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

Muhammad Haseeb [1] **Salaar Masood** [1] **Muhammad Abdullah Sohail** [1] **Mohammad Fatim Shoaib** [1] **Muhammad Tahir** [1]

Abstract

Federated Averaging (FedAvg) often degrades under non-IID client data, but it remains unclear whether this degradation reflects the loss of client-learned representations or a failure to use representations that are still present. We study this question mechanistically in sparse client-trained vision models, using dense-model controls to test whether the observed effects depend on sparsity. Our analysis combines class-specific circuit discovery, linear probing of frozen representations, head-only finetuning, and sparse feature dictionaries. Across CNN and ResNet models on CIFAR-10 and Fashion-MNIST, severe label skew can drive some per-class accuracies near zero even when class-specific internal structure remains recoverable. Linear probes substantially outperform the aggregated classifier, head-only finetuning partially restores accuracy, and USAE transfer reveals a largely shared feature basis between IID and non-IID models. Together, these diagnostics suggest that, in our setting, non-IID FedAvg degradation is not fully explained by representational erasure; it also reflects misalignment between preserved internal structure and the final prediction pathway. We release code, trained checkpoints, extracted circuits, and experiment logs at the following repository: FedMI.

1. Introduction

Federated Learning (FL) trains a shared model from decentralized client data, typically using Federated Averaging (FedAvg) (McMahan et al., 2017; Kairouz et al., 2021). At each round, clients train local copies of the current global model and the server averages their returned parameters. FedAvg is simple and communication-efficient, but it degrades sharply under non-IID client data, especially label skew (Zhao et al., 2018; Hsieh et al., 2020). Existing explanations emphasize client drift, objective inconsistency, and weight divergence (Zhao et al., 2018; Karimireddy et al., 2020; Wang et al., 2020); these characterize the optimization failure, but leave the internal effect of aggregation largely unobserved. Thus, even when FedAvg succeeds or fails in aggregate, it remains difficult to tell whether client-learned structure was erased, partially preserved, or preserved but poorly used by the final prediction pathway.

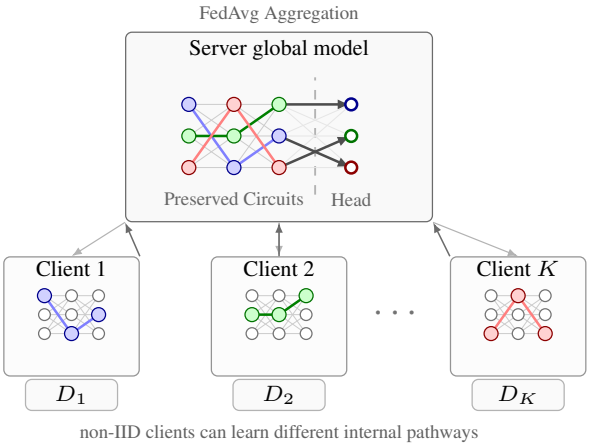


Figure 1. FedAvg under non-IID data. Clients train on local datasets D_k and return updates to the server; highlighted units schematically indicate client-specific internal pathways. In the aggregated global model, these functional features are successfully preserved but become misaligned with the final classification head.

In this work, we use Mechanistic Interpretability (MI) tools to investigate how FedAvg behaves under non-IID client data. MI studies neural networks by localizing the internal components and representations that implement model behavior (Olah et al., 2020; Elhage et al., 2021). We use this perspective to examine how aggregation changes the class-relevant structure learned by local clients, and whether that structure remains recoverable in the global model after FedAvg.

[1]Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan. Correspondence to: Muhammad Haseeb <26100253@lums.edu.pk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

We analyze models trained under IID and Dirichlet non-IID partitions (Zhao et al., 2018; Hsieh et al., 2020) using three mechanistic diagnostics. First, we use circuit discovery (Gao et al., 2025; Conmy et al., 2023) to measure how class-specific subnetworks diverge across clients and how much of their structure is preserved after aggregation. Second, we use linear probes (Alain & Bengio, 2017) and head-only finetuning (Oh et al., 2022) to test whether class information remains accessible in frozen global representations. Third, we apply Universal Sparse Autoencoders (USAEs), building on sparse dictionary learning (Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2024), to test whether the IID and non-IID global models share the same underlying feature basis.

Empirically, we observe that non-IID heterogeneity affects both circuit structure and readout alignment. Client circuits become less similar as heterogeneity increases, and the global model only partially preserves local circuit topology. At the same time, this same global model retains recoverable class-relevant structure: extracted circuits remain highly sufficient even for classes with near-zero global accuracy, linear probes and head-only finetuning recover substantial performance under severe skew. Further, USAE analysis indicates that IID and non-IID models share most of their feature basis. These results suggest that non-IID FedAvg degradation involves not only weakened internal structure, but also misalignment between retained representations and the model's default prediction pathway.

2. Related Work

2.1. Federated Optimization Under Non-IID Data

FedAvg remains the canonical optimization procedure for FL, but its local-update-then-average structure is sensitive to statistical heterogeneity (McMahan et al., 2017; Kairouz et al., 2021). Under non-IID client data, local objectives can drift away from the global objective, producing unstable convergence and degraded global accuracy (Zhao et al., 2018; Hsieh et al., 2020). A large body of work mitigates this problem by modifying the local objective or the aggregation rule. FedProx adds a proximal term to limit local deviation from the global model (Li et al., 2020); SCAFFOLD uses control variates to correct client drift (Karimireddy et al., 2020); FedNova normalizes local updates to reduce objective inconsistency (Wang et al., 2020); and MOON uses contrastive representation regularization to stabilize local training (Li et al., 2021). These methods improve optimization behavior, but generally evaluate success through aggregate loss or accuracy. Our work instead asks how non-IID aggregation changes the internal structure of the learned model.

2.2. Classifier and Readout Mismatch in Federated Learning

Several FL methods suggest that the classifier head is especially vulnerable under heterogeneous label distributions. FedRS studies label-distribution skew and argues that the last classification layer is more sensitive to missing or imbalanced classes than lower feature layers (Li & Zhan, 2021). Personalized FL methods make a related architectural separation: FedPer keeps shared lower layers while personalizing upper layers (Arivazhagan et al., 2019), FedRep learns a shared representation with client-specific heads (Collins et al., 2021), and FedBABU updates the model body during federated training while adapting the head afterward (Oh et al., 2022). These methods improve optimization behavior, but leave open how non-IID aggregation changes the internal structure of the learned model.

2.3. Model Merging and Weight-Space Averaging

FedAvg can also be viewed as a repeated form of weight-space model merging, where locally trained models with a shared initialization are averaged into a single global model. Recent model-merging work shows that parameter averaging and task-vector arithmetic can combine capabilities when models remain sufficiently aligned in weight space or lie in compatible loss basins (Wortsman et al., 2022; Matena & Raffel, 2022; Ilharco et al., 2023). At the same time, work on mode connectivity, model re-basing, and interference-aware merging emphasizes that permutation symmetries, representation alignment, and conflicting parameter directions can strongly affect whether independently trained models can be merged successfully (Frankle et al., 2020; Ainsworth et al., 2023; Yadav et al., 2023). These results provide a useful lens on FedAvg: non-IID clients may learn partially incompatible internal structures before aggregation. We use this perspective to study FedAvg at the level of circuits, probes, and sparse feature dictionaries rather than only post-merge accuracy.

