ADAPT-MFL: PROTOTYPE-GUIDED FEDERATED MEDICAL REPORT GENERATION WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated medical report generation has the potential to improve radiologists' productivity. Recently, large language models have advanced report generation through their strong language modeling and generation capabilities. However, training these models in real-world scenarios encounters two primary obstacles: privacy requirements that restrict data centralization and the prevalence of incomplete image–report pairs across institutions. Although federated learning enables collaborative training without centralizing patient records, learning from clients with missing modalities requires appropriate local supervision and effective integration of heterogeneous updates. To address these challenges, we propose Adapt-MFL, an Adaptive Decoupled Augmentation with Prototype Tuning for Multimodal Federated Learning under missing modality conditions. We first introduce a trainable prototype dictionary as a semantic bridge between modalities. The shared prototypes are optimized through cross-modal assignment alignment on paired clients and provide learned semantic anchors for visual adaptation on image-only clients. On text-only clients, local prototype copies are further optimized jointly with decoder adapters, serving as adaptive surrogate prefixes for report generation. Additionally, we devise a modular decoupled aggregation scheme that separately integrates visual and decoder updates from their contributing clients, with dynamic weighting based on feature–prototype similarity. The global prototype dictionary is updated exclusively by paired clients, separating local text-prefix adaptation from shared cross-modal anchor learning. Finally, we apply low-rank adaptation to the visual encoder, cross-modal projection, and language decoder to enable parameter-efficient fine-tuning. Extensive experiments conducted on two publicly available medical datasets, IU X-Ray and MIMIC-CXR, demonstrate that the proposed method achieves superior performance compared to state-of-the-art frameworks.

1 INTRODUCTION

Automated medical report generation aims to translate medical images into descriptions of clinical findings, offering a means of reducing radiologists' reporting workload. Recent advances in large language models (LLMs) have created new opportunities for this task by combining visual representations with powerful language modeling and generation capabilities (Wang et al., 2023; Jin et al., 2024; Li et al., 2024). Adapting these models to medical report generation, however, typically requires paired images and reports to establish the correspondence between visual findings and their textual descriptions.

In real-world clinical settings, obtaining such training data across institutions faces two primary obstacles. First, privacy requirements restrict the centralization of sensitive patient records. Second, participating institutions may have different modality availability: some provide paired images and reports, whereas others can contribute only images or only reports. Federated learning enables institutions to collaboratively train a shared model while keeping raw data local (Rieke et al., 2020). Nevertheless, restricting federated training to clients with complete image–report pairs leaves potentially useful visual observations and reporting patterns in unimodal collections unexploited.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Training data	Test NLL ↓	SacreBLEU ↑	ROUGE-L ↑
Paired-only	2.097 ± 0.028	9.622 ± 4.755	0.306 ± 0.052
Paired + text-only	1.868 ± 0.015	10.400 ± 5.448	0.277 ± 0.026

Table 1: Preliminary IU X-Ray diagnostic using a corrected CLIP–GPT-2 baseline, separate from the Adapt-MFL evaluation. Values are means $\pm$ sample standard deviations over three seeds. NLL improves in every seed, whereas SacreBLEU and ROUGE-L changes are not consistent across seeds.

The observations in Table 1 suggest that unpaired reports provide useful learning signals, but lower prediction loss alone does not ensure higher report-generation scores. Moreover, naively incorporating image-only data can degrade generation performance. Effectively incorporating incomplete data into a shared generator therefore requires addressing distinct supervision gaps: image-only clients lack target reports, whereas text-only clients lack the corresponding visual conditions. Consequently, neither can directly optimize the standard image-to-report training objective. Prototype-based methods offer shared latent references for connecting heterogeneous clients and modalities (Bao et al., 2026; Le et al., 2025). For conditional report generation, these references must support complementary local objectives that provide training signals without report targets and surrogate visual conditions when images are unavailable, enabling both types of unimodal clients to contribute to a common image-conditioned generator.

