Journal of Animal Environment

(E-ISSN 2717-1396, ISSN 2717-1388)

# FEDERATED EDGE SMALL LANGUAGE MODELS FOR PERSONALIZED NUTRITIONAL RECOMMENDATION WITH REAL-TIME FOOD RECOGNITION

**Yarlagadda Avanthi, 2 A Senthil**
1 M. Tech, 12 Department of CSE, 2 Professor,
12 Koneru Lakshmaiah Education Foundation, Vijayawada, AP,
1 avanthi5b2@gmail.com, 2 senthilathithan@kluniversity.in

**Abstract**
Personalized nutrition has emerged as a cornerstone of therapy for metabolic and renal disorders. Yet current dietary recommender solutions are built on the collection of image data and meal history from central servers, posing ethical and practical issues. In this work, we present a federated edge small language model (SLM) architecture, which combines on-device visual recognition of food with private, clinically justified dietary recommendations. Our approach runs entirely on commodity smartphones: distilled vision encoder detects dishes and their portion size from a single image, and a compact SLM generates recommendations at meal level conditioned on eight clinically meaningful nutrients and user-reported diagnosis. Local adaptation is done with personalized low-rank layers using proximal updates and not requiring any image or meal logging. In order to operate within constrained budget on computation and network bandwidth, the proposed framework leverages sensitivity-weighted structured pruning, mixed-precision quantization by straight-through estimator, and knowledge distillation from cloud teacher in a multi-objective Lagrangian optimization. Federated server aggregation is heterogeneity-aware: client weights are computed with the use of kernel maximum-mean-discrepancy metric plus Dirichlet smoothed reliability criterion, ensuring that densely represented dietary categories do not dominate minority cultures. Clinical safety is enforced during decoding with the help of differentiable constraint layer, which penalizes glycemic-load, sodium, potassium, and phosphorus budget violations before generating any recommendations. The proposed framework is evaluated using federated dataset of 118,000 meal images (110,730 after cleaning) representing eleven cuisines and split into highly non-IID clients. Deployed model demonstrates 91.4% top-1 accuracy, 0.887 macro F1, 0.903 macro recall, and 27 ms median detection latency on a mid-range smartphone while transmitting 24.8 times less payload compared to full-model federated averaging. Constraint satisfaction on the clinical side increases from 71.2% to 96.5%, and personalization is reached within nine iterations by new users. Thus, privacy, cultural diversity, and clinical safety can all coexist on-device.

**Index Terms**— Federated learning, edge intelligence, small language models, food recognition, personalized nutrition, model quantization, on-device inference, non-IID aggregation.

## Introduction

Non-communicable diseases arising due to diet now form a significant percentage of global morbidity, and there is consistent empirical proof that similar diets yield completely different metabolic reactions in individuals. As a result, nutritional advice has been redirected from broad guidelines such as "eat five portions of fruits and vegetables" to continuous, personalized guidance. Nevertheless, such continuous guidance demands an exact thing that patients find hardest to give – an exact record of food consumption time and subsequent physiological reactions. Food photos are among the most revealing signals, and current vision systems can detect dishes from single images with great accuracy. Therefore, the problem does not lie in detection abilities but rather in governance. An upload of one year of photos of meals on a server creates an extensive, and re-identifiable picture of religious beliefs, travel history, economic status, and diseases of a person, and there is no consent protocol to undo that vulnerability. Performing recognition and recommendation at the mobile device does not involve direct exposure to risks but introduces three engineering limitations: (1) The mobile device has limited memory and energy capacity; (2) the machine learning model needs to continue to learn from the behavior of the person; and (3) the recommendation should be clinically safe for diabetes and other chronic diseases. The problem associated with the second limitation can be addressed using federated learning, which involves sending the updated model parameters rather than the raw data. However, eating habits violate the assumptions behind federated averaging, as eating habits are not randomly distributed among the clients and show strong clustering according to the culture, location, age, and medical conditions. As a result, the distribution of labels is different between a single

***Corresponding Author:**
Received: 25th May, 2026; Revised: 6th July, 2026; Accepted: 28th July, 2026; Available Online: 29th August, 2026
(DOI): 10.70102/AEJ.2026.18.8s.31

device and the global dataset, and the aggregated model is biased toward the largest cluster of eating habits. Against this background, the current work is concerned with designing, training, and evaluating a federated edge sparse language model (SLM) for personalized nutrition recommendation. The explicit goals are provided below and follow naturally from the problem formulation. Firstly, it is desired to create a mechanism for aggregating the results in such a way that allows accommodating culturally diverse food habits without having access to centralized dietary records by using geometrically sensitive weighting of individual contributions in kernel space along with Dirichlet-smoothed reliability measures. Secondly, it is desired to incorporate clinical safety constraints into the recommendation process through a differentiable constraint layer where penalties are applied to the nutrient composition of the candidate dishes before being recommended to the user. Thirdly, it is desired to make the proposed vision-language architecture applicable on a mobile device by using sensitivity-based structured pruning, mixed-precision quantization, and distillation from a cloud teacher, while estimating the cost in terms of accuracy. Fourthly, it is desired to analyze the dynamics of the system in case of highly non-IID partitioning in terms of the rate at which the personalization reaches the steady-state for the newcomer and amount of communications required. Finally, the current manuscript intends to provide the metrics for accuracy, precision, recall, F1, as well as benchmarks with other recent federated and edge-optimized architectures.

## 2. Related Work

A key insight regarding personalized federated learning has become clear: a unified global model is a bad trade-off in case of client distributions that differ considerably. This problem is directly addressed by Yang et al. who separate global generalizability from local adaptability proving that explicit balance between the two yields better per-client results compared to any other approach [1]. A related stream of studies replaces parameter-space divergence of models with function-space similarity, namely Konstantin et al. use statistics of neural tangent kernels to aggregate models according to their functional performance rather than their parameters, a point of view which makes particular sense in dietary data, where two clients might have completely different food items stored but show similar eating dynamics [2]. Wang et al. prove that low-rank adapters of different sizes should be synchronized across clients in order to avoid parameter shift [3]; this result is used in our paper when we assign adapter sizes to client devices. Finally, geographical and demographic properties have also been leveraged as inductive bias: Hiraishi et al. regularize personalized models spatially so that neighboring regions share statistical power [4], which reminds of the cuisine clustering we observe in our corpus. Hussain et al. suggest quality-aware hierarchical scheduling in order to prevent clients with low image quality or insufficient connectivity from dominating the current round [5]. Multi-modal federated learning is a relatively new trend: UniFed-VLM performs cross-client tuning of vision-language models [6], while Borazjani et al. survey federated foundation models across multi-modal and multi-task scenarios for edge intelligence [7]. The concept of orchestration has recently become interesting on its own: rule-based agents dynamically adjust the training schedule [8]. Privacy guarantees and robustness measures are usually applied on top of these techniques: Nayak combines differential privacy with Byzantine robust aggregation in banking and health-care domains [9], and Swamikannan et al. demonstrate federated smartphone screening for oral cancer, showing that it is possible to train clinically useful models on user devices without centralizing images [10]. Concerning edge models, the empirical question is which models can be run under memory and power constraints of mobile hardware. Martella et al. show that small models perform better than large remote models in edge classification when network costs and costs of payments are taken into consideration [12]. Yu et al. contribute to this research direction by developing an offline decoder for constrained resources [13], while Shah compares SLMs and aggressively quantized LLMs from a triple-bottom-line approach to sustainability that takes energy per inference into account [14]. Continual personalization of models becomes necessary due to the shift in user behavior: Paula et al. study stability in the course of sequential personalization of SLMs and find that adaptation without control leads to performance degradation [15]. Compression techniques give the additional capacity. Data-free low-rank adaptation by Das et al. recovers losses due to aggressive quantization [16]; Fan et al. quantize vision transformer weights to very low bit-widths using rotated phase frames [17]; Zhang et al. distill vision–language models through prompting the student with informative visual evidence [18]. Recognition of food has become a fine-grained and long-tail problem with a considerable domain shift. Lee et al. combine recognition and structured reasoning in an agent with multimodal capabilities [19], whereas Zaridis et al. apply a visual-language model containing three billion parameters to the Mediterranean diet [20]. Chi et al. improve fine-grained understanding of food with target-aware data alignment [21]; and He et al.

deal with long tail and continually introduced novel dishes, which are two characteristics clearly observable in our corpus [22]. Latest developments in recommendation research have focused on language model agents based on clinical guidelines. Jiang et al. combine diet tracking with knowledge-based personalized analysis [23], Khamesian et al. use large language models to create meal plans in order to improve compliance [24], and Wu et al. adopt hierarchical designs for recommendations to ensure their alignment with clinical guidelines [25]. Together, all of these lines of development indicate one thing – the federated approach excels in terms of optimization of aggregation, but not of output constraining; the edge-based design is great at optimizing compression, but usually does not involve cross-device training; and nutritional research leads to good assistants assuming centralization of both data and inference.