2.4. Mechanistic Interpretability, Circuits, and Sparse Features

Mechanistic Interpretability aims to explain model behavior by identifying the internal components and computations that implement it (Olah et al., 2020; Elhage et al., 2021). Circuit analysis studies sparse subsets of model components sufficient for particular behaviors (Olah et al., 2020; Wang et al., 2023), with work on transformer circuits, automated circuit discovery, and circuit faithfulness establishing increasingly systematic tools for localizing mechanisms (Elhage et al., 2021; Conmy et al., 2023; Hanna et al., 2024). Sparse models can further make circuit extraction more tractable; recent work shows that weight-sparse transformers can expose interpretable task circuits (Gao et al., 2025).

In parallel, Sparse Autoencoders (SAEs) address superposition by decomposing dense activations into sparse feature directions (Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2024). We bring these tools into the federated setting, using circuits to study structural preservation, probes to test activation-level accessibility, and USAEs to compare IID and non-IID feature spaces, with the goal of distinguishing representational erasure from preserved-but-misaligned structure after FedAvg.

3. Methodology

3.1. Federated Learning Setup

We evaluate FedAvg (McMahan et al., 2017) on two datasets, CIFAR-10 and Fashion-MNIST, and two architectures, SimpleCNN and ResNet10. SimpleCNN consists of three convolutional blocks with channel widths $[64, 128, 256]$, max-pooling after each block, and a final linear classification head. ResNet10 consists of four sequential BasicBlocks with widths $[64, 128, 192, 256]$, skip connections within each block, global average pooling, and a final linear head. Detailed layer specifications, output sizes, and circuit-gated modules for both architectures are reported in Appendix A.

Unless otherwise noted, all experiments use $N = 10$ clients, 5 local epochs per round, batch size 256, Adam with learning rate 10^{-3}, and 100 communication rounds. At each round, clients initialize from the current global model, perform local training, and the server aggregates the resulting parameters by uniform averaging.

To vary statistical heterogeneity, we partition the training data across clients using a Dirichlet distribution with concentration parameter $\alpha \in \{0.5, 0.2, 0.05\}$, alongside an IID baseline.

3.2. Sparse Local Training

Client models are trained under a global magnitude-based sparsity schedule that anneals the target sparsity from 0 to $\rho = 99.9\%$ during local training (Gao et al., 2025). Let T_{local} denote the number of local optimization steps in a client update and let $T_{\text{ramp}} = \lceil 0.5 T_{\text{local}} \rceil$. At local step t, we set

$$\rho_t = \rho \cdot \min\left(\frac{t}{T_{\text{ramp}}}, 1\right), \tag{1}$$

so sparsity increases linearly over the first half of local training and is then held fixed at ρ. The magnitude threshold τ_t is chosen so that exactly $\lceil (1-\rho_t)|\theta| \rceil$ scalar weights are retained.

Thus, our main FedAvg experiments operate in a sparse-training regime. We use this setting deliberately because weight-sparse models yield compact, interventionally testable circuits, and prior bridge experiments suggest that sparse models can expose mechanisms relevant to dense models (Gao et al., 2025). Importantly, we also run dense-model controls in Appendix G, where local pruning is disabled and the same post-hoc analyses are repeated. The same qualitative circuit divergence, representational recovery, and feature-sharing patterns persist there, so sparsity is used as an interpretability-enabling lens rather than as a necessary condition for the paper's central finding.

To avoid fully pruned convolutional filters, we impose a minimum-alive constraint. If an output filter is left with fewer than $j = 4$ active weights after magnitude pruning, we replace the corresponding mask by its top-j weights in magnitude (Gao et al., 2025):

$$m'_o = \begin{cases} \mathbf{1}\{|w_o| \geq \text{top}_j(|w_o|)\} & \text{if } \|m_o\|_0 < j \\ m_o & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{1}\{\cdot\}$ denotes an elementwise binary indicator mask, $\|m_o\|_0$ denotes the number of active weights in the original mask, and $\text{top}_j(|w_o|)$ denotes the magnitude of the jth largest weight in filter o.

3.3. Circuit Discovery

Following the circuit-finding methodology of Gao et al. (2025), we define the **circuit** for class c as the minimal sub-network of parameters that is sufficient to correctly predict class c. Circuits are discovered via differentiable binary gating: for each eligible convolutional and linear layer, we attach per-output-channel gate scalars $g_{l,j} \in \mathbb{R}$, where l indexes the gated layer and j indexes an output channel or linear unit within that layer. Gates are initialised to 2.0. During the forward pass, layer l's output is multiplied channel-wise by a straight-through binary estimator:

$$\hat{g}_{l,j} = \underbrace{\mathbf{1}\{g_{l,j} > 0\}}_{\text{binary, forward}} - \underbrace{\sigma(g_{l,j}) - \sigma(g_{l,j})}_{\text{zero}} + \underbrace{\sigma(g_{l,j})}_{\text{gradient path}}, \tag{3}$$

so that the backward pass receives gradients through $\sigma(g_{l,j})$ while the forward pass uses a hard binary gate (Bengio et al., 2013). Gate parameters are optimised jointly over the cross-entropy loss and an ℓ_0 sparsity penalty:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_0 \sum_l \sum_j \sigma(g_{l,j}), \tag{4}$$

with λ_0=0.01 and T=300 optimisation steps using Adam at learning rate 0.1. After optimisation, gate $g_{l,j}$ is deemed active if $g_{l,j} > 0$, and the set of active output-channel indices per layer constitutes the circuit for class c.

To ensure comparability across clients, circuits are extracted using a **balanced global discovery pool** drawn uniformly from the full training dataset (20% per class, up to 1,024 samples per class).

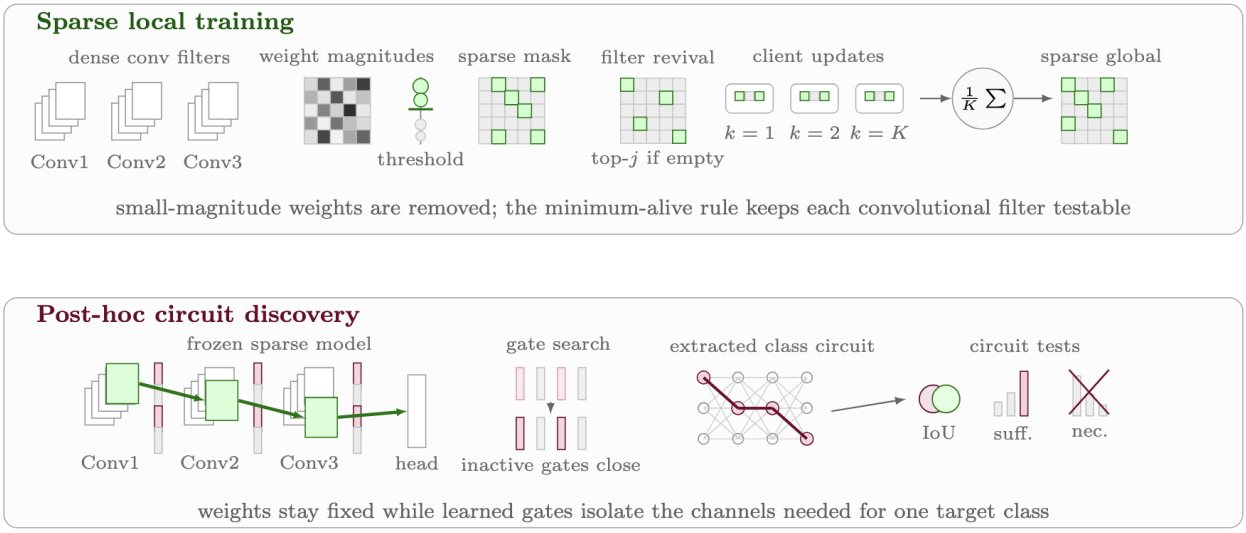


Figure 2. Sparse training and circuit discovery. During sparse local training, magnitude pruning removes small weights while a minimum-alive constraint revives the top-j weights of any dead filter; sparse client models are then aggregated by FedAvg. Circuit discovery is performed after training by freezing the model and learning channel gates that isolate a class-specific circuit, which is evaluated by IoU, sufficiency, and necessity.

