This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

ProMoE-FL: Prototype-conditioned Mixture of Experts for Multimodal Federated Learning with Missing Modalities

Aavash Chhetri[1], Bibek Niroula[1], Eduard Vazquez[2], Yash Raj Shrestha[3], Prashnna Gyawali[4], Loris Bazzani[5], and Binod Bhattarai[1,6,7]

[1] NepAl Applied Mathematics and Informatics Institute for research, Nepal
[2] Fogsphere (Redev.AI Ltd), UK
[3] University of Lausanne, Switzerland
[4] West Virginia University, USA
[5] University of Verona, Italy
[6] University College London, UK
[7] University of Aberdeen, UK
binod.bhattarai@abdn.ac.uk

Abstract. In this paper, we address the problem of multimodal federated learning with missing modality. Existing methods utilize an additional public dataset or perform naive feature synthesis that is based solely on the available modality. To address these limitations, we propose ProMoE-FL, a Prototype-conditioned Mixture-of-Experts framework for robust missing-modality feature synthesis in multimodal federated learning. ProMoE-FL builds a global client-aware prototype bank that captures clinically meaningful modality priors across institutions. Our Mixture of Experts is conditioned on these prototypes and modality indices to enable direction-aware expert routing for dynamically synthesizing missing features. We perform extensive quantitative and qualitative evaluations on four public chest X-ray datasets (MIMIC-CXR, NIH Open-I, PadChest, and CheXpert) and demonstrate that ProMoE-FL consistently outperforms state-of-the-art methods in both homogeneous as well as the more challenging heterogeneous settings. Code available at: https://github.com/bhattarailab/ProMoE-FL.

Keywords: Multimodal Federated Learning · Missing Modality · Prototype Learning.

1 Introduction

Clinicians inherently integrate information from a variety of sources and modalities such as pathology reports, laboratory data, omics data, X-Ray, Magnetic Resonance Imaging (MRI), and Computed Tomographhy (CT) to form holistic and context-aware clinical judgements. This has motivated the recent development of models that make use of highly multimodal data to computationally mirror such integrative clinical reasoning [1,2,29]. In healthcare, multimodal data

is inherently fragmented across medical centers due to privacy regulations and data governance constraints, posing significant challenges to centralized training paradigms. Federated Learning (FL) [21] addresses these challenges and provides a privacy-preserving approach to collectively training shared healthcare models without any exchange or distribution of patient data residing in medical centers. Recent efforts have therefore explored multimodal learning under the federated setting [25,24,5]. However, most of the existing literature on multimodal federated learning in healthcare assumes that all participating institutions have access to complete sets of modalities [31,7]. In real-world settings, multimodal clinical data are frequently incomplete across centres owing to disparities in equipment availability, acquisition protocols, and data management resources. This challenge calls for novel methodologies for dealing with missing modalities. In centralized multimodal learning, the challenge of missing modalities is addressed through prompt-based [28,33], generative [6,18,8], dropout-based [16] methods, or utilize specialized architectures [35,32,20] to model inter-modal dependencies. Such centralized approaches are impractical in FL scenarios, as their architectural designs do not consider the strict privacy, communication efficiency, and heterogeneity requirements inherent to practical FL deployments.

More recently, a growing body of FL work has begun to address the challenge of missing modalities, predominantly in non-medical domains [34,17,3,30], while a limited number of studies were proposed for the medical domain [23,22,26,27]. CAR-MFL [23] handles missing modalities by training with additional public multimodal data to compensate for incomplete client modalities. However, this strategy assumes the availability of representative public data, which is often unrealistic in healthcare due to privacy constraints and domain gaps between public and private datasets. FeatImp [22] performs feature-level imputation by training separate networks to learn modality-specific mappings to reconstruct missing features. This approach assumes stable cross-modal relationships across clients and synthesizes features solely from the observed modality, without explicitly modeling complementary information inherent to the missing modality. PmcmFL [3] substitutes class-level prototypes as features of missing modalities. Although prototype-based and closer to our formulation, it assigns identical representations to all samples within a class which limits instance-level diversity and fails to capture intra-class variability, which is critical in medical data.

