PATHWAYFED: DISENTANGLING CLIENT-SPECIFIC KNOWLEDGE PATHWAYS FOR FEDERATED MULTI-LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Multi-Label Learning (FML) enables multiple clients to collaboratively learn from heterogeneous multi-label data while preserving data privacy. However, variations in local label spaces and label co-occurrence patterns cause different clients to encode specialized knowledge through entangled feature-to-class pathways, resulting in implicit conflicts and the dilution of informative evidence during aggregation. To address this challenge, we propose **PathwayFed**, a novel framework that makes client-specific knowledge pathways more distinguishable and disentangles their contributions during federated aggregation. On the client side, PathwayFed refines knowledge pathways through dimensional expansion and an orthogonal sparse transformation, producing finer-grained and less-overlapping pathway expressions that make meaningful client-specific knowledge easier to identify and quantify. On the server side, we introduce a matched knowledge routing mechanism that reinforces strongly expressed client-specific knowledge while preserving its pathway association, followed by conflict-aware projection to suppress conflicting components and calibrate the compatibility of the routed knowledge pathways. Extensive experiments on PASCAL VOC 2007, MS-COCO 2014, and NUS-WIDE-81 demonstrate that PathwayFed outperforms state-of-the-art methods, particularly under severe heterogeneity.

1 INTRODUCTION

Federated multi-label learning (FML) enables multiple clients to collaboratively learn from heterogeneous multi-label data while preserving data privacy (McMahan et al., 2017). Compared with conventional single-label federated learning, multi-label scenarios involve diverse label spaces, visual concepts, and label co-occurrence patterns across clients (Wang et al., 2026; Kou et al., 2026). Such heterogeneity allows different clients to acquire complementary but unevenly distributed local knowledge during training (Guo et al., 2025; Huang et al., 2021; Zhang et al., 2021). Although clients optimize a shared prediction model, their local data emphasize different semantic concepts and correlations. Consequently, different clients may convey class-related knowledge through different feature-to-class routes, making effective model aggregation particularly challenging.

A key difficulty is that client knowledge is not uniformly conveyed through the routes from features to class predictions. We refer to these feature-to-class routes as *client-specific knowledge pathways*, or simply *pathways*. Under heterogeneous local distributions, pathway-specific contributions can vary substantially in both magnitude and compatibility across clients. Different clients can therefore exhibit distinct pathway patterns, even when they share the same feature representation. Such heterogeneous pathway usage is consistent with prior observations that neural representations can distribute semantic information unevenly across shared feature coordinates (Bau et al., 2017; Fong & Vedaldi, 2018; Pach et al., 2025). In heterogeneous multi-label learning, variations in local label spaces and label co-occurrence patterns can further cause client-specific pathway contributions to become entangled during aggregation, making informative and incompatible contributions difficult to distinguish. Directly aggregating such pathways may therefore dilute useful client-specific knowledge and introduce implicit conflicts (Hu et al., 2023; Li et al., 2023; Zheng et al., 2025). Figure 1 further visualizes this effect: client-conditioned pathway-relevant evidence differs across clients,

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.

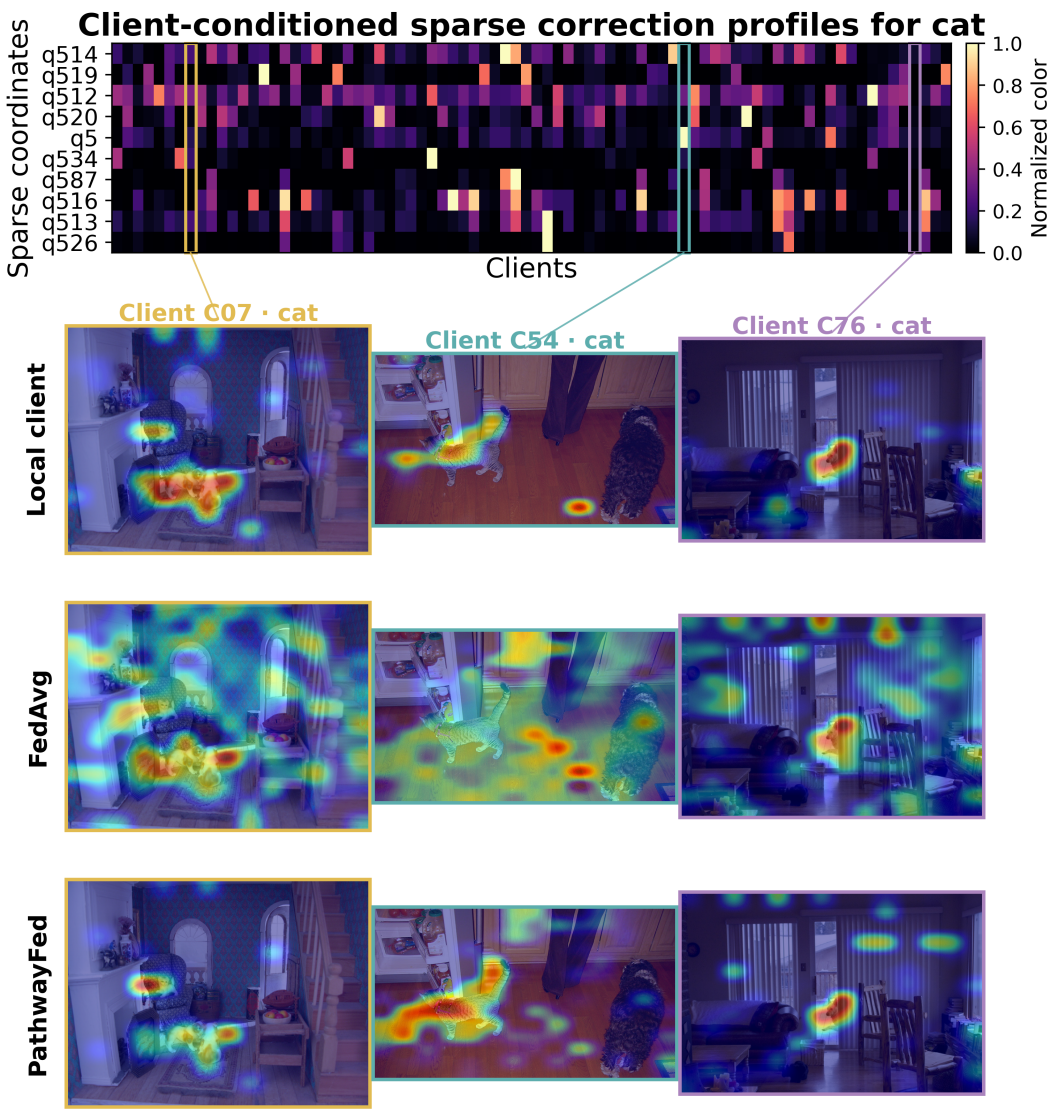


Figure 1: Qualitative visualization of client-specific sparse pathway-relevant evidence. Top: client-conditioned profiles of the ten strongest auxiliary sparse coordinates, with three representative clients highlighted. Bottom: sparse-branch CAMs compare the local model, sample-count-weighted FedAvg, and PathwayFed from the same communication round. FedAvg produces more diffuse spatial evidence, whereas PathwayFed better preserves concentrated local patterns. CAMs are independently normalized, so the comparison concerns spatial organization rather than absolute magnitude. The visualized sparse evidence is obtained by applying the fixed orthogonal sparse transformation to features from the frozen CLIP vision encoder.

while ordinary averaging produces more diffuse spatial evidence and PathwayFed better preserves concentrated client-specific patterns.

To address this challenge, we propose PathwayFed, a federated multi-label learning framework that disentangles knowledge pathways and enables pathway-aware aggregation. On the client side, PathwayFed constructs finer-grained unit-to-class pathways via dimensional expansion and an orthogonal sparse transformation, yielding complementary, less-overlapping pathway representations that facilitate the identification and quantification of client-specific knowledge. On the server side, matched knowledge routing preserves pathway associations while reinforcing strongly expressed client-specific knowledge, followed by conflict-aware projection to suppress incompatible compo-

nents and calibrate pathway compatibility. Together, these components preserve informative client-specific knowledge while reducing harmful aggregation conflicts.

We evaluate **PathwayFed** on three widely used multi-label benchmarks, including PASCAL VOC 2007, MS-COCO 2014, and NUS-WIDE-81. **PathwayFed** remains competitive under mild heterogeneity and becomes increasingly advantageous as client specialization strengthens.

Our contributions are summarized as follows:

- We propose PathwayFed, a novel framework for federated multi-label learning that disentangles knowledge pathways across heterogeneous clients, mitigating implicit conflicts and the dilution of informative evidence during aggregation.
- We refine knowledge pathways through dimensional expansion and an orthogonal sparse transformation, yielding finer-grained and less-overlapping pathway representations. We further introduce matched knowledge routing and conflict-aware projection to reinforce useful client-specific knowledge, preserve pathway associations, and suppress incompatible components during aggregation.
- Extensive experiments demonstrate the effectiveness of PathwayFed across multiple benchmark datasets. Notably, its performance gains over existing methods become increasingly pronounced as the degree of client heterogeneity increases.

2 Related Work

Federated Multi-Label Learning and Client Heterogeneity. Federated multi-label learning must accommodate clients with different visual concepts, label supports, and label co-occurrence patterns. Recent methods employ prompt adaptation, semantic conditioning, or correlation-aware learning (Wang et al., 2026; Kou et al., 2026), often with pretrained representations such as CLIP (Radford et al., 2021). More broadly, heterogeneous and personalized federated learning recognizes that clients acquire specialized knowledge from their local distributions (Guo et al., 2025; Zhu et al., 2025). This work examines how such knowledge is expressed through feature-to-class pathways and how its client-specific associations can be preserved during aggregation.

Disentanglement and Conflict-Aware Federated Aggregation. Neural representations can distribute semantic information across correlated and overlapping feature dimensions (Pach et al., 2025). Representation-disentanglement studies seek more separable or interpretable components, but primarily examine representation quality within a single model. Federated optimization, meanwhile, has progressed from global client weighting and optimization corrections (Li et al., 2020; Karimireddy et al., 2020; Reddi et al., 2021) toward client-level (Tang, 2024), layer-wise (Huang & Freris, 2025; Liu et al., 2024), and element-wise selective aggregation (Hu et al., 2023). Conflict-aware methods further suppress inconsistent local updates (Li et al., 2023; Zhang et al., 2024; Zheng et al., 2025). Although these approaches improve representation separability, aggregation granularity, or conflict handling, they do not explicitly preserve the association between client-specific knowledge and the feature-to-class pathways through which it is expressed. PathwayFed addresses this gap by refining pathway expressions and combining pathway-aware routing with conflict-aware projection.