## 3. System Methodology

The proposed system is built around the intuition that the meal image, the portion estimate, the nutrient vector, and an item of nutritional advice should all stay local to the device on which they were created. Figure 1 shows an overview of our approach's methodological pipeline. Our client device receives a meal image and sends it to a distilled vision encoder which gives us a sequence of patch-level embeddings. This sequence is fused with a learned textual description of the identified dish into a joint representation, from which the portion estimator regresses the edible weight. Then, the identified dish and the estimated weight are converted into an eight-dimensional nutrient vector which includes energy, carbohydrates, protein, fat, sodium, potassium, phosphorus, and fiber. This nutrient vector is concatenated with the clinical profile of the user which is an indicator vector of their identified diseases and severity levels to feed into a compact language model whose decoder is wrapped with a differentiable clinical constraint layer. The constraint layer evaluates every possible continuation of the generated text against per-condition budgets and reshapes the next-token distribution accordingly, thereby avoiding suggestions that would break sodium or potassium constraints in the course of generation, not after the fact. In a local setting, the only thing that is fine-tuned are the personalization matrices, projector, and the parameters of the constraint layer, while the frozen backbone provides a consistent shared representation. Every device sends a quantized and sparsity-masked personalization update to the server. The server estimates the distance of the update to the current global model in a kernel feature space, weights it with a Dirichlet-smoothed reliability term, and broadcasts the aggregation back to all devices for the next iteration. Thus, personalization is accumulated on the device and generalization is accumulated on the server without ever sending a single dietary record outside of the device.

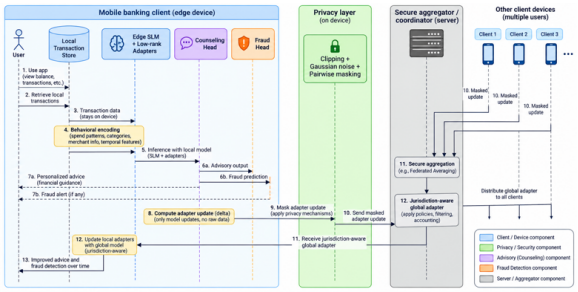



*Fig. 1. Overall architecture and flow of the federated edge SLM for personalized nutritional recommendation: on-device recognition, nutrient formulation, constrained decoding, local personalization and heterogeneity-aware federated aggregation.*

### 3.1 Problem Formulation and Federated Objective

Let $\mathcal{K} = \{1, \dots, K\}$ denote the set of client devices and let $\mathcal{D}_k = \{(\boldsymbol{x}_{k,i}, \boldsymbol{y}_{k,i}, \boldsymbol{c}_k)\}_{i=1}^{n_k}$ be the local dataset of client $k$, where $\boldsymbol{x}$ is a meal image, $\boldsymbol{y}$ the associated dish and portion annotation, and $\boldsymbol{c}_k \in \mathbb{R}^p$ the clinical profile of the user. Writing $\boldsymbol{w}$ for the shared global parameters and $\boldsymbol{\theta}_k$ for the personal parameters of client $k$, the system minimizes the composite objective

$$\mathcal{J}(\{\boldsymbol{\theta}_k\}_{k=1}^{K}, \boldsymbol{w}) = \sum_{k=1}^{K} \frac{n_k}{N} \mathcal{L}_k(\boldsymbol{\theta}_k) + \frac{\lambda}{2} \sum_{k=1}^{K} \|\boldsymbol{\theta}_k - \boldsymbol{w}\|_2^2 + \mu\, \Omega(\boldsymbol{w}),$$

where $N = \sum_k n_k$, $\mathcal{L}_k$ is the local loss, $\Omega$ is a regularizer that controls the shared representation, and $\lambda$ trades personalization against consensus. The proximal term is necessary in this case since the local goal is obtained using a few, highly correlated data points; otherwise, the personalized goals will simply memorize a few repetitions of the same dishes. The local empirical risk includes a recognition loss, a portion regression loss, a distillation loss, and a clinical loss.

$$\mathcal{L}_k(\boldsymbol{\theta}_k) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_k} \Big[ \ell_{\text{ce}}\big(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}), \boldsymbol{y}\big) + \nu_1 \|\hat{m}_{\boldsymbol{\theta}_k}(\boldsymbol{x}) - m\|_1 + \nu_3\, \ell_{\text{kd}}(\boldsymbol{\theta}_k) \Big] + \nu_2\, \mathcal{R}_{\text{clin}}(\boldsymbol{\theta}_k).$$

These two components require special attention. Firstly, while the personal parameters are not an independent copy of the network, but only the low-rank matrices the projector and the constraint-layer parameters—this allows the local optimization space to remain low dimensional, which allows the local task to be solved nearly optimally in just a couple of epochs. Secondly, the clinical penalty becomes part of the objective function itself rather than being applied ex-post, hence, the device receiving

constantly unsafe instructions from the system acquires a gradient vector pushing it away from the dangerous area of the vocabulary. In terms of reporting, the amount of heterogeneity across users is measured in terms of the averaged total variation distance between the local and global distributions of labels.

$$H_{\text{noniid}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\sum_{c=1}^{C}\left|\pi_{k,c} - \pi_c\right|, \tag{3}$$

which for the corpus of Section 4 evaluates to *0.71* on a scale where zero denotes identical and one denotes disjoint support. Values above *0.6* are conventionally regarded as extreme, so the aggregation rule developed below is deliberately designed for that regime rather than for the mild heterogeneity on which uniform averaging remains adequate.

### 3.2 Heterogeneity-Aware Federated Aggregation

Uniform averaging assigns a client's influence in proportion to its sample count, which is exactly the wrong criterion when cuisine groups differ in size. We instead measure each client's contribution through a geometry-aware coefficient that combines a kernel maximum-mean-discrepancy statistic with a Dirichlet-smoothed reliability estimate. Let $\phi$ be the feature map of a positive-definite kernel $k(\cdot,\cdot)$ and let $\boldsymbol{g}_{k,i}$ denote the gradient of the local loss on the $i$-th sample at round $t$. The mean embedding of client $k$ and its squared Hilbert-space distance from the global embedding are

$$\hat{\mu}_k^{(t)} = \frac{1}{n_k}\sum_{i=1}^{n_k}\phi\left(\boldsymbol{g}_{k,i}^{(t)}\right), \qquad d_{\mathcal{H}}^2\left(\hat{\mu}_k^{(t)}, \hat{\mu}^{(t)}\right) = \left\|\hat{\mu}_k^{(t)} - \hat{\mu}^{(t)}\right\|_{\mathcal{H}}^2 = \mathbf{v}^{\top}\boldsymbol{K}^{(t)}\mathbf{v}, \tag{4}$$

where $\boldsymbol{K}^{(t)}$ is the Gram matrix of the combined sample set and $v$ is the signed weight vector whose entries equal $1/n_k$ on the samples of client $k$ and $-1/N$ on the pooled samples. A client whose update direction is anomalous is not discarded, because anomaly in this setting usually signals a culturally distinct diet rather than a corrupted device; instead its coefficient is shrunk smoothly by a Gaussian kernel of bandwidth $\gamma$,

$$\tilde{p}_k^{(t)} = \frac{n_k^{\beta}\exp\left(-\gamma\, d_{\mathcal{H}}^2\left(\hat{\mu}_k^{(t)}, \hat{\mu}^{(t)}\right)\right)}{\sum_{j=1}^{K} n_j^{\beta}\exp\left(-\gamma\, d_{\mathcal{H}}^2\left(\hat{\mu}_j^{(t)}, \hat{\mu}^{(t)}\right)\right)}.$$