To address these limitations, we propose ProMoE-FL, a prototype-conditioned Mixture-of-Experts framework for missing modality synthesis in federated learning. The method generates modality-specific features by conditioning on available modalities and a global prototype bank of missing modalities. This enables privacy-preserving and structured feature reconstruction without external data. A shared, routing-based Mixture-of-Experts (MoE) architecture allows experts to be reused across modalities, preventing combinatorial parameter growth. Furthermore, by accumulating client-specific prototypes that summarize local modality distributions, the global prototype bank serves as a client-aware prior in the latent space. This guides feature synthesis to account for distributional variability across clients, thereby improving robustness under the non-independently

and identically distributed (i.e., non-IID) data characteristics prevalent in real-world federated settings. Our key contributions are the following:

- We propose ProMoE-FL, a prototype-conditioned feature generation framework for missing modality feature synthesis in federated multimodal learning.
- We introduce client-aware prototype attention to address cross-client heterogeneity and employ a MoE architecture for direction-aware synthesis within a shared parameterization.
- We conduct extensive experiments under heterogeneous settings that mirror real-world clinical data silos, demonstrating improved robustness. We further support our findings with qualitative analyses.

2 Method

We formulate our method for a typical multimodal federated setting with K clients, each with its private dataset D_k containing n_k samples. The n-th data sample in D_k is represented by the tuple $(\{X_m^{(n)}\}_{m=1}^{M_k}, Y^{(n)})$, where $Y^{(n)}$ denotes the label related to C classes and M_k is the number of modalities available at the k-th client. All clients share common architecture with the global model, which is defined as $\{\boldsymbol{w} = \{f_e^m\}_{m\in\mathcal{M}},\ f_c\}$, where $\mathcal{M}$ denotes the complete set of modalities, f_e^m is the encoder for modality m, and f_c is the classifier. We denote the latent feature representation extracted by the modality specific encoder as $h_m^{(n)} = f_e^m(X_m^{(n)}), \quad m \in \mathcal{M}$. The objective of federated training is to minimize the data-weighted empirical risk across all the clients, which is expressed as:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum_{k=1}^{K} \frac{|D_k|}{|D|} \mathcal{L}_{task}^{(k)}(\boldsymbol{w}), \tag{1}$$

with the local loss at client k:

$$\mathcal{L}_{task}^{(k)}(\boldsymbol{w}) = \frac{1}{|D_k|} \sum_{n=1}^{|D_k|} \mathcal{L}_{task}\Big(f_c(\bigoplus_{m\in\mathcal{M}} f_e^m(X_m^{(n)})), Y^{(n)}\Big), \tag{2}$$

where, $\oplus$ denotes a feature fusion operator, implemented as concatenation of modality-specific latent representations, and $D = \bigcup_{k=1}^{K} D_k$. In practical multi-center clinical deployments, multimodal systems frequently face missing modalities. A common approach to handle missing modality is zero-filling. For example, if the image modality is unavailable while text is present, the objective reduces to $\mathcal{L}_{task}\big(f_c(0 \oplus h_T^{(n)}), Y^{(n)}\big)$. Naively optimizing it distorts the joint feature space and biases the global model toward frequently observed modalities, thereby under-representing sparse ones.

Fig. 1 presents the overall framework of ProMoE-FL. For unimodal clients, a feature synthesis network reconstructs representations of missing modalities. While our approach operates in the feature space similarly to [22], ProMoE-FL differs in both design and robustness. Learning modality mappings solely

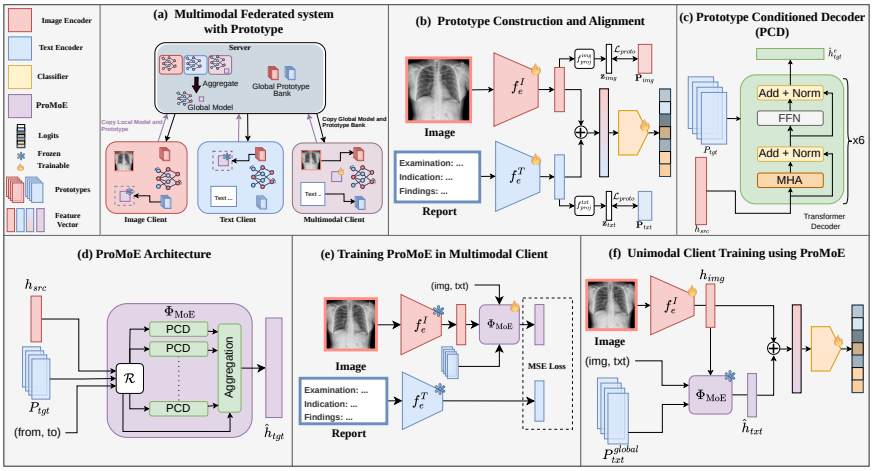

Fig. 1. Overview of the proposed ProMoE-FL framework.