3 Method

3.1 Problem Setup and Overview

Federated multi-label formulation. For K clients, client i owns a private multi-label dataset

$$\mathcal{D}_i = \{(x_{i,n}, \mathbf{y}_{i,n})\}_{n=1}^{N_i}, \qquad \mathbf{y}_{i,n} \in \{0,1\}^C, \tag{1}$$

where N_i denotes the number of local samples. The input space is $\mathcal{X} \subset \mathbb{R}^d$ and the label space is $\mathcal{Y} = \{0,1\}^C$, where C is the number of labels. For an instance $x \in \mathcal{X}$, its label vector is

$$\mathbf{y} = (y_1, \ldots, y_C)^\top \in \{0,1\}^C, \tag{2}$$

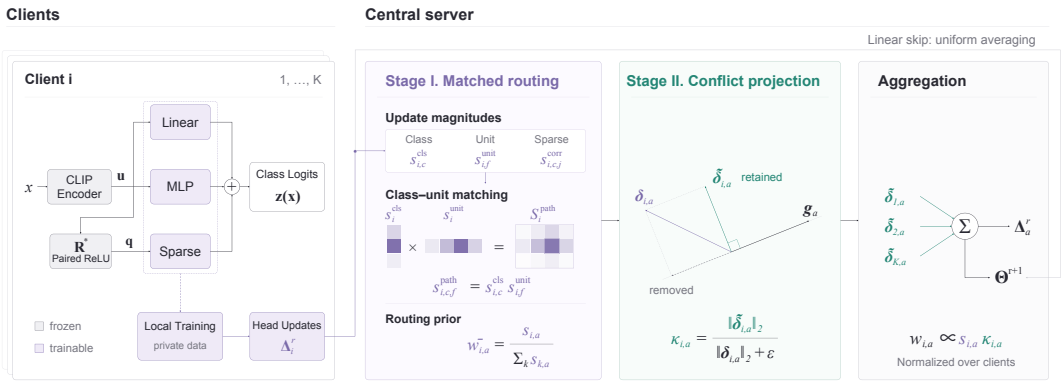


Figure 2: **Overview of PathwayFed: refine, route, and calibrate.** Dimensional expansion refines coarse feature-to-class expressions into finer-grained unit-to-class pathways, while the orthogonal sparse branch provides complementary sparse and less-overlapping pathway-relevant expressions. Matched routing reinforces strongly expressed client-specific contributions while preserving their pathway associations, and conflict-aware projection suppresses incompatible components before aggregation.

where $y_c = 1$ indicates that label c is relevant to x. Clients can differ in their visual distributions, label support, and label co-occurrence patterns. Their local multi-label objective is

$$\mathcal{L}_i(\mathbf{\Theta}) = \frac{1}{N_i} \sum_{n=1}^{N_i} \ell_{\mathrm{ml}} \left(f_{\mathbf{\Theta}}(x_{i,n}), \mathbf{y}_{i,n} \right), \tag{3}$$

where ℓ_{ml} denotes a multi-label classification loss.

At communication round r, the server broadcasts the parameter collection $\mathbf{\Theta}^r$. Client i performs local optimization, returns $\mathbf{\Theta}_i^{r+1}$, and thereby provides the ordinary local update $\mathbf{\Delta}_i^r = \mathbf{\Theta}_i^{r+1} - \mathbf{\Theta}^r$. Conventional federated aggregation combines these updates using client- or model-level rules, which can mix client-specific pathway contributions too coarsely. PathwayFed instead models how strongly each client contributes along different pathway-related components and aggregates these contributions selectively.

Client-specific pathway contributions are generally non-uniform across both pathways and clients, with substantial variation in update magnitude and compatibility. Accordingly, as illustrated in Figure 2, dimensional expansion and the fixed orthogonal sparse transformation refine pathway-relevant expressions, matched routing preserves strongly expressed client–pathway associations, and conflict-aware projection suppresses incompatible components before aggregation. Both server-side stages use only ordinary local model updates and require no additional client messages.

3.2 Pathway Refinement via Dimensional Expansion and Orthogonal Sparse Transformation

A frozen CLIP encoder (Radford et al., 2021) maps image x to $\mathbf{u}(x) \in \mathbb{R}^d$. Clients optimize a lightweight prediction head with expanded representation

$$\mathbf{h}_i(x) = \mathrm{ReLU}(\mathbf{W}_{e,i}\mathbf{u}(x) + \mathbf{b}_{e,i}), \tag{4}$$

where $\mathbf{h}_i(x) \in \mathbb{R}^m$, $\mathbf{W}_{e,i} \in \mathbb{R}^{m \times d}$, and $\mathbf{b}_{e,i} \in \mathbb{R}^m$, with $m > d$. The m dimensions of $\mathbf{h}_i(x)$ are *feature units*. For $f \in \{1, \ldots, m\}$, $h_{i,f}(x)$ is the scalar activation of unit f. For $c \in \{1, \ldots, C\}$, the complete class-logit vector is

$$\mathbf{z}_i(x) = \mathbf{W}_{\mathrm{skip},i}\mathbf{u}(x) + \mathbf{W}_{o,i}\mathbf{h}_i(x) + \mathbf{b}_{o,i} + \beta \mathbf{W}_{\mathrm{corr},i}\mathbf{q}(x), \tag{5}$$

where $\mathbf{z}_i(x) \in \mathbb{R}^C$, $\mathbf{W}_{\mathrm{skip},i} \in \mathbb{R}^{C \times d}$, $\mathbf{W}_{o,i} \in \mathbb{R}^{C \times m}$, $\mathbf{b}_{o,i} \in \mathbb{R}^C$, $\mathbf{q}(x) \in \mathbb{R}^{2d}$, $\mathbf{W}_{\mathrm{corr},i} \in \mathbb{R}^{C \times 2d}$, and β is a fixed configuration-specific correction scale (0.2 for COCO and 0.1 for NUS-WIDE in the reported launches). Independent sigmoids yield multi-label predictions.

Dimensional expansion maps $\mathbf{u}(x) \in \mathbb{R}^d$ to $\mathbf{h}_i(x) \in \mathbb{R}^m$ with $m > d$, spreading class-related information over a larger set of learnable feature units and thereby refining coarse feature-to-class expressions into finer-grained unit-to-class pathways. It does not itself guarantee semantic independence. Accordingly, our notion of disentanglement concerns client-specific pathway contributions during aggregation rather than semantic independence of individual feature units. Since $\mathbf{W}_{o,i} \in \mathbb{R}^{C\times m}$, the scalar logit for class c is

$$z_{i,c}(x) = \mathbf{W}_{\text{skip},i}[c,:]\mathbf{u}(x) + \sum_{f=1}^{m} W_{o,i}[c,f]h_{i,f}(x) + b_{o,i}[c] + \beta\mathbf{W}_{\text{corr},i}[c,:]\mathbf{q}(x). \tag{6}$$

We define the activation-based contribution of unit f to class c as

$$a_{i,c,f}(x) = |W_{o,i}[c,f]h_{i,f}(x)|\,. \tag{7}$$

The contribution from feature unit f to class c, mediated by $W_{o,i}[c,f]$, defines a *unit-to-class knowledge pathway*. The diagnostic in Section 4.5 uses $a_{i,c,f}(x)$ to interpret pathway distributions. This diagnostic does not transmit activation statistics.

The auxiliary branch applies an *orthogonal sparse transformation* constructed offline. Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be the public frozen-feature mean. Offline, an orthogonal transformation matrix $\mathbf{R}^\star \in \mathbb{R}^{d\times d}$ is obtained by sparsity optimization under an exact orthogonality constraint:

$$\mathbf{R}^\star = \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}_d} \mathbb{E}_{\mathbf{u}}\left[\left\|\mathbf{R}^\top(\mathbf{u}-\boldsymbol{\mu})\right\|_1\right]. \tag{8}$$

Using column vectors, $\mathbf{v}(x) = \mathbf{R}^{\star\top}(\mathbf{u}(x) - \boldsymbol{\mu}) \in \mathbb{R}^d$ and

$$\mathbf{q}(x) = \begin{bmatrix} \text{ReLU}(\mathbf{v}(x)) \\ \text{ReLU}(-\mathbf{v}(x)) \end{bmatrix}. \tag{9}$$

Paired ReLU preserves signed information: $\mathbf{v}(x) = \mathbf{q}_{1:d}(x) - \mathbf{q}_{d+1:2d}(x)$ and $\mathbf{u}(x) = \mathbf{R}^\star\mathbf{v}(x) + \boldsymbol{\mu}$. The ℓ_1 objective encourages sparsity, while orthogonality is a hard constraint. $\mathbf{R}^\star$ is learned offline and fixed throughout federation. Thus $q_j(x)$, $j \in \{1,\ldots,2d\}$, is an auxiliary sparse coordinate rather than a feature unit. By contrast, the $h_{i,f}(x)$ coordinates correspond to the feature units analyzed in Section 4.5. The orthogonal sparse transformation re-expresses complementary pathway-relevant information over sparse, decorrelated directions. The expanded branch provides finer-grained formal unit-to-class pathways, while the correction term $\beta\mathbf{W}_{\text{corr},i}\mathbf{q}(x)$ supplies complementary sparse routed coordinates. Together, the two branches make pathway-relevant information more selectively and less redundantly expressed across routed components. Details appear in Appendix C.