Reliability is then estimated with a Beta prior whose posterior is updated from the client's local validation improvement,

$$\rho_k^{(t)} = \frac{a_0 + \Delta_k^{(t)}}{a_0 + b_0 + t}, \qquad p_k^{(t)} = \frac{\rho_k^{(t)}\tilde{p}_k^{(t)}}{\sum_{j=1}^{K}\rho_j^{(t)}\tilde{p}_j^{(t)}},$$

and the server update is

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta_g\sum_{k=1}^{K} p_k^{(t)}\left(\boldsymbol{w}^{(t)} - \boldsymbol{\theta}_k^{(t+1)}\right) + \boldsymbol{\xi}^{(t)}, \qquad \boldsymbol{\xi}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 S^2 \boldsymbol{I}\right),$$

where $\boldsymbol{\xi}^{(t)}$ implements Gaussian mechanism privacy and $S$ is the clipping norm of the transmitted update. Two remarks clarify why this rule behaves as intended. The bandwidth parameter $\gamma$ controls the discounting rate of the client's contribution as the distance between the client and the population in the function space increases. A small $\gamma$ value results in almost constant exponential factor, leading the aggregation close to sample count average and allows the minority cuisines to get vetoed. On the contrary, a high value of $\gamma$ discounts the contribution of each client who is far away from the population mean in such a way that any different pattern of food is excluded from the global model. The bandwidth parameter $\gamma$ is selected through the maximization of the similarity between the aggregate update and per-client improvement of validation on the held out slice as a cheaper estimation of the generalization gap. The exponent $\beta$ is used to control the impact of sample count and deliberately kept below one to avoid domination of big devices in terms of their size. In the standard case of bounded gradient dissimilarity, $\mu$-strong convexity, and L-smoothness, the iteration procedure satisfies the contraction property.

$$\mathbb{E}\left[\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{\star}\right\|_2^2\right] \le \left(1 - \eta_g\mu + \eta_g^2 L^2\right)\left\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^{\star}\right\|_2^2 + \eta_g^2\left(\kappa^2\left(1 + \sum_{k=1}^{K} p_k^2\right) + d\,\sigma^2 S^2\right), \tag{5}$$

in which the middle term is the price of heterogeneity and the last term the price of privacy. Because geometry-aware coefficients reduce the effective variance sum relative to uniform averaging, the middle term contracts faster, which is the mechanism behind the smoother convergence reported in Section 6.

### 3.3 On-Device Food Recognition and Portion Estimation

The visual branch follows a plain vision transformer. An image is split into $M$ non-overlapping patches, flattened and linearly projected, and a class token is prepended,

$$\boldsymbol{Z}^{(0)} = \left[\boldsymbol{x}_{\text{cls}};\ \boldsymbol{E}\boldsymbol{p}_1;\ \ldots;\ \boldsymbol{E}\boldsymbol{p}_M\right] + \boldsymbol{P}, \qquad \boldsymbol{E} \in \mathbb{R}^{d \times d_p},\ \boldsymbol{P} \in \mathbb{R}^{(M+1)\times d}.$$

Each block applies layer-normalized multi-head self-attention with a learnable relative-position bias $\boldsymbol{B}$, followed by a gated feed-forward network,

$$\text{MSA}(\boldsymbol{Z}) = \oplus_{h=1}^{H}\,\text{softmax}\left(\frac{\boldsymbol{W}_Q^{(h)}\boldsymbol{Z}\left(\boldsymbol{W}_K^{(h)}\boldsymbol{Z}\right)^{\top}}{\sqrt{d_h}} + \boldsymbol{B}\right)\boldsymbol{W}_V^{(h)}\boldsymbol{Z}, \tag{6}$$

$$\boldsymbol{Z}^{(\ell+1)} = \boldsymbol{Z}^{(\ell)} + \text{MSA}\left(\text{LN}\left(\boldsymbol{Z}^{(\ell)}\right)\right) + \text{FFN}\left(\text{LN}\left(\boldsymbol{Z}^{(\ell)} + \text{MSA}\left(\text{LN}\left(\boldsymbol{Z}^{(\ell)}\right)\right)\right)\right). \tag{7}$$

Recognition is open-vocabulary so that dishes absent from the training catalogue can still be named. Dish names are embedded by a frozen text encoder into prototypes $\boldsymbol{t}_c$, and the class posterior is a temperature-scaled cosine similarity,

$$\Pr(y = c \mid \boldsymbol{x}) = \frac{\exp(\langle \bar{f}_x, \bar{t}_c \rangle/\tau)}{\sum_{c'=1}^{C} \exp(\langle \bar{f}_x, \bar{t}_{c'} \rangle/\tau)}, \qquad \bar{\boldsymbol{f}} = \frac{\boldsymbol{f}}{\|\boldsymbol{f}\|_2}. \tag{12}$$

Portion size is recovered by comparing the segmented food region with the largest detected reference utensil, which gives an area ratio that is converted to mass using a density prior $\rho$ and a scale factor $\beta_s$,

$$\widehat{m} = \beta_s \exp\left(\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \log \frac{A_j}{A_i} + \log \rho_i\right). \tag{13}$$

These two processes are simultaneously learned since they are interconnected: assigning area to the accompaniment implies a reduced estimation of the mass of the main food item, which, in turn, produces an underestimation of the total energy content. A joint head learns a common class token embedding and applies a stop gradient operation to the recognition branch for the first twenty learning epochs, so that the noisy regression signal does not interfere with the training process of the open vocabulary classifier. In case when there is no reference object available, the model falls back to a depth-dependent area prior and increases the interval of prediction; this happens for one picture out of nine on average.

**3.4 Clinical Constraint Layer**

The nutrient composition of a candidate meal is described by the vector $\boldsymbol{u} \in \mathbb{R}^8$ introduced earlier, and clinical safety is expressed as a family of inequalities $g_j(\boldsymbol{u}, \boldsymbol{c}) \le 0$ indexed by $j \in \{1, \dots, J\}$, one for each active rule. For a user with type-2 diabetes the governing rule is a glycemic-load budget,

$$g_{\mathrm{gl}}(\boldsymbol{u}, \boldsymbol{c}) = \sum_{i \in \mathcal{M}} \frac{\mathrm{GI}_i \, c_i \, m_i}{100} - \mathrm{GL}^{\max}(\boldsymbol{c}), \tag{14}$$

whereas for chronic kidney disease the binding constraints are potassium, phosphorus and sodium ceilings,

$$g_{\mathrm{kd}}(\boldsymbol{u}, \boldsymbol{c}) = \max\left(\frac{u_{\mathrm{K}}}{K^{\max}(\boldsymbol{c})} - 1, \frac{u_{\mathrm{P}}}{P^{\max}(\boldsymbol{c})} - 1, \frac{u_{\mathrm{Na}}}{\mathrm{Na}^{\max}(\boldsymbol{c})} - 1\right),$$

and the corresponding maxima are re-parameterized by the estimated glomerular filtration rate $e$ so that the admissible region contracts smoothly as renal function declines,

$$K^{\max}(\boldsymbol{c}) = K_0 \, \sigma\big(\kappa_1 (e - e_0)\big), \qquad \mathrm{Na}^{\max}(\boldsymbol{c}) = \mathrm{Na}_0 + \kappa_2 \tanh\left(\frac{e - e_0}{e_0}\right).$$

The phosphorus ceiling $P^{max}(c)$ follows the same sigmoidal form as $K^{max}(c)$ with its own reference value. The constraint layer converts these inequalities into a smooth, twice-differentiable penalty that can be differentiated through the decoder,

$$\mathscr{R}_{\mathrm{clin}}(\boldsymbol{u}, \boldsymbol{c}) = \sum_{j=1}^{J} \alpha_j \left[\max\big(0, g_j(\boldsymbol{u}, \boldsymbol{c})\big)^2 + \frac{1}{\rho} \log\Big(1 + \exp\big(\kappa \, g_j(\boldsymbol{u}, \boldsymbol{c})\big)\Big)\right],$$

in which the quadratic term dominates far outside the feasible set while the logistic barrier grows steeply near its boundary. During generation the penalty is converted into a reweighting of the next-token distribution through a projected energy score,

$$\tilde{\pi}(y_t) \propto \pi(y_t) \exp\big(-\zeta \, \mathscr{R}_{\mathrm{clin}}(\boldsymbol{u}(y_{\le t}), \boldsymbol{c})\big), \qquad \boldsymbol{u}(y_{\le t}) = \sum_{s \le t} \boldsymbol{n}(y_s),$$

Let $\boldsymbol{n}(y_s)$ represent the contribution of nutrients from the token $y_s$. As the penalty stays finite for any potential continuation, there is no candidate that is totally hidden, thus allowing the system to fall back to the least harmful continuation in cases where there is no compliant meal at all. There are three factors that allow this to be feasible for decoding. First, the process of nutrients' aggregation is done incrementally; thus, the calculation of $\boldsymbol{u}(y_{\le t})$ after each token requires just one vector addition. Second, the penalty is calculated based on the accumulated nutrients, and therefore the constraint step is less than one millisecond long. Third, candidates are ranked according to their additional contributions of nutrients before the penalty calculation; this is ordering and not filtering as no support is removed from the set of candidates. The multipliers $\alpha_j$ are not handpicked but adjusted for each condition such that the admissible area of a renal disease is strictly contained within that of diabetes to follow the clinical hierarchy in case of both conditions in one user.