Circuit Interventions. After a class circuit is extracted, we evaluate its functional role with two interventions on the frozen model. **Sufficiency** keeps only the active gated channels in the circuit and masks the remaining eligible channels; the reported score is the target-class accuracy of this circuit-only model. **Necessity** performs the complementary intervention: active circuit channels are ablated from the otherwise unmodified model, and the reported score is the remaining target-class accuracy. A high sufficiency score together with a low necessity score indicates that the extracted circuit is capable of supporting the class prediction and that removing it substantially disrupts that prediction.

Structural Similarity. Two circuits C_a and C_b are compared via the mean per-layer **Intersection-over-Union**:

$$\mathrm{IoU}(C_a, C_b) \;=\; \frac{1}{|L|}\sum_{l\in L}\frac{|C_a^l \cap C_b^l|}{|C_a^l \cup C_b^l|}, \tag{5}$$

where L is the set of eligible layers. We summarize circuit structure with mean IoU scores, where "mean" denotes averaging this per-layer IoU over the relevant classes, clients, or rounds for a given comparison. We use two complementary IoU measurements. **Inter-client consistency** is the mean IoU across all client pairs and all classes, measuring whether different clients learn similar class circuits. **Local-to-global preservation** is the mean IoU between each client's local circuit and the corresponding circuit extracted from the aggregated global model after each communication round, measuring how much client circuit structure survives FedAvg.

3.4. Linear Probing

Linear probing separates representation quality from the global model's default classifier: if a probe succeeds while the original head fails, then class information is still present in the frozen features but is not being used effectively by the learned readout. We train a linear probe on frozen penultimate-layer activations of the global model using the full training set, for 50 epochs with Adam at learning rate 10^{-4} and batch size 128. We compare probe accuracy against the global model's own classification accuracy; the resulting gap serves as a diagnostic for representational accessibility versus classifier misalignment.

3.5. Classification Head Finetuning

To test whether part of the accuracy degradation is localized to the final readout, we freeze the backbone of the global model and finetune only the final linear classification head on a balanced IID subset of 200 training samples (20 per class). The head is optimized for 5 epochs using AdamW with learning rate 10^{-3}, weight decay 5×10^{-4}, batch size 16, and cosine annealing. We compare post-finetuning accuracy against the original global model across heterogeneity levels; the IID model serves as a control.

3.6. Universal Sparse Autoencoder

Probes and head finetuning test label accessibility, but they do not directly show whether IID and non-IID models organize their features in a shared basis. To compare feature structure between IID and non-IID global models, we train

a USAE (Bricken et al., 2023; Cunningham et al., 2024) jointly on the penultimate-layer activations of both models. The USAE uses model-specific encoders and decoders together with a shared dictionary of $K = 2048$ concepts ($\sim 8\times$ expansion). Activations are extracted from both models on the same held-out test loader and stored as paired activation batches $(A_{\text{IID}}, A_{\text{nonIID}})$. During USAE training, each minibatch randomly chooses either model as the source side, so the dictionary is learned from both IID and non-IID activations rather than from one model alone. We treat a concept as shared if it activates non-trivially for both models over this paired activation pool. To assess cross-model compatibility, we evaluate concept transfer by encoding activations with the source encoder, decoding them into the target activation space, and classifying them with the target head.

4. Results

We evaluate global models trained under IID and Non-IID client partitions on CIFAR-10 with the CNN backbone, with concentration parameter $\alpha \in \{0.5, 0.2, 0.05\}$. Our analysis follows a diagnostic chain: we first measure how Non-IID training affects circuit structure within clients, across clients, and after aggregation. We then test whether class-relevant structure remains recoverable despite degraded global accuracy, using circuit extraction, linear probing, head-only finetuning, and feature subspace analysis. Appendix B–E report the same analyses for CIFAR-10 with ResNet and Fashion-MNIST with both backbones.

4.1. Circuit Stability and Divergence

We examine circuits from two complementary angles: whether they stabilize *within* each client across rounds (intra-client), and whether clients learn similar circuits *across* each other (inter-client).

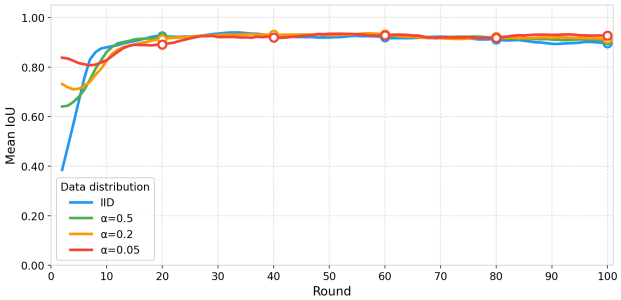


Figure 3. Intra-client circuit stability across communication rounds. Consecutive-round circuit overlap remains high across all client data distributions, stabilizing around IoU $\approx$ 0.85–0.92 for both IID and highly skewed non-IID partitions.

As shown in Figure 3, local training dynamics remain stable regardless of the client data distribution. Even under extreme non-IID conditions ($\alpha = 0.05$), the structural overlap between consecutive rounds rapidly climbs and plateaus near 0.90. This demonstrates that the failure of non-IID FedAvg is not rooted in local training instability or noise. Instead, local clients successfully converge and lock into highly stable computational pathways for the specific data they observe.

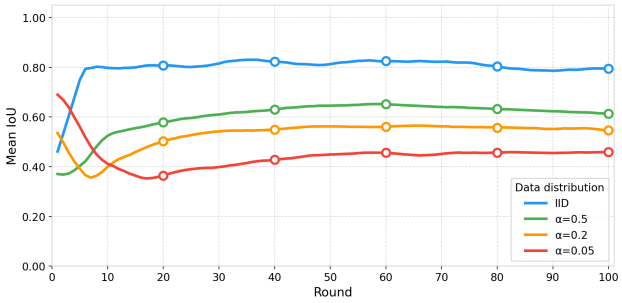


Figure 4. Inter-client circuit consistency across communication rounds. Under IID partitions, clients converge to similar class-specific circuits (IoU $\approx$ 0.80). As heterogeneity increases, client circuits diverge, stabilizing near IoU $\approx$ 0.45 at $\alpha = 0.05$.

