and B.

5.2 Overall Source Inference Performance

Comparison to State-of-the-Arts. Table 1 reports source inference performance under different heterogeneity levels and local epochs. Compared with Loss-based SIA, SpecSIA improves Mean ASR by 2.54%, 51.88%, and 11.61% on CIFAR-10, MNIST, and Synthetic, while reducing Temporal Std by 28.13%, 32.55%, and 54.34%, respectively. These gains indicate that SpecSIA better preserves client-specific evidence as global aggregation progresses, with more stable attribution across

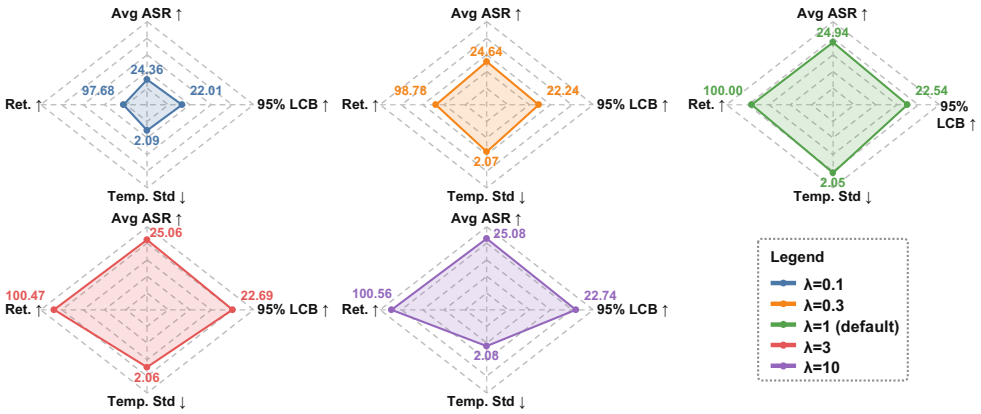


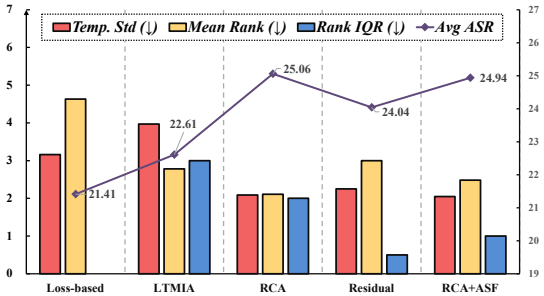
Figure 5: Sensitivity analysis with respect to γ. The figure reports Avg ASR, ASR Retention, Temporal Std, and 95% LCB under different γ values. The default setting is $\gamma = 1$.

rounds. Compared with LTMIA, SpecSIA achieves stronger effectiveness and temporal reliability on MNIST and Synthetic. On CIFAR-10, LTMIA attains higher ASR and LCB, while SpecSIA reduces Temporal Std by 45.87%. Overall, the results show that SpecSIA maintains reliable source attribution across varying heterogeneity and local-training settings.

Training Process Curve. Figure 4 shows the evolution of cross-dataset Mean ASR throughout federated training. For $\alpha = 0.5$ and 1, all methods are relatively competitive in the early rounds, but the performance of Loss-based SIA and LTMIA gradually degrades as communication proceeds, whereas SpecSIA maintains substantially stronger late-round attribution. This widening gap supports our motivation that repeated aggregation progressively obscures client-specific evidence, while SpecSIA better isolates the contribution introduced by the current client update. When $\alpha = 10$, the gap narrows as client distributions become more homogeneous, making source-specific evidence inherently weaker and less distinguishable across clients.

5.3 Diagnostic Analysis

Component Analysis. We examine the contributions of the components of SpecSIA. The configuration without either component corresponds to Loss-based SIA, while enabling both components gives the complete SpecSIA. RCA improves Avg ASR and 95% LCB over Loss-based SIA, indicating that measuring client-specific contribution provides stronger attribution evidence. ASF reduces Temporal Std, showing that suppressing shared dominant directions improves temporal stability across communication rounds. Figure 6 reports the component ablation over 27 FedAvg configurations. Introducing RCA increases Avg ASR from 21.40 to 25.08 and 95% LCB from 19.83 to 22.75, while reducing Temporal Std from 3.16 to 2.09. Adding ASF retains comparable attack effectiveness and further reduces Temporal Std to 2.05, indicating improved temporal stability.