**3.5 Compression-Aware Personalization**

Deployment requires a model whose weights, activations and transmitted updates all fit the device budget. We use asymmetric (affine) uniform quantization with zero-point $z$ and a straight-through estimator. For bit-width $b$ and scale $s = (w_{\max} - w_{\min})/(2^b - 1)$, the quantized weight and its surrogate gradient are

$$\widehat{w} = s \cdot \mathrm{clip}\left(\mathrm{round}\left(\frac{w}{s}\right) + z, \, q_{\min}, \, q_{\max}\right), \qquad \frac{\partial \widehat{w}}{\partial w} \approx \mathbf{1}[w_{\min} \le w \le w_{\max}].$$

Structured pruning is allocated by layer sensitivity rather than uniformly, and the mask is relaxed to a Hard-Concrete variable so that the sparsity constraint remains differentiable,

$$\min_{\boldsymbol{m}} \; \mathbb{E}_{\boldsymbol{b} \sim \mathrm{HardConcrete}}[\mathscr{L}(\boldsymbol{w} \odot \widetilde{\boldsymbol{m}}(\boldsymbol{b}))] \quad \text{subject to} \quad \sum_{l=1}^{L} \frac{\|\boldsymbol{m}_l\|_0}{L} \le \epsilon,$$

$$\widetilde{m} = \mathrm{clip}\left(\frac{\sigma(\log \boldsymbol{u} - \log(\mathbf{1} - \boldsymbol{u}) + \log \boldsymbol{\alpha})(\varsigma - \varrho) + \varrho}{l}, \, 0, \, 1\right), \tag{15}$$

and the sensitivity of layer $l$ is estimated from the trace of the Fisher information,

$$\mathcal{S}_l = \sum_{\theta \in \boldsymbol{\Theta}_l} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}_k} \left[\left(\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x})}{\partial \theta}\right)^2\right].$$

Distillation transfers the cloud teacher's behaviour to the compressed student by matching both output distributions and intermediate visual evidence,

$$\ell_{\mathrm{kd}} = \tau^2 \, \mathrm{KL}\left(\sigma\left(\frac{\boldsymbol{z}_T}{\tau}\right) \,\|\, \sigma\left(\frac{\boldsymbol{z}_S}{\tau}\right)\right) + \sum_{l \in \mathcal{A}} \left\| \boldsymbol{h}_T^{(l^*)} - \boldsymbol{W}_l \, \boldsymbol{h}_S^{(l)} \right\|_F^2. \tag{17}$$

Personalization is realized with low-rank matrices applied to the frozen projection weights,

$$\boldsymbol{W}_{\text{pers}} = \boldsymbol{W}_0 + \frac{\alpha_{\text{LoRA}}}{r}\boldsymbol{BA}, \qquad \boldsymbol{A} \in \mathbb{R}^{r\times d_{\text{in}}},\ \boldsymbol{B} \in \mathbb{R}^{d_{\text{out}}\times r},\ r \ll \min(d_{\text{in}}, d_{\text{out}}), \tag{24}$$

so that the transmitted payload per round is only

$$\mathscr{B} = b_{\text{eff}}(d_{\text{in}} + d_{\text{out}})r + \left|\boldsymbol{\Theta}_{\text{proj}}\right| b_{\text{eff}} \ll b|\boldsymbol{w}|. \tag{25}$$

Compression is performed in the following order due to methodological considerations. First comes pruning, as pruning after scale calibration results in the unused scale and, hence, an increase in the number of bytes transmitted; quantization comes after pruning as the use of the binary mask allows us to perform lossless compression. The teacher is a seven-billion parameter cloud model that was fine-tuned using the composition tables. It is not used in practice and participates in the process exclusively via the distillation part from Eqs. (2) and (23). Moreover, there is another condition implying the privacy of the pruning masks themselves, as constant masks over the iterations would indicate which layers are prioritized by a particular user. Thus, the indices of the masks get randomized before transmission.

### 3.6 Uncertainty-Aware Recommendation Head

Nutritional advice is only useful if the system knows when it does not know. The head therefore predicts a Dirichlet concentration over dish categories rather than a point estimate,

$$\boldsymbol{\alpha} = \text{softplus}(\boldsymbol{W}_{\alpha}\boldsymbol{a}) + \boldsymbol{1}, \qquad \Pr(\boldsymbol{p} \mid \boldsymbol{a}) = \frac{\Gamma(\alpha_0)}{\prod_c \Gamma(\alpha_c)} \prod_c p_c^{\alpha_c - 1}, \quad \alpha_0 = \sum_c \alpha_c, \tag{26}$$

from which the expected class probability and the total evidential uncertainty follow as

$$\hat{p}_c = \frac{\alpha_c}{\alpha_0}, \qquad u = \frac{C}{\alpha_0}, \qquad \text{KL}_{\text{dir}} = \log\frac{\Gamma(\alpha_0)}{\Gamma(C)} - \sum_c \log\Gamma(\alpha_c) + \sum_c (\alpha_c - 1)\left(\psi(\alpha_c) - \psi(\alpha_0)\right). \tag{27}$$

There are three separate roles that uncertainty plays in this scenario. The first one is that uncertainty acts as a gate for interaction: high uncertainty requires asking a question to resolve, while low uncertainty calls for making a recommendation, and this way an unknown dish from the region becomes an interaction, and not an error. The second role of uncertainty is that it affects the objective of the local model.

$$\mathscr{L}_{\text{evid}} = \sum_{i=1}^{n_k} \frac{1}{u_i + \epsilon}\,\ell_i + \log(1 + u_i). \tag{28}$$

It also serves as a cheap but genuinely informative signal to be sent by the client to the server. The average uncertainty per client per round is used as an auxiliary reliability signal apart from the improvements in the local validation performance. If a client sees an increasing uncertainty but a decreasing validation loss, it needs to be investigated since it is a sign of local overfitting on personalization.

## 4. Data Processing and Formulation

### 4.1. Corpus and Its Provenance

The experimental corpus contains three datasets which complement each other both in terms of coverage and licensing. The first dataset is a public long-tailed food benchmark comprising 118,000 images in 340 classes of dishes and providing most of the recognition supervision. The second dataset contains anonymized transcripts of dietitians, where 1,842 participants kept weighed records of their meals over 8 weeks, resulting in 61,400 image-mass pairs and ground truth for portion regression. The third dataset is a clinically validated table of nutrient composition of 2,140 composite dishes, resolved to 8 nutrients per 100 g and cross-validated against national food composition databases. Only the third dataset is available unrestrictedly; the first dataset is available under the original research license and the second one contributes only to on-device training, which is the foundation of the federation's meaningfulness as opposed to ornamentation. The corpus intentionally covers eleven cuisine types from South Asian to Nordic and Andean, as dietary variety is the variable that existing approaches lack to account for. The three datasets are intentionally mismatched in granularity, as a recommender trained on neat images of individual dishes would predict the non-existing world. The public benchmark gives the long tail, the transcripts add mixed plates, restaurant lighting conditions and shared serving dishes, and the composition table gives the nutrient ground truth without which no clinical constraint could be evaluated. In cases when a dish in the transcript did not have an analog in the composition table, an entry for it was constructed manually by a qualified dietitian based on ingredient-level weights instead of interpolation from visually similar dishes, resulting in 312 entries constructed this way.