from multimodal clients can degrade performance under client heterogeneity, where cross-modal relationships vary across sites and available modalities provide incomplete priors. To mitigate this, we introduce a client-aware global prototype bank encoding federation-wide modality priors, conditioning the synthesis process on target-modality prototypes. Furthermore, to ensure scalability without combinatorial parameter growth, the synthesis module is implemented as a Mixture-of-Experts architecture that dynamically routes modality-specific transformations through expert subnetworks.

Prototype Construction and Alignment. In heterogeneous federated settings, modality embeddings often diverge across clients due to distribution shifts and inconsistent modality availability [19]. As a result, mappings learned only from multimodal clients may be insufficient for reliable cross-modal synthesis. Inspired by [17,3], we introduce modality-specific, client-aware prototypes that act as global priors of the missing modality to complement the available modality and improve robustness to cross-client heterogeneity. For each modality m, every client k maintains a set of C learnable prototypes $\mathbf{P}_m^k = \{\mathbf{p}_1^{(m,k)}, \ldots, \mathbf{p}_C^{(m,k)}\}$ with $\mathbf{p}_c^{(m,k)} \in \mathbb{R}^d$. For each sample with at least one positive label, the modality-specific target centroid $\mathbf{c}_m^{(n)}$ is obtained by aggregating the corresponding class prototypes weighted by the multi-hot label vector and normalizing the result to unit norm. For each modality m, we obtain a projected representation $\mathbf{z}_m^{(n)} = f_{\text{proj}}^m(\mathbf{h}_m^{(n)}) \in \mathbb{R}^d$ using a modality-specific projection head implemented as a Fully Connected (FC) layer, and it is aligned with its corresponding centroid via $\mathcal{L}_{proto}^{(m)} = \frac{1}{|\mathcal{I}|} \sum_{n \in \mathcal{I}} \|\mathbf{z}_m^{(n)} - \mathbf{c}_m^{(n)}\|_2^2$, where $\mathcal{I} = \{n \mid \sum_{c=1}^{C} y_c^{(n)} > 0\}$. As illustrated in Fig. 1(b), we jointly optimize the client parameters $\mathbf{w}^k$ together with modality-specific projection heads and prototypes, $\{f_{\text{proj}}^m, \mathbf{P}_k^m\}_{m \in \mathcal{M}_k}$, where $\mathcal{M}_k \subseteq \mathcal{M}$

denotes the set of modalities available on client k. This is done via the following full training objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{m \in \mathcal{M}_k} \mathcal{L}_{\text{proto}}^{(m)}, \tag{3}$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss (Eq. 2), $\mathcal{L}_{\text{proto}}^{(m)}$ is the prototype alignment loss for modality m, and λ controls the relative importance of the prototype loss. The prototypes are updated through gradient-based optimization each communication round, which encourages alignment with class representations while maintaining diversity, hence, improving robustness in heterogeneous settings.

Missing Modality Feature Synthesis. For unimodal clients, we propose a *Prototype-Conditioned Decoder* (PCD) ϕ, illustrated in Fig. 1(c), to enable cross-modal feature synthesis. The PCD is implemented as a Transformer decoder where the source feature h_i of available modality i attends over the global prototype bank P_j of the target modality j received from the server to retrieve context necessary for synthesizing the target modality feature. Importantly, P_j is constructed via accumulation of prototypes from clients rather than aggregation (e.g., averaging). Consequently, the decoder can dynamically attend to the most relevant target prototypes during synthesis, enabling adaptive complementary knowledge retrieval across heterogeneous domains.

