FEDCONSULT: TURNING CLIENT HETEROGENEITY INTO ON-DEMAND EXPERTISE FOR PERSONALIZED FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables decentralized clients to collaboratively train models without sharing private data. Existing methods combine collaborative knowledge sharing with local adaptation to improve predictions under statistical heterogeneity, yet personalization gains remain client-bound, leaving substantial performance gaps across clients on specific classes. To address these limitations, we propose FedConsult, a federated expertise-transfer framework that distills locally acquired strengths into reusable visual specialty adapters. These adapters let each client complement its predictions with others' expertise on a sample-by-sample basis, improving weak-class recognition while preserving local personalized strengths. Specifically, FedConsult identifies each client's regions of specialty competence from discrepancies between its global and personalized prediction paths, and distills visual knowledge with similar prediction-improvement patterns into specialty adapters in a shared feature space. Recipient clients subsequently use response-aware sparse routing to selectively invoke complementary expertise within their personalized pathways. Experiments on both natural and medical image classification tasks show that FedConsult outperforms the strongest baselines by up to 3.96 and 9.13 percentage points in overall and client-specific weak-class accuracy, respectively. Beyond viewing heterogeneity solely as an optimization challenge, FedConsult turns the resulting capability diversity into transferable expertise, closing local capability gaps and supporting more reliable personalized federated systems.

1 INTRODUCTION

Federated learning enables distributed clients to collaboratively develop predictive models without centralizing their training data (McMahan et al., 2017). In practice, clients often hold heterogeneous data collected from different populations and environments; for example, medical institutions differ in patient populations, disease distributions, and imaging conditions (Sheller et al., 2020). Such heterogeneity creates uneven learning opportunities, leading personalized models to develop different strengths and weaknesses across classes and input patterns. A pattern that one client fails to recognize may be well captured by another, even when the latter has lower overall predictive performance. These complementary capabilities present an opportunity for collaboration: clients could draw on expertise acquired elsewhere to address their own capability gaps. Realizing this opportunity requires making locally acquired strengths accessible beyond their source models and selectively using them to improve predictions at the receiving client.

Existing research in federated learning focuses on mitigating the challenges of statistical heterogeneity, with less emphasis on making the distinct capabilities developed under such heterogeneity reusable across clients. One line of work integrates locally learned knowledge into a common global model through parameter aggregation (McMahan et al., 2017) and optimization strategies designed to address heterogeneous client updates (Li et al., 2020; Karimireddy et al., 2020; Gao et al., 2022). However, consolidating diverse and potentially conflicting local knowledge into a single predictor may obscure the distinctive strengths of individual clients. Another line of work pursues personalization through local model components (Arivazhagan et al., 2019; Collins et al., 2021), adaptation objectives (Dinh et al., 2020; Li et al., 2021a), or client relationships (Smith et al., 2017; Ghosh et al., 2020). These methods improve adaptation to local distributions, but the resulting capabilities remain closely

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.

coupled to source-specific representations and decision rules. Knowledge that is effective locally may therefore offer limited benefit when directly incorporated into a different recipient's predictor. Although these approaches advance shared learning and local adaptation, they do not explicitly establish how locally acquired strengths can be extracted for independent reuse and selectively applied across heterogeneous clients. This gap motivates examining both how expertise is represented for transfer and how its suitability is determined at the receiving client.

Turning this opportunity into effective collaboration requires addressing two connected challenges. The first is to identify and separate locally acquired expertise. Expert combination offers a way to exploit distinct client capabilities (Xie et al., 2025), but treating a source model as a single expert can leave different prediction-correction patterns coupled. Moreover, strong local performance alone does not distinguish the capabilities gained through personalization from those already available in the shared model. Improvements over a common reference provide a more direct basis for identifying these gains, while decomposing their correction patterns can make distinct specialties independently reusable. The second challenge is to determine when each specialty is useful to a recipient. Query-dependent source selection makes collaboration responsive to the input (Wang et al., 2026), yet source relevance alone does not establish the benefit of incorporating its knowledge into another client's personalized predictor. A correction that reduces prediction loss at its source may have a different effect under the recipient's representations and decision rules.

To address these limitations, we propose FedConsult, a personalized federated learning framework that transforms locally demonstrated strengths into reusable specialty experts. FedConsult identifies specialty regions through reductions in predictive loss achieved by a personalized model relative to a global reference, grounding expertise in the observed benefit of local adaptation. It groups these improvements by prediction-correction direction and distills each group into a compact visual adapter that operates on shared intermediate features. This common interface makes distinct specialties available across clients while keeping personalized prediction components local. To account for recipient-dependent utility, each client evaluates the adapters through its complete personalized predictor and learns a sparse router from their resulting prediction changes. The original personalized prediction remains an explicit routing option, allowing consultation to combine relevant expertise while retaining the option to rely on the local model. Experiments on natural and medical image classification benchmarks show improvements in both overall accuracy and client-specific weak-class accuracy, with further analyses supporting the contribution of complementary expertise transferred across clients.

The main contributions are summarized as follows:

- We propose FedConsult, a personalized federated learning framework that makes locally acquired expertise reusable across clients through compact visual specialty adapters, enabling sample-dependent collaboration within each recipient's personalized predictor.
- We develop an expertise-transfer mechanism that decomposes demonstrated personalization gains into distinct correction patterns and evaluates their utility through recipient-side predictive responses. Response-aware sparse routing combines suitable specialties with an explicit option to retain the original prediction.
- Extensive experimental results demonstrate that FedConsult improves overall and client-specific weak-class accuracy across natural and medical image classification benchmarks under heterogeneous client distributions. Ablation and source-attribution analyses further support the effectiveness of specialty decomposition and selective cross-client transfer.

2 Related Work

Aggregation algorithms. Aggregation algorithms consolidate knowledge learned from clients' local data into a single global model. Representative approaches average client updates (McMahan et al., 2017), control local drift (Li et al., 2020; Karimireddy et al., 2020; Gao et al., 2022), supplement training with generated data (Ye et al., 2024), mitigate update conflicts (Zheng et al., 2025a;b), or adapt aggregation to distribution profiles (Li et al., 2026). Given distributional differences across clients, integrating diverse and potentially conflicting local knowledge into a shared model may make it difficult to retain each client's distinctive strengths. FedConsult extends the server's role beyond aggregation: it also maintains a specialty expert bank that stores locally acquired specialties as separate

adapters and distributes them across clients. This preserves distinct specialties as independently reusable units alongside the shared model.

Personalized federated learning. Personalized FL adapts client models to local distributions through private components (Arivazhagan et al., 2019; Collins et al., 2021; Li et al., 2021b), regularized objectives (Dinh et al., 2020; Li et al., 2021a), or meta-learning (Fallah et al., 2020). Further work refines shared–private parameters or representations (Zhu et al., 2024; Tamirisa et al., 2024; Jiang et al., 2024; Yang et al., 2024), addresses distribution shift or provides client-specific guidance (Tang et al., 2024; Scott et al., 2024; Liu et al., 2025), and coordinates classifiers or query-conditioned transfer (Xu et al., 2023; Alballa et al., 2025). However, learned expertise often depends on the source client's distribution, representations, and decision rules. Knowledge effective locally may therefore provide limited benefit to heterogeneous recipients, restricting reuse beyond its source model. Recent approaches seek to broaden access to client capabilities: FedDRM routes queries to suitable client models (Wang et al., 2026), while dFLMoE combines local and received expert heads (Xie et al., 2025). Their collaboration units remain client models or expert heads, without explicitly separating distinct local correction patterns into dedicated, independently reusable carriers. FedConsult distills these patterns into fine-grained visual specialty adapters in a shared feature space. By evaluating their responses through each recipient's personalized predictor, it enables clients to selectively draw on others' expertise to complement their own predictions.