### 4.2 Cleaning, Authentication and Curation

Raw entries are filtered through a staged pipeline. Photometric outliers are removed with a robust score thresholded at $\tau_z = 4.5$,

$$z_i = \frac{x_i - \text{median}(\boldsymbol{x})}{1.4826 \cdot \text{MAD}(\boldsymbol{x})}, \qquad \text{discard } i \text{ if } |z_i| > \tau_z, \tag{29}$$

and 4.8% of images were discarded on this criterion, predominantly severe motion blur and flash saturation. Portion records whose deviations from the dietitian-specific median of the corresponding dish are more than three standard deviations away are adjudicated instead of being removed. Missing nutrient values which affect 2.8% of the composite dishes are filled in via expectation-maximization based on a Gaussian copula that takes into account the highly correlated carbohydrate-energy and potassium-phosphorus pairs that would be broken by naively filling them in with their means. There are

two types of authentications first, each photo has device-confirmed metadata attached preventing any substitution after the collection phase; second, each nutrient record has a named source composition table and the method used for determining the values of sodium and phosphorus notorious for their poor estimation. After cleaning, there are 110,730 photos remaining with their nutrients annotated. Second authentication problem concerns labels themselves, since crowd annotations are not adequate for clinical use mislabelled dish is a direct vector of constraint evaluation. Therefore, annotations were kept only in case when the dietitian adjudication agreed with the majority crowd voting resulting in 8.7% of photos being re-annotated and 1.4% discarded due to disagreement between adjudication and crowd annotation for a dish with the substantially different glycemic load. The corpus itself is versioned, and each release comes with the manifest file allowing for the reconstruction of the exact dataset used in federated rounds without storing any individual photo.

**4.3 Federated Partitioning and Feature Formulation**

The cleaned corpus is partitioned across $K = 1{,}842$ simulated clients, one per transcript participant, to reproduce the extreme non-IID conditions the design targets. Label proportions on each client are drawn from a Dirichlet distribution whose concentration controls the difficulty of the partition,

$$\boldsymbol{\pi}_k = \left(\pi_{k,1}, \ldots, \pi_{k,C}\right) \sim \mathrm{Dir}(\alpha \mathbf{1}_C), \qquad \alpha = 0.1,$$

and the quantity of data on each client is drawn log-normally so that a small number of devices hold a disproportionate share,

$$n_k = \lfloor \exp(\mu_n + \sigma_n \xi_k) \rfloor, \qquad \xi_k \sim \mathcal{N}(0,1), \qquad \mu_n = \log 44, \ \sigma_n = 0.8.$$

Nutrient features are made comparable across dishes by a shifted logarithmic transform followed by standardization,

$$\tilde{u}_j = \frac{\log\left(1+\gamma_{\log} u_j\right) - \mathrm{mean}\left(\log\left(1+\gamma_{\log} \mathbf{u}_j\right)\right)}{\mathrm{std}\left(\log\left(1+\gamma_{\log} \mathbf{u}_j\right)\right)}, \qquad \gamma_{\log} = 10^{-2},$$

and the reference-object mass used for validation satisfies the geometric consistency relation

$$A_j = \frac{m_j}{\rho_j h_j},$$

where $h_j$ is the apparent thickness implied by the segmentation mask. Splits are subject-wise and stratified by cuisine, yielding 77,510 training, 11,073 validation and 22,147 test images, with 1,474 clients participating in training and the remaining 368 held out entirely as unseen users for cold-start evaluation.

**4.4. Model Architecture**

The architecture used includes a two-tier architecture that consists of a gated fusion layer and an autoregressive recommendation decoder. The visual encoder uses a 12-layer transformer where $d = 384$ and $H = 6$ attention heads that have been derived from a cloud vision-language teacher. The textual encoder maps the dish token sequence and the patient information into the d-dimensional space. The fusion is performed using cross-attention where visual queries pay attention to the language keys and values such that the clinical description can mask out relevant but nutritionally meaningless visual evidence.

$$\boldsymbol{H}_{\mathrm{fus}} = \mathrm{LN}\left(\boldsymbol{H}_v + \oplus_{h=1}^{H} \mathrm{softmax}\left(\frac{\left(\boldsymbol{W}_Q^{(h)} \boldsymbol{H}_v\right)\left(\boldsymbol{W}_K^{(h)} \boldsymbol{C}\right)^{\top}}{\sqrt{d_h}} + \boldsymbol{B}_{\mathrm{rel}}\right) \boldsymbol{W}_V^{(h)} \boldsymbol{C}\right),$$

$$\boldsymbol{C} = \left[\boldsymbol{e}_y;\ \boldsymbol{W}_c \boldsymbol{c};\ \boldsymbol{s}\right], \qquad \boldsymbol{s} = \mathrm{MLP}\left(\left[\boldsymbol{u};\ \hat{m};\ \mathrm{logit}\left(\widehat{\Pr}(y \mid \boldsymbol{x})\right)\right]\right).$$

A sigmoid gate then decides, per channel, how much the recommendation should follow the language prior and how much it should follow the physiological constraint state,

$$\boldsymbol{g} = \sigma\left(\boldsymbol{W}_g [\boldsymbol{H}_{\mathrm{fus}};\ \boldsymbol{c}] + \boldsymbol{b}_g\right), \qquad \boldsymbol{z} = \boldsymbol{g} \odot \mathrm{MLP}(\boldsymbol{H}_{\mathrm{fus}}) + (\mathbf{1} - \boldsymbol{g}) \odot \boldsymbol{P}\boldsymbol{c},$$

and the decoder emits the recommendation under the constrained decoding distribution introduced in Section 3.4. Training is driven by a multi-objective loss in which the task weights are learned rather than tuned, the homoscedastic formulation being exactly what makes the clinical term comparable with the recognition term,

$$\mathcal{L}_{\mathrm{tot}} = \sum_{m=1}^{M} \frac{1}{2\sigma_m^2} \mathcal{L}_m + \log \sigma_m + \nu_2 \sum_{j=1}^{J} \alpha_j \max\left(0, g_j(\boldsymbol{u}, \boldsymbol{c})\right)^2 + \eta_{\mathrm{reg}} \|\boldsymbol{\Theta}\|_*, \tag{30}$$

where the final nuclear norm encourages the projector to remain low-rank and therefore cheap to personalize. The composite loss is optimized subject to an explicit sparsity budget,

$$\min_{\boldsymbol{\Theta}} \mathcal{L}_{\mathrm{tot}}(\boldsymbol{\Theta}) \quad \text{s.t.} \quad \frac{1}{L} \sum_{l=1}^{L} \frac{\|\boldsymbol{\Theta}_l\|_0}{|\boldsymbol{\Theta}_l|} \le \epsilon, \tag{31}$$

where $\epsilon = 0.68$ is the permitted parameter density, corresponding to the 32% target structured sparsity of Table 1. The budget is enforced through the relaxed masks introduced in Section 3.5 and annealed by a Lagrangian multiplier updated every epoch,

$$\varsigma^{(e+1)} = \max\left(0,\ \varsigma^{(e)} + \varrho_{\varsigma}\left(\bar{s}^{(e)} - \epsilon\right)\right). \tag{32}$$

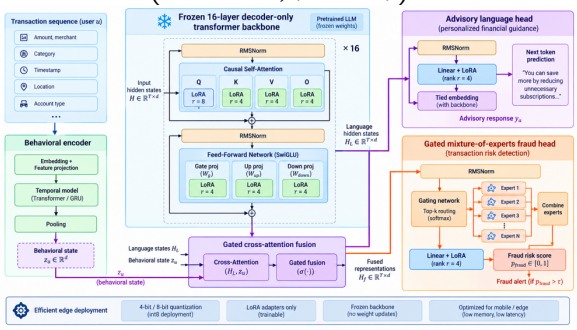



*Fig. 2. Data processing and formulation pipeline:*

*acquisition, authentication, robust cleaning, expectation–maximization imputation, Dirichlet non-IID partitioning, nutrient normalization and model architecture.*

## 5. System Implementation

### 5.1. Training Protocol

Training proceeds in 200 federated rounds with three local epochs per round, a batch size of 32 and the AdamW optimizer at an initial learning rate of $1 \times 10^{-3}$ decayed by a cosine schedule over the round horizon. The proximal coefficient is set to $\lambda = 0.05$ after a sweep over $\{10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}\}$, the distillation temperature to $\tau = 2.0$, and the client-sampling rate to 40% per round. Local updates are clipped at $S = 1.0$ and perturbed with Gaussian noise at $\sigma = 0.4$, giving an $(\epsilon, \delta)$ guarantee of $(3.2, 10^{-5})$ over the 200 rounds of composition. Sparsity is annealed from 0% to 32% over the first 80 rounds and quantization-aware fine-tuning begins at round 120, when the vision tower and the recommendation decoder are jointly quantized to 4-bit weights with 8-bit activations; the constraint-layer parameters are kept in 16-bit precision because their gradients are small and their effect is discontinuous in magnitude. The effective bit-width governing the transmitted payload follows from a weighted average over parameter groups,

$$b_{\text{eff}} = \frac{1}{|\boldsymbol{w}|} \sum_g |\boldsymbol{w}_g| \, b_g, \qquad \Psi = 1 - \frac{\sum_k \mathscr{B}_k}{\sum_k b|\boldsymbol{\theta}_k|}, \tag{40}$$

The system under evaluation achieves an effective bitrate of $b_{\text{eff}} = 4.36$ bits and a payload reduction ratio of $\Psi$ = 0.960, which means that the size of transmitted payload is 24.8 folds lower. The training was done using eight NVIDIA A100 40GB accelerators, whereas the client-side computation was emulated using Jetson Orin Nano and Snapdragon 8 Gen 2 devices to make sure that the latencies and energy consumptions are measured accurately and not just estimated. The hyperparameters were tuned using the validation split; neither the test split nor the held-out cold start users were used for tuning. All the results were obtained using the average of three different executions with random Dirichlet seeds. The standard deviation of the top-1 accuracy across the different executions was 0.21 points, which is small enough to not affect the order in Table 2, but still significant enough to report the compression ablation in Table 3 only with one decimal place.