3.3 Matched Knowledge Routing

Matched routing assigns component-wise aggregation weights according to client-specific update magnitudes while preserving client–pathway correspondence. The resulting routing scores quantify contribution strength rather than correctness or reliability. For the expansion and classifier parameters, $\Delta\mathbf{W}_{e,i}[f,:] \in \mathbb{R}^d$, $\Delta b_{e,i}[f] \in \mathbb{R}$, $\Delta\mathbf{W}_{o,i}[c,:] \in \mathbb{R}^m$, and $\Delta b_{o,i}[c] \in \mathbb{R}$. Define

$$s_{i,f}^{\text{unit}} = \left\|[\Delta\mathbf{W}_{e,i}[f,:], \Delta b_{e,i}[f]]\right\|_2, \tag{10}$$

$$s_{i,c}^{\text{cls}} = \left\|[\Delta\mathbf{W}_{o,i}[c,:], \Delta b_{o,i}[c]]\right\|_2, \tag{11}$$

$$s_{i,c,f}^{\text{path}} = s_{i,c}^{\text{cls}} s_{i,f}^{\text{unit}}. \tag{12}$$

The factorized score $s_{i,c,f}^{\text{path}}$ combines class-side and unit-side update magnitudes. A pathway receives a large routing score only when both are strong. For the auxiliary correction branch, $\Delta W_{\text{corr},i}[c,j] \in \mathbb{R}$ and $s_{i,c,j}^{\text{corr}} = |\Delta W_{\text{corr},i}[c,j]|$. For a routed component a, let $s_{i,a} \in \mathbb{R}_{\geq 0}$ be its assigned score. We use $s_{i,c}^{\text{cls}}$ for $\mathbf{b}_{o,i}[c]$, $s_{i,f}^{\text{unit}}$ for $\mathbf{W}_{e,i}[f,:]$ and $b_{e,i}[f]$, $s_{i,c,f}^{\text{path}}$ directly for the complete client–class–unit pathway parameter $W_{o,i}[c,f]$, and $s_{i,c,j}^{\text{corr}}$ for $W_{\text{corr},i}[c,j]$. The skip branch uses uniform averaging in the implementation. Thus, not every parameter is assigned a full client–class–unit score. For the participant set $\mathcal{P}_r$,

$$\bar{w}_{i,a} = \frac{s_{i,a}}{\sum_{k\in\mathcal{P}_r} s_{k,a}}. \tag{13}$$

When all scores vanish, we use uniform averaging.

3.4 Conflict-Aware Projection and Final Aggregation

Matched routing determines component-wise aggregation relevance, but a highly weighted local update may still oppose the round-level optimization direction. Conflict-aware projection therefore removes opposing components and quantifies the retained update through a compatibility coefficient. For routed component a, let $\boldsymbol{\delta}_{i,a} \in \mathbb{R}^{d_a}$ be the vectorized local update, where d_a is its dimension. The server forms an inclusive round-level prototype

$$\mathbf{g}_a = \sum_{k \in \mathcal{P}_r} \boldsymbol{\delta}_{k,a}. \tag{14}$$

It then computes

$$\widetilde{\boldsymbol{\delta}}_{i,a} = \begin{cases} \boldsymbol{\delta}_{i,a}, & \langle \boldsymbol{\delta}_{i,a}, \mathbf{g}_a \rangle \geq 0, \\ \boldsymbol{\delta}_{i,a} - \dfrac{\langle \boldsymbol{\delta}_{i,a}, \mathbf{g}_a \rangle}{\|\mathbf{g}_a\|_2^2 + \epsilon} \mathbf{g}_a, & \text{otherwise.} \end{cases} \tag{15}$$

The scalar retained-energy compatibility is

$$\kappa_{i,a} = \frac{\|\widetilde{\boldsymbol{\delta}}_{i,a}\|_2}{\|\boldsymbol{\delta}_{i,a}\|_2 + \epsilon}. \tag{16}$$

For scalar components, a sign conflict has no orthogonal residual, so the projected update and compatibility are zero. Routing and compatibility jointly determine

$$w_{i,a} = \frac{s_{i,a}\kappa_{i,a}}{\sum_{k \in \mathcal{P}_r} s_{k,a}\kappa_{k,a}}, \qquad \boldsymbol{\Delta}_a^r = \sum_{i \in \mathcal{P}_r} w_{i,a} \widetilde{\boldsymbol{\delta}}_{i,a}. \tag{17}$$

The server applies $\boldsymbol{\Theta}_a^{r+1} = \boldsymbol{\Theta}_a^r + \boldsymbol{\Delta}_a^r$ component-wise. If $\|\mathbf{g}_a\|_2$ or every projection-adjusted score vanishes, it falls back to Eq. (13).

4 Experiments

4.1 Experimental Setup

Benchmarks and client partitions. We evaluate PASCAL VOC 2007 (20 labels), MS-COCO 2014 (80 labels), and NUS-WIDE-81 (81 labels). Following the heterogeneity evaluation setting of FedMPT (Wang et al., 2026), we construct clients by KMeans clustering frozen CLIP ViT-B/16 image features. For VOC, COCO, and NUS-WIDE, we use $K \in \{2, 4, 8, 12, 16, 20\}$, $K \in \{8, 16, 32, 48, 64, 80\}$, and $K \in \{8, 16, 32, 49, 65, 81\}$, respectively. For each dataset, these six values correspond to $t \in \{10, 20, 40, 60, 80, 100\}\%$, with larger t indicating more specialized client partitions. Evaluation remains on a shared set. All clients participate in every round. We report global mAP, class-wise F1 (CF1), and overall F1 (OF1).

Implementation details. All experiments use a frozen CLIP ViT-B/16 encoder, asymmetric multi-label loss, 100 communication rounds, and three seeds (34–36). The fixed $R^\star$ is constructed once from unlabeled public ImageNet training features without benchmark test images or labels. Full details are provided in the appendix.

Compared methods. We compare with Fed-DualCoOp (Sun et al., 2022), Fed-SCPNet, Fed-MaPLe, FedPGP, Fed-TCP, FedTPG (Qiu et al., 2024), Fed-PosCoOp, FedAWA (Shi et al., 2025), Fed-RAM (Tan et al., 2025), FedMVP (Singha et al., 2025), FedMPT (Wang et al., 2026), and Fed-Harmony (Kou et al., 2026). The main table reports representative and recent methods, while the complete benchmark comparison is provided in Appendix J.

4.2 Main Results under Heterogeneity

Table 1 shows a consistent heterogeneity-dependent pattern. PathwayFed remains competitive under mild partitioning and becomes increasingly advantageous as client specialization strengthens. Across the three benchmarks, its performance remains comparatively stable under strong heterogeneity, whereas several prior methods degrade more substantially. This pattern is consistent with the

Table 1: Comparison under the heterogeneity axis. t increases from left to right. The complete benchmark comparison is provided in Appendix J. The final row of each dataset gives the signed difference from the strongest prior method in the full comparison. Positive values are green and negative values are red.

Methods	Venue	t=10%			20%			40%			60%			80%			100%			Avg.		
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
PASCAL VOC 2007																						
Fed-DualCoOp (Sun et al., 2022)	NeurIPS'22	87.67	81.32	81.91	85.98	78.84	76.16	85.73	78.33	76.44	84.46	77.13	75.25	83.76	75.44	73.89	82.02	75.00	74.73	84.94	77.68	76.40
FedTPG (Qiu et al., 2024)	ICLR'24	88.11	81.13	78.05	85.23	81.26	77.73	85.95	79.05	78.54	84.23	77.37	78.68	84.45	77.29	76.13	82.72	76.67	75.87	85.12	78.80	77.50
FedAWA (Shi et al., 2025)	CVPR'25	84.75	78.57	80.41	83.17	76.06	78.35	81.77	75.11	79.99	79.48	72.26	79.24	78.81	70.49	77.68	78.22	70.73	70.15	81.03	73.87	77.64
Fed-RAM (Tan et al., 2025)	CVPR'25	88.05	81.50	81.98	86.24	81.27	80.70	85.50	79.74	80.01	85.68	79.24	78.90	85.53	79.56	72.84	83.04	77.00	68.99	85.67	79.72	77.24
FedMVP (Singha et al., 2025)	ICCV'25	87.14	80.24	82.30	86.86	79.32	80.22	86.64	79.00	79.93	85.61	79.37	80.65	84.87	79.23	75.60	82.43	79.56	77.92	85.59	79.45	79.44
FedMPT (Wang et al., 2026)	CVPR'26	90.13	84.33	86.82	90.12	84.91	83.23	90.01	84.42	84.31	90.00	83.88	84.56	89.61	83.72	80.36	87.19	81.31	82.43	89.51	83.76	83.62
FedHarmony (Kou et al., 2026)	CVPR'26	90.31	86.29	87.88	90.46	86.03	87.84	90.69	85.95	87.59	90.48	86.08	87.24	89.90	84.94	86.60	89.21	84.52	85.73	90.17	85.64	87.15
PathwayFed	Ours	90.27	85.83	87.56	90.32	86.04	87.83	90.55	86.07	87.88	90.55	86.08	87.93	90.49	86.05	87.91	90.47	85.96	87.86	90.44	86.01	87.83
Δ vs. full prior best	–	-0.04	-0.46	-0.32	-0.14	+0.01	-0.01	-0.14	+0.12	+0.29	+0.07	+0.00	+0.69	+0.59	+1.11	+1.31	+1.26	+1.44	+2.13	+0.27	+0.37	+0.68
MS-COCO 2014																						
Fed-DualCoOp (Sun et al., 2022)	NeurIPS'22	61.88	58.34	62.15	58.39	54.11	59.82	55.95	51.99	57.17	53.80	49.55	58.33	52.28	48.34	57.86	51.46	48.57	56.95	55.63	51.82	58.71
FedTPG (Qiu et al., 2024)	ICLR'24	63.91	59.39	64.70	61.34	56.05	55.72	60.95	55.47	51.31	60.54	54.52	54.18	54.53	51.63	53.92	50.88	53.18	52.97	58.69	55.04	55.47
FedAWA (Shi et al., 2025)	CVPR'25	65.58	60.72	61.20	62.09	59.73	60.50	61.47	57.58	59.14	60.14	56.67	59.46	59.01	52.69	58.77	57.57	49.65	57.73	60.98	56.17	59.47
Fed-RAM (Tan et al., 2025)	CVPR'25	66.06	60.98	62.89	63.08	57.97	64.45	59.97	57.97	64.45	60.85	56.95	62.73	60.88	53.00	56.25	55.64	50.37	55.04	61.08	56.21	60.97
FedMVP (Singha et al., 2025)	ICCV'25	65.76	58.06	64.01	63.19	56.75	65.48	62.26	57.20	63.51	60.60	53.42	62.66	59.56	51.84	59.51	58.48	50.42	55.30	61.64	54.62	61.75
FedMPT (Wang et al., 2026)	CVPR'26	67.37	62.88	67.57	65.91	60.38	67.11	65.31	59.10	66.07	63.86	58.46	64.32	63.09	57.19	63.68	62.37	56.96	62.80	64.65	59.16	65.26
FedHarmony (Kou et al., 2026)	CVPR'26	69.06	61.67	70.58	68.82	60.77	70.76	67.47	58.02	70.16	66.13	55.12	69.54	65.05	53.26	69.06	63.00	51.34	68.60	66.59	56.70	69.78
PathwayFed	Ours	68.88	64.90	70.05	69.74	65.02	70.49	69.77	64.09	70.59	69.56	63.59	70.53	69.45	63.28	70.55	69.33	63.64	70.46	69.46	64.09	70.44
Δ vs. full prior best	–	-0.18	+3.23	-0.53	+0.92	+4.25	-0.27	+2.30	+6.07	+0.43	+3.43	+8.47	+0.99	+4.40	+10.02	+1.49	+6.33	+12.30	+1.86	+2.87	+7.39	+0.66
NUS-WIDE-81																						
Fed-DualCoOp (Sun et al., 2022)	NeurIPS'22	51.02	45.89	69.00	49.97	44.36	65.30	46.25	41.62	67.02	42.69	37.11	66.10	40.65	35.69	60.28	39.40	34.54	59.04	45.00	39.87	64.46
FedTPG (Qiu et al., 2024)	ICLR'24	53.01	50.69	73.32	52.78	47.66	72.50	50.00	46.72	69.52	50.40	44.21	70.47	49.92	43.89	67.10	49.46	42.55	65.46	50.93	45.95	69.73
FedAWA (Shi et al., 2025)	CVPR'25	54.67	47.61	74.51	53.91	45.52	70.33	50.40	43.74	71.60	52.87	43.22	71.51	52.75	43.02	66.89	50.71	41.13	66.53	52.55	44.04	70.23
Fed-RAM (Tan et al., 2025)	CVPR'25	56.35	48.89	76.07	54.05	47.89	76.40	53.51	45.69	74.33	52.25	40.69	74.38	52.61	42.12	72.98	51.20	40.69	70.35	53.33	44.33	74.09
FedMVP (Singha et al., 2025)	ICCV'25	53.20	48.37	77.80	53.32	46.39	77.10	51.90	45.21	75.14	52.98	43.43	75.02	51.33	42.79	74.55	51.08	40.16	72.90	52.30	44.39	75.42
FedMPT (Wang et al., 2026)	CVPR'26	58.36	51.56	78.98	57.27	49.29	77.93	57.03	48.14	76.73	56.52	46.88	76.94	56.08	45.73	76.60	54.87	45.03	76.81	56.69	47.77	77.33
FedHarmony (Kou et al., 2026)	CVPR'26	51.45	40.52	71.50	59.33	50.27	80.14	58.12	48.49	79.91	56.73	46.82	79.57	55.35	45.53	79.10	53.90	43.65	78.68	55.81	45.88	78.15
PathwayFed	Ours	60.12	57.81	80.05	60.40	58.03	80.28	60.40	57.74	80.35	60.19	57.38	80.30	60.10	57.18	80.23	59.95	57.12	80.18	60.19	57.54	80.23
Δ vs. full prior best	–	+1.76	+6.25	+1.07	+1.07	+8.74	+0.14	+2.28	+9.60	+0.44	+3.67	+10.50	+0.73	+4.02	+11.45	+1.13	+5.08	+12.09	+1.50	+3.50	+9.77	+2.90