However, a single PCD may be insufficient to capture complex, non-linear relationships between heterogeneous modality pairs, especially in highly multimodal scenarios in the medical domain [7]. Thus, we propose a Mixture of Experts Φ_{MoE}, illustrated in Fig. 1(d) where we define E PCDs $\phi_e(h_i, P_j)$ as independent experts $\{\phi_e\}_{e=1}^{E}$. In order to redirect the data flow via the right experts, a Modality-Aware Router $\mathcal{R}$ computes a gating distribution conditioned on the instance context and the modality indices. Therefore, the network synthesizes the bottleneck feature of the missing modality j as:

$$\hat{h}_j^{(n)} = \Phi_{\text{MoE}}(h_i^{(n)}, P_j; i, j) = \frac{\sum_{e=1}^{E} \mathcal{R}(h_i^{(n)}, P_j, i, j)_e \; \phi_e(h_i^{(n)}, P_j)}{\left\| \sum_{e=1}^{E} \mathcal{R}(h_i^{(n)}, P_j, i, j)_e \; \phi_e(h_i^{(n)}, P_j) \right\|_2}. \tag{4}$$

This formulation allows Φ_{MoE} to adaptively select the optimal combination of prototype-conditioned experts for any missing modality scenario. Although the proposed framework is formulated for an arbitrary set of modalities $\mathcal{M}$, for clarity of exposition we specialize the following derivations to the two-modality case, $\mathcal{M} = \{I, T\}$ used in our experiments, namely Image (I) and Text (T), without loss of generality. For training unimodal clients, as shown in Fig. 1(f), the task-calibrated loss is hence, defined as:

$$\mathcal{L}_{task} = \mathcal{L}\Big(fc\Big(h_i^{(n)} \oplus \hat{h}_j^{(n)}\Big), Y^{(n)}\Big), \qquad \hat{h}_j^{(n)} = \Phi_{\text{MoE}}(h_i^{(n)}, P_j; i, j) \tag{5}$$

where, $i, j \in \mathcal{M},\ i \neq j$.

Federated Training in ProMoE-FL. At the start of each communication round, the server dispatches Φ_{MoE} , $\boldsymbol{w}$, and prototype bank $\{P_m^{global}\}_{m \in \mathcal{M}}$ to all

clients. Each client performs local training for N epochs, optimizing the training objective $\mathcal{L}$ as defined in Eq. 3. On multimodal clients, we construct a pool of paired features $\mathcal{H}_k = \{(h_I^{(n)}, h_T^{(n)}) \mid (x_I^{(n)}, x_T^{(n)}) \in D_k\}$. As illustrated in 1(e), this pool is then used to train the feature synthesis network Φ_{MoE}^k using a symmetric reconstruction objective:

$$\mathcal{L}(\Phi_{MoE}^k) = \frac{1}{|\mathcal{H}_k|} \sum_{(h_I^{(n)}, h_T^{(n)}) \in \mathcal{H}_k} \sum_{\substack{i,j \in \mathcal{M}_k \\ i \neq j}} \left\| \Phi_{MoE}^k(h_i^{(n)}, P_j; i, j) - h_j^{(n)} \right\|_2^2. \quad (6)$$

At the end of each communication round, each client transmits its updated model parameters $\boldsymbol{w}^k$, the modality-specific projection heads, and the local prototype sets $\mathbf{P}_m^k$ to the server. Multimodal clients additionally communicate Φ_{MoE}^k. The resulting communication overhead is negligible, as the prototype bank, projection, and modality-aware routing mechanism together account for only approximately 0.3% of the total model parameters. All the communicated parameters are aggregated via FedAvg [21] due to its simplicity and effectiveness as demonstrated by prior studies [22,23,26]. In contrast to the model parameters, the server accumulates all client prototypes into a global prototype bank: $\mathbf{P}_m^{global} = \bigcup_{k=1}^{K} \mathbf{P}_m^k$ for each modality $m \in \mathcal{M}$.