RCA	ASF	Avg ASR↑	Temp. Std↓	95% LCB↑
✗	✗	21.40	3.16	19.83
✓	✗	**25.08**	2.09	**22.75**
✓	✓	24.96	**2.05**	22.53

Figure 6: Component and cross-configuration analysis. The upper-right figure evaluates performance and reliability across configurations, while the lower-right table reports the component ablation.

Cross-Configuration Analysis. We further evaluate cross-configuration consistency across Loss-based SIA, LTMIA, RCA, Residual, and SpecSIA, ranking methods by Mean ASR within each federated configuration. As shown in Figure 6, RCA achieves the highest Avg ASR and best Mean

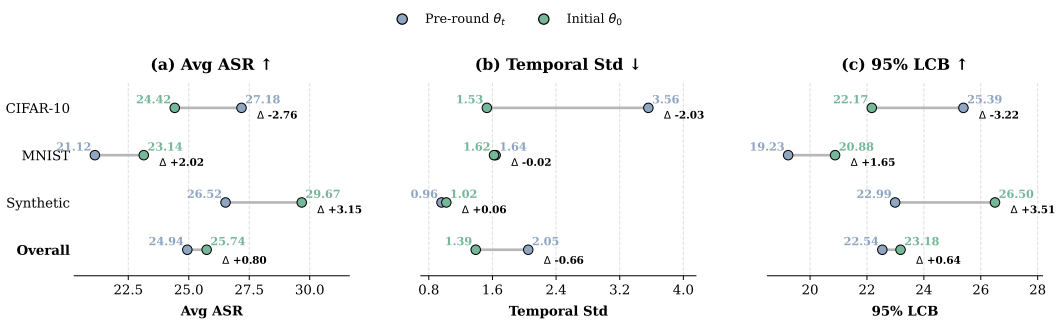


Figure 7: Gradient-anchor analysis over 27 FedAvg configurations, comparing the pre-round anchor θ_t and initial anchor θ_0 in Avg ASR, Temporal Std, and 95% LCB. Δ denotes the change from θ_t to θ_0.

Rank, while SpecSIA attains the lowest Temporal Std and reduces Rank IQR from 2.00 to 1.00 relative to RCA. Although Loss-based SIA and Residual obtain smaller Rank IQR, their worse Mean Rank and lower Avg ASR indicate that rank dispersion should be interpreted jointly with attack effectiveness. Overall, RCA contributes most of the effectiveness gain, whereas ASF improves temporal and cross-configuration stability with only a marginal reduction in Avg ASR.

Sensitivity to Spectral Calibration We perturb the adaptive calibration by setting $\lambda_t^{(\gamma)} = \gamma\widehat{\lambda}_t$ with $\gamma \in \{0.1, 0.3, 1, 3, 10\}$. Figure 5 shows that SpecSIA is robust to spectral calibration, retaining at least 97.68% of the default ASR with Temporal Std between 2.05 and 2.09. As γ increases, the retention factor $\rho_j = \lambda_t^{(\gamma)}/(\lambda_t^{(\gamma)} + \mu_j)$ increases, leading to weaker spectral attenuation and slightly higher Avg ASR and 95% LCB, suggesting that dominant directions may still contain useful source-discriminative information. We nevertheless use $\gamma = 1$, which directly adopts the adaptive scale $\widehat{\lambda}_t$, achieves nearly peak ASR, and yields the lowest Temporal Std without introducing an additional tuned multiplier. Dataset-wise sensitivity results are provided in Appendix F.

Gradient Anchor Analysis. Figure 7 compares the two gradient anchors across datasets and metrics. Table 2 summarizes their definitions and server views. Using the initial model θ_0 improves Avg ASR and 95% LCB on MNIST and Synthetic, but degrades both metrics on CIFAR-10. Overall, the initial anchor increases Avg ASR from 24.94 to 25.74 and 95% LCB from 22.54 to 23.18, while reducing Temporal Std from 2.05 to 1.39.

Table 2: Two Gradient-anchor definitions and server views. Results see Fig. 7.

Anchor	Definition	Server View
Pre-round	$g_{z,t} = \nabla_\theta \ell(\theta_t; z)$	Current round
Initial	$g_{z,0} = \nabla_\theta \ell(\theta_0; z)$	Initial state

The initial-model anchor captures information accumulated throughout the training trajectory, whereas the pre-round anchor focuses on attribution relative to the current communication round. Although θ_0 achieves higher overall Avg ASR and 95% LCB, using it would introduce additional historical model-state information that is not assumed by the conventional loss-based baseline. To ensure a fair comparison under the same current-round server view, we therefore retain θ_t as the default anchor rather than selecting the empirically stronger θ_0 setting. Even under this more restrictive choice, SpecSIA still outperforms the loss-based attack. Configuration-level gradient-anchor results are provided in Appendix G.