These modality-specific learning objectives also raise a challenge for server-side aggregation. Clients with different modalities optimize different parts of the model: image-only learning updates the visual pathway, whereas text-only learning updates the language pathway. Averaging each parameter block over all clients would include unchanged copies from clients that did not train that block, diluting the corresponding updates. Moreover, prototypes adapted using text alone receive no new evidence about image–report correspondence. Directly incorporating these local prototypes into the shared dictionary would mix text-specific conditioning with cross-modal anchors learned from paired observations. Effective collaboration therefore requires specifying both how each client learns and how its updates enter the global model.

To address these challenges, we propose Adapt-MFL, a federated framework for LLM-based medical report generation under missing modality conditions. We introduce a learnable prototype dictionary that serves as a semantic bridge between paired and unimodal clients. Paired clients jointly optimize image-conditioned report generation and cross-modal prototype assignment alignment, establishing shared anchors for subsequent unimodal learning. Image-only clients hold the dictionary fixed and adapt their visual representations through a prototype anchoring objective. Text-only clients initialize a trainable local dictionary from the global prototypes and use retrieved prototype vectors as surrogate visual prefixes. Jointly optimizing these local prefixes and decoder adapters enables the available reports to provide a generation training signal without requiring corresponding images.

We further devise a modular decoupled aggregation scheme to integrate these complementary updates. Visual adapters are aggregated over paired and image-only clients, while decoder adapters are aggregated over paired and text-only clients. Within each group, feature–prototype cosine similarity dynamically weights unimodal contributions according to their local prototype compatibility. The global dictionary is updated only by paired clients; text-only clients contribute through their decoder updates, keeping local prefix adaptation separate from global cross-modal anchor learning. Low-rank adaptation (Hu et al., 2022) is applied to the visual encoder, cross-modal projection, and language decoder, enabling parameter-efficient training while keeping the pretrained backbone weights frozen. At inference, the shared generator produces reports from actual image representations.

Our main contributions are as follows:

- We propose Adapt-MFL, a federated LLM framework specifically designed for medical report generation with missing modalities, enabling collaborative learning across institutions with paired image–report data, images only, or reports only.
- We introduce a trainable prototype matrix that serves as a semantic bridge between modalities. Paired clients optimize shared prototypes through cross-modal assignment alignment,

providing learned anchors for image-only visual adaptation. Text-only clients further optimize local prototype copies jointly with decoder adapters to learn adaptive conditioning prefixes for report generation.

- We design a dynamic weighting and modular decoupled aggregation strategy that integrates visual and decoder updates from their respective contributing clients according to feature–prototype compatibility, while restricting global prototype updates to paired clients. Combined with LoRA across the visual encoder, cross-modal projection, and language decoder, this strategy enables parameter-efficient federated fine-tuning under heterogeneous modality availability.

2 Related work

Medical report generation with large language models. Automated medical report generation (MRG) translates medical images into descriptions of clinical findings, with the potential to reduce radiologists' reporting workload. R2Gen uses a memory-driven Transformer to connect visual features with report generation (Chen et al., 2020). Recent approaches exploit pretrained language models and explicit semantic conditioning: R2GenGPT aligns visual representations with a frozen LLM (Wang et al., 2023), PromptMRG introduces diagnosis-driven prompts (Jin et al., 2024), and KARGEN and KACL incorporate medical knowledge into LLM-based reporting (Li et al., 2024; Sha et al., 2025). These advances strengthen image-to-report modeling, but learning the correspondence between visual findings and report language typically relies on paired images and reports. Privacy constraints complicate pooling such data across institutions, motivating federated training.

Federated medical report generation. Federated learning enables collaborative model training while keeping raw medical records local (Rieke et al., 2020). In MRG, Chen & Pan (2024) investigate multimodal federated report generation, while FedMRG combines low-rank communication with client-aware visual learning and generic/specialized decoder adapters (Che et al., 2026). LoRA supplies a compact trainable parameterization (Hu et al., 2022), while federated aggregation determines how these local adaptations contribute to the global model. These studies extend report generation to distributed institutions and address client heterogeneity. However, FedMRG's training formulation assumes local image–report pairs. Multimodal federated learning must also accommodate heterogeneous modality availability (Che et al., 2023): image-only institutions lack report targets, whereas text-only institutions lack the visual conditions required by a paired generation objective. Exploiting these clients therefore requires complementary local objectives and aggregation rules that account for which model components each modality can train.