Table 2: Component ablation under high heterogeneity. The final row reports the gain of full PathwayFed over dense uniform aggregation.

Orthogonal sparse transformation	Matched routing	Conflict projection	COCO $K = 80$	NUS-WIDE $K = 81$
–	–	–	65.05 ± 0.28	55.94 ± 0.22
–	✓	–	67.36 ± 0.27	58.55 ± 0.13
✓	–	–	65.82 ± 0.57	56.22 ± 0.10
✓	✓	–	68.06 ± 0.15	58.89 ± 0.11
✓	✓	✓	69.33 ± 0.31	59.95 ± 0.12
Δ vs. dense uniform baseline		–	$+\mathbf{4.28}$	$+\mathbf{4.01}$

intended role of pathway-aware aggregation: stronger specialization produces increasingly uneven pathway usage, making coarse aggregation more likely to mix informative client-specific contributions.

4.3 Ablation Study: Routing and Conflict Calibration

Table 2 examines the complementary roles of matched routing, the orthogonal sparse representation, and conflict calibration under high heterogeneity. Matched routing accounts for the largest component-wise improvement over dense uniform aggregation. Adding the orthogonal sparse branch yields a complementary improvement when matched routing is enabled, while conflict-aware projection provides a further improvement by suppressing incompatible components. Because dimensional expansion is not independently isolated, this ablation does not attribute the representation-side gain to expansion alone.

4.4 Mechanism Analysis under Increasing Heterogeneity

Our pathway view predicts that stronger client specialization increases cross-client pathway mismatch. We diagnose cross-client pathway mismatch post hoc using the class-related pathway dis-

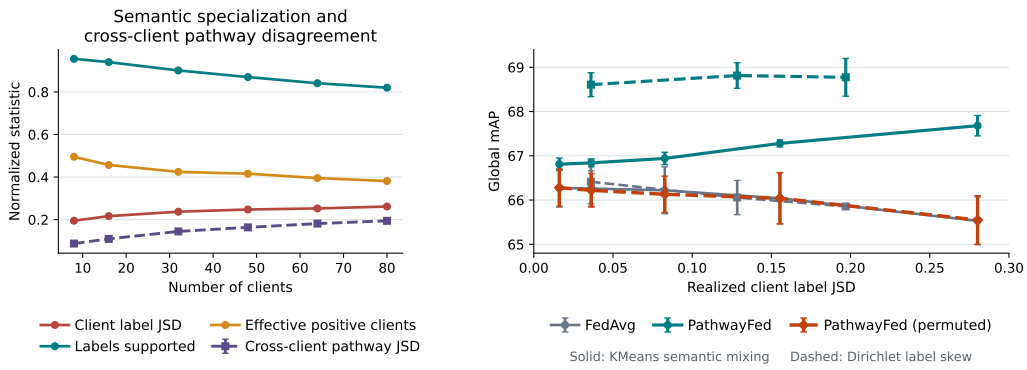


Figure 3: Mechanism controls on COCO. Left: increasing specialization raises cross-client pathway disagreement in FedHarmony. Right: semantic-mixing, route-permutation, and Dirichlet controls test the importance of client–pathway association.

Table 3: Client- and sample-scale-normalized routing controls on COCO K=80. Values are mean $\pm$ standard deviation over seeds 34–36. The raw-routing row reports the corresponding result from Table 1.

Method	mAP	CF1	OF1
PathwayFed (raw routing)	69.33 ± 0.31	63.64 ± 0.16	70.46 ± 0.07
PathwayFed (client-scale normalized)	69.41 ± 0.33	63.63 ± 0.14	70.44 ± 0.06
PathwayFed (sample-count normalized)	69.44 ± 0.32	63.55 ± 0.15	70.41 ± 0.06

tribution $\pi_{i,c,f}$, obtained by normalizing the positive-sample pathway contribution $|W_o[c, f]h_f(x)|$ over the valid classes of client i. Full definitions are provided in Appendix G.

Pathway disagreement increases with client specialization. FedHarmony serves as an external diagnostic baseline for characterizing heterogeneous pathway usage independently of PathwayFed's pathway-aware aggregation. Figure 3 summarizes the mechanism controls. Along the COCO KMeans axis, cross-client pathway JSD increases from 0.0874 ± 0.0016 at $K = 8$ to 0.1944 ± 0.0023 at $K = 80$ (three seeds), alongside increasing client-to-global label JSD and decreasing label support. Thus, the same global units exhibit increasingly different class-related pathway distributions across specialized clients. This disagreement is an observable manifestation associated with pathway entanglement, not an entanglement score.

Performance depends on preserving client–pathway correspondence. Figure 3 shows that, at $K = 80$, the advantage of PathwayFed grows with realized label JSD. Because permutation preserves the routing-weight distribution and non-uniformity while disrupting its association with the originating update, the resulting degradation cannot be attributed solely to generic non-uniform weighting. The same trend under Dirichlet label skew shows that the effect is not specific to KMeans.

Routing gains are robust to client- and sample-scale normalization. To control for global client update magnitude and local data volume, we normalize each routed-component score by either the norm of the client's complete local update or its local sample count. Table 3 shows that both normalized variants retain essentially the same mAP and OF1 as raw routing on COCO $K = 80$, with only minor CF1 variation, indicating that the routing advantage cannot be explained solely by global update scale or client data volume.

The relative gain of matched routing increases with pathway mismatch. Figure 4 interpolates between uniform and matched routing with $\lambda \in \{0, 0.5, 1\}$ while holding all other training choices fixed. The full-routing gain increases from -0.89 to $+2.63$ mAP on COCO and from -0.21 to $+3.02$ on NUS-WIDE along the heterogeneity axis. This behavior is consistent with, but does not by itself causally prove, the pathway-entanglement explanation.

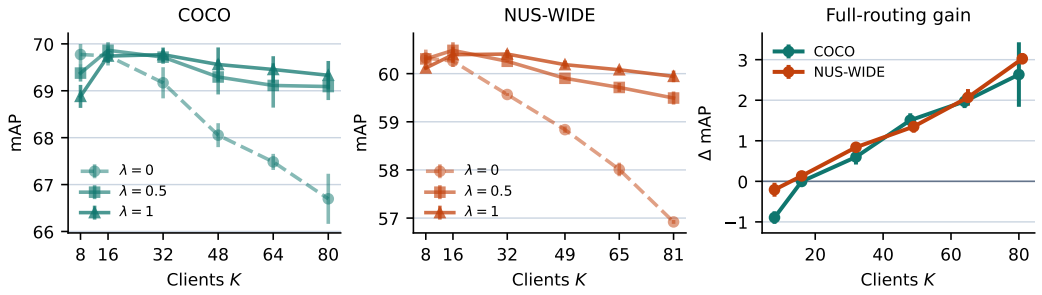


Figure 4: Routing-strength controls across the KMeans heterogeneity axis. The gain of full matched routing over uniform aggregation increases with client specialization and pathway mismatch.

Table 4: Within-client unit-to-class pathway concentration of the final aggregated model on COCO $K = 80$.

Method	Pathway Entropy ↓	Top-1 Share ↑	Top-3 Share ↑	Effective Classes ↓
FedHarmony	0.826 ± 0.002	0.148 ± 0.002	0.312 ± 0.002	25.76 ± 0.10
PathwayFed	$\mathbf{0.768 \pm 0.005}$	$\mathbf{0.193 \pm 0.004}$	$\mathbf{0.380 \pm 0.006}$	$\mathbf{21.65 \pm 0.21}$

4.5 Analysis of Unit-to-Class Pathway Concentration

Complementing the cross-client analysis in Section 4.4, we evaluate within-client pathway concentration using $\boldsymbol{\pi}_{i,:,f}$ on each client-conditioned distribution.