6 Conclusion

We presented SpecSIA, a passive and reference-free source inference attack that attributes target samples through normalized current-round client contributions and adaptive spectral filtering. Experiments across heterogeneous federated configurations show that SpecSIA maintains competitive attack effectiveness while improving temporal reliability and cross-configuration consistency. Our results further show that reliable source inference should be evaluated beyond average attack success, considering both temporal and configuration-level variation.

Reproducibility Statement

We provide the complete SpecSIA algorithm, federated training configurations, attack settings, random seeds, target-sampling procedure, and evaluation metric definitions in Appendices A and B. Additional derivations and configuration-level experimental results are provided in the subsequent appendices.

AI Use Statement

We used generative AI tools to assist with writing and language polishing, literature retrieval and summarization, research brainstorming, and the refinement of the manuscript structure. Generative AI was also used to support the development and refinement of methodological descriptions, mathematical derivations and proofs, experimental analysis, and interpretation of results. In addition, AI tools assisted with the preparation and refinement of figures and code used for analysis and visualization.

All AI-assisted mathematical arguments, methodological decisions, experimental interpretations, references, figures, code, and manuscript text were independently reviewed and verified by the authors. All reported numerical results were obtained from the authors' own implementations and experimental runs. The authors take full responsibility for the final content of this work.

References

Haiyang Chen, Xiaolong Xu, Xiang Zhu, Xiaokang Zhou, Fei Dai, Yansong Gao, Xiao Chen, Shuo Wang, and Hongsheng Hu. Where does this data come from? enhanced source inference attacks in federated learning. In *IJCAI*, pp. 4815–4823, 2025.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947, 2020.

Xinlong He, Yang Xu, Sicong Zhang, Weida Xu, and Jiale Yan. Enhance membership inference attacks in federated learning. *Comput. Secur.*, 136:103535, 2024. doi: 10.1016/j.cose.2023.103535.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. Source inference attacks in federated learning. In *ICDM*, pp. 1102–1107. IEEE, 2021. doi: 10.1109/ICDM51629.2021.00129.

Hongsheng Hu, Xuyun Zhang, Zoran Salcic, Lichao Sun, Kim-Kwang Raymond Choo, and Gillian Dobbie. Source inference attacks: Beyond membership inference attacks in federated learning. *IEEE TDSC*, 21(4):3012–3029, 2024. doi: 10.1109/TDSC.2023.3321565.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning.

In *ICML*, volume 119, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

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

Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. Fedscale: Benchmarking model and system performance of federated learning at scale. In *ICML*, pp. 11814–11827, 2022.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jiaxin Li, Marco Arazzi, Antonino Nocera, and Mauro Conti. Subject data auditing via source inference attack in cross-silo federated learning. *JISA*, 90:104034, 2025.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, volume 2, pp. 429–450, 2020. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *IEEE S&P*, pp. 691–706. IEEE, 2019.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE S&P*, pp. 739–753. IEEE, 2019. doi: 10.1109/SP.2019.00065.

Truc Nguyen, Phung Lai, Khang Tran, NhatHai Phan, and My T. Thai. Active membership inference attack under local differential privacy in federated learning. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5714–5730. PMLR, 2023.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE S&P*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41.

Jia Song, Jianting Yuan, Guanxin Chen, Yipeng Liu, and Nan Yang. LTMIA: A loss trajectory-based membership inference attack method in federated learning. *JISA*, 97:104364, 2026. doi: 10.1016/j.jisa.2025.104364.

Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *AISec*, pp. 1–11, 2019.

Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Comput. Secur.*, 110:102402, 2021.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *NeurIPS*, volume 33, pp. 7611–7623, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/564127c03caab942e503ee6f810f54fd-Abstract.html.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM TIST*, 10(2):12:1–12:19, 2019. doi: 10.1145/3298981.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

Table 3: Summary of the main notation used in SpecSIA.