3 Experiments and Results

Datasets and Setups. We perform a thorough experimentation on challenging datasets in both homogeneous and heterogeneous setups: MIMIC-CXR [14,11], NIH Open-I [9], PadChest [4], and CheXpert [13]. We use the validation and test splits provided by MIMIC-CXR for the evaluation of the global model, following the protocol outlined in [23]. We experimented with a variety of client configurations, denoted as **I:T:M**, where I, T, and M represent the numbers of image-only, text-only, and multimodal clients, respectively. The homogenous setup comprises 10 clients, each with training data for 810 patients sampled from the same dataset MIMIC-CXR. In the heterogeneous setup, multimodal clients include 2 clients from NIH Open-I, while any additional multimodal clients are sampled from PadChest. All image-only and text-only clients are drawn from CheXpert and PadChest, respectively.

Implementation Details. We employ pretrained Resnet50 [12] and BERT-base [10] architecture as the image and text encoders, respectively. Their outputs are projected to 256-dimensional features and L2-normalized prior to concatenation. Optimization is performed using Adam [15] with a learning rate of $1e^{-4}$. λ is tuned from $\{0.1, 0.5, 1, 5, 10\}$. Each client trains locally for 3 epochs per communication round over 30 rounds in total. Performance is measured using macro AUC (Average area under the Receiver Operating Characteristic Curve). The results are averaged over three random seeds.

Baselines. We compare our method with two simple yet effective baselines: Zero-filling which replaces missing modalities with zero vectors and Uniform-filling which uniformly samples feature vectors (FedAvg [21] + Zero/Unif.). We

Table 1. AUC↑ Performance in Homogeneous and Heterogeneous partitions. Client configurations are denoted by I:T:M, where I, T and M denote the number of Image-only, Text-Only and Multimodal Clients Respectively. Additionally, experiments conducted with public data (Pub.) are also included.

	Method	Partitions (I:T:M)							
		4:0:6	0:4:6	2:2:6	6:0:4	0:6:4	3:3:4	8:0:2	0:8:2	4:4:2
Homogeneous	FedAvg [21] + Zero	86.64$_{\pm 0.46}$	87.95$_{\pm 0.53}$	87.40$_{\pm 0.54}$	82.99$_{\pm 1.03}$	87.01$_{\pm 0.33}$	84.48$_{\pm 0.61}$	79.69$_{\pm 1.61}$	86.32$_{\pm 0.52}$	80.47$_{\pm 0.56}$