3 METHOD

3.1 FRAMEWORK OVERVIEW

FedConsult converts predictive gains from personalized federated training into visual specialty adapters that clients can selectively reuse. Each adapter encodes a distinct pattern of local predictive improvement as a compact transformation of shared intermediate features, allowing its knowledge to integrate into a recipient's personalized prediction path. This interface connects source-side specialty learning with recipient-side consultation while keeping the trained global and personalized model components fixed.

To construct these adapters, FedConsult identifies samples where personalization reduces predictive loss relative to the global reference and groups them by the direction of their prediction corrections. Each group supervises an adapter through classification and distillation objectives, and the server collects the resulting adapters into a federated specialty expert bank. To determine which specialties benefit a given input, each recipient evaluates the candidate adapters through its complete personalized predictor. A locally trained, response-aware sparse router uses the resulting prediction changes to combine relevant candidate responses, while retaining the original personalized prediction as an explicit routing option. The overall workflow is illustrated in Figure 1.

3.2 CLIENT-SPECIFIC SPECIALTY REGION IDENTIFICATION

We consider a central server and K clients, where client i holds $\mathcal{D}_i = \{(x_n, y_n)\}_{n=1}^{N_i}$ over C classes. Following DualFed (Zhu et al., 2024), Stage 1 learns a global–personalized predictor: the global path provides a federation-wide reference, while the personalized path captures client-specific adaptation. Let $A(x) \in \mathbb{R}^{D\times H\times W}$ be a shared visual map, f_θ the global tail, h_g the global head, and (r_i, h_i) client i's personalized projector and head. Their logits are

$$z(x) = f_\theta(A(x)), \quad \ell_i^{\mathrm{glo}}(x) = h_g(z(x)), \quad \ell_i^{\mathrm{per}}(x) = \ell_i^{\mathrm{glo}}(x) + h_i(r_i(z(x))). \tag{1}$$

Both paths use classification objectives, keeping the global reference independently predictive. The global representation and head are aggregated by client sample size, whereas (r_i, h_i) remain local. After convergence, all base-model parameters are frozen, fixing both the reference and each client's personalized predictor.

For a source sample $(x, y) \in \mathcal{D}_i$, FedConsult measures the benefit of local personalization over the common reference using cross-entropy $\mathcal{L}_{\mathrm{CE}}$:

$$u_i(x, y) = \mathcal{L}_{\mathrm{CE}}(\ell_i^{\mathrm{glo}}(x), y) - \mathcal{L}_{\mathrm{CE}}(\ell_i^{\mathrm{per}}(x), y), \qquad \mathcal{S}_i = \{(x, y) \in \mathcal{D}_i : u_i(x, y) > 0\}. \tag{2}$$

Thus, $\mathcal{S}_i$ is client i's empirical specialty region: the samples for which personalization increases the ground-truth likelihood.

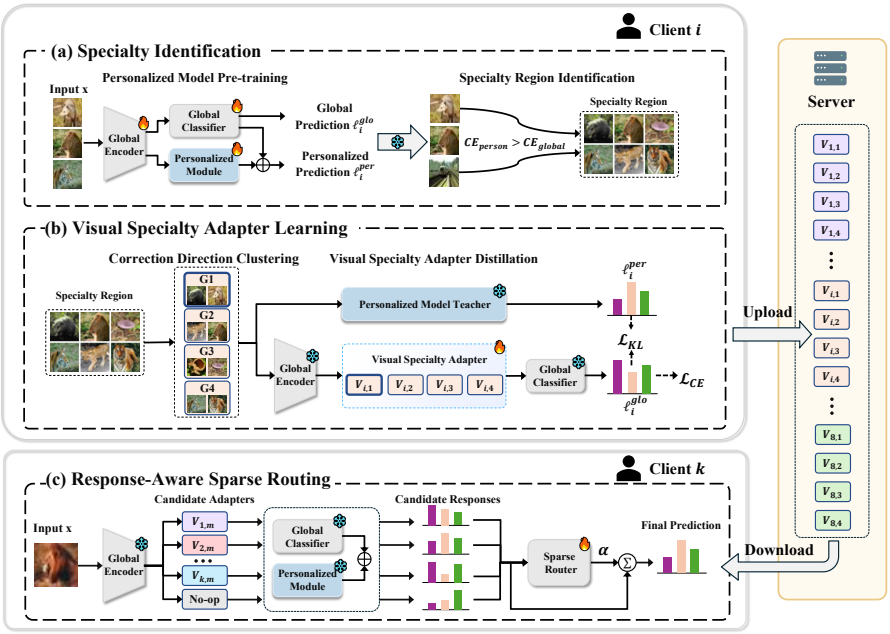


Figure 1: The workflow of FedConsult.

3.3 Visual Specialty Adapter Learning

A specialty region may contain distinct or conflicting correction patterns. FedConsult separates them by prediction-improvement direction. For $(x, y) \in \mathcal{S}_i$, let $\delta_i(x) = \ell_i^{\mathrm{per}}(x) - \ell_i^{\mathrm{glo}}(x)$ and $H_C = I_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top$. The normalized correction direction is

$$d_i(x) = \frac{H_C\delta_i(x)}{\|H_C\delta_i(x)\|_2}, \tag{3}$$

where I_C is the identity and $\mathbf{1}_C$ the all-ones vector; centering removes class-independent shifts. With $w_i(x, y) = \max(u_i(x, y), \epsilon)$, $\epsilon > 0$, weighted k-means obtains assignments $a_i(x) \in \{1, \ldots, M\}$ and centers $\mu_{i,m} \in \mathbb{R}^C$:

$$\min_{a_i, \{\mu_{i,m}\}} \sum_{(x,y)\in\mathcal{S}_i} w_i(x, y) \left\| d_i(x) - \mu_{i,a_i(x)} \right\|_2^2. \tag{4}$$

Utility weighting prioritizes stronger demonstrated improvements. We use $M = 4$ to preserve multiple correction roles with a compact expert set.

Each group is encoded by a channel-wise FiLM adapter (Perez et al., 2018) at the shared Layer3 visual map. For $\gamma_{i,m}, \beta_{i,m} \in \mathbb{R}^D$ and modulation strength $\eta > 0$,

$$V_{i,m}(A(x)) \equiv \widetilde{A}_{i,m}(x) = \big(1 + \eta \tanh \gamma_{i,m}\big) \odot A(x) + \eta \tanh \beta_{i,m}, \tag{5}$$

where $\odot$ is channel-wise multiplication, scale and bias are broadcast spatially, and $\tanh$ bounds the modulation. Operating in the shared feature space makes the learned visual correction interpretable by each recipient's downstream personalized path. Only $(\gamma_{i,m}, \beta_{i,m})$ are trainable. For samples $\mathcal{S}_{i,m}$ assigned to mode m, define $\ell_{i,m}^V(x) = h_g(f_\theta(V_{i,m}(A(x))))$. We define $q_i^{\mathrm{per}}(x) = \sigma(\ell_i^{\mathrm{per}}(x)/T)$ and $q_{i,m}^V(x) = \sigma(\ell_{i,m}^V(x)/T)$, where σ is softmax and $T > 0$ the temperature. The adapter minimizes

$$\mathcal{L}_V^{i,m} = \mathbb{E}_{(x,y)\in\mathcal{S}_{i,m}} \Big[\mathcal{L}_{\mathrm{CE}}(\ell_{i,m}^V(x), y) + \lambda_{\mathrm{KD}} T^2 \,\mathrm{KL}\big(q_i^{\mathrm{per}}(x) \,\|\, q_{i,m}^V(x)\big) \Big], \tag{6}$$

where KL is Kullback–Leibler divergence and λ_{KD} controls distillation. CE anchors the class target, while the soft personalized response is distilled into a reusable transformation of the shared visual map. Each client uploads its M adapters, and the server indexes every source–mode pair as an expert in the federated bank $\mathcal{V}$.