Symbol	Definition
$z = (x, y)$	Target sample with input x and label y
K	Number of clients
i, t	Client index and communication round
$\ell(\theta; z)$	Loss of model θ on target z
θ_t	Pre-round global model at round t
$\theta_{i,t}$	Locally trained model of client i
$\Delta\theta_{i,t}$	Client update, $\theta_t - \theta_{i,t}$
$g_{z,t}$	Target gradient, $\nabla_\theta \ell(\theta_t; z)$
$\hat{g}_{z,t}$	Normalized target gradient
$b_{i,t}$	Normalized client update
B_t	Normalized client-update matrix $[b_{1,t}, \ldots, b_{K,t}]$
$\mu_{j,t}$	Normalized energy of the j-th spectral direction
$\hat{\lambda}_t$	Adaptive spectral calibration
$\rho_{j,t}$	Retention factor of the j-th spectral direction
$\widetilde{B}_t$	Spectrally filtered client-update matrix
$\widetilde{b}_{i,t}$	Filtered representation of client i
$s_{i,t}(z)$	Source-attribution score for client i
$\hat{i}_t(z)$	Predicted source client

Table 4: Model and federated-training settings.

Dataset	Model	K	Batch Size	Learning Rate
CIFAR-10	CifarCnn	10	64	0.01
MNIST	MnistCnn	10	64	0.01
Synthetic	MLP (hidden width 200)	10	64	0.01

A Implementation and Reproducibility Details

A.1 SpecSIA Algorithm

Algorithm 1 summarizes the complete SpecSIA procedure. The attack uses only the target sample, the current pre-round global model, and the client updates available in the current communication round.

A.2 Dataset, Federated Learning, and Attack Settings

We evaluate CIFAR-10, MNIST, and Synthetic with $K = 10$ clients. All clients participate in every communication round. The client models and basic optimization settings are summarized in Table 4.

Local optimization uses SGD with momentum 0.9 and no weight decay. We vary the number of local epochs as

$$E \in \{1, 5, 10\},$$

and control client-data heterogeneity using

$$\alpha \in \{0.5, 1, 10\}.$$

Each dataset therefore contains nine (α, E) configurations. Every configuration is trained for 200 communication rounds and repeated with five random seeds

$$\{42, 43, 44, 45, 46\}.$$

The random seed controls model initialization, client-data partitioning, and target sampling.

Algorithm 1 SpecSIA Source Attribution

Require: Target sample z, pre-round model θ_t, client updates $\{\Delta\theta_{i,t}\}_{i=1}^{K}$

1: Compute the target gradient
$$g_{z,t} \leftarrow \nabla_\theta \ell(\theta_t; z)$$
2: Normalize the target gradient
$$\widehat{g}_{z,t} \leftarrow \frac{g_{z,t}}{\|g_{z,t}\|_2}$$
3: **for** $i = 1, \ldots, K$ **do**
4: Normalize the client update
$$b_{i,t} \leftarrow \frac{\Delta\theta_{i,t}}{\|\Delta\theta_{i,t}\|_2}$$
5: **end for**
6: Construct
$$B_t = [b_{1,t}, \ldots, b_{K,t}]$$
7: Compute the Gram matrix
$$G_t = B_t^\top B_t$$
8: Compute the positive normalized spectral energies
$$\mu_{j,t} = \frac{\lambda_j(G_t)}{K}$$
9: Set
$$\widehat{\lambda}_t = \mathrm{median}\{\mu_{j,t} : \mu_{j,t} > 0\}$$
10: Compute the filtered client representations
$$\widetilde{B}_t = B_t \left[K\widehat{\lambda}_t (G_t + K\widehat{\lambda}_t I)^{-1}\right]$$
11: **for** $i = 1, \ldots, K$ **do**
12: Compute
$$s_{i,t} = \widehat{g}_{z,t}^\top \widetilde{b}_{i,t}$$
13: **end for**
14: **return**
$$\widehat{i}_t(z) = \arg\max_{i \in \mathcal{K}} s_{i,t}$$

For source inference, we sample 100 known training members without replacement from each client partition, yielding 1,000 target samples per federated run. All attack methods evaluated within the same run share the same target samples and ground-truth source assignments.

