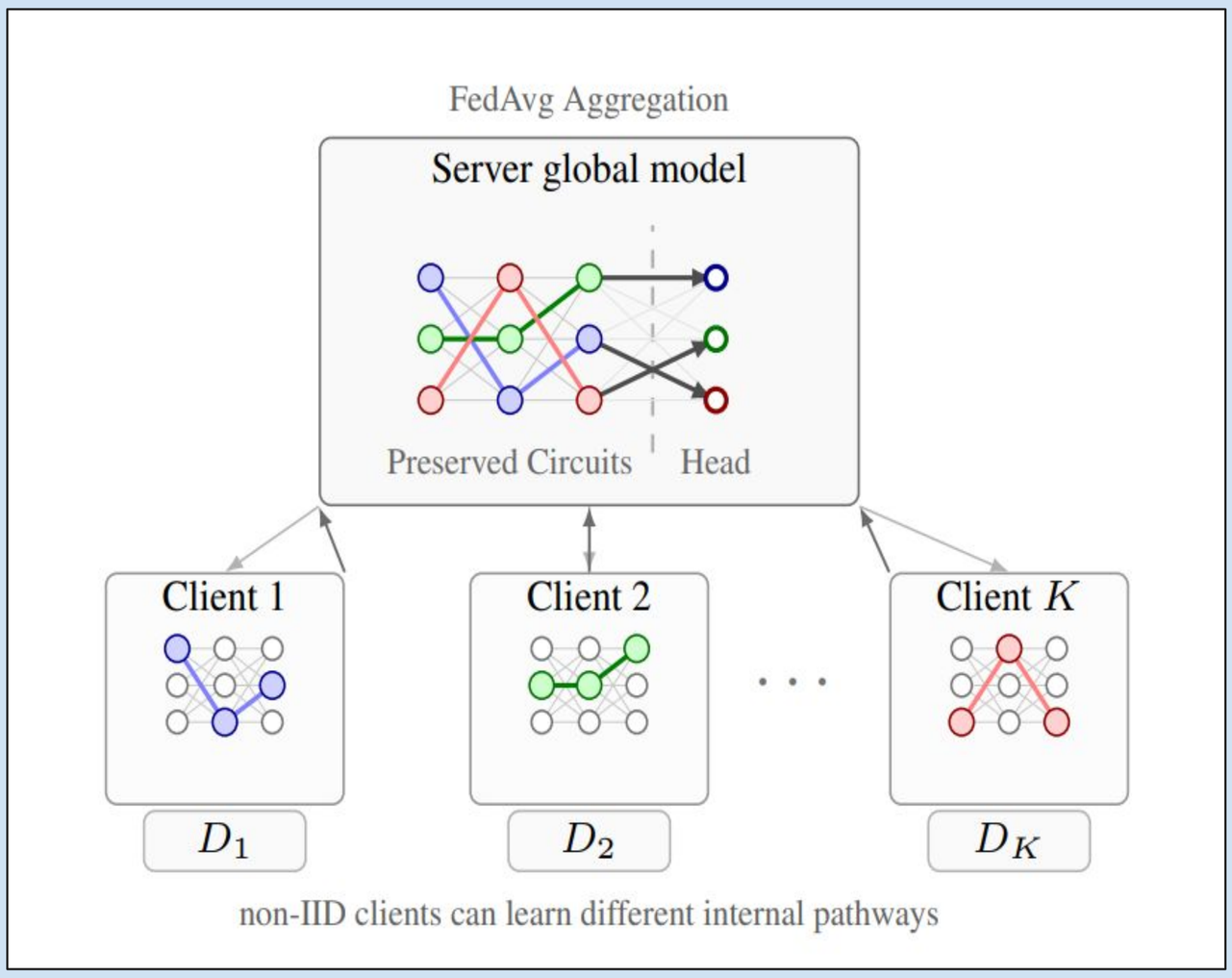


Non-IID FedAvg doesn't erase representations—it misaligns them.