3.4 Response-Aware Sparse Routing

Adapter utility depends on both the recipient and sample. Target client t therefore evaluates local and external experts through its frozen personalized predictor, with $p_t^{\mathrm{per}}(x) = \sigma(\ell_t^{\mathrm{per}}(x))$ as a no-op candidate. For expert $e = (i, m)$, let $z_{t,e}(x) = f_\theta(V_{i,m}(A(x)))$. Its target-side response is

$$p_{t,e}(x) = \sigma(h_g(z_{t,e}(x)) + h_t(r_t(z_{t,e}(x)))) \,. \tag{7}$$

A scorer shared across experts combines context $c_t(x) = g_{\mathrm{ctx}}(z(x))$, the current belief, and response change $\Delta p_{t,e}(x) = p_{t,e}(x) - p_t^{\mathrm{per}}(x)$: $s_{t,e}(x) = g_{\mathrm{exp}}([c_t(x); p_t^{\mathrm{per}}(x); \Delta p_{t,e}(x)])$. The no-op score is $s_{t,\mathrm{noop}}(x) = g_{\mathrm{noop}}([c_t(x); p_t^{\mathrm{per}}(x)])$, where $[\,;\,]$ denotes concatenation. Sparse consultation then gives

$$\begin{aligned} \boldsymbol{\alpha}_t(x) &= \mathrm{sparsemax}([s_{t,\mathrm{noop}}(x), \{s_{t,e}(x)\}_{e\in\mathcal{V}}]/\tau_r) \,, \\ p_t^\star(x) &= \alpha_{t,\mathrm{noop}}(x) p_t^{\mathrm{per}}(x) + \sum_{e\in\mathcal{V}} \alpha_{t,e}(x) p_{t,e}(x), \end{aligned} \tag{8}$$

where $\tau_r > 0$ is the routing temperature. The shared scorer keeps router size independent of the bank size and is implemented with lightweight MLPs, layer normalization, and GELU. Sparsemax assigns zero weight to irrelevant candidates, while the no-op retains the original personalized prediction.

With the personalized model and expert bank frozen, the target-local router minimizes

$$\mathcal{L}_R^t = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\Big[-\log[p_t^\star(x)]_y + \lambda_r b_t(x,y)\big(1 - \alpha_{t,\mathrm{noop}}(x)\big)\Big], \tag{9}$$

Here $[p_t^\star]_y$ is the target-class probability and $b_t(x, y)$ equals one only when the personalized prediction is correct but every expert response is wrong. Thus, $\lambda_r \geq 0$ discourages harmful consultation in precisely these cases. The router remains local, allowing each recipient to consult the shared bank according to its own capability gaps.

4 Experiments

4.1 Experimental Settings

Baselines and Datasets. Eleven mainstream federated learning methods are selected for comparison, including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), FedSAM (Qu et al., 2022), FedPAC (Xu et al., 2023), FedALA (Zhang et al., 2023), FedAS (Yang et al., 2024), DualFed (Zhu et al., 2024), ConFree (Zheng et al., 2025a), cwFedAvg (Lee & Choi, 2025), FedFew (Guo et al., 2026), and FedDRM (Wang et al., 2026). We conduct experiments on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), each containing 60,000 32×32 RGB images. CIFAR-10 comprises 10 classes with 6,000 images per class, whereas CIFAR-100 comprises 100 fine-grained classes with 600 images per class. Following the benchmark protocol of FedDRM (Wang et al., 2026), we combine the original training and test sets before constructing an eight-client federation. Each resulting client partition is then divided into local training and test sets with a 70/30 ratio.

Non-IID settings. Following FedDRM (Wang et al., 2026), we consider two complementary forms of label heterogeneity. For the Dirichlet setting, class proportions are sampled independently for each client from a Dirichlet distribution with concentration $\alpha = 0.3$ (Yurochkin et al., 2019), producing both label imbalance and unequal client sizes. For the pathological setting, we sort samples by class, divide them into label-homogeneous shards, and assign a fixed number of shards to every client (McMahan et al., 2017). We use five shards per client (5-SPC) on CIFAR-10 and 25 shards per client (25-SPC) on CIFAR-100. We further introduce client-level covariate shift using the FedDRM protocol (Wang et al., 2026): each client receives one distinct combination of gamma correction, hue adjustment, and saturation scaling, with factors drawn from $\{0.6, 1.4\}$, $\{-0.1, 0.1\}$, and $\{0.5, 1.5\}$, respectively.

Implementation details. All methods use an ImageNet-pretrained ResNet-18 backbone (Deng et al., 2009; He et al., 2016) and are trained for 800 communication rounds with 10 local update steps per round and a batch size of 128. All eight clients participate in every round. For FedConsult, the global–personalized predictor is optimized with AdamW using an initial learning rate of 10^{-3}, cosine decay, and weight decay 10^{-4}. After personalized training, each client learns four visual specialty

Table 1: Overall and client-specific weak-class accuracy (%) on CIFAR benchmarks.

Method	CIFAR-10				CIFAR-100			
	Dir-0.3		5-SPC		Dir-0.3		25-SPC	
	Acc	Weak	Acc	Weak	Acc	Weak	Acc	Weak
FedAvg	54.90	42.95	65.55	52.34	41.50	14.66	36.72	15.79
FedProx	54.48	49.52	66.22	52.65	41.40	13.20	37.01	15.85
FedSAM	61.66	49.58	69.91	54.19	43.72	15.18	40.11	18.78
FedALA	88.41	58.04	85.06	79.22	54.29	16.79	66.57	48.28
FedPAC	84.13	31.67	82.42	71.64	51.91	3.64	58.73	33.77
FedAS	85.40	36.61	83.55	73.61	53.21	5.04	60.21	35.63
DualFed	89.36	61.05	87.36	80.07	55.40	16.61	63.31	40.18
ConFree	84.25	25.76	83.47	73.09	50.96	5.36	59.81	35.20
cwFedAvg	85.38	61.80	83.41	77.36	50.52	19.67	63.11	44.53
FedFew	87.13	54.65	83.32	76.39	54.79	4.73	63.39	44.09
FedDRM	86.74	59.05	83.88	73.85	56.69	17.12	60.83	36.23
FedConsult (Ours)	**90.47**	**65.49**	**88.17**	**84.07**	**60.65**	**28.78**	**67.61**	**51.47**

adapters for 20 epochs, forming a 32-expert federation-wide pool; the recipient-specific sparse router is trained for 50 epochs. Adapter and router learning rates are 5×10^{-3} and 10^{-3}, respectively. Method-specific hyperparameters are tuned to achieve their best performance.

4.2 Experimental Results on Public Datasets

To evaluate predictive performance on local capability gaps under heterogeneous federated learning, we introduce client-specific weak-class accuracy (Weak). For each client, we rank the classes represented in its local training data by their test accuracy and select the bottom quartile. We calculate Weak by dividing the total number of correct test predictions by the total number of test samples across the selected client-class pairs, measuring performance on each client's least accurately recognized classes. We also report sample-weighted test accuracy (Acc) to evaluate overall predictive performance across clients.