Prototype learning with missing modalities. Shared prototypes offer semantic references across heterogeneous clients. FedProto exchanges class prototypes to regularize local models (Tan et al., 2022). PmcmFL uses prototype masking and contrastive learning for incomplete modalities (Bao et al., 2026), while MFCPL combines prototype regularization, contrast, and cross-modal alignment under severe modality missingness (Le et al., 2025). Beyond prototypes, CreamFL exchanges representations on public data and uses cross-modal ensembling and contrastive learning to connect heterogeneous clients (Yu et al., 2023). Extending these ideas to LLM-based MRG requires supporting conditional generation when either images or target reports are absent. Adapt-MFL assigns distinct roles to learnable latent prototypes: paired clients update global cross-modal anchors, image-only clients use fixed anchors, and text-only clients adapt local prototype prefixes for decoder training. The server aggregates adapters over their respective contributing clients and weights unimodal updates by prototype compatibility, while excluding text-only prototype updates from the global dictionary.

3 Method

Adapt-MFL first constructs local objectives from the supervision available at each client (Section 3.2), then aggregates their updates according to the model components they optimize (Section 3.3). Figure 1 summarizes this connection.

3.1 Problem formulation and decoupled architecture

We consider M clients partitioned into three disjoint sets: paired clients $\mathcal{C}_{\text{full}}$, image-only clients $\mathcal{C}_{\text{img}}$, and text-only clients $\mathcal{C}_{\text{txt}}$. Client c holds $n_c = |\mathcal{D}_c|$ local training examples:

$$\mathcal{D}_c = \begin{cases} \{(x_i, y_i)\}_{i=1}^{n_c}, & c \in \mathcal{C}_{\text{full}}, \\ \{(x_i, \ell_i)\}_{i=1}^{n_c}, & c \in \mathcal{C}_{\text{img}}, \\ \{(y_i, \ell_i)\}_{i=1}^{n_c}, & c \in \mathcal{C}_{\text{txt}}. \end{cases}$$

Here x_i denotes a radiographic image input, $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ is a report, and ℓ_i denotes its available disease labels. Image-only clients lack reports, whereas text-only clients lack the corresponding images. The local objectives below use the available images or reports without requiring disease labels to define prototype assignments. The shared model is trained collaboratively without centralizing raw records.

The model comprises five modules: a visual encoder E_{vis}, a cross-modal projection Proj, a text embedding layer E_{txt}, a language decoder D_{llm}, and a prototype dictionary $P \in \mathbb{R}^{K \times d}$. The projected visual sequence and its image-level representation are

$$V(x) = \text{Proj}(E_{\text{vis}}(x)) \in \mathbb{R}^{L_x \times d}, \qquad z_x = \frac{1}{L_x} \sum_{u=1}^{L_x} V(x)_u. \tag{1}$$

Prototype comparison uses the pooled vector z_x, while image-conditioned generation retains the visual token sequence $V(x)$. For a report, the frozen text embedding layer maps its non-padding tokens into the same d-dimensional space. We obtain its report-level representation by mean pooling, $z_y = \frac{1}{T} \sum_{t=1}^{T} E_{\text{txt}}(y_t)$. The dictionary rows $p_1, \ldots, p_K$ are learnable latent anchors shared across clients. They are not predefined disease classes.

Trainable parameter blocks. The pretrained backbone and text embeddings remain frozen. LoRA (Hu et al., 2022) adapts the linear layers of the visual encoder, projection, and language decoder, including attention and feed-forward projections where present. For an adapted layer,

$$Wu = W_0 u + \frac{a_{\text{L}}}{r_{\text{L}}} BAu, \tag{2}$$

where W_0 is frozen, r_{L} is the adapter rank, and a_{L} is its scaling coefficient. Let Φ_{vis} collect the A, B factors in the visual encoder and projection, and let Φ_{llm} collect those in the decoder. The server maintains $(\Phi_{\text{vis}}^r, \Phi_{\text{llm}}^r, P^r)$ at communication round r. Each participating client initializes its corresponding local blocks from these global parameters; the subsequent update policy depends on its modality availability.