For the conventional loss-based attack, the target is attributed to the client model with minimum prediction loss:

$$\widehat{i}_t^{\,\mathrm{loss}}(z) = \arg\min_{i \in \mathcal{K}} \ell(\theta_{i,t}; z). \tag{14}$$

For LTMIA, we use a sliding loss window of length 10. Under one-based round indexing, define

$$\mathcal{W}_t = \{\max(1, t-9), \ldots, t\}. \tag{15}$$

The LTMIA score is computed from the average target loss over the available window,

$$\bar{\ell}_{i,t}^{\,\mathrm{LTMIA}}(z) = \frac{1}{|\mathcal{W}_t|} \sum_{\tau \in \mathcal{W}_t} \ell(\theta_{i,\tau}; z), \tag{16}$$

and the source prediction is

$$\widehat{i}_t^{\mathrm{LTMIA}}(z) = \arg\min_{i \in \mathcal{K}} \overline{\ell}_{i,t}^{\mathrm{LTMIA}}(z). \tag{17}$$

The first nine communication rounds therefore use all currently available historical losses, while the window contains 10 rounds from the tenth round onward. All 200 communication rounds are retained when computing the reported attack metrics.

B Evaluation Metrics

Let $A_{m,c,s,t}$ denote the attack success rate of method m under configuration c, independent run s, and communication round t. For N target samples,

$$A_{m,c,s,t} = \frac{1}{N} \sum_{z} \mathbb{1}\left[\widehat{a}_{m,c,s,t}(z) = a(z)\right], \qquad N = 1000. \tag{18}$$

For each run, the mean attack success rate across $T = 200$ communication rounds is

$$\overline{A}_{m,c,s} = \frac{1}{T} \sum_{t=1}^{T} A_{m,c,s,t}. \tag{19}$$

The configuration-level Mean ASR is

$$\overline{A}_{m,c} = \frac{1}{S} \sum_{s=1}^{S} \overline{A}_{m,c,s}, \qquad S = 5. \tag{20}$$

Temporal reliability. For each run, we first compute the standard deviation over communication rounds,

$$\sigma_{m,c,s}^{\mathrm{round}} = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} \left(A_{m,c,s,t} - \overline{A}_{m,c,s}\right)^2}. \tag{21}$$

The reported Temporal Std is then averaged across the five independent runs:

$$\sigma_{m,c}^{\mathrm{temp}} = \frac{1}{S} \sum_{s=1}^{S} \sigma_{m,c,s}^{\mathrm{round}}. \tag{22}$$

Hence, Temporal Std measures cross-round variation within individual federated runs rather than pooling all rounds from different runs.

Statistical reliability. We compute the standard deviation of the five run-level mean ASRs as

$$s_{m,c}^{\mathrm{run}} = \sqrt{\frac{1}{S-1} \sum_{s=1}^{S} \left(\overline{A}_{m,c,s} - \overline{A}_{m,c}\right)^2}. \tag{23}$$

The 95% lower confidence bound is

$$\mathrm{LCB}_{m,c} = \overline{A}_{m,c} - 1.96 \frac{s_{m,c}^{\mathrm{run}}}{\sqrt{S}}. \tag{24}$$

Higher Mean ASR and LCB indicate stronger attack effectiveness and statistical reliability, respectively, while lower Temporal Std indicates stronger temporal reliability.

Cross-configuration consistency. For a fixed set of methods, let $r_{m,c}$ denote the rank of method m under configuration c according to Mean ASR, where rank 1 denotes the highest ASR. Across a configuration set $\mathcal{C}$, we report

$$\mathrm{MeanRank}_m = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} r_{m,c}, \tag{25}$$

and

$$\mathrm{RankIQR}_m = Q_{0.75}\left(\{r_{m,c}\}_{c \in \mathcal{C}}\right) - Q_{0.25}\left(\{r_{m,c}\}_{c \in \mathcal{C}}\right). \tag{26}$$

Lower Mean Rank indicates stronger average relative performance. Rank IQR characterizes the dispersion of the relative ranking across federated configurations and should therefore be interpreted jointly with Mean Rank.

Macro aggregation. When reporting performance over multiple configurations, we first compute the metric within each configuration and then average across configurations:

$$\mathrm{AvgASR}_m = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \overline{A}_{m,c}, \tag{27}$$

$$\mathrm{AvgTempStd}_m = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sigma^{\mathrm{temp}}_{m,c}. \tag{28}$$

For dataset-level results, $|\mathcal{C}| = 9$; for the aggregate component analysis across all three datasets, $|\mathcal{C}| = 27$.