As shown in Table 1, FedConsult outperforms all baselines in both Acc and Weak across the four heterogeneous settings. Compared with the best-performing baseline for each metric and setting, it improves Acc by 0.81 to 3.96 percentage points and Weak by 3.19 to 9.11 percentage points. Relative to the DualFed baseline, the Weak gains are particularly pronounced on CIFAR-100, reaching 12.17 and 11.29 percentage points under Dir-0.3 and 25-SPC, respectively. These improvements show that the benefits extend to locally weak classes alongside gains in overall accuracy, supporting the effectiveness of the proposed specialty transfer framework in addressing uneven predictive performance across local classes.

4.3 Experimental Results on Real Medical Datasets

We further evaluate FedConsult on three medical image classification tasks. Retina combines 2,390 fundus images from ACRIMA (Diaz-Pinto et al., 2019), RIM-ONE (Batista et al., 2020), and REFUGE (Orlando et al., 2020) to evaluate binary glaucoma recognition across distinct acquisition domains. Skin14 merges HAM10000 (Tschandl et al., 2018) and MSLDv2.0 (Ali et al., 2024) into a 36,656-image benchmark spanning 14 diagnostic categories for multi-class lesion classification. BreaKHis (Spanhol et al., 2015) comprises 7,909 breast histopathology images from 82 patients, spanning four benign and four malignant tumor subtypes, and is used for binary benign-versus-malignant classification.

The three source domains of Retina naturally form a three-client federation, with each domain treated as a center-specific client. For Skin14 and BreaKHis, we follow the CIFAR protocol with eight clients and class-wise Dirichlet partitioning at $\alpha = 0.3$. Each client uses a 70/30 local train–test split. Retina is trained for 100 communication rounds, whereas Skin14 and BreaKHis use 800 rounds. The batch size is 32 for all three datasets, and the remaining method-specific settings follow the benchmark experiments.

Table 2: Overall and client-specific weak-class accuracy (%) on medical datasets.

	Retina		Skin14		BreaKHis	
Method	Acc	Weak	Acc	Weak	Acc	Weak
FedAvg	89.69	85.49	73.59	41.77	90.06	74.92
FedProx	89.55	86.94	74.04	41.41	85.81	59.28
FedSAM	91.92	85.82	73.07	25.79	84.46	54.07
FedALA	88.02	83.58	82.99	29.33	97.51	79.00
FedPAC	89.69	80.97	82.22	28.59	97.19	35.94
FedAS	90.67	80.54	83.17	37.09	97.27	39.06
DualFed	93.71	90.69	82.57	29.97	97.39	86.13
ConFree	90.81	83.56	82.31	33.57	97.06	31.25
cwFedAvg	86.91	81.54	80.30	24.27	85.14	60.59
FedFew	91.09	86.19	79.58	17.54	97.56	53.13
FedDRM	91.78	86.58	83.55	38.40	97.36	42.19
FedConsult (Ours)	**95.13**	**93.28**	**84.01**	**44.07**	**97.85**	**95.26**

Table 3: Component ablation on CIFAR-100 Dir-0.3; checkmarks denote enabled components.

Configuration	Specialty Adapter	Specialty Region	Direction Clustering	Response-Aware Routing	No-op	Acc	Weak
Baseline						55.40	16.61
I	✓			✓		58.67	22.61
II	✓	✓		✓		59.19	24.39
III	✓	✓	✓	✓		60.33	28.08
IV	✓	✓	✓			57.08	20.44
V	✓	✓	✓	✓	✓	**60.65**	**28.78**

Table 2 shows that FedConsult consistently outperforms all baselines in both Acc and Weak across the three medical tasks. On Skin14 and BreaKHis, the overall improvements over the strongest competing methods are modest, while the weak-class advantages are substantially more pronounced. This result directly reflects the intended role of specialty transfer: filling local capability gaps that may be obscured by strong overall accuracy. By improving predictions for disease categories on which local models are least capable, FedConsult reduces uneven diagnostic performance and supports more consistent and reliable use of federated models in clinical settings.

4.4 Ablation Study

Table 3 examines the contributions of specialty construction and recipient-side routing through controlled variants. Generic Adapters learn visual adapters without identifying specialty regions or clustering directions. Adding region identification restricts adapter learning to samples with demonstrated personalization gains, while incorporating direction clustering separates these gains into distinct correction patterns. The resulting improvements in Acc and Weak support the value of both identifying useful local knowledge and decomposing it into complementary specialties.

Full w/o Response-Aware Router retains the specialty adapters but combines their predictions using a shared mixture. Its lower performance indicates that the benefit of these adapters depends on selecting them according to their effects within the recipient's predictor. The full model achieves the highest Acc and Weak by combining response-aware sparse routing with the No-op action, which retains the original personalized prediction as an explicit alternative to consultation. Together, these results support the joint contribution of specialty construction and selective consultation to overall and weak-class performance.

4.5 Visual Specialty Adapter Analysis

We assess visual specialty adapters in terms of transfer effectiveness, functional differentiation, and decomposition granularity. The carrier comparison in Table 4 evaluates classifier heads, complete personalized paths, and visual specialty adapters, transferring one unit from each external source client. Neither classifier heads nor complete personalized paths yield meaningful performance

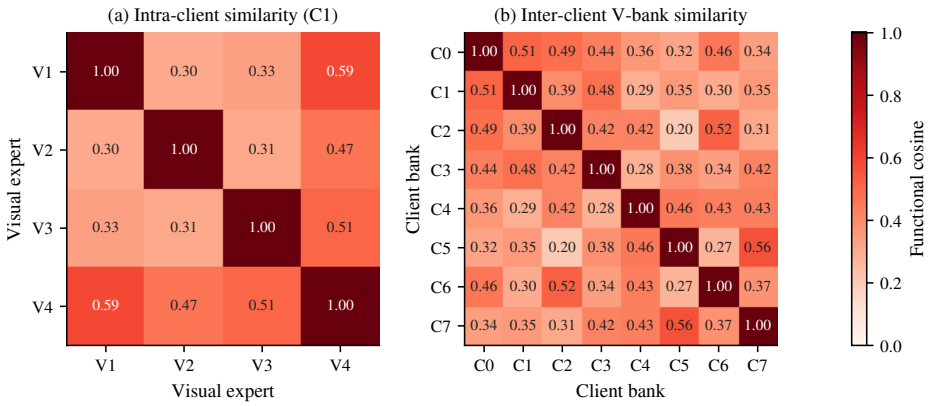


Figure 2: Functional similarity of visual specialty adapters within and across source adapter banks.

Table 4: Comparison of specialty carriers and visual-adapter granularity on CIFAR-100 Dir-0.3.