Mechanistic Evidence for Preserved-but-Misaligned Representations in Non-IID FedAvg

Muhammad Haseeb, Salaar Masood, Muhammad Abdullah Sohail, Mohammad Fatim Shoaib, Muhammad Tahir
* Lahore University of Management Sciences

1 TL;DR

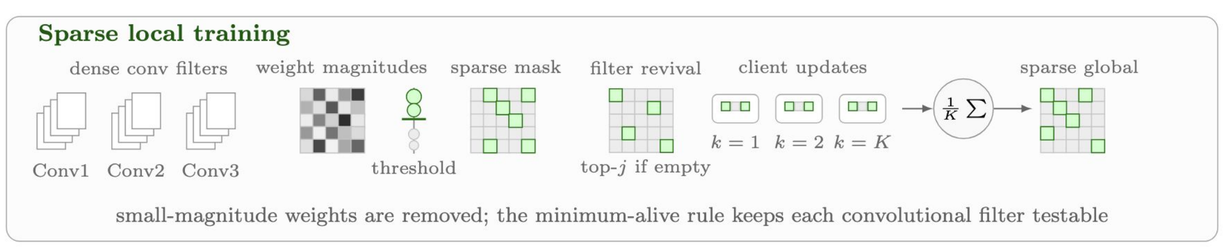
- ❖ **Method:** We use mechanistic tools (circuit discovery, probing, USAEs) to diagnose FedAvg on extreme non-iid datasets.
- ❖ **Finding:** Functional class circuits and linearly separable features survive aggregation, even for classes with near 0% global accuracy.
- ❖ **Implication:** Federated optimization failures are partly issues of coordinate misalignment, not just feature erasure.

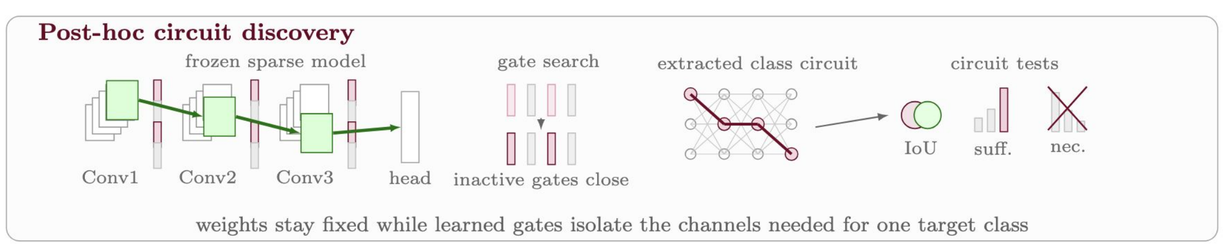
2 Background / Motivation

- **The Problem:** Federated Averaging (FedAvg) performance crashes when client data is highly heterogeneous (label skew/non-IID).
- **The Question:** Does this degradation happen because the model erases what clients learned (weight divergence), or does it misalign representations that are actually still there?
- **Our Approach:** Use Mechanistic Interpretability (circuit discovery, probes, USAEs) to peer inside sparse client-trained vision models.