	FedAvg [21] + Unif.	87.79$_{\pm 0.63}$	89.27$_{\pm 0.56}$	88.49$_{\pm 0.56}$	84.83$_{\pm 0.32}$	88.64$_{\pm 0.25}$	86.90$_{\pm 0.25}$	80.60$_{\pm 0.91}$	88.37$_{\pm 0.34}$	85.74$_{\pm 0.28}$
	FeatImp [22]	89.31$_{\pm 0.33}$	**89.52**$_{\pm 0.11}$	89.12$_{\pm 0.55}$	87.61$_{\pm 0.12}$	**89.30**$_{\pm 0.42}$	88.12$_{\pm 0.47}$	86.16$_{\pm 0.44}$	**88.98**$_{\pm 0.28}$	88.79$_{\pm 0.12}$
	PmcmFL [3]	84.54$_{\pm 0.17}$	83.76$_{\pm 0.92}$	84.78$_{\pm 0.24}$	81.72$_{\pm 0.38}$	84.21$_{\pm 0.46}$	83.07$_{\pm 0.21}$	77.03$_{\pm 0.70}$	81.19$_{\pm 0.50}$	82.56$_{\pm 0.24}$
	ProMoE-FL (ours)	**89.44**$_{\pm 0.20}$	89.17$_{\pm 0.43}$	**89.41**$_{\pm 0.20}$	**88.27**$_{\pm 0.75}$	89.13$_{\pm 0.29}$	**89.25**$_{\pm 0.27}$	**87.46**$_{\pm 0.47}$	88.51$_{\pm 0.58}$	**89.41**$_{\pm 0.62}$
Heterogeneous	FedAvg [21] + Zero	79.32$_{\pm 1.24}$	74.93$_{\pm 0.86}$	76.30$_{\pm 1.06}$	78.95$_{\pm 0.55}$	74.41$_{\pm 0.33}$	74.98$_{\pm 0.70}$	72.76$_{\pm 0.78}$	73.34$_{\pm 0.86}$	73.25$_{\pm 0.30}$
	FedAvg [21] + Unif.	78.55$_{\pm 1.88}$	76.58$_{\pm 0.63}$	77.74$_{\pm 1.00}$	77.35$_{\pm 0.64}$	76.48$_{\pm 1.01}$	77.81$_{\pm 0.77}$	71.59$_{\pm 0.28}$	75.63$_{\pm 1.07}$	76.69$_{\pm 0.81}$
	FeatImp [22]	80.19$_{\pm 0.92}$	76.09$_{\pm 0.37}$	78.42$_{\pm 0.62}$	80.93$_{\pm 0.53}$	77.14$_{\pm 0.59}$	78.56$_{\pm 1.49}$	77.15$_{\pm 1.34}$	76.78$_{\pm 0.65}$	80.06$_{\pm 1.45}$
	PmcmFL [3]	74.66$_{\pm 0.99}$	72.10$_{\pm 0.27}$	76.99$_{\pm 0.45}$	71.01$_{\pm 0.40}$	68.39$_{\pm 0.16}$	75.96$_{\pm 1.27}$	67.51$_{\pm 0.79}$	64.65$_{\pm 0.51}$	71.48$_{\pm 0.67}$
	ProMoE-FL (ours)	**81.91**$_{\pm 0.83}$	**78.28**$_{\pm 0.32}$	**79.88**$_{\pm 0.43}$	**83.02**$_{\pm 0.26}$	**77.96**$_{\pm 1.54}$	**80.29**$_{\pm 1.35}$	**79.68**$_{\pm 1.52}$	**78.18**$_{\pm 0.07}$	**80.96**$_{\pm 0.72}$
Pub.	CAR-MFL [23]	87.87$_{\pm 0.12}$	87.65$_{\pm 0.31}$	**88.95**$_{\pm 0.33}$	87.66$_{\pm 0.45}$	86.61$_{\pm 0.14}$	88.29$_{\pm 0.55}$	85.93$_{\pm 0.24}$	**86.32**$_{\pm 0.54}$	88.19$_{\pm 0.17}$
	ProMoE-FL (ours)	**89.16**$_{\pm 0.36}$	**87.76**$_{\pm 0.01}$	87.98$_{\pm 1.50}$	**89.79**$_{\pm 0.09}$	**88.10**$_{\pm 0.71}$	**88.86**$_{\pm 0.57}$	**89.28**$_{\pm 0.51}$	85.90$_{\pm 0.33}$	**89.19**$_{\pm 0.68}$