Relative improvement. For a baseline method b, relative ASR and LCB improvements are

$$\mathrm{ASRImprove} = \frac{\mathrm{ASR}_m - \mathrm{ASR}_b}{\mathrm{ASR}_b} \times 100\%, \tag{29}$$

$$\mathrm{LCBImprove} = \frac{\mathrm{LCB}_m - \mathrm{LCB}_b}{\mathrm{LCB}_b} \times 100\%, \tag{30}$$

while the relative reduction in Temporal Std is

$$\mathrm{StdReduction} = \frac{\mathrm{Std}_b - \mathrm{Std}_m}{\mathrm{Std}_b} \times 100\%. \tag{31}$$

C Additional Method Derivations

C.1 First-Order Marginal-Contribution Interpretation

Recall the update convention

$$\Delta\theta_{i,t} = \theta_t - \theta_{i,t}, \tag{32}$$

so that

$$\theta_{i,t} = \theta_t - \Delta\theta_{i,t}. \tag{33}$$

A Taylor expansion of the target loss around the common pre-round model θ_t gives

$$\ell(\theta_{i,t}; z) = \ell(\theta_t - \Delta\theta_{i,t}; z) \tag{34}$$

$$= \ell(\theta_t; z) - g_{z,t}^\top \Delta\theta_{i,t} + R_{i,t}, \tag{35}$$

where

$$g_{z,t} = \nabla_\theta \ell(\theta_t; z)$$

and $R_{i,t}$ denotes the higher-order remainder. Therefore,

$$\ell(\theta_t; z) - \ell(\theta_{i,t}; z) = g_{z,t}^\top \Delta\theta_{i,t} - R_{i,t}. \tag{36}$$

If the Hessian of the target loss is locally bounded by

$$\|\nabla_\theta^2 \ell(\theta; z)\|_2 \le L, \tag{37}$$

then the Taylor remainder satisfies

$$|R_{i,t}| \le \frac{L}{2} \|\Delta\theta_{i,t}\|_2^2. \tag{38}$$

Thus,

$$g_{z,t}^\top \Delta\theta_{i,t} \tag{39}$$

provides a first-order approximation of the target-loss reduction introduced by the current client update. Its decomposition

$$g_{z,t}^\top \Delta\theta_{i,t} = \|g_{z,t}\|_2 \|\Delta\theta_{i,t}\|_2 \cos\angle(g_{z,t}, \Delta\theta_{i,t}) \tag{40}$$

shows that the raw first-order contribution contains both update magnitude and directional target relevance. RCA removes the two magnitude terms and retains the normalized target–update alignment.

C.2 Spectral Properties of ASF

Let

$$B_t = U_t \Sigma_t V_t^\top \tag{41}$$

be the singular value decomposition of the normalized client-update matrix. Then

$$B_t^\top B_t = V_t \Sigma_t^2 V_t^\top. \tag{42}$$

Substituting this expression into the ASF transformation gives

$$\widetilde{B}_t = B_t \left[K\lambda_t (B_t^\top B_t + K\lambda_t I)^{-1} \right] \tag{43}$$

$$= U_t \Sigma_t \operatorname{diag}\left(\frac{K\lambda_t}{\sigma_{j,t}^2 + K\lambda_t} \right) V_t^\top. \tag{44}$$

Defining

$$\mu_{j,t} = \frac{\sigma_{j,t}^2}{K}, \tag{45}$$

the retention factor becomes

$$\rho_{j,t}(\lambda_t) = \frac{\lambda_t}{\lambda_t + \mu_{j,t}}. \tag{46}$$

Its derivative with respect to spectral energy is

$$\frac{\partial \rho_{j,t}}{\partial \mu_{j,t}} = -\frac{\lambda_t}{(\lambda_t + \mu_{j,t})^2} < 0. \tag{47}$$

Therefore, directions with larger spectral energy are attenuated more strongly.

The singular value also admits the interpretation

$$\sigma_{j,t}^2 = \|B_t^\top u_{j,t}\|_2^2 = \sum_{i=1}^{K} \left(u_{j,t}^\top b_{i,t}\right)^2. \tag{48}$$

Hence, a large $\mu_{j,t}$ indicates that the corresponding spectral direction contains substantial aggregate projection energy across the client-update population.

For the adaptive choice

$$\widehat{\lambda}_t = \mathrm{median}\{\mu_{j,t} : \mu_{j,t} > 0\}, \tag{49}$$

the relative retention factor becomes

$$\rho_{j,t} = \frac{1}{1 + \mu_{j,t}/\widehat{\lambda}_t}. \tag{50}$$

A direction at the median positive spectral energy is therefore retained with factor $1/2$.