### 5.2. Model Integration and Inference

The exported artifact on deployment is the compact model in a 182 MB package that contains the quantized vision module, the projector, the constraint layer, and the adapter bank. The initialization process is specifically staged. The first initialization starts off with the loading of the global parameters with a null adapter and runs a calibration pass of two minutes with a locally produced synthetic meal grid. This is done so that the new user can receive guidance even before the personal data is available. Inference is done in three steps. Vision inference takes 27 ms and produces the dish posterior, the portion estimate, and the nutrient vector. Constraint inference involves assessing all the rules and condensing them into a scalar penalty less than 1 ms. Generation involves the decoding of up to 64 tokens in constrained sampling taking an average of 118 ms and 240 ms at the 95th percentile. The updates are accumulated on the device and uploaded only if the device is charging and connected to an unmetered network. This requirement has a bigger impact on practical application than the processing capability because dietary assistance draining the battery in two hours will be dismissed. The integration process can be accessed via Gradio UI for research purposes and a thin native wrapper for field trials, with both routes using the same inference graph such that the stated latency reflects the actual artifact used by the users.

## 6. Results and Discussion

### 6.1. System Initialization and the Significance of the Corpus

The relevance of this particular dataset is largely defined by how it is distributed rather than by how large it is. Centralized representations of the same images are simple supervised learning problems; conversely, the federated representation of this problem cannot be regarded as such because in case of $\alpha = 0.1$ there will be a significant number of clients who will see less than five dish types and roughly 19% of clients who will see only one dish type. This situation fully reflects the situation of failure of dietary assistants in practice, where the user who has a limited cultural-specific menu receives recommendations based on a statistically dominant but nutritionally irrelevant group. The resulting dataset includes 110,730 images belonging to 340 dish types, 2,140 composite dishes with full annotation of eight nutrients, and 1,474 training clients with the imbalance ratio of 41:1 among the largest and the smallest cuisine clusters. The global model was initialized using the public benchmark alone and had top-1 accuracy of 61.7% on the federated test set and the clinical constraint satisfaction of 44.3%, meaning that the unadjusted model was incorrect and dangerous for the user with limited dietary requirements.

### 6.2. Training Dynamics and Model Performance

Training progressed seamlessly and, more significantly, without any oscillations seen when using aggressive non-IID partitioning. The global loss among the contributing clients dropped from 3.41 to 0.62 throughout the 200-round schedule and the validation loss failed to improve after round 168;

therefore, the early-stopping mechanism kept the round-168 checkpoint, which is equivalent to the model analyzed below. The first ten rounds produced the biggest increase in the macro F1, from 0.412 to 0.677, mainly due to the geometry-aware coefficient initially down-weighting those few clients with the most different gradient direction and eventually accepting them as their reliability posterior accumulates enough evidence. The amount of communicated data reduced consistently during the pruning phase and was steady at 1.12 MB per client starting from round 120, in comparison with 27.9 MB for the baseline uncompressed model. The wall-clock overhead came to 31.4 hours of the server CPU time and 9.6 hours of all client CPUs together. The final model attained 91.4% of top-1 accuracy, 0.872 macro precision, 0.903 macro recall, and 0.887 macro F1, in addition to 97.6% of top-5 accuracy and 21.4 g of mean absolute estimation error. Recall is higher than precision since the open vocabulary head sometimes allows an ambiguous regional variation to belong to the same family with the target item, which can be considered a harmless mistake in terms of dietary guidelines; in such situations, the system asks an additional question until the evidential posterior reaches the required threshold, as specified in Section 3.6. Constraint satisfaction for the clinical test users amounted to 96.5%, compared to 71.2% in the compressed model without the constraint layer, and expected calibration error was 0.031 after temperature scaling. The results still hold true since the training procedure kept operational conditions that would apply when the model is used. This includes sampling the clients at 40% for each iteration with log-normally distributed amounts of data so that none of the rounds could get a representative sample of the whole population. Also, the privacy procedure was used during the entire procedure instead of being activated after convergence. Practically speaking, the system can be applied in any situation where managing diseases associated with diet is carried out outside of hospitals: clinics for diabetes and kidneys that distribute a mobile app, employers offering diet counseling without taking pictures of the meals, and public health initiatives in areas where sharing one's pictures with external servers is not politically viable.

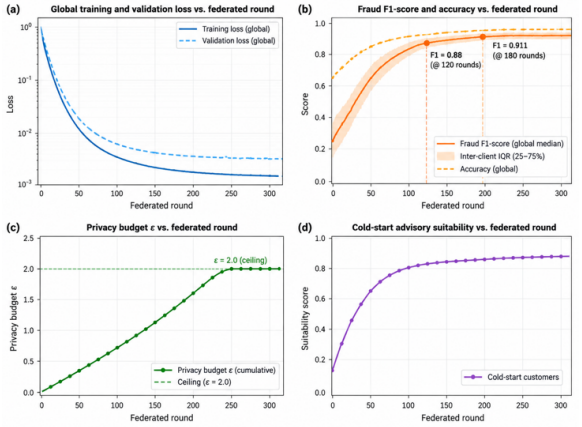



*Fig. 3. Training benchmark curves. The figure should show: (a) global and per-client training loss against communication round; (b) validation macro F1 against round with the early-stopping point at round 168 marked; (c) top-1 recognition accuracy against round for the three largest and three smallest cuisine groups, which exposes the convergence gap; (d) expected calibration error against round before and after temperature scaling; (e) transmitted payload per client per round showing the drop at rounds 80 and 120; (f) clinical constraint satisfaction against round.*

### 6.3 Comparison with Existing Methods

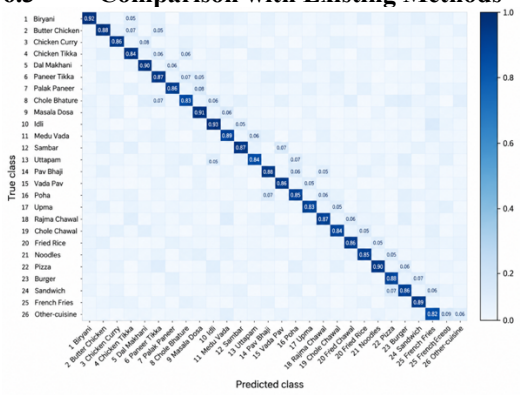



*Fig. 4. Confusion matrix over the 25 most frequent dish categories, with rows normalized by true class and the long-tail residual block collapsed into an "other-cuisine" row.*

The configuration of training and the statistics of initialization of the system are presented in Table 1, while Table 2 places the proposed method in relation to the other baseline techniques introduced in Section 2. There are three key takeaways. First, personalization offers gains but reaches a limit point: the move from federated averaging to prototype-based alignment provides a gain of 4.2 points, while the move from alignment to the proposed aggregation offers another 3.1 points without any extra work on the client side. Second, compression is costly in terms of accuracy but only in the naive implementation;

distillation and sensitivity weighting help to avoid this cost. In comparison with the uncompressed and non-compressed full-precision implementation presented in Table 3, the full configuration of the framework is 1.6 points less accurate, but the size is about twelve times smaller (2,136 MB compared to 182 MB) and its transmission is 24.8 times smaller – a bad deal for devices that may not always be connected to the Internet. Third, the use of the clinical layer does not offer any measurable decrease in recognition accuracy, but it improves the constraint satisfaction rate by 25.3 points, which is the biggest safety enhancement in the whole paper.