3.2 Learning from incomplete pairs through shared prototypes

Training begins with paired-client warm-up, which establishes the dictionary before unimodal clients participate. Subsequent rounds use the paired objective to learn cross-modal anchors and the two unimodal objectives to train the visual and language pathways with incomplete data.

3.2.1 Prototype assignments and paired supervision

Define cosine similarity as $s(a, b) = a^\top b / (\|a\|_2 \|b\|_2)$, with a small positive lower bound on each norm for numerical stability. For a representation z and a local dictionary P_c, the assignment probability of prototype k is

$$q_k(z; P_c) = \frac{\exp(s(z, p_{c,k})/\tau)}{\sum_{j=1}^{K} \exp(s(z, p_{c,j})/\tau)}, \qquad \tau > 0. \tag{3}$$

All cosine comparisons normalize their operands; the decoder receives the selected prototype vector itself as its conditioning embedding.

For a paired example (x, y), the client aligns the two assignment distributions using symmetric Kullback–Leibler divergence:

$$\mathcal{L}_{\text{align}}(x, y) = \frac{1}{2} \left[D_{\text{KL}}(q(z_x; P_c) \| q(z_y; P_c)) + D_{\text{KL}}(q(z_y; P_c) \| q(z_x; P_c)) \right]. \tag{4}$$

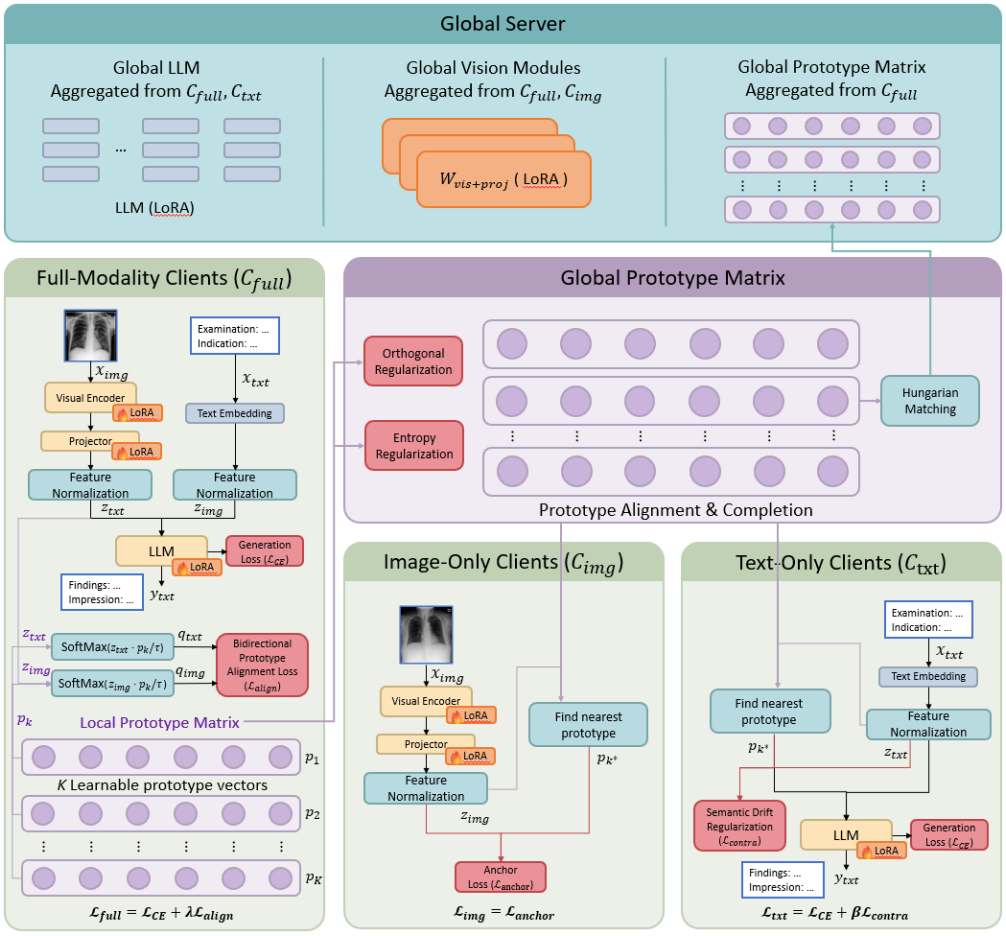

Figure 1: Adapt-MFL local training and aggregation. Paired clients learn cross-modal anchors. Image-only clients use fixed global anchors, whereas text-only clients adapt a local dictionary to condition report generation. Only paired-client prototype updates enter the global dictionary. Unimodal client contributions are weighted by feature–prototype cosine similarity.