Table 4 shows that PathwayFed exhibits lower normalized pathway entropy, larger top-1 and top-3 mass, and fewer effective classes, corresponding to a more concentrated and less diffuse unit-to-class pathway distribution. This post-hoc diagnostic does not establish semantic independence. Additional definitions and sparsity diagnostics are provided in the appendix.

5 Conclusion

We presented PathwayFed for addressing client-specific pathway entanglement in heterogeneous federated multi-label learning. Dimensional expansion refines coarse feature-to-class expressions into finer-grained unit-to-class pathways, while the orthogonal sparse transformation provides complementary sparse and less-overlapping pathway-relevant expressions. Matched routing reinforces strongly expressed client-specific contributions while preserving their client–pathway associations, and conflict-aware projection suppresses incompatible components and calibrates how much routed knowledge is retained. Across three benchmarks, PathwayFed remains competitive under mild heterogeneity, while its performance margin increases as client specialization strengthens. Controlled analyses show that cross-client pathway disagreement increases with specialization, performance depends on preserving client–pathway correspondence, and the relative gain of matched routing increases with pathway mismatch. Post-hoc concentration analysis further indicates less diffuse unit-to-class pathway structure. PathwayFed thus refines pathway expressions, preserves strongly expressed client-specific contributions through matched routing, and calibrates incompatible components before aggregation.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8730–8738, 2018.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, 2025.

Yi Hu, Hanchi Ren, Chen Hu, Jingjing Deng, and Xianghua Xie. An element-wise weights aggregation method for federated learning. In *IEEE International Conference on Data Mining Workshops*, pp. 190–197, 2023. doi: 10.1109/ICDMW60847.2023.00031.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-IID data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7865–7873, 2021.

Ziwen Huang and Nikolaos M. Freris. Reinforcement-learning-based layer-wise aggregation for personalized federated learning. *IEEE Internet of Things Journal*, 2025. doi: 10.1109/JIOT.2024.3502245.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

Zhiqiang Kou, Junxiang Wu, Wenke Huang, Wenwen He, Ming-Kun Xie, Changwei Wang, Yuheng Jia, Di Jiang, Yang Liu, Xin Geng, and Qiang Yang. Fedharmony: Harmonizing heterogeneous label correlations in federated multi-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jiachen Li, Yuchao Zhang, Yiping Li, Xiangyang Gong, and Wendong Wang. Gradient calibration for non-i.i.d. federated learning. In *Proceedings of the 2nd ACM Workshop on Data Privacy and Federated Learning Technologies for Mobile Edge Network*, pp. 119–124, 2023. doi: 10.1145/3615593.3615721.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020.

Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. FedLPA: One-shot federated learning with layer-wise posterior aggregation. In *NeurIPS*, 2024.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *Advances in Neural Information Processing Systems*, 38:95706–95742, 2025.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konecný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021.

Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30651–30660, 2025.

Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. Fedmvp: Federated multimodal visual prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *Advances in Neural Information Processing Systems*, 2022.

Hao Tan, Zichang Tan, Jun Li, Ajian Liu, Jun Wan, and Zhen Lei. Recover and match: Open-vocabulary multi-label recognition through knowledge-constrained optimal transport. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4650–4660, 2025.

Yitong Tang. Adapted weighted aggregation in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23763–23765, 2024. doi: 10.1609/aaai.v38i21.30557.

Xucong Wang, Pengkun Wang, Zhe Zhao, Liheng Yu, Shuang Wang, and Yang Wang. Fedmpt: Federated multi-label prompt tuning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17226–17236, 2026.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021.

Xinyu Zhang, Weiyu Sun, and Ying Chen. Tackling the non-IID issue in heterogeneous federated learning by gradient harmonization. *IEEE Signal Processing Letters*, 31:2595–2599, 2024. doi: 10.1109/LSP.2024.3430042.

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. FedCALM: Conflict-aware layer-wise mitigation for selective aggregation in deeper personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15444–15453, 2025. doi: 10.1109/CVPR52734.2025.01439.

Ran Zhu, Mingkun Yang, Shiqiang Wang, Jie Yang, and Qing Wang. Flick: Empowering federated learning with commonsense knowledge. In *Advances in Neural Information Processing Systems*, 2025.

A Reproducibility Details

Client construction. We cluster frozen CLIP ViT-B/16 image features with KMeans and assign each cluster to one client. The six t values are percentages of the dataset label count, matching the layout in FedMPT. All clients participate in every communication round.

Clean runner. The clean pipeline uses a frozen CLIP ViT-B/16 encoder at 224px, cached image features, asymmetric loss, and a 100-round budget. Seeds are 34, 35, and 36. Caching does not change the numerical feature representation because the encoder is frozen. The NUS-WIDE-81 image paths are rebuilt from the official image archive while labels use the 81-concept split.

Communication boundary. In matched routing, clients transmit only ordinary local model/head updates. No additional activation, label-count, sample-count, or pathway statistics are transmitted. The server calculates all routing weights from the received local head states. This is a protocol restriction on additional statistics, not a formal privacy guarantee against update inversion.

B Efficiency and Communication Details

Table 5: Current PathwayFed resource profile. The orthogonal transformation is an inference buffer, not a trainable or per-round state.

Method	Tunable parameters	Orthogonal transform buffer	Additional client messages
PathwayFed	0.577M	1.00 MiB	none

Table 5 summarizes this auxiliary profile of the current implementation. Performance results are reported in Table 1.

C Additional Interpretation of the Routing and Representation Ablation

The 2×2 ablation uses a common update-only aggregation primitive for both the raw representation and the orthogonal sparse transformation. The raw control assigns routing weights from the corresponding local router and classifier update magnitudes. The orthogonal model uses the same rule for its routed branches and correction weights. At $\lambda = 0$, all of these coordinate-wise weights are uniform. Therefore, this experiment changes neither the client messages nor the training protocol when it switches between uniform and matched routing. Matched routing is the aggregation-level mechanism that separates strongly expressed client-specific contributions while preserving their client–pathway associations. The orthogonal sparse branch provides complementary pathway refinement, and conflict-aware projection calibrates compatibility after routing. Because dimensional expansion is not independently varied, this ablation does not isolate its contribution.

D Offline Construction of the Orthogonal Sparse Transformation

Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be the public frozen-feature mean. We initialize $\mathbf{R} \in \mathbb{R}^{d \times d}$ with the covariance eigenbasis and minimize $\mathbb{E}_{\mathbf{u}}[\|\mathbf{R}^\top(\mathbf{u} - \boldsymbol{\mu})\|_1]$ subject to $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_d$. Optimization follows the orthogonal manifold, with QR retraction after each update. Thus the ℓ_1 objective encourages sparse projected features while orthogonality is an exact constraint, not a training regularization loss. The resulting $\mathbf{R}^\star$ is fixed for every client and communication round. It uses no client labels, client updates, or benchmark test samples, and supplies the complementary sparse and less-overlapping pathway-relevant representation used by the correction branch. Orthogonality applies to these representation-coordinate directions. This constraint does not establish that the formal unit-to-class pathways $h_f \rightarrow c$ through $W_o[c, f]$ are mutually orthogonal. Accordingly, each q_j is an auxiliary sparse coordinate rather than a formal feature unit or pathway.

E Routing-Strength Intervention

For the routing-strength intervention in Figure 4, the server replaces matched routing weights w_{update} with $w_\lambda = \lambda w_{\text{update}} + (1-\lambda) w_{\text{uniform}}$ for every routed parameter group. The values $\lambda \in \{0, 0.5, 1\}$ were independently trained under the same protocol. This intervention directly varies the strength of matched pathway-aware routing while holding the frozen representation, local objective, client split, and training budget fixed, and tests whether routing becomes more beneficial as pathway mismatch grows. It is not a representation-refinement intervention.

F Fixed-Capacity Semantic-Mixing Control

For the controlled intervention in Section 4.4, we retain the COCO $K = 80$ KMeans client assignment with probability γ. We place the remaining samples uniformly into the unfilled slots across clients, preserving the original capacity of every client exactly. This construction varies semantic specialization while holding the training examples, client count, capacity vector, test set, optimizer, and 100-round budget fixed, thereby controlling for explanations based merely on more clients or different client sizes. The realized label-distribution JSD values for $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$ are 0.0161, 0.0363, 0.0827, 0.1554, and 0.2802, respectively. These values characterize label-distribution heterogeneity, not pathway entanglement.

G Cross-Client Pathway Disagreement Across Heterogeneity Levels

Cross-client pathway disagreement asks how differently the same global feature unit is used across client-conditioned distributions. For every active client–unit pair, we use the distribution $\boldsymbol{\pi}_{i,:,f}$ defined in Section 4.5, embed it in the common 80-class support by assigning zero probability to unsupported classes, and average it over active clients $\mathcal{I}_f$. We report the normalized Jensen–Shannon divergence $D_f = [H(\bar{\boldsymbol{\pi}}_{:,f}) - |\mathcal{I}_f|^{-1} \sum_{i \in \mathcal{I}_f} H(\boldsymbol{\pi}_{i,:,f})] / \log 80$, where $\bar{\boldsymbol{\pi}}_{:,f} = |\mathcal{I}_f|^{-1} \sum_{i \in \mathcal{I}_f} \boldsymbol{\pi}_{i,:,f}$. A unit is included only when it is active (mean ReLU activation above 10^{-6} and nonzero pathway strength) for at least two clients. We then average D_f equally over valid global feature units. This statistic measures cross-client pathway disagreement, an observable manifestation associated with pathway entanglement. It is neither an entanglement score nor a measure of within-client pathway concentration, which is reported in Table 4.

As reported in Table 6, cross-client pathway JSD increases monotonically with K and mean client-label JSD (Spearman $\rho = 1.0$ over the six levels, reported descriptively). The trend is monotonic in each seed. In contrast, local pathway entropy can decline as a client becomes more specialized, so the two statistics capture different levels of the heterogeneous aggregation problem.

Fixed-K semantic-mixing diagnostic. With $K = 80$, capacities and training examples fixed, the completed seed-34 checkpoints show the same increase in cross-client pathway JSD as semantic mixing is reduced: 0.0815, 0.0875, 0.1015, 0.1254, and 0.1930 for $\gamma = 0, 0.25, 0.5, 0.75, 1$, respectively. This is an illustrative single-seed diagnostic and is interpreted descriptively.