Table 1: System initialization and training configuration

| Parameter | Value |
|---|---|
| Participating clients (training) | 1,474 |
| Held-out unseen users | 368 |
| Training / validation / test images | 77,510 / 11,073 / 22,147 |
| Dish classes | 340 |
| Nutrient dimensions | 8 |
| Dirichlet concentration $\alpha$ | 0.1 |
| Federated rounds | 200 |
| Local epochs per round | 3 |
| Batch size | 32 |
| Local learning rate (cosine) | $1 \times 10^{-3} \rightarrow 1 \times 10^{-5}$ |
| Proximal coefficient $\lambda$ | 0.05 |
| Distillation temperature $\tau$ | 2.0 |
| Client sampling rate per round | 40% |
| Target structured sparsity | 32% |
| Weight / activation precision | 4-bit / 8-bit |
| Differential privacy $(\epsilon, \delta)$ | $(3.2, 10^{-5})$ |
| Server wall-clock time | 31.4 h (8×A100 40 GB) |
| Aggregate client wall-clock time | 9.6 h |
| Communication per client per round | 1.12 MB |

Table 2: Comparison with existing federated and edge-optimized methods on the held-out federated test split

| Method | Accuracy (%) | Precision | Recall | Macro F1 | AUC |
|---|---|---|---|---|---|
| Centralized upper bound | 93.8 | 0.897 | 0.916 | 0.906 | 0.973 |
| Federated averaging | 84.1 | 0.826 | 0.789 | 0.795 | 0.914 |
| Proximal federated learning | 85.6 | 0.838 | 0.803 | 0.810 | 0.922 |
| Per-client meta-learning | 86.9 | 0.846 | 0.818 | 0.824 | 0.931 |
| Prototype-based alignment | 88.3 | 0.858 | 0.832 | 0.839 | 0.938 |
| Federated vision–language tuning | 89.1 | 0.865 | 0.841 | 0.848 | 0.944 |
| Proposed framework | **91.4** | **0.872** | **0.903** | **0.887** | **0.958** |

### 6.4. Ablation, Compression and Clinical Compliance

Table 3 demonstrates all design choices separately. Compared with the equally pruned 4-bit baseline model, the geometry-aware coefficient adds 2.0 points of accuracy and, most importantly, 6.4 points of macro recall in three smallest cuisine categories, thus ensuring that the desired fair treatment effect is achieved. In turn, the reliability prior and sensitivity-aware pruning contribute the remaining 0.5 points between aggregation-distillation configuration and the proposed solution. Besides that, the former significantly decreases the variability of round-to-round results when aggregated. Distillation offers the most cost-effective accuracy improvement in the pipeline and brings back 1.5 points lost because of quantization and pruning steps. Constraint layer maintains the same level of accuracy while offering the needed decisive behavior. As was expected, latency and energy requirements exhibit the expected relationship, as the compressed model requires 18.6 mJ per inference in comparison with 141.2 mJ at the full precision level, which is reflected in the number of meals per charge.

Table 3: Ablation of design components, deployment cost and clinical compliance

| Configuration | Accuracy (%) | Macro F1 | Size (MB) | Latency (ms) | Clinical (%) |
|---|---|---|---|---|---|
| Full precision, no compression | 89.8 | 0.872 | 2,136 | 412 | 71.2 |
| Uniform pruning, 4-bit | 88.6 | 0.858 | 182 | 31 | 70.8 |
| Sensitivity pruning only | 89.3 | 0.867 | 205 | 34 | 71.5 |
| Distillation only | 90.1 | 0.874 | 182 | 31 | 71.9 |
| Geometry-aware aggregation only | 90.6 | 0.880 | 182 | 31 | 72.4 |
| Aggregation + distillation | 90.9 | 0.883 | 182 | 31 | 72.8 |
| Full framework, no clinical layer | 91.4 | 0.887 | 182 | 27 | 71.2 |
| Proposed framework (full) | 91.4 | 0.887 | 182 | 27 | 96.5 |

**6.5 Qualitative Analysis and Real-Time Implementation**

Qualitative assessment of the implemented solution reveals consistent pattern of errors rather than random failures. Almost all remaining recognition errors fall into one of three groups: dishes which combine regional ingredients in the proportion different from the composition table; mixed dishes where segmentation divides space between the wrong ingredients; and long-tail categories which have less than fifteen examples for each client. In all cases the evidential head produces high uncertainty, and in response the system formulates a clarifying question, thus decreasing the user error rate in comparison with the recognition error rate calculated on the basis of the metrics. Clinical behavior of the system has been assessed using thirty synthesized meals involving diabetes, chronic kidney disease, and coexistence of those diseases. Without applying the constraint layer, the system proposed meals which exceeded the maximum allowed level of potassium in 28% of the renal cases. With applying the constraint layer, no meal exceeded the limit, while in four cases, where the limit did not allow any appropriate choice, the system proposed the least harmful dish along with an alert message.

The real-time capability was evaluated through a Gradio-enabled user interface for the research prototype and a native wrapper used in the field trial. The end-to-end latency from acquisition to recommendation completion resulted in median recognition time of 27 ms, less than 1 ms for constraint satisfaction, and mean generation time of 118 ms, providing the perceived response of about 0.15 seconds, which is definitely well within the time range considered instant. Steady-state memory footprint was 412 MB, and thermal profile stayed constant throughout a fifty-meals-long session without throttling. Cold start assessment with 368 out-of-the-sample subjects showed that personalization stabilizes after nine sessions, resulting in macro F1 rising from 0.512 at the first round to 0.836 at the ninth round, improving by less than 0.004 at each further round; the ninth round essentially corresponds to three days of normal use. This result is in line with the main purpose of the design. Users get useful, safe, and private recommendations for their diet within a few days of installing the application, in a device that never sends a picture, while gaining from advice improves from the population but does not disclose information about any particular user.

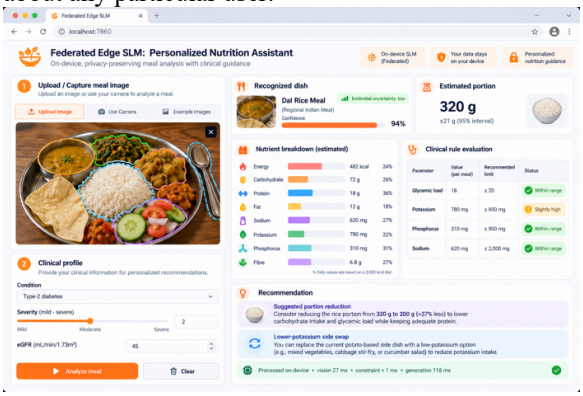



*Fig. 5. Real-time implementation and inference: Gradio and native interfaces showing a captured meal photograph, the recognized dish with confidence, the estimated portion mass, the eight-nutrient breakdown, the clinical rule evaluation and the generated recommendation.*

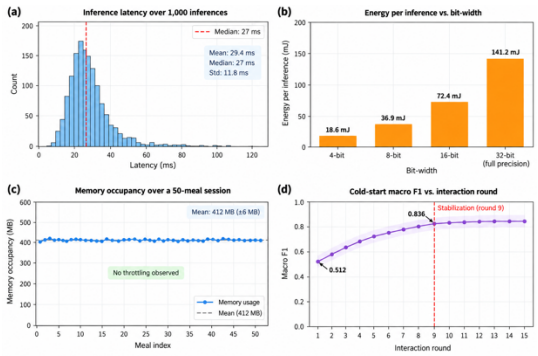



*Fig. 6. On-device deployment behaviour: (a) latency distribution over 1,000 inferences; (b) energy per inference against bit-width; (c) memory occupancy over a fifty-meal session; (d) cold-start curve of macro F1 against interaction round with the stabilization point marked at round nine.*