Finally, consider a uniform rescaling

$$\mu'_{j,t} = c\mu_{j,t}, \qquad c > 0. \tag{51}$$

The median scales accordingly,

$$\widehat{\lambda}'_t = c\widehat{\lambda}_t. \tag{52}$$

Therefore,

$$\frac{\widehat{\lambda}'_t}{\widehat{\lambda}'_t + \mu'_{j,t}} = \frac{\widehat{\lambda}_t}{\widehat{\lambda}_t + \mu_{j,t}}. \tag{53}$$

Thus, the adaptive filtering rule is invariant to uniform spectral rescaling and depends on relative, rather than absolute, spectral structure.

D Detailed Source Inference Results

The main paper reports all 27 combinations of dataset, heterogeneity level α, and local epoch E. Here we discuss several additional patterns visible across these configurations.

D.1 Effect of Local Training Intensity

Increasing the number of local epochs has different effects on endpoint-loss and contribution-based attribution. On MNIST with $\alpha = 0.5$, the Mean ASR of Loss-based SIA decreases from 20.43 at $E = 1$ to 14.09 at $E = 10$, while LTMIA decreases from 22.48 to 15.56. In contrast, SpecSIA remains substantially more stable, changing from 28.43 to 27.88.

A similar trend appears on Synthetic with $\alpha = 0.5$. Loss-based SIA decreases from 34.63 to 25.68, and LTMIA decreases from 34.93 to 26.32, whereas SpecSIA changes from 34.93 to 31.85.

On CIFAR-10, increased local training can improve source attribution but also substantially increase temporal variation. For example, under $\alpha = 1$, SpecSIA increases from 26.89 ASR at $E = 1$ to 29.83 at $E = 10$. Over the same settings, its Temporal Std increases from 2.15 to 3.45, compared with 1.75 to 6.69 for Loss-based SIA and 2.84 to 8.71 for LTMIA. This illustrates why average attack success and temporal reliability should be evaluated separately.

D.2 Effect of Data Heterogeneity

Absolute source inference performance generally decreases as α increases and client distributions become less heterogeneous. For example, at $E = 10$, the Mean ASR of SpecSIA changes from

$$34.61 \rightarrow 19.91$$

on CIFAR-10,

$$27.88 \rightarrow 12.36$$

on MNIST, and

$$31.85 \rightarrow 17.37$$

on Synthetic when α increases from 0.5 to 10.

This trend is observed across the evaluated attacks and indicates that stronger client distribution similarity weakens the available source-specific signal.

D.3 Comparison with LTMIA

LTMIA and SpecSIA exhibit different effectiveness–reliability trade-offs. On CIFAR-10, LTMIA obtains the highest dataset-level Mean ASR and LCB, while SpecSIA reduces Temporal Std by 45.87% relative to LTMIA. On MNIST and Synthetic, SpecSIA improves both Mean ASR and Temporal Std relative to LTMIA. These results indicate that trajectory-based loss aggregation does not uniformly dominate current-round contribution attribution across datasets.

E Additional Component Analysis

E.1 Residual Attribution

Residual attribution is implemented as a hard projection baseline. Let

$$d_i = \theta_t - \theta_{i,t} \tag{54}$$

denote the update direction of candidate client i, and let

$$g_z = \nabla_\theta \ell(\theta_t; z) \tag{55}$$

be the target gradient at the common pre-round model.

Table 5: Dataset-wise cross-configuration ranking of the component methods. Lower Mean Rank is better; Rank IQR measures rank dispersion.

Dataset	Method	Mean Rank ↓	Rank IQR ↓
CIFAR-10	Loss-based	4.33	1.00
	LTMIA	**1.00**	0.00
	RCA	3.22	1.00
	Residual	2.78	1.00
	RCA+ASF	3.67	2.00
MNIST	Loss-based	5.00	0.00
	LTMIA	4.00	0.00
	RCA	**1.67**	2.00
	Residual	2.56	1.00
	RCA+ASF	1.78	0.00
Synthetic	Loss-based	4.56	1.00
	LTMIA	3.33	1.00
	RCA	**1.44**	1.00
	Residual	3.67	2.00
	RCA+ASF	2.00	1.00

For candidate client i, we collect the updates of all other clients as

$$D_{-i} = [d_j]_{j \neq i}. \tag{56}$$

Let

$$P_{-i} = D_{-i} D_{-i}^{\dagger} \tag{57}$$

denote the orthogonal projector onto the subspace spanned by the other client updates. The residual target gradient is

$$r_i(z) = (I - P_{-i}) g_z. \tag{58}$$

The candidate score is

$$s_i^{\mathrm{res}}(z) = \frac{r_i(z)^{\top} d_i}{\|r_i(z)\|_2}, \tag{59}$$

and the predicted source is

$$\widehat{i}^{\mathrm{res}}(z) = \arg\max_{i \in \mathcal{K}} s_i^{\mathrm{res}}(z). \tag{60}$$

In implementation, the projection is constructed through an SVD rather than explicitly forming a parameter-space projection matrix. The computation uses all trainable model parameters and is performed in double precision.