Protocol	Transfer unit	Candidates	Parameters/source	Acc	Weak
Carrier control	Classifier head	7	51,300	55.57	18.18
	Personalized path	7	315,748	55.45	17.86
	Visual specialty adapter	7	512	**58.95**	**26.76**
Granularity control	Client-level M1	8	512	59.32	24.93
	Pattern-level M2	16	1,024	60.13	**29.17**
	Pattern-level M4	32	2,048	**60.65**	28.78
	Pattern-level M8	64	4,096	60.51	27.22
	Class-level adapter	620	$512m_i$	59.42	27.54

gains, indicating that directly transferring these personalized components is insufficient to exploit cross-client expertise in the evaluated settings. This result is consistent with the dependence of their predictive capabilities on source-specific representations and decision mappings. In contrast, visual specialty adapters achieve effective transfer by encoding local expertise as transformations of shared intermediate features, which are processed by the recipient's own personalized predictor. These findings support the shared feature interface as an effective basis for reusing local expertise across heterogeneous clients, while also offering substantially lower parameter requirements. To examine functional differentiation, Figure 2 characterizes each adapter through a class-conditioned functional signature comprising its true-label margin, incorrect-to-correct responses, and sign-reversed correct-to-incorrect responses. The intra-client similarities in panel (a) indicate that adapters derived from different correction directions exhibit distinct functional profiles, providing evidence that the decomposition captures variation in predictive behavior within a source client. The inter-client similarities in panel (b) additionally reveal source-dependent structure in these profiles. Together, these results suggest that the adapter banks preserve functional diversity both within and across clients.

The granularity comparison in Table 4 further examines the relationship between specialty decomposition and transfer performance. Increasing the number of adapters per client from one to a moderate level improves predictive performance, supporting the utility of representing distinct local correction patterns as independently selectable specialties. Further increases yield no additional benefit in the evaluated settings, indicating that transfer performance does not improve monotonically with decomposition granularity. The M4 configuration achieves the best observed balance between predictive performance and adapter count, supporting a compact decomposition of local expertise.

4.6 Response-Aware Sparse Routing Analysis

We examine how response-aware sparse routing affects predictions at different levels of local predictive ability.

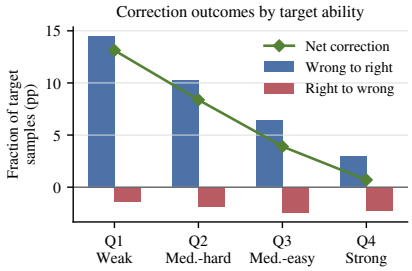


Figure 3: Prediction transitions for response-aware sparse routing.

Figure 3 groups client-class pairs by their original personalized accuracy and reports the corresponding prediction transitions. The weakest group exhibits the largest net gain from incorrect-to-correct transitions over correct-to-incorrect transitions. This gain decreases as local accuracy increases but remains positive in the strongest group, indicating that consultation primarily benefits weak classes while also providing a net improvement for classes with strong local performance.

4.7 Cross-Client Specialty Transfer

We examine whether cross-client specialties complement local weaknesses while supporting established strengths. Figure 4 groups positive leave-one-source-out contributions by source clients' class-specific accuracy ranks. Sources ranked 2 and 3 account for the largest combined share across all recipient ability groups, while both the highest-ranked and lower-ranked sources also contribute. These contributions extend to locally strong classes, suggesting that specialty consultation can address capability gaps and further improve predictions where recipients already perform well. The distribution also indicates that useful expertise is spread across clients and that source accuracy alone does not determine transfer value.

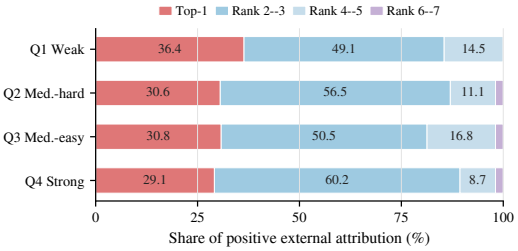


Figure 4: Cross-client specialty attribution by source ability rank.

Table 5: Overall and weak-class accuracy (%) under different candidate pools.

Candidate Pool	Acc	Weak
w/o adapters	55.40	16.61
Local4 only	58.38	23.66
External28 real	60.27	27.54
External28 random	57.39	22.93
External28 averaged	55.39	17.32
All32	**60.65**	**28.78**

Table 5 further distinguishes expertise transfer from capacity expansion. Local adapters improve upon the original personalized model, indicating that additional local adaptation contributes to performance. Learned external adapters nevertheless outperform the local-only, random, and averaged alternatives, while combining local and external adapters achieves the best performance. These results support the value of structured cross-client expertise beyond additional capacity and suggest that external consultation can complement.

5 Conclusion

In this paper, we propose FedConsult, a personalized federated learning framework that makes locally acquired predictive strengths available for selective reuse across heterogeneous clients. Its core design distills distinct patterns of personalization gains into compact visual specialty adapters and assesses their utility through each recipient's own predictor. This connects expertise transfer to its effect on individual predictions, enabling clients to combine complementary knowledge while retaining their original prediction as an explicit option. Across natural image classification settings and three medical imaging benchmarks, FedConsult improves both overall and client-specific weak-class accuracy, showing that collaboration benefits extend to locally underperforming classes. These results support the effectiveness of converting heterogeneous local capabilities into reusable expertise for personalized federated prediction.

AI USE STATEMENT

Generative AI tools were used to aid or polish writing, support research ideation or execution, and draft sections of the paper. The authors take full responsibility for the content of the manuscript, including any text generated or polished by these tools. We have ensured that AI-assisted text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human-subject or animal experimentation was conducted. All datasets used in this study, including CIFAR-10, CIFAR-100, ACRIMA, RIM-ONE, REFUGE, Skin14, and BreaKHis, were previously released and used in accordance with their access and usage guidelines. No personally identifiable information was used, and no experiment raised additional privacy or security concerns. We took care to avoid biased or discriminatory outcomes and remain committed to transparency and research integrity.

REPRODUCIBILITY STATEMENT

The main paper specifies the learning problem, model components, objectives, federated protocol, datasets, partitioning schemes, and evaluation metrics. Appendix A provides the training procedure, data processing steps, model and adapter configurations, optimization settings, hyperparameters, computational setup, and additional implementation details required to reproduce the experiments. All benchmark datasets are identified and cited in the paper so that the reported evaluation settings can be reconstructed.

REFERENCES

Norah Alballa, Wenxuan Zhang, Ziquan Liu, Ahmed Mohamed Abdelmoniem Sayed, Mohamed Elhoseiny, and Marco Canini. Query-based knowledge transfer for heterogeneous learning environments. In *International Conference on Learning Representations*, volume 2025, pp. 51458–51489, 2025.

Shams Nafisa Ali, Md Tazuddin Ahmed, Tasnim Jahan, Joydip Paul, SM Sakeef Sani, Nawsabah Noor, Anzirun Nahar Asma, and Taufiq Hasan. A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity. *Biomedical Signal Processing and Control*, 98:106742, 2024.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis and Stereology*, 39(3):161–167, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18(1):29, 2019.

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568, 2020.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10102–10111. IEEE, 2022.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19586–19597, 2020.

Ping Guo, Tiantian Zhang, Xi Lin, Xiang Li, Zhi-Ri Tang, and Qingfu Zhang. Few-for-many personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17515–17524, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Meirui Jiang, Anjie Le, Xiaoxiao Li, and Qi Dou. Heterogeneous personalized federated learning by local-global updates mixing via convergence rate. In *International Conference on Learning Representations*, volume 2024, pp. 57054–57082, 2024.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

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

Gyuejeong Lee and Daeyoung Choi. Class-wise federated averaging for efficient personalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Mohan Li, Dario Fenoglio, Martin Gjoreski, and Marc Langheinrich. Federated learning with profile mapping under distribution shifts and drifts. In *International Conference on Learning Representations*, 2026.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021a.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021b.

Qiqi Liu, Jiaqiang Li, Yuchen Liu, Yaochu Jin, Lingjuan Lyu, Xiaohu Wu, and Han Yu. Personalized federated learning under local supervision. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4069–4079. IEEE, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pp. 18250–18280. PMLR, 2022.