Inter-client divergence. Under IID conditions, all clients converge to similar circuit structures and maintain this throughout training. As heterogeneity increases, circuits diverge between clients. Under extreme skew ($\alpha = 0.05$), clients rely on substantially different circuits to represent the same class, with structural overlap stabilizing near IoU ≈ 0.45.

Together, these results show that federated training produces stable local circuits regardless of data heterogeneity, but what clients converge to depends heavily on their local class distribution, causing clients to stabilize around different class-specific circuits.

4.2. Local vs Global Circuits

Given that Non-IID clients learn divergent circuits, we next ask how FedAvg integrates these conflicting structures. We measure the IoU between each client's local circuit and the corresponding circuit extracted from the aggregated global model using the local-to-global preservation metric defined in Section 3.3.

Figure 5 shows that, under IID conditions, the global model preserves local circuit topology nearly faithfully. Under Non-IID conditions, preservation degrades but does not collapse. The key observation is that local-to-global preservation (IoU ≈ 0.55) remains higher than inter-client circuit overlap (IoU ≈ 0.45) even under extreme skew. This means FedAvg does not simply average circuits into noise. The global model inherits partial structure from clients rather

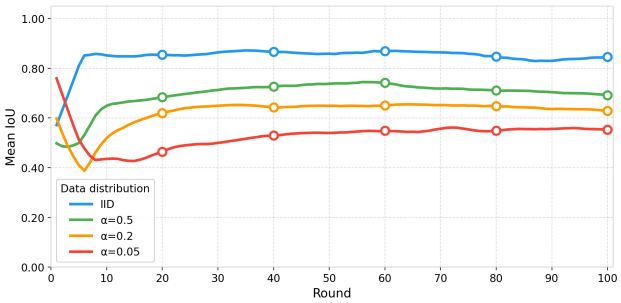


Figure 5. Local-to-global circuit preservation after FedAvg. The IID global model closely preserves client circuit structure (IoU ≈ 0.85). Under non-IID partitions, preservation drops but does not collapse: even at $\alpha = 0.05$, the global model retains roughly 0.55 IoU with local client circuits.

than converging to any single local solution. Appendix B shows the same partial preservation pattern across the remaining dataset and backbone pairs.

4.3. Functional Circuits Persist Even for Near-Zero Accuracy Classes

Section 4.2 shows that the global circuit only partially overlaps with local clients' circuits under Non-IID conditions. We now ask a stronger question: does the global model retain working circuits for classes it can no longer classify? Table 1 reports per-class accuracy, circuit sufficiency, and necessity at the final communication round, round 100, under extreme Non-IID ($\alpha = 0.05$).

Table 1. Per-class global accuracy, circuit sufficiency, and necessity at round 100 under extreme Non-IID ($\alpha = 0.05$) for CIFAR-10 with CNN. Overall global accuracy is 43.05%. Nodes denotes active gated channels in the extracted global circuit.

Class	Global Acc	Circuit Suff.	Necessity	Nodes
Airplane	0.5%	99.9%	0.0%	43
Automobile	43.3%	100.0%	0.0%	47
Bird	82.1%	100.0%	28.0%	42
Cat	0.0%	100.0%	0.0%	41
Deer	0.0%	100.0%	0.0%	32
Dog	53.4%	100.0%	0.1%	26
Frog	69.3%	100.0%	0.0%	24
Horse	20.8%	100.0%	0.0%	30
Ship	81.0%	100.0%	0.0%	36
Truck	80.1%	100.0%	0.0%	46

Circuit sufficiency is near-perfect across all ten classes regardless of global accuracy. Classes at 0% global accuracy still yield circuits with 100% target-class sufficiency. Necessity results confirm that these circuits are not redundant: ablating the circuit nodes collapses accuracy to near zero for nine out of ten classes, with Bird being the only exception where backup pathways retain 28% accuracy. Together, sufficiency and necessity establish that the discovered circuits are the primary computational pathway for each class in the global model. The fact that these circuits exist and are causally responsible for class prediction, yet the model scores 0% on several classes end-to-end, confirms that the failure is in how the model routes inputs through these circuits during normal inference, not in whether the circuits exist. Appendix B.1 shows the same pattern across the remaining dataset and backbone pairs.

4.4. Linearly Separable Global Representations

Circuit discovery searches the weight space for functional sub-networks. A complementary question is whether the same latent structure is visible from the activation side: do the global model's penultimate-layer representations contain enough class-discriminative geometry that a simple linear classifier can exploit it, even when the model's own head cannot? To test this, we train a single linear layer on the frozen penultimate-layer activations of the global model (Alain & Bengio, 2017) and compare its test accuracy to the global model's default classification accuracy.

Table 2. Global model accuracy vs. linear probe accuracy on CIFAR-10. The probe is a single linear layer trained on frozen features from the global model's penultimate layer for 50 epochs with Adam at learning rate 10^{-4} and batch size 128. Delta is probe accuracy minus global accuracy.

Config	Global Acc	Probe Acc	Delta
IID	68.3%	68.5%	+0.2%
$\alpha = 0.5$	63.8%	67.5%	+3.7%
$\alpha = 0.2$	55.5%	66.9%	+11.4%
$\alpha = 0.05$	43.1%	64.7%	+21.7%

Under IID conditions, probe and global accuracy are nearly identical, as expected. Under Non-IID conditions, a gap emerges and widens monotonically with heterogeneity. At $\alpha = 0.05$, the gap reaches 21.7 points. Probe accuracy remains stable across all heterogeneity levels, dropping only 3.8 points from IID to $\alpha = 0.05$, while default global accuracy drops 25.2 points over the same range. The backbone continues to produce class-discriminative features that a fresh linear classifier can exploit, even as the global model's own head fails. Appendix C shows this pattern holds across all four architecture and dataset combinations, with probe gains between 11 and 22 points under extreme skew.

4.5. Head Realignment

The linear probe shows that class-discriminative features remain linearly separable in the frozen backbone. A direct test is whether this is exploitable in practice: we finetune only the classification head on 200 balanced samples for 5 epochs while keeping the backbone frozen. The IID model

serves as a control: its head should already be well-aligned, so finetuning on 200 samples should yield negligible change.

Table 3. Global model accuracy before and after head-only finetuning on 200 balanced IID samples for 5 epochs. The backbone is frozen throughout. Delta is post-finetune minus pre-finetune accuracy.

CONFIG	PRE-FT	POST-FT	DELTA
IID	68.3%	65.9%	−2.4%
$\alpha = 0.5$	63.8%	64.0%	+0.2%
$\alpha = 0.2$	55.5%	60.9%	+5.4%
$\alpha = 0.05$	43.1%	53.3%	+10.2%

The IID model is almost unaffected (-2.4%), confirming the control. For Non-IID models, recovery scales monotonically with heterogeneity, reaching $+10.2$ points at $\alpha = 0.05$. However, the recovery is partial as it is still 15 points lower than IID baselines. This suggests that backbone geometry also degrades under Non-IID aggregation. Appendix D shows the same pattern across all four model families, with gains ranging from $+8.7$ to $+9.9$ points under severe heterogeneity.