E.2 Dataset-Wise Component Results

Table 5 further separates the cross-configuration ranking analysis by dataset. Ranks are computed over the nine (α, E) configurations within each dataset.

The dataset-wise results show that the effect of ASF is not identical across datasets. LTMIA is consistently strong on CIFAR-10, whereas RCA and RCA+ASF occupy the leading ranks on MNIST and Synthetic. In particular, RCA+ASF has a concentrated rank distribution on MNIST while remaining among the stronger methods on Synthetic.

Table 6: Dataset-wise sensitivity to spectral calibration. All values are averaged over the nine (α, E) configurations of each dataset.

Dataset	γ	Avg ASR ↑	Temp. Std ↓	95% LCB ↑
CIFAR-10	0.1	26.82	3.61	25.00
	0.3	26.99	3.58	25.18
	1	27.18	3.56	25.39
	3	27.25	3.55	25.46
	10	27.23	3.56	25.42
MNIST	0.1	20.79	1.63	18.96
	0.3	20.94	1.63	19.09
	1	21.12	1.64	19.23
	3	21.17	1.65	19.29
	10	21.19	1.65	19.34
Synthetic	0.1	25.48	1.02	22.06
	0.3	25.98	0.99	22.46
	1	26.52	0.96	22.99
	3	26.76	0.97	23.31
	10	26.82	1.02	23.44

These results also illustrate why Rank IQR should not be interpreted in isolation. For example, a method may obtain a small Rank IQR because it is consistently low-ranked. Mean Rank and Rank IQR are therefore reported jointly.

F Additional Spectral Calibration Results

The main paper reports the aggregate sensitivity of SpecSIA to

$$\lambda_t^{(\gamma)} = \gamma \widehat{\lambda}_t. \tag{61}$$

Table 6 reports the corresponding results separately for each dataset.

The response to γ is smooth on all three datasets. Across the full tested range, changes in Avg ASR are modest relative to the differences observed across attack methods and federated configurations. Temporal Std is similarly stable.

The default adaptive calibration $\gamma = 1$ is therefore interpreted as a robust reference operating point rather than as a uniquely optimal hyperparameter. The slightly higher ASR obtained for larger γ values is consistent with the broad performance plateau reported in the main paper.

G Additional Gradient Anchor Results

The main paper compares the default pre-round gradient anchor

$$g_{z,t} = \nabla_\theta \ell(\theta_t; z) \tag{62}$$

with a fixed initial-model anchor

$$g_{z,0} = \nabla_\theta \ell(\theta_0; z). \tag{63}$$

To make the dataset-dependent trade-off more explicit, Table 7 reports the relative change obtained by replacing the pre-round anchor with the initial anchor.

The anchor effect is clearly dataset-dependent. On CIFAR-10, the fixed initial anchor substantially reduces temporal variation but sacrifices both Mean ASR and LCB. On MNIST, the initial anchor improves all three criteria. On Synthetic, it improves Mean ASR and LCB while slightly increasing Temporal Std.

Table 7: Relative change of the initial-model anchor with respect to the default pre-round anchor. Positive Std Reduction indicates lower temporal variation.

Dataset	**ASR Change ↑**	**Std Reduction ↑**	**LCB Change ↑**
CIFAR-10	-10.15%	57.02%	-12.68%
MNIST	9.56%	1.22%	8.58%
Synthetic	11.88%	-6.25%	15.27%
Overall	3.21%	32.20%	2.84%

Aggregated across the three datasets, the initial anchor improves Mean ASR and LCB and reduces Temporal Std. Nevertheless, we retain the pre-round model θ_t as the default anchor because it preserves the current-round attack model defined in Section 3. The initial anchor requires retaining θ_0 and is therefore treated as an optional stateful extension rather than part of the default SpecSIA threat model.