Table 6: FedHarmony cross-client pathway disagreement on the COCO KMeans axis. Values are mean $\pm$ standard deviation over seeds 34–36.

K	Mean client-label JSD	Cross-client pathway JSD ↑
8	0.1414	0.0874 ± 0.0016
16	0.1551	0.1095 ± 0.0023
32	0.1703	0.1446 ± 0.0023
48	0.1773	0.1636 ± 0.0020
64	0.1848	0.1814 ± 0.0026
80	0.1942	0.1944 ± 0.0023

Table 7: Partial-client-participation robustness on COCO $K{=}80$ under the current conflict-projection configuration. Values are global mAP means $\pm$ standard deviations over three seeds.

Method	$p = 20\%$	$p = 50\%$
FedHarmony-style	62.36 ± 0.83	62.84 ± 0.31
PATHWAYFED	$\mathbf{68.28 \pm 0.56}$	$\mathbf{68.93 \pm 0.39}$

H ROUTE-PERMUTATION NEGATIVE CONTROL

For every routed class, feature, and correction coordinate, standard matched routing derives non-negative client weights from the corresponding local update magnitudes. The negative control independently permutes these weights along the client axis for each coordinate before aggregation. It consequently preserves the distribution of weights, effective client count, concentration, and all local models, while breaking the association between a routing weight and the client that produced its update. It does not use labels, sample counts, or activations. This control tests whether performance depends on preserving the correct client–pathway association rather than merely applying concentrated or otherwise non-uniform weights.

I PARTIAL-CLIENT-PARTICIPATION ROBUSTNESS

The main results use full participation to isolate semantic heterogeneity. We also evaluate the practical setting in which each client is independently available with probability p in a round. Both methods retain their original aggregation rules, while PATHWAYFED derives all routing weights only from submitted local head updates and sends no labels, sample counts, activations, or auxiliary routing statistics. Table 7 shows that PathwayFed retains a 5.92-point advantage at 20% availability and a 6.09-point advantage at 50% availability. When only a subset of specialized clients is present, matched routing can still selectively preserve their strongly expressed contributions. This robustness result does not by itself establish pathway disentanglement.

J FULL BENCHMARK COMPARISON

Table 8 reports the complete benchmark comparison.

K ADDITIONAL ROUTING-STRENGTH CONTROLS

Figure 5 reports all three metrics across the heterogeneity axis, and Table 9 summarizes the endpoint gains.

L COMPLEMENTARY CLIENT-SPECIFIC ROUTED UPDATE-SUBSPACE ANALYSIS

For each client, we concatenate the normalized routed-block updates recorded every ten rounds and compute an SVD. The top-k energy is $\sum_{r=1}^{k} \sigma_r^2 / \|\mathbf{X}_i\|_F^2$. Client-specific and pooled-global energies are the corresponding squared-Frobenius projection ratios onto their top-5 bases. Table 10 and Figure 6 summarize the joint routed MLP diagnostic over seeds 34–36. These diagnostics characterize the concentration and distinctness of client-specific routed-update geometry. They provide complementary evidence for heterogeneous client-specific contributions, but the update subspaces are not formal unit-to-class pathways, and low overlap or large Grassmann distance does not establish pathway independence or semantic disentanglement.

M UNIT-TO-CLASS PATHWAY CONCENTRATION DIAGNOSTICS

This post-hoc representation-level diagnostic measures how concentrated the class-related pathway mass of each active feature unit is within a client distribution. It provides complementary evi-

Table 8: Complete benchmark comparison under the common image-feature KMeans heterogeneity axis. The Δ rows are computed against the strongest prior method in this full comparison.

Methods	Venue	t=10%			20%			40%			60%			80%			100%			Avg.	
		mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
PASCAL VOC 2007																						
Fed-DualCoOp	NeurIPS'22	87.67	81.32	81.91	85.98	78.84	76.16	85.73	78.33	76.44	84.46	77.13	75.25	83.76	75.44	73.89	82.02	75.00	74.73	84.94	77.68	76.40
Fed-SCPNet	CVPR'23	83.37	78.74	77.86	81.15	74.68	76.05	80.03	74.05	74.09	79.89	66.04	73.45	76.51	74.29	71.13	76.48	74.06	70.64	79.57	73.64	73.87
Fed-MaPLe	CVPR'23	84.22	78.00	82.36	87.87	81.27	80.70	84.36	76.35	75.82	81.82	73.73	73.55	80.71	70.65	72.04	76.08	68.13	70.89	82.51	74.69	75.89
FedPGP	ICML'24	85.47	77.97	79.81	84.51	76.76	77.66	84.47	75.99	73.97	79.89	72.72	67.48	76.60	68.71	66.52	78.16	69.30	68.35	81.52	73.58	72.30
Fed-TCP	CVPR'24	81.56	76.02	79.23	81.15	72.41	79.50	80.65	78.93	80.44	77.32	78.38	78.89	76.50	76.64	73.50	76.42	71.52	76.48	78.93	75.65	78.01
FedTPG	ICLR'24	88.11	81.13	78.05	85.23	81.26	77.73	85.95	79.05	78.54	84.23	77.37	78.68	84.45	77.29	76.13	82.72	76.67	75.87	85.12	78.80	77.50
Fed-PosCoOp	WACV'25	87.42	80.67	79.51	84.71	82.77	77.86	82.53	80.90	76.97	81.65	80.31	76.43	80.71	80.17	75.79	79.90	78.26	75.08	82.82	80.51	76.94
FedAWA	CVPR'25	84.75	78.57	80.41	83.17	76.06	78.35	81.77	75.11	79.99	79.48	72.26	79.24	78.81	70.49	77.68	78.22	70.73	70.15	81.03	73.87	77.64
Fed-RAM	CVPR'25	88.05	81.50	81.98	86.24	81.27	80.70	85.50	79.74	80.01	85.68	79.24	78.90	85.53	79.56	72.84	83.04	77.00	68.99	85.67	79.72	77.24
FedMVP	ICCV'25	87.14	80.24	82.30	86.86	79.32	80.22	86.64	79.00	79.93	85.61	79.37	80.65	84.87	79.23	75.60	82.43	79.56	77.92	85.59	79.45	79.44
FedMPT	CVPR'26	90.13	84.33	86.82	90.12	84.91	83.23	90.01	84.42	84.31	90.00	83.88	84.56	89.61	83.72	80.36	87.19	81.31	82.43	89.51	83.76	83.62
FedHarmony	CVPR'26	90.31	86.29	87.88	90.46	86.03	87.84	90.69	85.95	87.59	90.48	86.08	87.24	89.90	84.94	86.60	89.21	84.52	85.73	90.17	85.64	87.15
PathwayFed	Ours	90.27	85.83	87.56	90.32	86.04	87.83	90.55	86.07	87.88	90.55	86.08	87.93	90.49	86.05	87.91	90.47	85.96	87.86	90.44	86.01	87.83
Δ vs. full prior best	–	-0.04	-0.46	-0.32	-0.14	+0.01	-0.01	-0.14	+0.12	+0.29	+0.07	+0.00	+0.69	+0.59	+1.11	+1.31	+1.26	+1.44	+2.13	+0.27	+0.37	+0.68
MS-COCO 2014																						
Fed-DualCoOp	NeurIPS'22	61.88	58.34	62.15	58.39	54.11	59.82	55.95	51.99	57.17	53.80	49.55	58.33	52.28	48.34	57.86	51.46	48.57	56.95	55.63	51.82	58.71
Fed-SCPNet	CVPR'23	60.96	57.96	59.70	57.68	52.91	57.45	52.79	53.09	53.50	50.10	50.73	55.36	48.40	48.84	45.61	46.67	42.25	47.75	52.77	50.96	53.23
Fed-MaPLe	CVPR'23	64.83	61.28	64.45	62.98	58.40	62.16	61.80	58.25	57.43	60.81	57.83	55.89	55.83	51.76	52.54	47.24	49.97	50.13	58.92	56.25	57.10
FedPGP	ICML'24	63.10	61.09	65.36	59.86	58.35	64.78	58.21	57.52	62.16	58.18	55.79	60.87	56.52	52.36	60.85	54.53	53.19	53.54	58.40	56.38	61.26
Fed-TCP	CVPR'24	62.66	61.50	67.16	62.79	57.46	62.69	61.77	53.16	60.13	60.07	52.46	57.72	59.99	50.93	55.78	57.22	47.75	53.56	60.75	53.88	59.51
FedTPG	ICLR'24	63.91	59.39	64.70	61.34	56.05	55.72	60.95	55.47	51.31	60.54	54.52	54.18	54.53	51.63	53.92	50.88	53.18	52.97	58.69	55.04	55.47
Fed-PosCoOp	WACV'25	63.28	58.69	62.57	59.94	55.71	60.91	55.88	50.41	60.13	53.23	49.58	59.37	50.95	50.03	58.04	50.02	48.06	56.09	55.55	52.08	59.52
FedAWA	CVPR'25	65.58	60.72	61.20	62.09	59.73	60.50	61.47	57.58	59.14	60.14	56.67	59.46	59.01	52.69	58.77	57.57	49.65	57.73	60.98	56.17	59.47
Fed-RAM	CVPR'25	66.06	60.98	62.89	63.08	57.97	64.45	59.97	57.97	64.45	60.85	56.95	62.73	60.88	53.00	56.25	55.64	50.37	55.04	61.08	56.21	60.97
FedMVP	ICCV'25	65.76	58.06	64.01	63.19	56.75	65.48	62.26	57.20	63.51	60.60	53.42	62.66	59.56	51.84	59.51	58.48	50.42	55.30	61.64	54.62	61.75
FedMPT	CVPR'26	67.37	62.88	67.57	65.91	60.38	67.11	65.31	59.10	66.07	63.86	58.46	64.32	63.09	57.19	63.68	62.37	56.96	62.80	64.65	59.16	65.26
FedHarmony	CVPR'26	69.06	61.67	70.58	68.82	60.77	70.76	67.47	58.02	70.16	66.13	55.12	69.54	65.05	53.26	69.06	63.00	51.34	68.60	66.59	56.70	69.78
PathwayFed	Ours	68.88	64.90	70.05	69.74	65.02	70.49	69.77	64.09	70.59	69.56	63.59	70.53	69.45	63.28	70.55	69.33	63.64	70.46	69.46	64.09	70.44
Δ vs. full prior best	–	-0.18	+3.23	-0.53	+0.92	+4.25	-0.27	+2.30	+6.07	+0.43	+3.43	+8.47	+0.99	+4.40	+10.02	+1.49	+6.33	+12.30	+1.86	+2.87	+7.39	+0.66
NUS-WIDE-81																						
Fed-DualCoOp	NeurIPS'22	51.02	45.89	69.00	49.97	44.36	65.30	46.25	41.62	67.02	42.69	37.11	66.10	40.65	35.69	60.28	39.40	34.54	59.04	45.00	39.87	64.46
Fed-SCPNet	CVPR'23	49.56	42.11	65.63	47.61	39.98	59.46	43.31	40.92	62.04	39.28	36.51	60.45	38.61	39.98	58.46	38.49	39.02	53.23	42.81	39.75	59.88
Fed-MaPLe	CVPR'23	54.78	46.69	72.32	53.68	42.13	71.12	51.34	40.69	73.32	51.62	43.40	72.53	50.98	43.68	72.70	51.20	40.99	70.59	52.27	42.93	72.10
FedPGP	ICML'24	55.92	46.48	75.91	53.41	43.27	75.28	53.78	43.07	72.56	49.64	44.82	72.79	48.45	42.58	72.47	43.20	42.72	71.73	50.73	43.82	73.46
Fed-TCP	CVPR'24	54.46	48.23	70.42	51.93	46.57	70.20	50.91	45.65	69.96	48.28	42.47	65.44	47.20	41.75	66.80	48.89	40.50	63.17	50.28	44.20	67.67
FedTPG	ICLR'24	53.01	50.69	73.32	52.78	47.66	72.50	50.00	46.72	69.52	50.40	44.21	70.47	49.92	43.89	67.10	49.46	42.55	65.46	50.93	45.95	69.73
Fed-PosCoOp	WACV'25	52.88	47.96	70.38	50.64	45.89	70.20	48.77	42.69	68.66	42.39	39.15	68.04	40.62	35.12	64.33	40.08	32.76	60.11	45.90	40.60	66.95
FedAWA	CVPR'25	54.67	47.61	74.51	53.91	45.52	70.33	50.40	43.74	71.60	52.87	43.22	71.51	52.75	43.02	66.89	50.71	41.13	66.53	52.55	44.04	70.23
Fed-RAM	CVPR'25	56.35	48.89	76.07	54.05	47.89	76.40	53.51	45.69	74.33	52.25	40.69	74.38	52.61	42.12	72.98	51.20	40.69	70.35	53.33	44.33	74.09
FedMVP	ICCV'25	53.20	48.37	77.80	53.32	46.39	77.10	51.90	45.21	75.14	52.98	43.43	75.02	51.33	42.79	74.55	51.08	40.16	72.90	52.30	44.39	75.42
FedMPT	CVPR'26	58.36	51.56	78.98	57.27	49.29	77.93	57.03	48.14	76.73	56.52	46.88	76.94	56.08	45.73	76.60	54.87	45.03	76.81	56.69	47.77	77.33
FedHarmony	CVPR'26	51.45	40.52	71.50	59.33	50.27	80.14	58.12	48.49	79.91	56.73	46.82	79.57	55.35	45.53	79.10	53.90	43.65	78.68	55.81	45.88	78.15
PathwayFed	Ours	60.12	57.81	80.05	60.40	58.03	80.28	60.40	57.74	80.35	60.19	57.38	80.30	60.10	57.18	80.23	59.95	57.12	80.18	60.19	57.54	80.23
Δ vs. full prior best	–	+1.76	+6.25	+1.07	+1.07	+8.74	+0.14	+2.28	+9.60	+0.44	+3.67	+10.50	+0.73	+4.02	+11.45	+1.13	+5.08	+12.09	+1.50	+3.50	+9.77	+2.90

