Explainable Federated Learning via Global-Local Attribution Alignment

Dawood Wasif[1], Terrence J. Moore[2], Chang-Tien Lu[1], Jin-Hee Cho[1]

[1] Virginia Tech, [2] US Army Research Laboratory

Paper | Code | Webpage

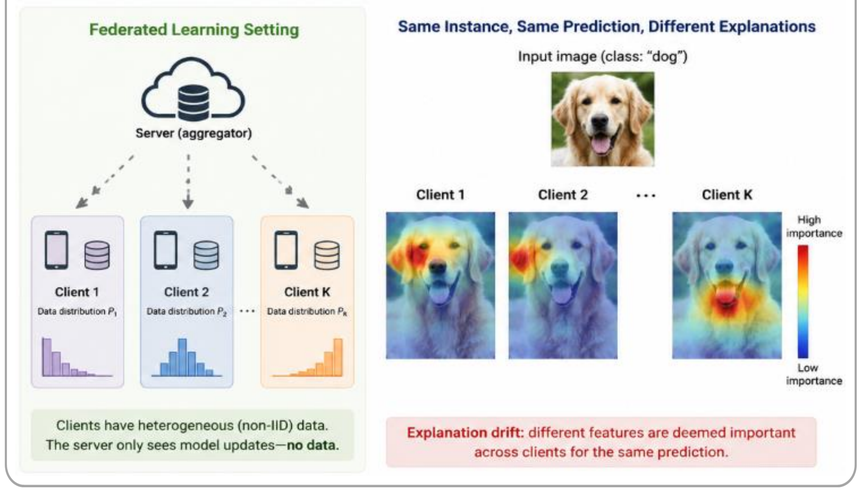
Problem & Motivation

Federated learning (FL) trains across private data silos under privacy and regulatory constraints. As FL reaches healthcare, finance, and mobile platforms, knowing *why* a model predicts becomes essential, not optional.

Centralized XAI struggles in FL because heterogeneous clients, hidden raw inputs, and leakage risks can cause **explanation drift**, where identical predictions rely on different features across clients.

Three tensions, at once:

- **Privacy:** no raw data, gradients, or dense maps may cross the network.
- **Heterogeneity:** non-IID clients diverge in reasoning, not just accuracy.
- **Bandwidth:** high-dimensional attribution objects scale poorly per round.



🎯 **Goal:** Develop a federated explanation framework that produces explanations that are **locally faithful, globally consistent, privacy-preserving, and communication-efficient**.

Key Contributions

***xFedAlign* decouples task learning from explanation coordination.** Task models are trained normally with FedAvg, while explanations are aligned separately in a compact group-level space.

- Coordinate explanation semantics independently of task optimization
- Use client-side surrogates to translate black-box model behavior into privacy-aware attribution artifacts for the Global Explanation Prior.
- Reduce explanation drift with minimal extra communication overhead.
- **Validate across vision, text, and tabular tasks**, showing improved explanation consistency, fidelity, privacy, and robustness.

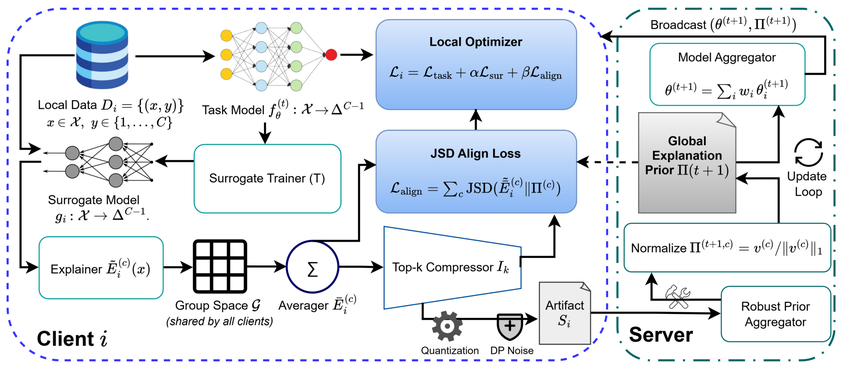
Proposed Methodology

***xFedAlign* keeps FedAvg task training unchanged, while aligning explanations through a compact Global Explanation Prior.**