4.6. Shared Feature Structure Across IID and Non-IID Models

The linear probe shows that class-discriminative features survive aggregation. We now ask whether these are the *same* features present in the IID model, or different features that happen to be linearly decodable. To test this, we train a USAE (Bricken et al., 2023; Cunningham et al., 2024) jointly on paired penultimate-layer activations of the IID global model (68.3% accuracy) and the extreme Non-IID global model ($\alpha = 0.05$, 43.1% accuracy), producing a shared concept dictionary of $K = 2048$ entries with model-specific encoders and decoders.

Concept overlap. Of the 2048 dictionary entries, 2021 (98.7%) activate non-trivially on inputs from both models over the paired held-out activation pool; this overlap is computed globally over the paired test-loader activations, not as a per-class average. The shared and full dictionaries differ by at most 0.003 accuracy in every condition, confirming that the 27 model-specific concepts carry negligible functional information. The two models share nearly the same feature basis.

Cross-model reconstruction. Table 4 reports accuracy under USAE round-trips: activations encoded by the source model, decoded into the target model's activation space, and classified by the target head. Same-model entries bound the maximum achievable accuracy given reconstruction loss; cross-model entries test whether features transfer.

Two patterns stand out in Table 4.

First, shared and full dictionaries perform nearly identically in every condition. The 27 model-specific concepts make no meaningful difference.

Table 4. USAE concept-transfer accuracy on CIFAR-10. Same-model entries are reconstruction upper bounds; cross-model entries test concept transfer.

DICTIONARY	SOURCE → TARGET	ACCURACY
SHARED (2021)	IID → IID	0.6421
	IID → NON-IID	0.3645
	NON-IID → IID	0.5899
	NON-IID → NON-IID	0.4054
FULL (2048)	IID → IID	0.6426
	IID → NON-IID	0.3650
	NON-IID → IID	0.5921
	NON-IID → NON-IID	0.4083

Second, when Non-IID activations are translated into the IID model's activation space through the USAE, the IID head still performs well, close to its own same-model upper bound. The Non-IID head fails regardless of concept source: it scores 0.4054 on its own activations and drops further to 0.3645 when fed IID activations, both below its unmodified accuracy of 43.1%. This is consistent with the probe and head finetuning results: the features are present, but the Non-IID classification head cannot use them. Appendix E shows that the fraction of shared concepts remains above 91% across all four model families.

Further cross-reconstruction analysis using SVCCA confirms that this shared feature basis is preserved in a misaligned coordinate system and we can recover accuracy even when using the original unmodified IID head on Non-IID backbone. We show this in Appendix F

5. Discussion

Reframing weight divergence. Prior work attributes Non-IID failure to weight divergence: conflicting updates overwrite learned representations during averaging (Zhao et al., 2018; Karimireddy et al., 2020). Our results do not support this as the primary mechanism. Across four independent analyses, we find that substantial class-relevant structure remains recoverable from the global backbone after aggregation. What changes is the structural relationship between the backbone and the classification head. Existing methods that reduce update divergence (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020) address a symptom of this problem. Our findings suggest the root cause is representational misalignment rather than representational loss.

Implications for algorithm design. Two directions follow from these results. First, post-aggregation head realignment on a small balanced set directly targets the identified failure mode and requires no changes to the federated training protocol. Second, circuit-aware aggregation — aligning

neurons that implement the same functional circuit across clients before averaging — could reduce structural shift at its source. Neuron permutation methods for model merging (Ainsworth et al., 2023) are applicable here, with the IoU metric used in this paper providing a natural objective for measuring alignment quality. More broadly, mechanistic tools can expose failures invisible to aggregate accuracy: a model may appear to have forgotten a class while still containing a recoverable circuit for it.

Limitations. Our analysis is conducted on CIFAR-10 and Fashion-MNIST with CNN and ResNet architectures. Whether these findings extend to transformer architectures, language models, or privacy-constrained FL settings remains open. Circuit discovery operates at the channel level and does not capture sub-channel structure. The linear probe confirms linear separability but does not rule out non-linear representational damage. USAE round-trip numbers should be read against same-model upper bounds due to reconstruction loss. The residual 15-point gap to IID performance after head finetuning at $\alpha = 0.05$ indicates that backbone degradation also contributes and head realignment alone is not a complete solution.

6. Conclusion

We studied FedAvg under non-IID data through a mechanistic lens, asking whether poor global accuracy reflects destroyed client knowledge or a failure to express knowledge that remains internally present. Across circuit discovery, linear probing, head-only finetuning, and USAE feature analysis, we find consistent evidence for the latter: even when the global model performs poorly on some classes, class-relevant circuits, linearly decodable representations, and shared sparse features remain recoverable from the aggregated model. This pattern appears across dataset–backbone pairs and is further supported by dense-model controls.

These results suggest that non-IID degradation should not be understood only as statistical drift or representational collapse. In our experiments, FedAvg can preserve useful internal structure while disrupting its alignment with the final prediction pathway. This reframes part of the non-IID problem as one of mechanistic preservation and expression, and points toward federated methods that diagnose and repair where recoverable knowledge becomes disconnected from behavior.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017.

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3259–3269. PMLR, 2020.

Gao, L., Rajaram, A., Coxon, J., Govande, S. V., Baker, B., and Mossing, D. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*, 2025.

Hanna, M., Pezzelle, S., and Belinkov, Y. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Mechanistic Interpretability Workshop at ICML*, 2024. URL https://openreview.net/forum?id=grXgesr5dT.

Hsieh, K., Phanishayee, A., Mutlu, O., and Gibbons, P. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning (ICML)*, pp. 4387–4398. PMLR, 2020.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, pp. 5132–5143. PMLR, 2020.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10713–10722, 2021.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020.

Li, X.-C. and Zhan, D.-C. FedRS: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 995–1005, 2021. doi: 10.1145/3447548.3467254.

Matena, M. S. and Raffel, C. A. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, 2017.

Oh, J., Kim, S., and Yun, S.-Y. FedBABU: Toward enhanced representation for federated image classification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=HuaYQfggn5u.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7611–7623, 2020.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022.

Yadav, P., Tam, D., Choshen, L., Raffel, C. A., and Bansal, M. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Confere
html.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Appendix

Table of Contents

This appendix provides implementation details and additional results supporting the analyses in the main text. Each entry links to the corresponding appendix section.

Appendix A Model Architecture Details 11
Appendix B Additional Circuit Similarity Results 12
Appendix B.1 Per-Class Circuit Sufficiency and Necessity 12
Appendix C Additional Linear Probe Results 14
Appendix D Additional Head-Only Finetuning Results 15
Appendix E Additional USAE Results 16
Appendix F Feature Alignment via SVCCA Intervention 17
Appendix G Generalization to Dense Models and Justification for Sparse Training 17

Note: Appendix B contains the additional circuit-overlap figures and per-class sufficiency and necessity measurements for the remaining dataset–backbone pairs. Appendices C–E report the corresponding probing, finetuning, and concept-transfer results. Appendix G provides the dense-model control experiments confirming that our findings are not artifacts of sparse training.

Table 5. Model architectures used in our experiments. Output sizes are shown for CIFAR-10 inputs ($3 \times 32 \times 32$). For Fashion-MNIST, the same architectures are used with single-channel input; the corresponding final spatial sizes are 3×3 for the CNN and 1×1 for the ResNet. Circuit discovery gates convolutional layers but excludes the final classifier and ResNet downsample projections.