Jonathan Scott, Hossein Zakerinia, and Christoph Lampert. Pefll: Personalized federated learning by learning to learn. In *International conference on learning representations*, volume 2024, pp. 32406–32436, 2024.

Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):12598, 2020.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7): 1455–1462, 2015.

Rishub Tamirisa, Chulin Xie, Wenxuan Bao, Andy Zhou, Ron Arel, and Aviv Shamsian. Fedselect: Personalized federated learning with customized selection of parameters for fine-tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23985–23994. IEEE, 2024.

Xueyang Tang, Song Guo, Jie Zhang, and Jingcai Guo. Learning personalized causally invariant representations for heterogeneous federated clients. In *International Conference on Learning Representations*, volume 2024, pp. 10016–10037, 2024.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Zijian Wang, Xiaofei Zhang, Xin Zhang, Yukun Liu, and Qiong Zhang. Beyond aggregation: Guiding clients in heterogeneous federated learning. In *International Conference on Learning Representations*, volume 2026, pp. 149981–150004, 2026.

Luyuan Xie, Tianyu Luan, Wenyuan Cai, Guochen Yan, Zhaoyu Chen, Nan Xi, Yuejian Fang, Qingni Shen, Zhonghai Wu, and Junsong Yuan. dflmoe: Decentralized federated learning via mixture of experts for medical data analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10203–10213. IEEE, 2025.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *International Conference on Learning Representations*, 2023.

Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11986–11995. IEEE, 2024.

Rui Ye, Yaxin Du, Zhenyang Ni, Yanfeng Wang, and Siheng Chen. Fake it till make it: Federated learning with consensus-oriented generation. In *International Conference on Learning Representations*, volume 2024, pp. 7058–7083, 2024.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pp. 7252–7261. PMLR, 2019.

Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. FedALA: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11237–11244, 2023.

Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. ConFREE: Conflict-free client update aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22875–22883, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15444–15453. IEEE, 2025b.

Guogang Zhu, Xuefeng Liu, Jianwei Niu, Shaojie Tang, Xinghao Wu, and Jiayuan Zhang. Dualfed: enjoying both generalization and personalization in federated learning via hierachical representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11060–11069, 2024.

A Reproducibility Details

This appendix specifies the complete training pipeline, parameter lifecycle, model configuration, data construction, and optimization settings used in our experiments. The implementation follows the three components described in Section 3: client-specific specialty region identification, visual specialty adapter learning, and response-aware sparse routing.

A.1 End-to-End Training Algorithm

Algorithm 1 summarizes the complete FedConsult procedure. Stage 1 first establishes a stable global–personalized predictor for every client. The resulting model is then frozen before specialty discovery, adapter learning, and recipient-side router training. This separation ensures that the later stages transfer already-formed local strengths rather than jointly changing the underlying personalized predictors.

Algorithm 1 End-to-end training procedure of FedConsult

Require: Client datasets $\{\mathcal{D}_i\}_{i=1}^{K}$; adapters per client M; specialty bank $\mathcal{V} \leftarrow \varnothing$
Stage 1: global–personalized federated training
for each communication round **do**
  The server broadcasts the current global parameters.
  for each participating client i in parallel **do**
    Update the global and personalized paths on $\mathcal{D}_i$.
    Upload only the updated global parameters; retain (r_i, h_i) locally.
  end for
  Aggregate the uploaded global parameters by client sample size.
end for
Freeze the global parameters and every client's personalized parameters.
Stage 2: specialty discovery and adapter learning
for each source client i in parallel **do**
  Compute $u_i(x, y)$ and form $\mathcal{S}_i$ using Equation (2).
  Compute correction directions $d_i(x)$ and obtain M weighted direction groups using Equation (4).
  for each group $m \in \{1, \ldots, M\}$ **do**
    Initialize $V_{i,m}$ and update only $(\gamma_{i,m}, \beta_{i,m})$ using Equation (6).
  end for
  Upload $\{V_{i,m}\}_{m=1}^{M}$ to the server.
end for
The server indexes the uploaded adapters to form $\mathcal{V}$ and synchronizes the bank to all clients.
Stage 3: recipient-side response-aware routing
for each recipient client t in parallel **do**
  Freeze its base predictor and all adapters in $\mathcal{V}$.
  for each local router-training sample x **do**
    Evaluate the no-op response $p_t^{\mathrm{per}}(x)$ and every candidate response $\{p_{t,e}(x) : e \in \mathcal{V}\}$.
    Compute response-aware scores and Sparsemax weights $\boldsymbol{\alpha}_t(x)$ using Equation (8).
    Update only the local context encoder and scoring networks using Equation (9).
  end for
  Retain the trained router locally.
end for
return Frozen personalized predictors, synchronized specialty bank $\mathcal{V}$, and local routers.

Table 6 makes explicit which parameters are optimized, frozen, communicated, or retained locally in each stage.

A.2 Architecture and Trainable Components

Base predictor. The global–personalized predictor follows the construction described in Section 3.2. For all later stages, the shared encoder and tail, global head, and client-specific projector and head are

Table 6: Parameter lifecycle in FedConsult.

Stage	Optimized	Uploaded/synchronized	Retained locally
Base training	Global and personalized paths	Global path	(r_i, h_i)
Specialty discovery	None	None	Samples, utilities, groups
Adapter learning	$(\gamma_{i,m}, \beta_{i,m})$	Specialty adapters	Frozen base model
Router training	Context and scoring MLPs	None	Recipient router

frozen. The adapter input is the shared Layer3 activation $A(x) \in \mathbb{R}^{256\times H\times W}$, while the downstream feature $z(x)$ has dimension 512.

Visual specialty adapters. Each adapter is a channel-wise FiLM transformation with a learned scale $\gamma_{i,m} \in \mathbb{R}^{256}$ and bias $\beta_{i,m} \in \mathbb{R}^{256}$. It therefore contains 512 trainable parameters. We use $M = 4$ adapters per source client and modulation strength $\eta = 0.1$. With eight clients, the complete bank contains 32 adapters, 16,384 scalar parameters, or 64 KiB in FP32; each source uploads 8 KiB of adapter parameters once after specialty learning.

Response-aware router. The router first maps the 512-dimensional frozen feature through LayerNorm–Linear$(512, 64)$–GELU to obtain the sample context. For each candidate expert, the scorer receives the 64-dimensional context, the C-dimensional original personalized probabilities, and the C-dimensional response difference. A scorer shared by all experts applies LayerNorm–Linear$(64 + 2C, 128)$–GELU–Linear$(128, 1)$. The no-op scorer uses LayerNorm–Linear$(64 + C, 128)$–GELU–Linear$(128, 1)$. Sharing the expert scorer keeps the number of router parameters independent of the number of candidate adapters.

The candidate set contains all synchronized local and external adapters plus one no-op action. It therefore comprises 33 actions in the eight-client experiments and 13 actions for the three-client Retina experiment. Sparsemax converts candidate scores into sparse weights, and the final output is the weighted mixture of the corresponding probability vectors. During this stage, only the context mapping and the two scorers are trained; the base predictor and specialty bank remain fixed.

A.3 Datasets and Federated Partitions

Table 7 summarizes the federated construction of all datasets. Acc is computed by pooling correct predictions across clients and weighting them by local test-set size. Weak is computed over the bottom quartile of class-wise test accuracies among classes represented in each client's local training data, as defined in Section 4.

Table 7: Dataset and federated-partition details.