Table 9: Endpoint gains of full matched routing over uniform routing under the current conflict-projection configuration. Values are mean ± standard deviation across seeds 34–36. Negative gains at low K indicate that pathway-selective routing is not beneficial before client knowledge becomes sufficiently specialized.

Dataset	K	Δ mAP	Δ CF1	Δ OF1
COCO	8	-0.89 ± 0.16	-0.08 ± 0.68	-0.51 ± 0.14
COCO	80	$+2.63 \pm 0.80$	$+4.61 \pm 0.24$	$+0.66 \pm 0.10$
NUS-WIDE	8	-0.21 ± 0.17	-0.09 ± 0.25	-0.14 ± 0.05
NUS-WIDE	81	$+3.02 \pm 0.14$	$+7.41 \pm 0.21$	$+0.51 \pm 0.02$

dence about the pathway structure produced by PathwayFed. It does not measure semantic independence or monosemanticity of individual feature units. Unlike the cross-client pathway disagreement in Appendix G, which compares pathway usage across client distributions, this diagnostic measures within-client pathway concentration. For final global checkpoints, we evaluate comparable $512 \rightarrow 1024$ ReLU heads on identical fixed client-local COCO K=80 partitions. For every valid class, pathway contribution is the positive-sample mean $|W_{o,i}[c, f]h_{i,f}(x)|$, normalized over valid classes. Metrics use all active units with mean ReLU activation above 10^{-6}. Displayed heatmap units are selected separately and do not enter quantitative selection. PathwayFed also yields a sparser active-unit structure. Importantly, the pathway-concentration differences in Table 11 are computed only among active units, so they are not obtained by averaging in inactive units. Dead units are reported separately as a sparsity diagnostic. The representative client and its top-20 classes are selected exclusively from the partition via median client-to-global label JSD and local positive support, respectively. These statistics characterize less diffuse unit-to-class pathway structure, not semantically independent feature units.

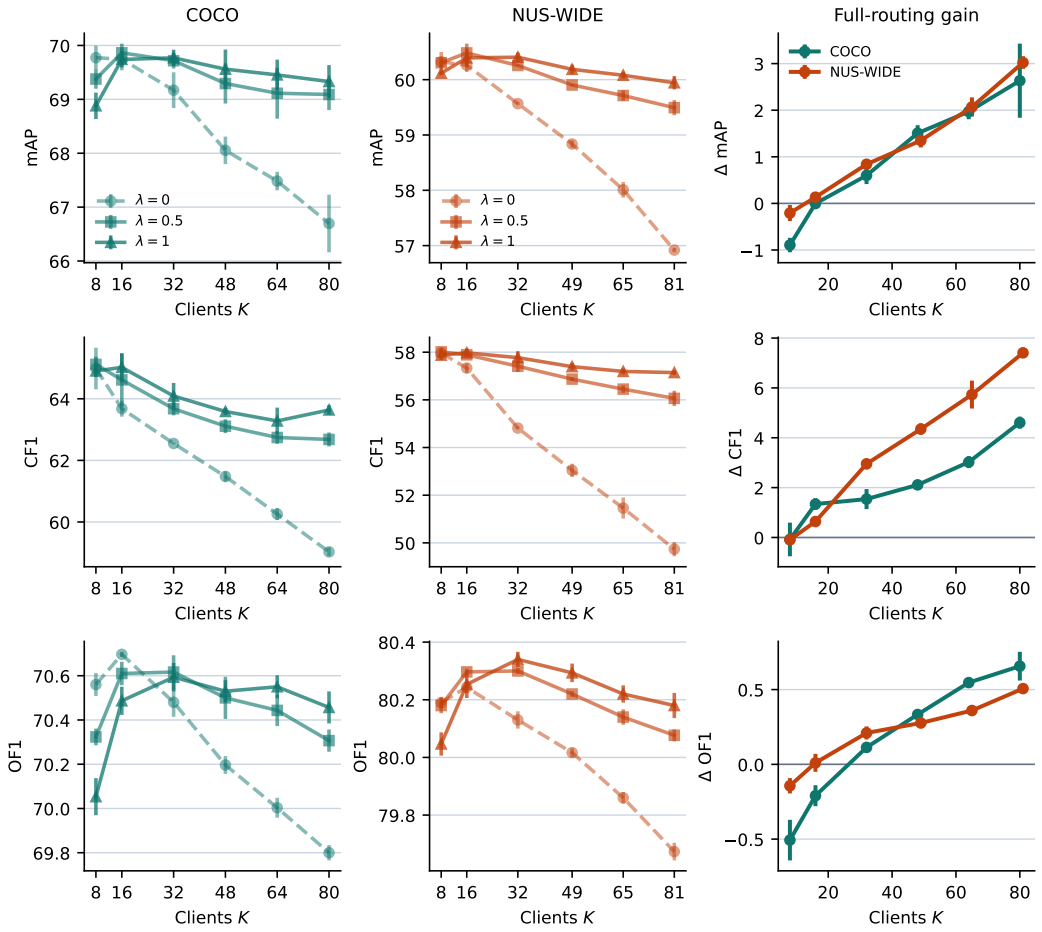


Figure 5: Full three-seed routing-strength controls under the clean pipeline (mean ± standard deviation). Rows report mAP, CF1, and OF1. Columns report COCO, NUS-WIDE, and the gain of full matched routing over uniform routing. The main paper shows the mAP row.

Table 10: Three-seed server-side routed update-subspace diagnostic on COCO K=80. Values are mean ± standard deviation across seeds. These geometric update subspaces are not formal unit-to-class pathways.

Metric	Value
Effective rank	2.04 ± 0.01
Top-5 update energy	$98.18 \pm 0.02\%$
Client-specific top-5 energy	$98.19 \pm 0.02\%$
Pooled-global top-5 energy	$9.15 \pm 0.05\%$
Pairwise top-5 overlap	0.0059 ± 0.0001
Grassmann distance	3.403 ± 0.001

N High-Heterogeneity Route-Permutation Control

We further test whether routing requires the correct association between a client update and its pathway-specific weight. Under the maximally specialized COCO K=80 KMeans split, we independently permute the routing weights along the client axis for every routed coordinate. This preserves each coordinate's weight distribution, concentration, non-uniformity, and all local updates, but assigns weights to different originating clients. Table 12 reports the current conflict-projection configuration over the same three seeds. Destroying the association between each routing weight and the client update from which it was estimated reduces mAP by 2.67 points. This supports the

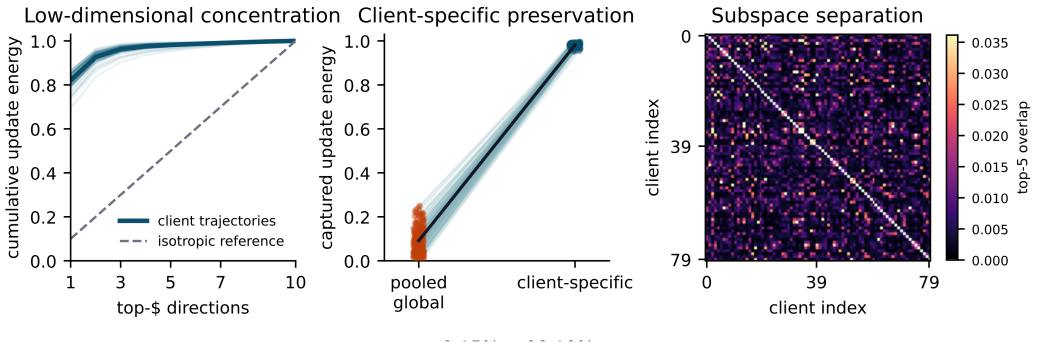


Figure 6: Complementary geometric analysis of client-specific routed update subspaces on COCO $K{=}80$. Left: cumulative update energy. Middle: client-specific versus pooled-global top-5 preservation. Right: pairwise top-5 subspace overlap. These diagnostics describe update geometry rather than semantic pathway independence.