The paired-client objective combines report generation from the observed visual condition with prototype assignment alignment:

$$\mathcal{L}_{\text{full}}(x, y) = -\sum_{t=1}^{T} \log p_{\Phi_{c,\text{llm}}}(y_t \mid V_c(x), y_{<t}) + \lambda_{\text{align}} \mathcal{L}_{\text{align}}(x, y). \tag{5}$$

Here the dependence of $V_c(x)$ on $\Phi_{c,\text{vis}}$ is implicit, and $y_{<t}$ contains the preceding ground-truth report tokens during teacher-forced training. The coefficient $\lambda_{\text{align}} \geq 0$ balances generation and alignment. Paired clients optimize $(\Phi_{c,\text{vis}}, \Phi_{c,\text{llm}}, P_c)$. Generation supervises the visual and language adapters, while assignment alignment supervises the visual adapters and dictionary against the frozen text representations. The alignment objective encourages agreement between soft assignments without requiring a predefined clinical label for each prototype.

3.2.2 Image-only data: visual adaptation without report targets

Image-only clients lack report sequences and therefore do not optimize a supervised report-generation loss. They initialize $\Phi_{c,\text{vis}}$ from the server and retain a fixed copy of P^r throughout the round. For an image x, its mean-pooled representation in Equation 1 selects a pseudo-target:

$$k_x^* = \arg\max_{k \in \{1,\ldots,K\}} s(z_x, p_k^r), \qquad \mathcal{L}_{\text{img}} = -\lambda_{\text{anchor}} \log q_{k_x^*}(z_x; P^r). \tag{6}$$

Here $\lambda_{\text{anchor}} > 0$ scales the anchoring objective. The index is held fixed for the corresponding gradient step, and gradients flow through z_x to the visual encoder and projection adapters. The prototype dictionary, text embedding layer, and decoder remain frozen.

3.2.3 TEXT-ONLY DATA: REPORT GENERATION FROM LOCAL PROTOTYPE PREFIXES

Local dictionary initialization and retrieval. At the start of round r, a text-only client initializes a trainable local dictionary $P_c^{r,0} = P^r$ and decoder adapters $\Phi_{c,\text{llm}}^{r,0} = \Phi_{\text{llm}}^r$. Its visual modules and text embedding layer remain frozen. At local optimization step s, the report representation z_y selects a row of the current local dictionary:

$$k_y^{*,r,s} = \arg\max_{k\in\{1,\dots,K\}} s(z_y, p_{c,k}^{r,s}), \qquad \widetilde{z}_{x,c}^{r,s} = p_{c,k_y^{*,r,s}}^{r,s}. \tag{7}$$

The selected d-dimensional vector is inserted as a single surrogate visual-prefix token before the report tokens. The full-report embedding is used for training-time retrieval; the decoder receives the retrieved prefix and the causal report history, with the report tokens providing the generation targets. The discrete index is treated as fixed during each backward pass, while the selected prototype vector remains connected to the computational graph. Retrieval is recomputed with the current dictionary at subsequent local steps.

Joint decoder and prototype optimization. Suppressing round and step indices for readability, the generation objective is

$$\mathcal{L}_{\text{txt-gen}}(y) = -\sum_{t=1}^{T} \log p_{\Phi_{c,\text{llm}}}(y_t \mid \widetilde{z}_{x,c}, y_{<t}). \tag{8}$$

The selected prototype serves as the positive anchor in a text-to-prototype assignment term, giving the combined local objective

$$\mathcal{L}_{\text{txt}}(y) = \mathcal{L}_{\text{txt-gen}}(y) - \beta \log q_{k_y^*}(z_y; P_c). \tag{9}$$

The coefficient $\beta \geq 0$ controls the assignment term, and the text representation remains frozen. The generation term updates the decoder adapters and the selected prototype, whereas the assignment term updates the positive and competing dictionary rows. The latter influences decoder learning indirectly through the adapted prefixes used in subsequent local steps.