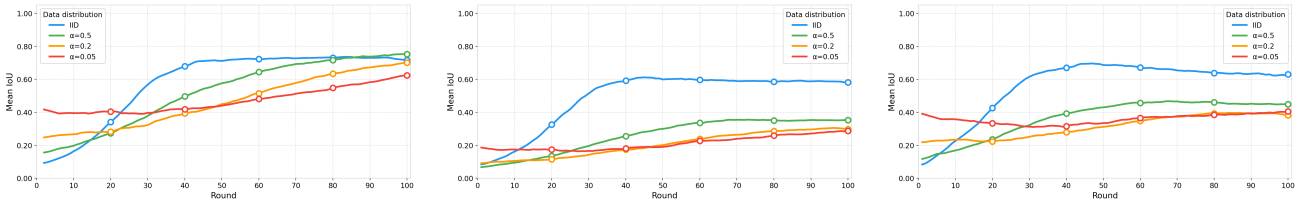
MODEL	LAYER	CHANNELS	KERNEL / STRIDE / PADDING	OUTPUT SIZE	GATED?
CNN	CONV1	$3 \to 64$	$3/1/1$	$64 \times 16 \times 16$	YES
	CONV2	$64 \to 128$	$3/1/1$	$128 \times 8 \times 8$	YES
	CONV3	$128 \to 256$	$3/1/1$	$256 \times 4 \times 4$	YES
	FC	$4096 \to 10$	–	10	NO
RESNET	CONV1	$3 \to 64$	$7/2/3$	$64 \times 8 \times 8$	YES
	BLOCK1.0.CONV1	$64 \to 64$	$3/1/1$	$64 \times 8 \times 8$	YES
	BLOCK1.0.CONV2	$64 \to 64$	$3/1/1$	$64 \times 8 \times 8$	YES
	BLOCK2.0.CONV1	$64 \to 128$	$3/2/1$	$128 \times 4 \times 4$	YES
	BLOCK2.0.CONV2	$128 \to 128$	$3/1/1$	$128 \times 4 \times 4$	YES
	BLOCK3.0.CONV1	$128 \to 192$	$3/2/1$	$192 \times 2 \times 2$	YES
	BLOCK3.0.CONV2	$192 \to 192$	$3/1/1$	$192 \times 2 \times 2$	YES
	BLOCK4.0.CONV1	$192 \to 256$	$3/2/1$	$256 \times 1 \times 1$	YES
	BLOCK4.0.CONV2	$256 \to 256$	$3/1/1$	$256 \times 1 \times 1$	YES
	FC	$256 \to 10$	–	10	NO

A. Model Architecture Details

Table 5 summarizes the layer structure of the two architectures used in our experiments. Output sizes are shown for CIFAR-10 inputs ($3 \times 32 \times 32$). For Fashion-MNIST, the same architectures are used with single-channel input; the corresponding final spatial sizes are 3×3 for the CNN and 1×1 for the ResNet. Circuit discovery gates convolutional layers but excludes the final classifier and ResNet downsample projections.

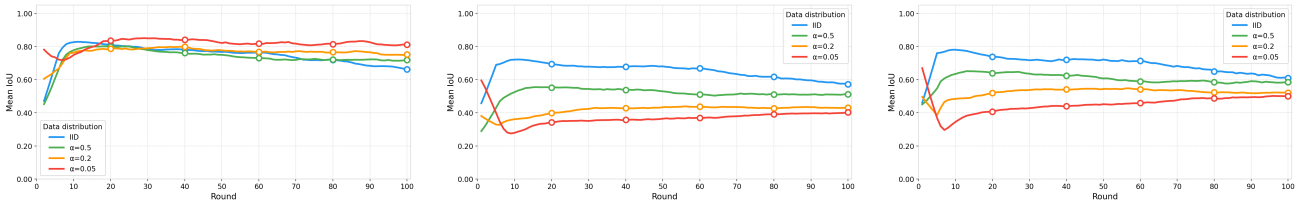
B. Additional Circuit Similarity Results

Figures 6–8 show intra-client circuit stability, inter-client circuit consistency, and local-to-global circuit preservation for the remaining dataset and architecture pairs. Each figure shows all three metrics together for a given dataset–backbone combination.



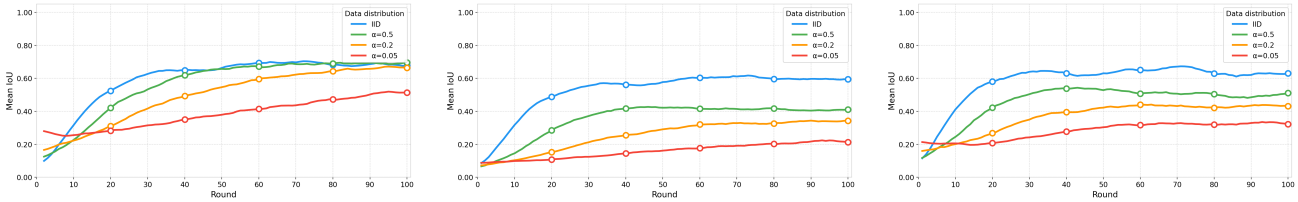
(a) Intra-client stability. *(b)* Inter-client consistency. *(c)* Local-to-global preservation.

Figure 6. Circuit similarity trends on CIFAR-10 with ResNet backbone.



(a) Intra-client stability. *(b)* Inter-client consistency. *(c)* Local-to-global preservation.

Figure 7. Circuit similarity trends on Fashion-MNIST with CNN backbone.



(a) Intra-client stability. *(b)* Inter-client consistency. *(c)* Local-to-global preservation.

Figure 8. Circuit similarity trends on Fashion-MNIST with ResNet backbone.

B.1. Per-Class Circuit Sufficiency and Necessity

Tables 6–8 extend the per-class circuit sufficiency and necessity analysis from Table 1 to the remaining dataset and backbone pairs under $\alpha = 0.05$. Circuit sufficiency is the target-class sufficiency score of the extracted class-specific circuit; necessity measures the remaining accuracy when the circuit is ablated; nodes denotes the number of active gated channels in the extracted circuit.

Notably, we observe a distinct cross-family difference in these results: CIFAR-10 circuits are almost universally necessary (ablation degrades accuracy to near 0%), with only the Bird class acting as an exception across both architectures due to backup pathways. In contrast, Fashion-MNIST displays the opposite trend for its best-performing classes. High-accuracy classes like Bag, Ankle boot, and Dress retain substantial accuracy after their main circuits are ablated, suggesting the model learned distributed, redundant representations rather than purely sparse pathways for those concepts under non-IID training. This discrepancy is likely due to the relative simplicity of the Fashion-MNIST task; because the features are comparatively easy to learn, the models have ample capacity to develop multiple, overlapping backup pathways. Consequently, even when the primary class-specific circuit is ablated, these secondary pathways can easily compensate and successfully classify the inputs.

Table 6. Per-class global accuracy, circuit sufficiency, and necessity for CIFAR-10 with ResNet backbone under $\alpha = 0.05$. Overall global accuracy is 34.18%.