1. **Task-Model Training:** Clients train the task model locally and the server aggregates updates normally using the Fed-Avg-style protocol.
2. **Surrogate Explanations:** Each client fits a lightweight **surrogate** g_i to mimic its local task model and produce group-level attributions.
3. **Global Explanation Prior Alignment:** Clients send only compressed top-k attribution artifacts. The server aggregates them into a Global Explanation Prior **Π**, which guides explanation alignment in the next round. The overall training objective is as follows:

$$\min_{\theta,\{g_i\},\Pi} \underbrace{\sum_i w_i \,\mathbb{E}\left[\ell(f_\theta(x), y)\right]}_{\text{task risk}} + \alpha \underbrace{\sum_i w_i \,\mathrm{KL}(p_\theta \,\|\, q_{g_i})}_{\text{surrogate fidelity}} + \beta \underbrace{\sum_i w_i \sum_c \mathrm{JSD}(\tilde{\bar{E}}_i^{(c)} \,\|\, \Pi^{(c)})}_{\text{explanation alignment}}$$

xFedAlign Framework



Clients train task models locally, use lightweight surrogates to extract group-level explanations, and share only compressed top-k explanation artifacts, and receive a Global Explanation Prior **Π** to align explanations.

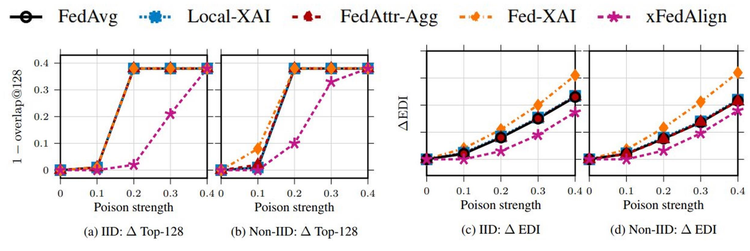
- **Step 1**: Server broadcasts task model θ(t) and prior **Π**(t).
- **Step 2**: Clients update the local task model on private data.
- **Step 3**: Surrogates generate normalized group-level attributions.
- **Step 4:** Clients compress explanations into hardened top-k artifacts.
- **Step 5**: Server robustly aggregates artifacts into **Π**(t+1).
- **Step 6:** Updated prior aligns client explanations in the next round.

Benchmark Results Across Modalities

		MNIST		AG News		UCI Adult	
Split	**Method**	Acc↑	EDI↓	Acc↑	EDI↓	Acc↑	EDI↓
IID	FedAvg	.986	–	.739	–	.845	–
	Local-XAI	.982	.020	**.755**	.302	.844	.003
	FedAttr-Agg	.983	.077	.729	.421	.844	.016
	Fed-XAI	.904	**.001**	.722	.096	.830	.008
	xFedAlign	**.986**	.000	.753	**.009**	**.855**	**.004**
Non-IID	FedAvg	.930	–	.707	–	.797	–
	Local-XAI	.927	.110	.706	.263	.751	.059
	FedAttr-Agg	.930	.159	.706	.416	**.842**	.127
	Fed-XAI	.449	.300	.593	.211	.790	.056
	xFedAlign	**.930**	**.076**	**.710**	**.166**	.799	**.004**

Across image, text, and tabular datasets, ***xFedAlign*** preserves strong Accuracy (Acc↑) while consistently lowering the Explanation Drift Index (EDI↓): which indicates more consistent explanations across clients, especially under Non-IID data heterogeneity.

Robustness to Poisoning



xFedAlign stays more stable under attribution poisoning, with lower top-k changes and EDI drift compared to baselines.

Conclusion

xFedAlign enables faithful and consistent federated explanations without changing standard task training. By aligning compact explanation artifacts through a Global Explanation Prior, it improves consistency, privacy, and robustness while keeping communication overhead low.

Acknowledgments: *Supported by NSF Award No. 2107450 and ARO Grant No. W911NF-24-2-0241.*