3 Methodology

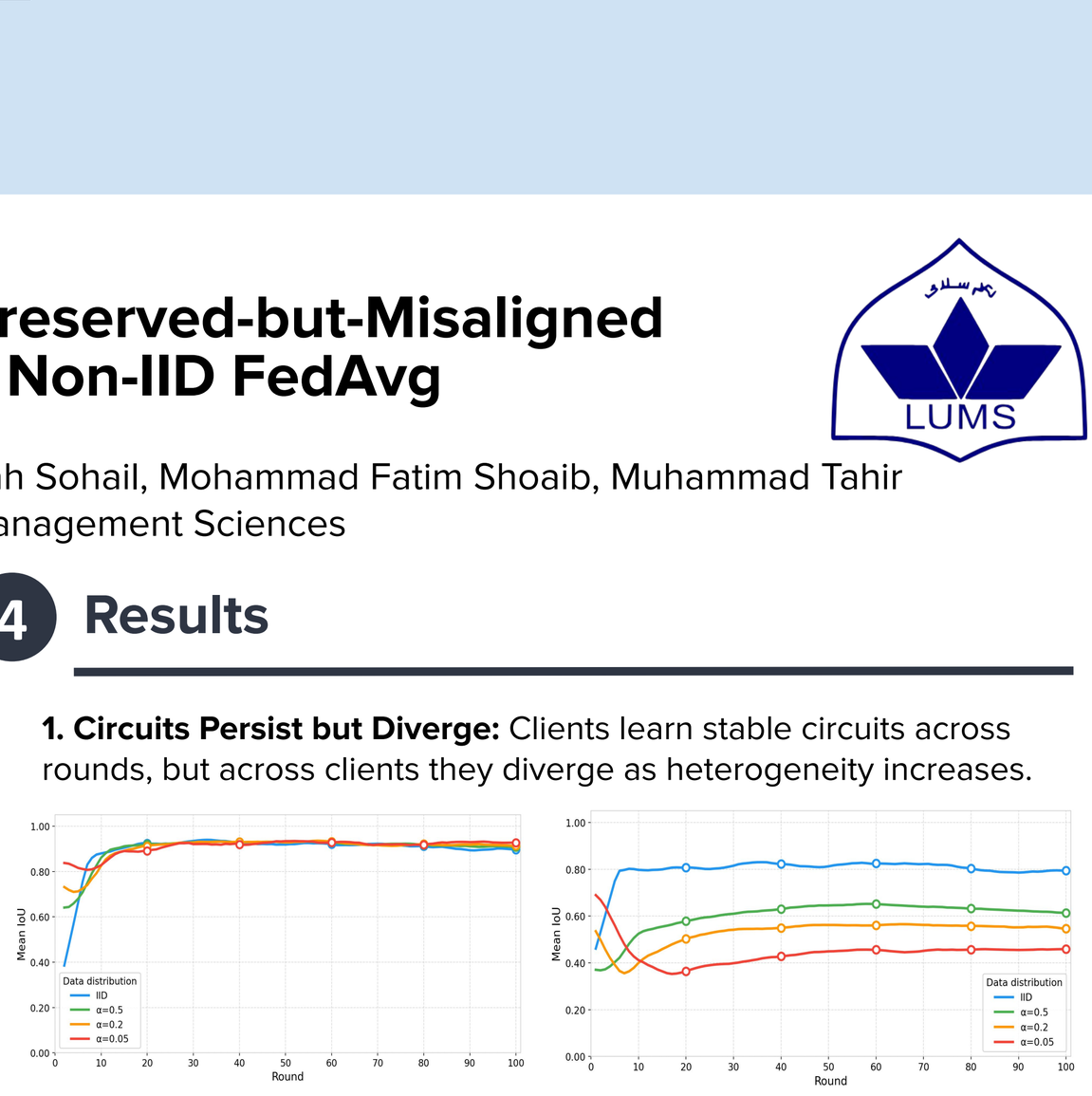
- **Setup:** Train sparse CNN/ResNet models on CIFAR-10/Fashion-MNIST across varying Dirichlet skew (α).
- **Circuit Discovery:** Identify sub-networks sufficient for class prediction.
- **Linear Probing & Head Fine-tuning:** Test if features remain accessible using linear layers.
- **Universal Sparse Autoencoders (USAEs):** Compare feature basis across IID/non-IID models.



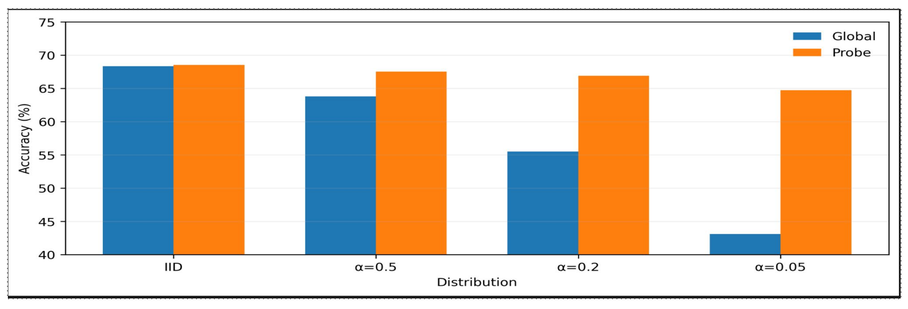


4 Results

1. Circuits Persist but Diverge: Clients learn stable circuits across rounds, but across clients they diverge as heterogeneity increases.



2. Features Remain Separable: Even when the global model entirely fails on a class (0% accuracy), a linear probe is able to discover class-discriminative features, recovering up to +21.7% accuracy.



3. Shared Concept Space: USAEs reveal that IID and non-IID models share 98.7% of their underlying concept dictionary.

Target Head	Source → Target	Sparse	Dense
IID Target	IID → IID	0.6421	0.7020
	Non-IID → IID	0.5899	0.6760
Non-IID Target	Non-IID → Non-IID	0.4054	0.4770
	IID → Non-IID	0.3645	0.4630

5 Limitations and Discussion

- **Implications:**
  ➔ We should reframe "weight divergence" as "representational misalignment."
  ➔ Fixes should target post-aggregation head realignment.
- **Limitations:**
  ➔ This work is tested only on Vision models (CNN/ResNet). Does this exact mechanistic phenomenon extend to Transformers/LLMs?
  ➔ Probing tests linear boundaries only. Does the same phenomenon apply to non-linear damage?

PAPER

CODE

haseeb099m@gmail.com