CLASS	GLOBAL ACC	CIRCUIT SUFF.	NECESSITY	NODES
AIRPLANE	10.8%	100.0%	0.0%	66
AUTOMOBILE	27.0%	100.0%	0.0%	32
BIRD	79.4%	100.0%	17.3%	24
CAT	0.0%	97.5%	0.0%	88
DEER	0.5%	99.9%	0.0%	66
DOG	43.0%	100.0%	0.3%	23
FROG	44.3%	100.0%	0.0%	37
HORSE	7.9%	100.0%	0.0%	51
SHIP	52.8%	100.0%	0.0%	44
TRUCK	76.1%	100.0%	0.0%	32

Table 7. Per-class global accuracy, circuit sufficiency, and necessity for Fashion-MNIST with CNN backbone under $\alpha = 0.05$. Overall global accuracy is 76.74%.

CLASS	GLOBAL ACC	CIRCUIT SUFF.	NECESSITY	NODES
T-SHIRT	42.1%	98.5%	0.0%	47
TROUSER	93.3%	100.0%	0.0%	19
PULLOVER	52.0%	99.8%	0.0%	40
DRESS	86.8%	99.3%	34.0%	48
COAT	44.6%	99.9%	0.0%	46
SANDAL	74.6%	100.0%	0.4%	47
SHIRT	92.0%	99.8%	49.1%	38
SNEAKER	87.0%	100.0%	0.1%	26
BAG	96.5%	99.4%	81.3%	30
ANKLE BOOT	98.5%	100.0%	85.8%	26

Table 8. Per-class global accuracy, circuit sufficiency, and necessity for Fashion-MNIST with ResNet backbone under $\alpha = 0.05$. Overall global accuracy is 72.39%.

CLASS	GLOBAL ACC	CIRCUIT SUFF.	NECESSITY	NODES
T-SHIRT	65.7%	100.0%	0.0%	36
TROUSER	91.6%	99.8%	0.0%	36
PULLOVER	64.5%	100.0%	0.0%	42
DRESS	80.3%	100.0%	38.7%	25
COAT	40.7%	99.6%	0.0%	53
SANDAL	63.8%	100.0%	0.0%	46
SHIRT	76.3%	100.0%	0.1%	32
SNEAKER	45.1%	100.0%	0.0%	41
BAG	96.7%	100.0%	58.8%	12
ANKLE BOOT	99.2%	100.0%	0.2%	19

C. Additional Linear Probe Results

Table 9 reports linear probe results for CIFAR-10 with a ResNet backbone. Tables 10 and 11 report the corresponding Fashion-MNIST results.

Table 9. Global model accuracy vs. linear probe accuracy on CIFAR-10 with ResNet backbone. Same probe setup as Table 2.

CONFIG	GLOBAL ACC	PROBE ACC	DELTA
IID	59.7%	54.0%	−5.7%
$\alpha = 0.5$	56.8%	53.5%	−3.3%
$\alpha = 0.2$	49.7%	50.1%	+0.4%
$\alpha = 0.05$	34.2%	49.4%	+15.2%

Table 10. Global model accuracy vs. linear probe accuracy on Fashion-MNIST with CNN backbone. Same probe setup as Table 2.

CONFIG	GLOBAL ACC	PROBE ACC	DELTA
IID	90.9%	91.2%	+0.3%
$\alpha = 0.5$	90.0%	90.8%	+0.8%
$\alpha = 0.2$	88.8%	90.3%	+1.5%
$\alpha = 0.05$	76.7%	89.8%	+13.1%

Table 11. Global model accuracy vs. linear probe accuracy on Fashion-MNIST with ResNet backbone. Same probe setup as Table 2.

CONFIG	GLOBAL ACC	PROBE ACC	DELTA
IID	89.7%	86.5%	−3.3%
$\alpha = 0.5$	88.5%	86.8%	−1.7%
$\alpha = 0.2$	87.1%	85.7%	−1.4%
$\alpha = 0.05$	72.4%	83.4%	+11.0%

While the end-to-end trained head slightly outperforms the linear probe in the balanced IID ResNet10 case, the critical finding is the significant shift in the accuracy delta from −5.7% (IID) to +15.2% ($\alpha = 0.05$). This shows that discriminative features remain accessible in the backbone even as the default prediction pathway fails.

D. Additional Head-Only Finetuning Results

Tables 12–14 report head-only finetuning results for the additional dataset and backbone pairs. As in Table 3, the backbone is frozen throughout and only the final classification head is updated on a small balanced IID subset.

Table 12. Head-only finetuning on CIFAR-10 with ResNet backbone. The backbone is frozen and only the classification head is finetuned on 200 balanced IID samples for 5 epochs.

CONFIG	PRE-FT	POST-FT	DELTA
IID	59.7%	59.4%	−0.3%
$\alpha = 0.5$	56.8%	56.6%	−0.3%
$\alpha = 0.2$	49.7%	52.2%	+2.4%
$\alpha = 0.05$	34.2%	42.9%	+8.7%

Table 13. Head-only finetuning on Fashion-MNIST with CNN backbone. The backbone is frozen and only the classification head is finetuned on 200 balanced IID samples for 5 epochs.

CONFIG	PRE-FT	POST-FT	DELTA
IID	90.9%	90.1%	−0.8%
$\alpha = 0.5$	90.0%	89.6%	−0.4%
$\alpha = 0.2$	88.8%	88.5%	−0.2%
$\alpha = 0.05$	76.7%	86.3%	+9.6%

Table 14. Head-only finetuning on Fashion-MNIST with ResNet backbone. The backbone is frozen and only the classification head is finetuned on 200 balanced IID samples for 5 epochs.

CONFIG	PRE-FT	POST-FT	DELTA
IID	89.7%	89.6%	−0.2%
$\alpha = 0.5$	88.5%	88.6%	+0.1%
$\alpha = 0.2$	87.1%	87.4%	+0.3%
$\alpha = 0.05$	72.4%	82.2%	+9.9%

E. Additional USAE Results

Tables 15–17 report USAE concept-transfer results for the additional dataset and backbone pairs. Same-model entries are reconstruction upper bounds, while cross-model entries test transfer between IID and non-IID activation spaces.

Table 15. USAE concept-transfer accuracy on CIFAR-10 with ResNet backbone. Same-model entries are reconstruction upper bounds; cross-model entries test concept transfer.

DICTIONARY	SOURCE → TARGET	ACCURACY
SHARED (1868)	IID → IID	0.5830
	IID → NON-IID	0.3312
	NON-IID → IID	0.5533
	NON-IID → NON-IID	0.3454
FULL (2048)	IID → IID	0.5909
	IID → NON-IID	0.3353
	NON-IID → IID	0.5533
	NON-IID → NON-IID	0.3454

Table 16. USAE concept-transfer accuracy on Fashion-MNIST with CNN backbone. Same-model entries are reconstruction upper bounds; cross-model entries test concept transfer.

DICTIONARY	SOURCE → TARGET	ACCURACY
SHARED (1985)	IID → IID	0.8865
	IID → NON-IID	0.7344
	NON-IID → IID	0.8761
	NON-IID → NON-IID	0.7627
FULL (2048)	IID → IID	0.8889
	IID → NON-IID	0.7361
	NON-IID → IID	0.8774
	NON-IID → NON-IID	0.7566

Table 17. USAE concept-transfer accuracy on Fashion-MNIST with ResNet backbone. Same-model entries are reconstruction upper bounds; cross-model entries test concept transfer.