compare with state-of-the-art (SOTA) methods: FeatImp [22] and PmcmFL [3], representing feature-level imputation and prototype-based approaches, respectively. One of the advantages of our ProMoE-FL is that it does not rely on public data like CAR-MFL [23]. However, ProMoE-FL can be easily adapted by incorporating a virtual client that uses public data during training. In this way, we can fairly compare with public-data-based SOTA, CAR-MFL [23].

Quantitative Results. Table 1 reports results across client configurations under both homogeneous and heterogeneous settings. In the homogeneous case, our method achieves the best performance in 6 configurations, with only a modest average gap of 0.37% in the remaining 3. Under heterogeneous distributions, which represent a more realistic federated scenario, ProMoE-FL consistently outperforms all baselines across every configuration. Notably, PmcmFL shows the largest degradation, likely due to its reliance on a single class-wise prototype for missing modalities, which limits adaptability under distributional shifts. Finally, in the heterogenous settings, we compare our method with CAR-MFL[23], SOTA for Public dataset based approach. The results demonstrate that our method achieves competitive performance. We also ablate the mixture-of-experts component in the 8:0:2 heterogeneous setting. A single PCD achieves an AUC of 78.98, whereas the MoE variant attains 79.68, indicating a performance gain from incorporating expert routing.

Qualitative Results. Figure 2(a–b) shows t-SNE analyses of synthesized features from the test set of ProMoE-FL, FeatImp and the ground truth. FeatImp demonstrates mild centroid offsets and slightly more diffuse feature dispersion, indicating comparatively less precise alignment in the latent space. In contrast, ProMoE-FL shows a tighter integration between synthesized and ground-truth features, with a more precise alignment of their class centroids. These observations suggest that ProMoE-FL provides a more refined and structurally faithful

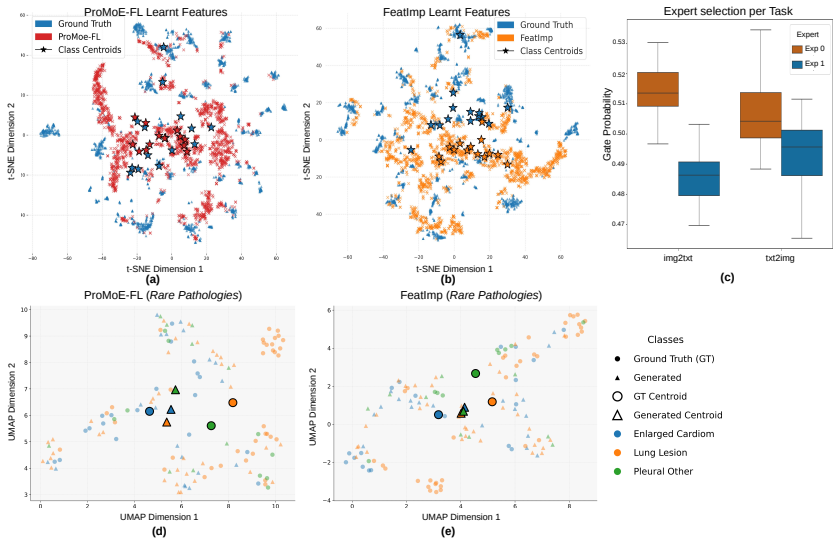

Fig. 2. Qualitative Analysis of the methods. (a–b) t-SNE of learnt features with class centroids. (c) Gating distributions across experts. (d–e) UMAP of rare pathological classes.

semantic mapping. As shown in Figure 2(c), expert utilization remains balanced across both synthesis directions, with gating values centered near 0.5. This suggests stable multi-expert collaboration rather than collapse to a single expert, supporting the use of weighted gating instead of sparse Top-k routing.

Clinical Relevancy. The UMAP visualizations in Figure 2(d–e) highlight performance on rare pathological classes. While FeatImp captures the broader pathological structure, it exhibits a degree of mode collapse where synthesized centroids for different pathologies converge toward a single point. For conditions such as Enlarged Cardiomediastinum, Fracture, and Lung Lesion, ProMoE-FL maintains clear separation between class centroids while remaining closely aligned with the ground truth. This ability to prevent feature overlap is vital in a medical context, as it ensures that synthesized data from missing-modality patients retains the discriminative signals necessary for reliable diagnosis. Moreover, PadChest [4] contains Spanish radiology reports, whereas the other datasets are in English. To the best of our knowledge, this is the first work to consider linguistic domain heterogeneity in federated learning with missing modalities, reflecting realistic cross-lingual clinical settings.

4 Conclusion

In this work, we introduced ProMoE-FL, a prototype-conditioned Mixture-of-Experts framework for multimodal federated learning under missing modality

constraints. By conditioning the synthesis on a client-aware global prototype bank of missing modality, ProMoE-FL enables robust cross-modal feature synthesis addressing data heterogeneity in multimodal federated setups without relying on public data. Extensive experiments across various client configurations show consistent improvements over state-of-the-art methods, particularly in heterogeneous (non-IID) settings, where conventional imputation approaches struggle to generalize. We demonstrate strong alignment between our synthesized and ground-truth feature distributions, preserving diagnostic structures, including rare pathologies such as lung lesions, enlarged cardiomediastinum, and fractures. By moving beyond simple feature imputation toward adaptive prototype-conditioned synthesis, ProMoE-FL provides a practical approach to addressing data heterogeneity in real-world clinical federated learning.