Dataset	Images	Classes	Clients	Client construction	Local split	Rounds	Batch
CIFAR-10 Dir-0.3	60,000	10	8	Dirichlet ($\alpha = 0.3$) + covariate shift	70/30	800	128
CIFAR-10 5-SPC	60,000	10	8	Five shards/client + covariate shift	70/30	800	128
CIFAR-100 Dir-0.3	60,000	100	8	Dirichlet ($\alpha = 0.3$) + covariate shift	70/30	800	128
CIFAR-100 25-SPC	60,000	100	8	Twenty-five shards/client + covariate shift	70/30	800	128
Retina	2,390	2	3	Acquisition domain	70/30	100	32
Skin14	36,656	14	8	Image-level Dirichlet ($\alpha = 0.3$)	70/30	800	32
BreaKHis	7,909	2	8	Image-level Dirichlet ($\alpha = 0.3$)	70/30	800	32

CIFAR benchmarks. For each CIFAR dataset, the original training and test images are first combined and then partitioned among eight clients. In the Dirichlet setting, class proportions are sampled with concentration $\alpha = 0.3$. In the pathological setting, class-sorted data are divided into homogeneous shards, with five shards assigned to each CIFAR-10 client and 25 to each CIFAR-100 client. Every client partition is subsequently divided into local training and test subsets at a 70/30 ratio. Client-level covariate heterogeneity is introduced by assigning each client one combination of gamma correction, hue adjustment, and saturation scaling, using factors from $\{0.6, 1.4\}$, $\{-0.1, 0.1\}$, and $\{0.5, 1.5\}$, respectively.

Medical image datasets. Retina combines ACRIMA, RIM-ONE, and REFUGE, with each acquisition domain forming one client for binary glaucoma classification. The source-domain partition preserves naturally occurring differences between centers. Skin14 follows the Skin-Lesions-14 benchmark, which merges the HAM10000 dermoscopic dataset (Tschandl et al., 2018) and MSLDv2.0 (Ali et al., 2024). We use the processed public release available on Kaggle, containing 36,656 images from 14 skin-lesion classes, and divide it among eight clients using class-wise image-level Dirichlet sampling with $\alpha = 0.3$. BreaKHis is used for binary benign-versus-malignant classification and follows the image-level Dirichlet construction used in the dFLMoE setting, also with eight clients and $\alpha = 0.3$. No additional synthetic covariate transformation is applied to the medical datasets.

Preprocessing. The covariate-shift CIFAR pipeline uses a random resized crop at 32×32 (scale range $[0.9, 1.0]$), followed by the fixed client-specific color transformation, conversion to a tensor, and normalization with mean and standard deviation $(0.5, 0.5, 0.5)$. Retina images are resized to 96×96 and normalized with the corresponding domain-specific statistics. Skin14 images are resized to 28×28, while BreaKHis images are resized to 224×224; both use ImageNet normalization. The same preprocessing is used by the frozen base predictor, adapter learning, and router training within each dataset.

A.4 Training and Hyperparameter Configurations

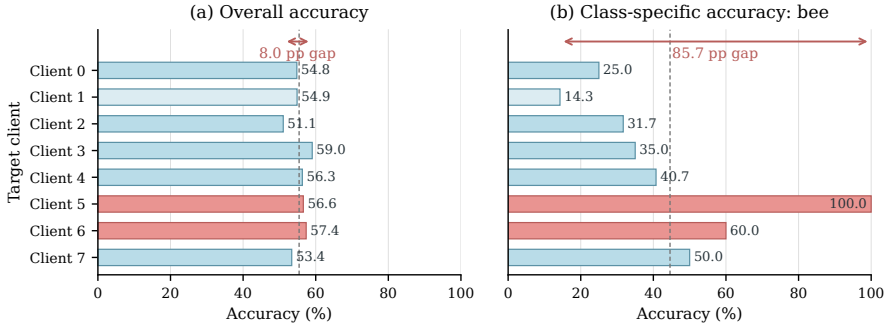
Table 8 reports the FedConsult settings used across experiments. We report method-specific parameters for FedConsult; baseline configurations follow their corresponding implementations and the tuning protocol stated in the main paper.

Table 8: FedConsult training and optimization settings.

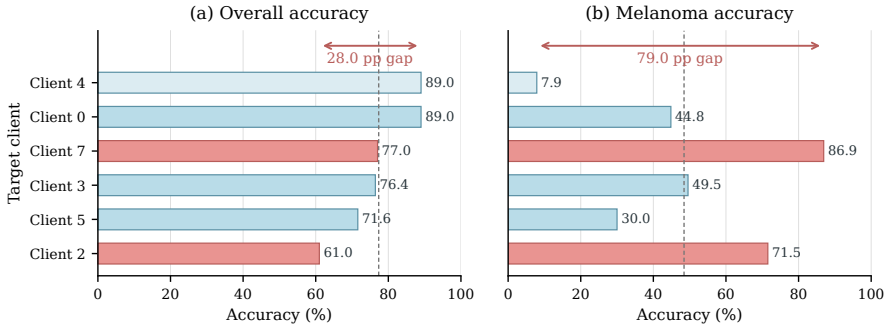
Component	Hyperparameter	Value
Base predictor	Optimizer	AdamW
	Initial learning rate / schedule	10^{-3} / cosine decay
	Weight decay	10^{-4}
	Local updates per round	10
	Client participation	All clients
Specialty adapters	Adapters per client M	4
	Training epochs / batch size	20 / 256
	Optimizer learning rate	5×10^{-3}
	Weight decay	10^{-4}
	Modulation strength η	0.1
	Distillation weight λ_{KD}	0.5
	Distillation temperature T	2
Sparse router	Training epochs / batch size	50 / 256
	Response-evaluation batch size	512 (CIFAR); 64 (medical)
	Optimizer learning rate	10^{-3}
	Weight decay	10^{-4}
	Context / hidden dimensions	64 / 128
	Routing normalization / temperature	Sparsemax / 2
	Risk weight λ_r	0.1
	Gradient-norm clipping	5

A.5 Computational Environment

Experiments were executed with PyTorch on NVIDIA GeForce RTX 4090 GPUs with 24 GB of memory.



(a) CIFAR-100: overall and bee accuracy.



(b) Skin14: overall and melanoma accuracy.

Figure 5: Personalized clients with similar overall performance can exhibit substantially different capabilities on a specific class.

B Additional Experiments and Analyses

This section provides additional evidence for the motivation and design of FedConsult. Unless stated otherwise, mechanism analyses use CIFAR-100 Dir-0.3, the same global–personalized base predictor, and the evaluation metrics defined in the main paper.

B.1 Empirical Motivation: Uneven Client Capabilities

The overall performance of a personalized client does not fully characterize its competence on individual classes. Figure 5 contrasts client-level overall accuracy with accuracy on one representative class. On CIFAR-100, the spread in overall accuracy is only 8.0 percentage points, whereas accuracy for the *bee* class differs by 85.7 points. The Skin14 example exhibits the same phenomenon: the overall gap among the displayed clients is 28.0 points, while their melanoma accuracy spans 79.0 points. The Skin14 panel includes the clients for which melanoma is represented in the corresponding local test data.

The baseline results in Table 1 further reveal a difference between unified aggregation and personalization. On CIFAR-100 Dir-0.3, FedAvg, FedProx, and FedSAM obtain Acc values of 41.40–43.72 and Weak values of 13.20–15.18. Several personalized baselines substantially improve Acc but exhibit much larger class-specific deficits: FedPAC, FedAS, and FedFew reach 51.91–54.79 Acc while attaining only 3.64–5.04 Weak. Thus, local adaptation can produce a strong personalized predictor without ensuring that its gains extend to every locally relevant class. This contrast motivates transferring complementary expertise specifically to the capability gaps left by personalization.