Local adaptation and global participation. At the end of the round, the client uploads its decoder-adapter increment and contribution weight. Its adapted dictionary is not uploaded or included in global prototype aggregation. The local dictionary is discarded and reinitialized from the new server dictionary at the next participation round. Thus, text-only data affect the global model through decoder updates, while the global dictionary continues to receive cross-modal supervision from paired clients. For image-to-report inference, the global decoder is conditioned on the actual visual sequence $V(x)$; neither a reference report nor a text-client local dictionary is required.

3.3 INTEGRATING LOCAL UPDATES INTO THE SHARED GENERATOR

The local objectives produce three distinct forms of adaptation: visual updates grounded in images, decoder updates supervised by reports, and prototype updates with either paired or text-only supervision. We aggregate adapters according to the pathway they optimize, while restricting the global dictionary to paired-client updates. This retains a shared source of cross-modal anchors across rounds and lets text-only data contribute through the language pathway.

Client contribution weights. After local training, each unimodal client computes a mean cosine similarity over its n_c local training examples:

$$S_c^r = \begin{cases} \dfrac{1}{n_c}\displaystyle\sum_{i=1}^{n_c} \max_k s(z_{x_i,c}^{r,\text{end}}, p_k^r), & c \in \mathcal{C}_{\text{img}}, \\ \dfrac{1}{n_c}\displaystyle\sum_{i=1}^{n_c} \max_k s(z_{y_i}, p_{c,k}^{r,\text{end}}), & c \in \mathcal{C}_{\text{txt}}. \end{cases} \tag{10}$$

Image-only clients compare their final visual representations with the fixed server dictionary. Text-only clients compare their frozen report representations with the final local dictionary, using the same nearest-prototype rule as local generation. We set

$$\alpha_c^r = \begin{cases} 1, & c \in \mathcal{C}_{\text{full}}, \\ \max(0, S_c^r), & c \in \mathcal{C}_{\text{img}} \cup \mathcal{C}_{\text{txt}}. \end{cases} \tag{11}$$

Clipping at zero gives nonnegative aggregation weights. Paired clients receive a reference weight of one. These scores measure local feature–prototype compatibility, rather than the improvement a client update would produce in global report generation. In particular, the text-client score is evaluated against its adapted local dictionary.

Parameter increments and contributor sets. Let $\mathcal{A}^r$ be the participating clients, and write $\mathcal{A}_b^r = \mathcal{A}^r \cap \mathcal{C}_b$ for $b \in \{\text{full}, \text{img}, \text{txt}\}$. For an eligible adapter block $b \in \{\text{vis}, \text{llm}\}$, client c sends

$$\Delta\Phi_{c,b}^r = \Phi_{c,b}^{r,\text{end}} - \Phi_b^r.$$

Paired clients additionally send their prototype increment $\Delta P_c^r = P_c^{r,\text{end}} - P^r$. All local dictionaries start with the server's row indexing, and their rows retain those indices during local optimization. Each adapter block is normalized over clients that optimize it, excluding frozen copies from its aggregation denominator. Text-client dictionary increments remain local because their training does not observe image–report correspondence. The server updates its three blocks as

$$P^{r+1} = P^r + \frac{\sum_{c \in \mathcal{A}_{\text{full}}^r} n_c \Delta P_c^r}{\sum_{c \in \mathcal{A}_{\text{full}}^r} n_c}, \tag{12}$$

$$\Phi_{\text{vis}}^{r+1} = \Phi_{\text{vis}}^r + \frac{\sum_{c \in \mathcal{A}_{\text{full}}^r \cup \mathcal{A}_{\text{img}}^r} \alpha_c^r n_c \Delta\Phi_{c,\text{vis}}^r}{\sum_{c \in \mathcal{A}_{\text{full}}^r \cup \mathcal{A}_{\text{img}}^r} \alpha_c^r n_c}, \tag{13}$$

$$\Phi_{\text{llm}}^{r+1} = \Phi_{\text{llm}}^r + \frac{\sum_{c \in \mathcal{A}_{\text{full}}^r \cup \mathcal{A}_{\text{txt}}^r} \alpha_c^r n_c \Delta\Phi_{c,\text{llm}}^r}{\sum_{c \in \mathcal{A}_{\text{full}}^r \cup \mathcal{A}_{\text{txt}}^r} \alpha_c^r n_c}. \tag{14}$$

If a block has no positive-weight contributor, its previous global value is retained. Aggregation operates componentwise on the LoRA factors A and B through their increments, rather than averaging the composed matrices BA. Frozen backbone weights are shared by all clients and are not transmitted each round. The resulting adapter blocks and paired-client prototype aggregate are broadcast for the next round.