DICTIONARY	SOURCE → TARGET	ACCURACY
SHARED (1864)	IID → IID	0.8910
	IID → NON-IID	0.7048
	NON-IID → IID	0.8711
	NON-IID → NON-IID	0.7263
FULL (2048)	IID → IID	0.8938
	IID → NON-IID	0.7323
	NON-IID → IID	0.8729
	NON-IID → NON-IID	0.7262

F. Feature Alignment via SVCCA Intervention

To further investigate the geometric nature of representation misalignment, we apply Singular Vector Canonical Correlation Analysis (SVCCA; (Raghu et al., 2017)) to the final-layer activation spaces of IID and non-IID models. This analysis aims to determine if the class-discriminative information in non-IID backbones is preserved in a coordinate system that is simply misaligned with the global classification head.

Methodology here is simple. We utilize Singular Vector Canonical Correlation Analysis (SVCCA) to identify a shared linear basis between the IID and non-IID activation spaces. The procedure first applies SVD to the centered activations to whiten the subspaces and remove redundant dimensions, then computes the cross-covariance in this whitened space to find canonical directions that maximize correlation. To test for representational recoverability, we map non-IID test activations into this shared CCA space and reconstruct them back into the IID coordinate system using only the top-k most correlated components. These aligned features are then evaluated using the original, frozen IID classification head. Finally, these translated' features are passed directly into the original, frozen IID classification head to evaluate their recoverability.

Table 18. Accuracy recovery via SVCCA feature alignment on CIFAR-10. Baseline measures the accuracy of raw non-IID features passed to the original IID head. SVCCA-Aligned measures accuracy after mapping features into the IID coordinate system via the top-k components.

BACKBONE	BASELINE (RAW)	SVCCA-ALIGNED	COMPONENTS (k)
CIFAR-10 CNN	12.72%	66.26%	2048
CIFAR-10 RESNET	11.63%	50.30%	128

As shown in Table 18, the original IID classification head performs at near-random levels (12%) when fed raw non-IID features. However, after applying a linear SVCCA mapping, accuracy recovers to 66.26% for the CNN and 50.30% for the ResNet.

The recovery of accuracy using the frozen IID head confirms that the non-IID backbone retains class-relevant features. The initial failure of the head is attributable to a linear shift in the feature basis—a coordinate misalignment—that renders the stable local representations unreadable to the default classification pathway.

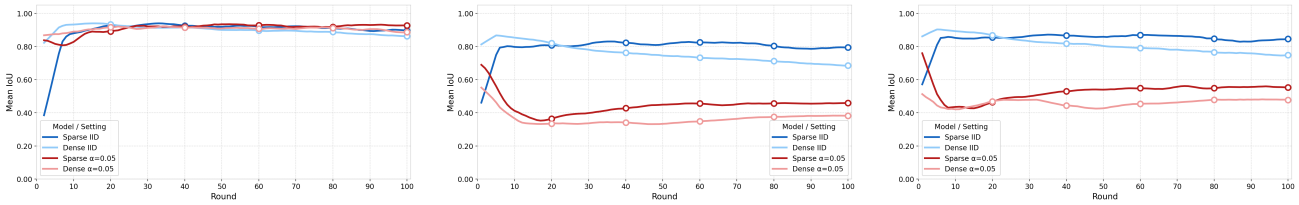
G. Generalization to Dense Models and Justification for Sparse Training

A potential concern is whether our findings, derived from weight-sparse models, generalize to standard, dense FedAvg. We use a sparse-training regime deliberately as a tool for mechanistic interpretability. Weight-sparse models expose more compact and causally validated circuits, making analysis more tractable and reliable (Gao et al., 2025). However, to confirm that our conclusions are not artifacts of sparsity, we conducted a parallel set of experiments using a **dense-model control**. In this control, clients train standard dense models (local magnitude pruning is disabled), which are then averaged by the server. All post-hoc analyses, including circuit discovery, linear probing, and USAE, were performed identically to the sparse-model experiments.

The federated learning setup for the dense control was identical to the sparse experiments, using 10 clients, 100 communication rounds, and 5 local epochs per round for both IID and extreme non-IID ($\alpha = 0.05$) partitions. We made two minor adjustments to account for the different optimization landscape of dense models. First, we used a slightly smaller learning rate of 10^{-4} (compared to 10^{-3} for sparse models) to ensure stable convergence. Second, post-hoc circuit discovery was run for 500 optimization steps (compared to 300 for sparse models) to ensure the gates fully converged on the more complex dense weight structure. All other post-hoc analyses, including linear probing and USAE, were performed identically to the sparse-model experiments.

Our results show that the core mechanistic phenomena persist in the dense setting, confirming that sparsity serves as an effective analytical lens rather than a necessary condition for the observed behaviors.

Circuit Stability and Divergence. As shown in Figure 9, circuit dynamics in dense models mirror those in sparse models. Intra-client stability remains high (IoU ≈ 0.90) for both IID and non-IID clients, confirming that local models converge to stable representations. Concurrently, inter-client consistency is high for IID clients (IoU ≈ 0.70) but collapses for non-IID clients (IoU ≈ 0.38). The core dynamic of stable but divergent circuits is not dependent on weight sparsity.



(a) Intra-client Stability (Dense) *(b)* Inter-client Consistency (Dense) *(c)* Local vs. Global Consistency (Dense)

Figure 9. Circuit similarity trends for dense models on CIFAR-10 with CNN backbone. The qualitative patterns are identical to the sparse-model results in the main paper. (a) Intra-client stability remains high for both IID and non-IID clients. (b) Inter-client consistency drops sharply under non-IID data. (c) The aggregated global model partially preserves local circuits.

Representational Recoverability. Table 19 confirms that the misalignment between the backbone and classification head persists in dense models. Under extreme Non-IID ($\alpha = 0.05$), the dense model's linear probe recovers $+15.9\%$ over global accuracy and head-only finetuning recovers $+10.8\%$, closely matching the sparse model's respective $+21.7\%$ and $+10.2\%$ gains. Both regimes show the same qualitative pattern: features remain substantially more capable than the global head suggests.

Table 19. Representational diagnostics for CIFAR-10 with CNN backbone, comparing sparse and dense models. Linear probe and Head-FT accuracy are measured on frozen backbone features. Delta rows report the gap between Non-IID probe/Head-FT accuracy and global accuracy.

Config	Model	Global Acc	Probe Acc	Head-FT Acc
IID	Sparse	68.3%	68.5%	65.9%
	Dense	74.2%	74.2%	62.7%
Non-IID ($\alpha = 0.05$)	Sparse	43.1%	64.7%	53.3%
	Dense	49.9%	65.8%	60.7%
Non-IID Delta	Sparse	–	**+21.7%**	**+10.2%**
	Dense	–	**+15.9%**	**+10.8%**

Shared Feature Basis. USAE analysis (Table 20) confirms a high degree of feature sharing in both model types. In both sparse and dense settings, Non-IID features transfer well to the IID head (sparse: 0.5899; dense: 0.6760), closely approaching each model's own same-model upper bound. The Non-IID head fails in both directions regardless of model type, confirming that the classification head is the primary bottleneck rather than the feature representations themselves.

Table 20. USAE concept-transfer accuracy for CIFAR-10 with CNN backbone, comparing sparse and dense models. Same-model entries are reconstruction upper bounds; cross-model entries test concept transfer.

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