B.2 Why Visual Specialty Adapters?

Table 4 first establishes that directly transferring a classifier head or a complete personalized path provides little benefit. Relative to the Stage 1 predictor at 55.40 Acc and 16.61 Weak, the transferred classifier head reaches 55.57 and 18.18, while the personalized path reaches 55.45 and 17.86. In

contrast, the visual specialty adapter improves the two metrics to 58.95 and 26.76, respectively. The comparison therefore identifies the visual adapter, rather than a conventional personalized component, as the effective carrier of cross-client specialty knowledge.

Classifier heads and complete personalized paths are closely coupled to the source client's locally adapted representation and decision function. Their direct transfer preserves this source-specific coupling, which limits their compatibility with a recipient whose distribution and personalized predictor differ. A visual specialty adapter instead encodes a bounded transformation of the federation-wide feature space. The transformed feature is interpreted by the recipient's own frozen personalized pathway, allowing source knowledge to supplement the target predictor without replacing its decision structure. The substantially stronger results of the 512-parameter adapter, compared with the 51,300-parameter head and 315,748-parameter personalized path, show that the gain follows from this transferable feature interface rather than parameter count.

B.3 Functional Diversity and Specialty Granularity

Following the analysis principle used for grouped prompt experts, we characterize each adapter by what it changes in prediction space rather than by parameter distance alone. For adapter e and class c, we compute three quantities over evaluation samples of class c: the mean change in the true-label margin, the fraction of wrong-to-right transitions, and the fraction of right-to-wrong transitions. The functional signature is

$$\phi_e = \left[\{\overline{\Delta m}_{e,c}\}_{c=1}^{C}; \{\rho_{e,c}^{\mathrm{W2R}}\}_{c=1}^{C}; \{-\rho_{e,c}^{\mathrm{R2W}}\}_{c=1}^{C}\right] \in \mathbb{R}^{3C}. \tag{10}$$

The negative sign assigns harmful right-to-wrong transitions the opposite orientation. Figure 2 reports cosine similarity between normalized signatures, separately for adapters from the same source and adapters from different sources.

The heterogeneous similarities within each source bank indicate that direction grouping does not merely reproduce four copies of one client-level correction. Differences across source banks additionally show that the functional organization remains client dependent. Across all 32 adapters, the signature matrix has an effective rank of 24.50, and 25 adapters yield a positive marginal contribution in leave-one-adapter-out analysis. These statistics provide complementary evidence that the bank contains multiple nonredundant response patterns.

The granularity results in Table 4 follow the same functional view. A single adapter can merge correction patterns that affect different classes or logits in incompatible directions. Increasing M separates these patterns and improves transfer, but an excessively large bank fragments the support available to each adapter and increases the routing burden. In our experiments, $M = 4$ offers the strongest Acc with competitive Weak using only 2,048 adapter parameters per source; we therefore use it throughout the main experiments.

B.4 Response-Aware Sparse Routing

Table 9 isolates the effect of candidate selection and weighting. Top-1 routing makes a discrete choice and cannot combine complementary responses. Dense Softmax assigns positive mass to every action, including candidates that need not help the current sample. Sparsemax instead learns a sample-dependent support by assigning exact zeros to unselected candidates. Uniform variants remove learned relative weights after the support is determined, while weighted variants retain them.

Table 9: Fusion rules for response-aware routing on CIFAR-100 Dir-0.3.

Fusion rule	Acc	Weak
Top-1 selection	58.59	25.64
Softmax (uniform)	55.87	18.10
Softmax (weighted)	59.91	27.85
Sparsemax (uniform)	60.17	27.30
Sparsemax (weighted)	**60.65**	**28.78**

Weighted Sparsemax achieves the best joint Acc and Weak. Its advantage over Top-1 shows that multiple specialties can be complementary for one sample, while its advantage over dense Softmax

supports excluding irrelevant responses rather than averaging across the entire bank. The uniform Sparsemax result also confirms that sparse support selection alone is beneficial, although adaptive weights are needed for the best overall result.

The no-op action and risk term protect competence already present in the recipient. The no-op competes with all adapter responses and allows the router to retain the original personalized probability vector. The risk penalty is activated only when the original personalized prediction is correct and every expert response is incorrect; it then discourages total expert mass. In the component ablation of Table 3, adding no-op to the complete specialty and routing design increases Acc from 60.33 to 60.65 and Weak from 28.08 to 28.78. Together with the transition analysis in Figure 3, this supports the intended behavior: consultation primarily repairs weak predictions without discarding the recipient's existing personalized competence.

B.5 Cross-Client Specialty Transfer

We use leave-one-source-out (LOSO) analysis to measure the marginal value of a source after routing has been learned. For each source client, its four adapters are masked, the remaining weights are renormalized, and the resulting accuracy is compared with that of the complete bank. This intervention differs from counting routing frequency: a frequently selected source may be replaceable, whereas a less frequent source can be indispensable for a specific recipient–class pair.

Figure 6 presents the positive LOSO credit from two complementary views. Panel (a) groups external contributions by the source client's accuracy rank for the corresponding class. Substantial credit is distributed across several rank groups rather than concentrated in the Top-1 source, showing that source accuracy alone does not determine transfer value. Panel (b) resolves the same credit by target–source pair. Its nonuniform rows show that recipients rely on different combinations of source clients, while multiple clients provide useful specialties across the system. Together, the two views indicate that effective expertise is distributed and complementary rather than dominated by one universally strongest source.

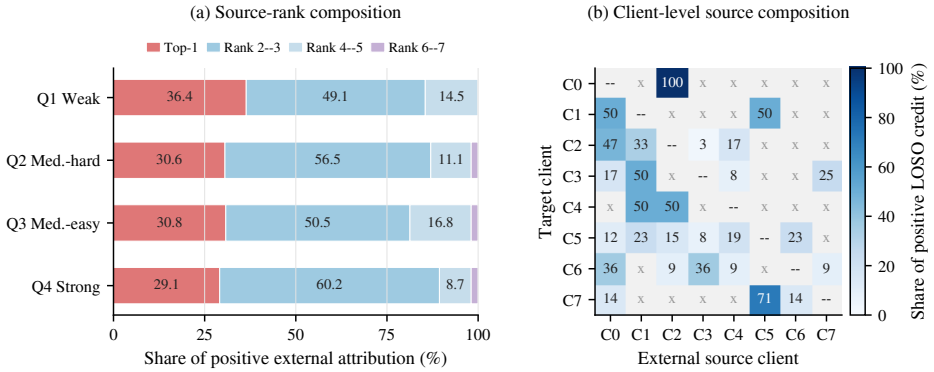


Figure 6: Positive external LOSO credit by source ability rank and client identity on CIFAR-100 Dir-0.3.

Routing statistics provide a second view of this behavior. External adapters receive 77.80% of the average probability mass and account for 61.87% of hard top-1 choices; local adapters and no-op retain the remaining mass and choices. The candidate-pool controls in Table 5 further show that real external adapters outperform local-only, random, and averaged alternatives. Taken together, the LOSO, source-ranking, source–target, and pool-control analyses indicate that personalized clients develop different but complementary strengths, including cases in which a client with lower aggregate accuracy provides useful expertise to a stronger recipient.