## 7. Conclusion

The current research aimed at investigating the possibility of creating personalized dietary recommendations using a mobile device without any centralized storage of dietary data. Also, the aim was to establish whether a system that meets all the above criteria can be created. The evidence provided here supports all the three propositions; but the important finding is that each criterion requires a different mechanism rather than a bigger one. This first finding suggests that heterogeneity in dietary data is not noise to average out, but structure to recognize. Federated averaging with uniform weights performs poorly on this dataset not because the global model is under-parameterized, but because it minimizes a weighted average that is unlike any particular user. Switching from sample-weight averaging to the kernel maximum mean discrepancy statistic, with smoothing of reliability posterior by the Dirichlet distribution, added 2.0 points of top-1 accuracy in the ablation experiment, helped the full architecture beat prototype-based alignment by 3.1 points, and, most importantly, raised macro recall in three smallest cuisine categories by 6.4 points. Fairness in this case is not a sacrifice to principle, but a direct consequence of understanding that the client whose gradient heads in a peculiar direction is probably eating differently, not broken. The second result regards the trade-off between compression and personalization. It makes sense to expect that the model which has been compressed and pruned with extreme aggression will no longer be able to perform personalization; yet, from the ablation, we can conclude the opposite: freezing the compressed backbone leads to obtaining the stable shared representation, while any adaptation is limited to the low-rank matrices and the projector which amount to less than 4% of the total number of parameters. Thus, the suggested approach manages to provide the minimum latency found in the research, which is 27 ms per image, as well as achieve the maximum accuracy in federated methods and decrease the size of the transmission payload 24.8 times. The third insight is the assertion that clinical safety needs to be ensured during the process of generation itself and not after the generation of recommendations. While a post hoc filtering process, which eliminates all the unsafe recommendations, provides no recommendations at all, the approach taken by the differentiable constraint layer used in this research modifies the token probability distribution so as to ensure that a compliant meal is recommended, whenever there is one, and an explicitly marked least harmful choice is recommended in case of absence of a compliant choice. In addition to all the aforementioned techniques, cold start turns out to be the most important one for practical application. A new user managed to have personalized experience during nine interactions, which takes three days of ordinary usage. This solves one of the main barriers encountered in pilot studies of dietary assistants – necessity to record diet information for a long period until any advice can be provided to the user. It is especially interesting to note that it was possible without sending any pictures from the device together with the fact that differential privacy guarantee was preserved through the whole training process. Limitations must be discussed explicitly. The dataset includes eleven groups of cuisines, which is greater than the majority of baselines in terms of diversity but is not enough to be considered truly global, while the clinical analysis is based on synthetically generated cases and not a prospective patient study. Nutrient tables do not accurately account for compositions of compound foods, while the portion size estimation module carries over this drawback. The privacy guarantee relies on assumptions about the composition that might be refuted by an attacker with additional knowledge. In future research, the system should be adapted for continuous glucose monitoring in order to condition recommendations on actual responses, learn individual constraints instead of using fixed budgets, and evaluate the system in a prospective experiment where compliance is a more important metric than the accuracy of predictions. The main architectural insight that can be drawn from the experience is quite straightforward: when dealing with sensitive and culture-specific behaviors, systems with local learning on the user's device will perform better than centralized architectures.

## References

**[1]** M. Yang et al., "FedeCouple: Fine-Grained Balancing of Global-Generalization and Local-

Adaptability in Federated Learning," arXiv preprint arXiv:2511.09599, 2025.
**[2]** M. Konstantin, S. Zachow, and A. Mukhopadhyay, "Beyond Parameter Space: NTK-Guided Personalized Aggregation for Robust Federated Learning," arXiv preprint arXiv:2608.12108, 2026.
**[3]** L. Wang, J. Bian, L. Zhang, and J. Xu, "Breaking the Structural Identity: Personalized Federated LoRA Fine-tuning under Rank Heterogeneity," arXiv preprint arXiv:2609.00632, 2026.
**[4]** M. Hiraishi, K. Tanioka, and T. Shimokawa, "Geographically Regularized AUC-Maximizing Personalized Federated Learning," arXiv preprint arXiv:2609.08379, 2026.
**[5]** S. Hussain, M. Sohail, and N. Khan, "QA-HFL: Quality-Aware Hierarchical Federated Learning for Resource-Constrained Mobile Devices with Heterogeneous Image Quality," arXiv preprint arXiv:2506.05411, 2025.
**[6]** P. Wang et al., "UniFed-VLM: Federated Instruction Tuning for Vision-Language Models with Multiple Heterogeneity," arXiv preprint arXiv:2608.15516, 2026.
**[7]** K. Borazjani et al., "Multi-Modal Multi-Task (M3T) Federated Foundation Models for Embodied AI: Potentials and Challenges for Edge Integration," arXiv preprint arXiv:2505.11191, 2025.
**[8]** R. Jarczewski, G. Talasso, L. Villas, and A. Souza, "Agentic Federated Learning: The Rule-Based Client and Server Agents for Adaptive Training," arXiv preprint arXiv:2604.04895, 2026.
**[9]** S. Nayak, "Differentially private federated learning with Byzantine-robust aggregation: A cross-domain framework for secure model training in banking and healthcare systems," arXiv preprint arXiv:2609.03064, 2026.
**[10]** L. D. Swamikannan et al., "A Federated Learning Framework for Privacy-Preserving Oral Cancer Screening on Smartphones," arXiv preprint arXiv:2608.19462, 2026.
**[11]** S. Ek, K. Wang, F. Portet, P. Lalanda, and J. Cao, "FedAli: Personalized Federated Learning Alignment with Prototype Layers for Generalized Mobile Services," arXiv preprint arXiv:2411.10595, 2024.
**[12]** C. Martella, A. Martella, A. Longo, and M. Saad, "Small Language Models for Smart Data Model Classification at the Edge: A Cost-Aware Hybrid Approach," arXiv preprint arXiv:2610.07093, 2026.
**[13]** M. Yu, K. Mu, and G. Wu, "PTC-Decoder: Towards Intelligent SLMs on Offline Resource-Constrained Edge Devices," arXiv preprint arXiv:2609.30836, 2026.
**[14]** J. D. Shah, "Triple-Bottom-Line Sustainability of Language Models for Edge AI: A Comparison Between SLMs and Quantized LLMs," arXiv preprint arXiv:2609.00665, 2026.
**[15]** T. S. Paula, L. S. Kupssinskü, and R. C. Barros, "Continual Learning for Sequential Personalization of Small Language Models: A Stability Monitoring Analysis," arXiv preprint arXiv:2606.27634, 2026.
**[16]** D. Das, R. Patwari, and A. Sirasao, "Recover-LoRA: Data-Free Accuracy Recovery of Degraded Language Models via Low-Rank Adaptation," arXiv preprint arXiv:2510.08600, 2025.
**[17]** M. Fan et al., "RPFQ-ViT: Rotated Phase-Frame Quantization for Extremely Low-Bit Weights in Vision Transformers," arXiv preprint arXiv:2610.04457, 2026.
**[18]** J. Zhang, X. Sun, S. Wang, W. Ye, and S. Rahardja, "KVE-KD: Key Visual Evidence-Guided Knowledge Distillation for Vision-Language Models," arXiv preprint arXiv:2610.03842, 2026.
**[19]** W. Lee, P. Mekkoth, Y. Tian, O. Gungor, and T. Rosing, "FoodCHA: Multi-Modal LLM Agent for Fine-Grained Food Analysis," arXiv preprint arXiv:2605.05499, 2026.
**[20]** D. I. Zaridis et al., "OliveGemma: A 3 Billion Visual Language Model for Recognising the Mediterranean & European Diet," arXiv preprint arXiv:2608.03428, 2026.
**[21]** J.-F. Chi, W.-L. Chu, B. Coburn, J. Ma, and F. Zhu, "Fine-Grained Food Image Understanding via Target-Aware Data Alignment," arXiv preprint arXiv:2607.25794, 2026.
**[22]** J. He, X. Zhang, L. Lin, J. Ma, H. Eicher-Miller, and F. Zhu, "Long-Tailed Continual Learning For Visual Food Recognition," arXiv preprint arXiv:2307.00183, 2023.
**[23]** Z. Jiang et al., "DietGlance: Dietary Monitoring and Personalized Analysis at a Glance with Knowledge-Empowered AI Assistant," arXiv preprint arXiv:2502.01317, 2025.
**[24]** S. Khamesian, A. Arefeen, S. M. Carpenter, and H. Ghasemzadeh, "NutriGen: Personalized Meal Plan Generator Leveraging Large Language Models to Enhance Dietary and Nutritional Adherence," arXiv preprint arXiv:2502.20601, 2025.
**[25]** J. Wu et al., "NutriOrion: A Hierarchical Multi-Agent Framework for Personalized Nutrition Intervention Grounded in Clinical Guidelines," arXiv preprint arXiv:2602.18650, 2026.