4 Experiments and Results

4.1 Datasets and Implementation

We use MIMIC-CXR (Johnson et al., 2019) for primary training and internal testing, and IU X-Ray (Demner-Fushman et al., 2016) for external zero-shot evaluation. The MIMIC-CXR training set is partitioned at the patient level into eight simulated clients: four retain paired images and reports, two retain images only, and two retain reports only. Within each client, 90% of the data are used for local training and 10% for local validation. These choices describe the main Adapt-MFL evaluation, where IU X-Ray is used for external zero-shot testing. The separate small-model diagnostic described in the Supplementary Material trains and tests on an IU X-Ray split and is not a zero-shot experiment.

The reported backbone is Qwen3-VL-8B. LoRA uses rank 64 in the visual encoder and language decoder, and the shared dictionary contains $K = 16$ prototypes of dimension $d = 4096$. Training comprises 50 federated rounds, including a five-round warm-up using paired clients only. Each participating client performs 100 local iterations per round with effective batch size 128. Optimization uses AdamW with learning rate 2×10^{-5}, $\lambda_{\text{align}} = 0.5$, and $\beta = 0.1$.

Evaluation reports BLEU-1 through BLEU-4, ROUGE-L, and clinical-label F1, recall, and precision. Higher scores are better. Text-overlap and label-based metrics measure different aspects of report quality; neither alone establishes clinical reliability.

Method	B1	B2	B3	B4	R-L	F1	Rec	Pre
MIMIC-CXR: internal evaluation								
R2GenGPT	39.40	24.94	17.01	12.27	27.22	26.60	23.80	35.70
PromptMRG	38.23	24.05	16.43	11.79	26.70	30.70	26.40	43.20
Fed-Prefix	38.46	24.10	16.43	11.90	26.77	26.90	24.10	35.50
Fed-VeRA	38.03	23.89	16.22	11.64	26.58	25.40	23.10	33.80
FedDAT	35.29	21.95	14.92	10.69	25.44	17.70	15.10	25.90
FedMRG	**40.19**	25.52	17.55	12.68	27.55	33.90	30.20	45.40
Adapt-MFL	36.20	**26.70**	**21.29**	**17.71**	**30.63**	47.82	44.68	56.06
IU X-Ray: external zero-shot evaluation								
R2GenGPT	38.68	22.22	13.52	8.73	26.43	13.19	12.80	14.34
PromptMRG	42.18	25.33	16.28	11.05	**30.61**	18.17	15.76	16.39
Fed-Prefix	37.68	21.63	13.19	8.53	25.82	13.07	12.76	14.08
Fed-VeRA	37.23	21.05	12.54	7.96	25.38	12.68	12.35	13.60
FedDAT	36.16	21.30	13.62	9.19	26.02	11.11	10.92	11.69
FedMRG	**43.55**	**26.38**	17.30	11.95	29.73	19.69	17.22	17.86
Adapt-MFL	34.15	26.10	**20.45**	**16.90**	28.80	35.20	32.95	33.40

Table 2: Performance on MIMIC-CXR and IU X-Ray evaluation. B1–B4 denote BLEU-1–4, R-L denotes ROUGE-L, Rec denotes recall, and Pre denotes precision. All metrics ↑. Bold marks the best text-overlap score in each block.

4.2 Comparison with existing methods

The comparison includes R2GenGPT (Wang et al., 2023), PromptMRG (Jin et al., 2024), Fed-Prefix based on prefix-tuning (Li & Liang, 2021), Fed-VeRA based on VeRA (Kopiczko et al., 2024), FedDAT (Chen et al., 2024), and FedMRG (Che et al., 2026). The Fed-Prefix and Fed-VeRA labels refer to federated adaptations; their citations identify the underlying adaptation methods. Table 2 reports the results of these complete systems; the comparison does not isolate the effect of aggregation from other modeling and training choices.