Table 11: Unit-to-class pathway concentration on COCO $K{=}80$ (seeds 34–36). Dead units are a sparsity diagnostic, not a concentration metric.

Method	Entropy ↓	Top-1 ↑	Top-3 ↑	Effective classes ↓	Dead units
FedHarmony	0.826 ± 0.002	0.148 ± 0.002	0.312 ± 0.002	25.76 ± 0.10	$5.13 \pm 0.40\%$
PathwayFed	0.768 ± 0.005	0.193 ± 0.004	0.380 ± 0.006	21.65 ± 0.21	$18.45 \pm 1.77\%$

importance of preserving client–pathway associations and rules out an explanation based only on generic non-uniform reweighting.

O Additional Routing Diagnostics

Figure 7 reports the current conflict-projection diagnostic using server-side routing statistics on COCO $K{=}80$ (seed 34). Matched routing increases mean signed coordinate coherence from 0.182 under uniform weighting to 0.366, while the mean effective client fraction is 0.353 and the mean top-client share is 0.110. These single-seed values describe routing coherence and concentration rather than an additional optimization objective or a reliability estimate.

As shown in Figure 8, under the current fixed-capacity semantic-mixing intervention, stronger specialization ($\gamma : 0 \rightarrow 0.75$) reduces the mean effective client fraction from 0.525 to 0.432 and increases the mean top-client share from 0.074 to 0.092. Signed coherence remains higher for matched routing than uniform aggregation in both conditions (0.345 versus 0.174 for mixed clients and 0.353 versus 0.171 for specialized clients), consistent with matched routing concentrating weight on clients with more strongly expressed pathway-specific updates while increasing coordinate-level coherence. These are single-seed diagnostics and are interpreted descriptively.

P Failure-Mode Illustration

Figure 9 illustrates how coarse aggregation can mix or dilute client-specific pathway contributions over shared coordinates. FedAvg assigns each client one scalar weight across parameter directions. When pathway-specific knowledge signals are concentrated in only a subset of clients, it averages strong informative updates with weak or opposing contributions from clients carrying little pathway signal on the same coordinate. FedCALM can suppress conflicting components through conflict-free updates and layer-wise selection, but its decision remains coarser than the pathway-specific structure considered here.

Table 12: Route-permutation control at high heterogeneity on COCO $K=80$. Values are global metrics (mean $\pm$ standard deviation over seeds 34–36).

Method	mAP	CF1	OF1
Matched routing	69.18 ± 0.36	63.23 ± 0.31	70.43 ± 0.06
Permuted routing weights	66.51 ± 0.55	58.77 ± 0.16	69.72 ± 0.05

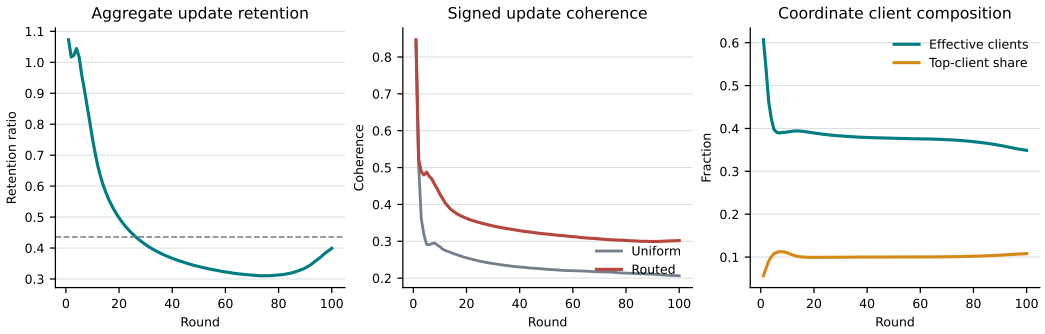


Figure 7: Per-round routing diagnostics from the current conflict-projection COCO $K=80$ run (seed 34). **Left:** matched routing improves signed coordinate coherence over uniform weighting. **Right:** the effective contributing-client fraction and top-client share quantify routing-weight concentration across rounds.

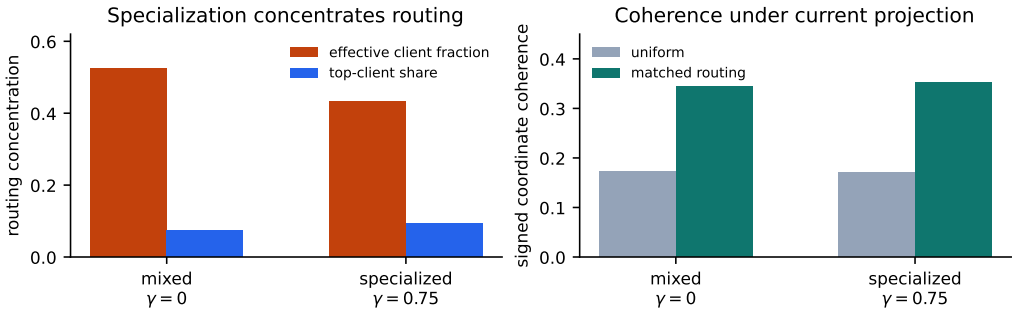


Figure 8: Current conflict-projection coordinate diagnostics for the COCO $K=80$ semantic-mixing control (seed 34). Stronger specialization concentrates routing on fewer clients, while matched routing remains more sign-coherent than uniform aggregation.

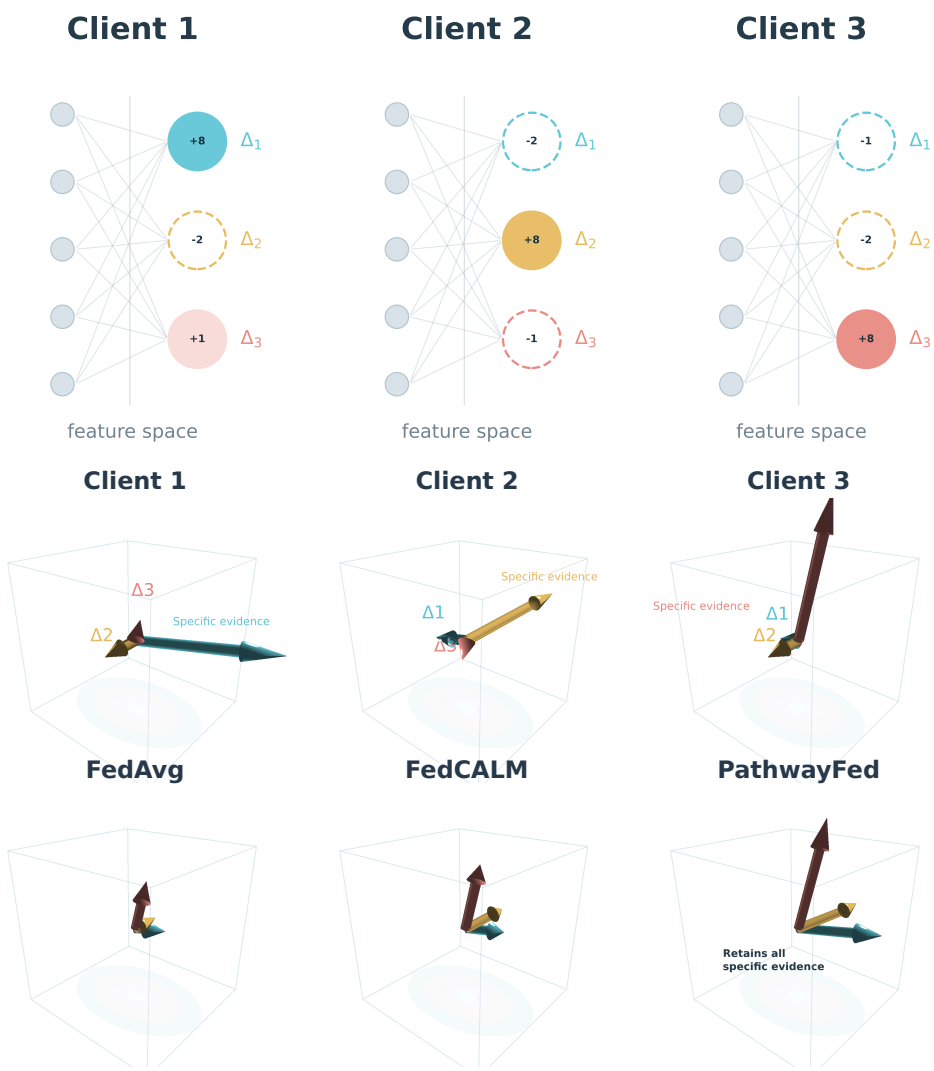


Figure 9: Preservation of pathway-specific local knowledge signal. **Top:** same-color second-layer units of local two-layer heads denote one shared coordinate Δ_j, and their values are ordinary local updates. Each client has strong pathway-specific local knowledge signal on a different coordinate, together with weak, partly opposing corrections elsewhere. **Bottom:** FedAvg dilutes the three local-knowledge-bearing directions through scalar averaging. FedCALM suppresses conflict through layer-wise mitigation, but leaves only weak, non-specialized coordinate updates. PathwayFed suppresses weak opposing corrections while preserving strong matched pathway-specific local knowledge signal on all three coordinates.

GENERATIVE AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used solely for limited language-related assistance, including translation, grammatical revision, and light stylistic refinement. They did not contribute to the conception of the research, formulation of hypotheses, design of experiments, development of the proposed method, analysis of data, interpretation of results, or creation of the substantive scientific content.