Acknowledgments. This work was supported as part of the "Swiss AI initiative" by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a168 on Alps.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acosta, J., Falcone, G., Rajpurkar, P., Topol, E.: Multimodal biomedical ai. Nature Medicine **28**(9), 1773–1784 (2022)
2. AlSaad, R., Abd-alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.A., Damseh, R., Sheikh, J.: Multimodal large language models in health care: Applications, challenges, and future outlook. J Med Internet Res **26**, e59505 (Sep 2024)
3. Bao, G., Zhang, Q., Miao, D., Gong, Z., Hu, L., Liu, K., Liu, Y., Shi, C.: Multimodal federated learning with missing modality via prototype mask and contrast. In: ICML (2024)
4. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66**, 101797 (2020)
5. Chen, J., Pan, R.: Medical report generation based on multimodal federated learning. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society **113**, 102342 (2024)
6. Chen, Y., Liu, C., Huang, W., Cheng, S., Arcucci, R., Xiong, Z.: Generative text-guided 3d vision-language pretraining for unified medical image segmentation (2023)
7. Chhetri, A., Niroula, B., Shrestha, P., Shrestha, Y.R., Anderson, L.A., Gyawali, P.K., Bazzani, L., Bhattarai, B.: Med-mmfl: A multimodal federated learning benchmark in healthcare. arXiv preprint arXiv:2602.04416 (2026)
8. Dai, R., Li, C., Yan, Y., Mo, L., Qin, K., He, T.: Unbiased missing-modality multimodal learning. In: ICCV (2025)
9. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)

10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation [Online] **101**(23), e215–e220 (2000)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
14. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lau, K., Adler, J., Sjölund, J.: A unified representation network for segmentation with missing modalities. arXiv preprint arXiv:1908.06683 (2019)
17. Le, H.Q., Thwal, C.M., Qiao, Y., Tun, Y.L., Nguyen, M.N., Huh, E.N., Hong, C.S.: Cross-modal prototype based multimodal federated learning under severely missing modality. Information Fusion **122**, 103219 (2025)
18. Lee, H., Lee, D.Y., Kim, W., Kim, J.H., Kim, T., Kim, J., Sunwoo, L., Choi, E.: Vision-language generative model for view-specific chest x-ray generation. In: CHIL (2024)
19. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR (2021)
20. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: AAAI (2021)
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 1273–1282. PMLR (2017)
22. Poudel, P., Chhetri, A., Gyawali, P., Leontidis, G., Bhattarai, B.: Multimodal federated learning with missing modalities through feature imputation network. In: MIUA (2025)
23. Poudel, P., Shrestha, P., Amgain, S., Shrestha, Y.R., Gyawali, P., Bhattarai, B.: CAR-MFL: Cross-Modal Augmentation by Retrieval for Multimodal Federated Learning with Missing Modalities . In: MICCAI (2024)
24. Qayyum, A., Ahmad, K., Ahsan, M.A., Al-Fuqaha, A., Qadir, J.: Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. IEEE Open Journal of the Computer Society **3**, 172–184 (2022)
25. Sachin, D.N., Annappa, B., Ambasange, S., Tony, A.E.: A multimodal contrastive federated learning for digital healthcare. SN Computer Science **4**, 1–12 (2023)
26. Saha, P., Mishra, D., Wagner, F., Kamnitsas, K., Noble, J.A.: Examining modality incongruity in multimodal federated learning for medical vision and language-based disease detection. arXiv preprint arXiv:2402.05294 (2024)
27. Saha, P., Mishra, D., Wagner, F., Kamnitsas, K., Noble, J.A.: Incongruent multimodal federated learning for medical vision and language-based multi-label disease detection. AAAI **39**(27), 28331–28339 (Apr 2025)

28. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with Fixed Set Pathology Recognition Through Report-Guided Contrastive Training, p. 690–700. Springer Nature Switzerland (2022)
29. Shrestha, P., Amgain, S., Khanal, B., Linte, C., Bhattarai, B.: Medical vision language pretraining: A survey. arXiv preprint arXiv:2312.06224 (2023)
30. Sun, G., Mendieta, M., Dutta, A., Li, X., Chen, C.: Towards multi-modal transformers in federated learning. In: ECCV. p. 229–246. Springer-Verlag, Berlin, Heidelberg (2024)
31. Thrasher, J., Devkota, A., Siwakotai, P., Chivukula, R., Poudel, P., Hu, C., Bhattarai, B., Gyawali, P.: Multimodal federated learning in healthcare: A review. arXiv preprint arXiv:2310.09650 (2023)
32. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: CVPR (2024)
33. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training, p. 101–111. Springer Nature Switzerland (2023)
34. Yu, Q., Liu, Y., Wang, Y., Xu, K., Liu, J.: Multimodal federated learning via contrastive representation ensemble. In: ICLR (2023)
35. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) ACL. pp. 2608–2618 (Aug 2021)