On MIMIC-CXR, Adapt-MFL achieves the highest listed BLEU-2, BLEU-3, BLEU-4, and ROUGE-L values. Relative to FedMRG, BLEU-4 increases by 5.03 points, from 12.68 to 17.71, and ROUGE-L increases by 3.08 points, from 27.55 to 30.63. BLEU-1 is lower, at 36.20 compared with 40.19. On IU X-Ray, Adapt-MFL achieves the highest listed BLEU-3 and BLEU-4 values, with a 4.95-point BLEU-4 improvement over FedMRG. It does not lead on BLEU-1, BLEU-2, or ROUGE-L. The observed advantage is therefore concentrated in selected higher-order text-overlap metrics.

Clinical-metric consistency. For the IU X-Ray entries of PromptMRG, FedMRG, and Adapt-MFL, F1 exceeds the arithmetic mean of precision and recall, which is inconsistent with conventional F1 evaluated on matching units and averaged using the same weights. The comparative conclusions therefore concern text-overlap metrics, without inferring clinical superiority from these entries.

4.3 Ablation study

Table 3 compares the complete method with variants removing prototype tuning or decoupled aggregation. The complete-method row matches the MIMIC-CXR results in Table 2. Removing either component reduces all three reported ablation metrics. Without prototype tuning, BLEU-4 decreases by 3.07 points and ROUGE-L by 4.28 points; without decoupled aggregation, the corresponding decreases are 1.45 and 2.46 points. These results support the utility of both components within the recorded configuration, but do not isolate the separate effects of contributor selection and dynamic weighting.

Model variant	BLEU-4	ROUGE-L	F1
Without prototype tuning	14.64	26.35	39.26
Without decoupled aggregation	16.26	28.17	44.16
Adapt-MFL	**17.71**	**30.63**	**47.82**

Table 3: Component ablations for Adapt-MFL. The complete-method values match the MIMIC-CXR comparison row. All metrics ↑.

5 Discussion and limitations

Adapt-MFL gives unpaired images and reports distinct routes into a shared conditional generator. The prototype dictionary connects these routes to paired-client training, but the resulting supervision remains indirect. Image self-assignment can reinforce an incorrect anchor. A report-derived prefix summarizes a latent assignment rather than reconstructing patient-specific visual evidence, and its single-token condition differs from the visual token sequence used at inference. The global model therefore continues to rely on paired clients for image–report correspondence.

The aggregation rule specifies where each update enters the model. Its cosine weights express local prototype compatibility and do not establish that a higher-weight update improves image-conditioned report generation. This distinction is especially relevant for text clients, whose local dictionaries are themselves optimized on the reports used to calculate the score. The reported experiments cover a simulated eight-client federation and external zero-shot evaluation; their conclusions concern the listed generation metrics and component ablations. Automatic report metrics do not establish clinical reliability.

6 Conclusion

Adapt-MFL addresses two connected challenges in federated medical report generation: learning from unpaired data and incorporating the resulting updates into a shared image-conditioned generator. Paired-client prototypes supply fixed anchors for image-only visual adaptation and trainable local prefixes for text-only decoder adaptation. The server aggregates visual and language adapters over their respective contributors, scales unimodal updates by prototype compatibility, and learns the global dictionary from paired clients alone. This connects the local surrogate objectives to the corresponding generation modules while separating text-prefix adaptation from global cross-modal anchor learning. Experiments on MIMIC-CXR and external IU X-Ray evaluation show gains in selected text-overlap metrics, with component ablations supporting the complete configuration.

AI use statement

Generative AI tools were used to assist with drafting and polishing the manuscript, including language editing, organization, and clarification of technical descriptions. The authors take responsibility for the final content of the paper.

References

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Chaochao Chen, Liang Hu, and Longbing Cao. Incomplete multimodal federated learning via masking and contrasting prototypes. *IEEE Transactions on Neural Networks and Learning Systems*, 37(9):4050–4062, 2026. doi: 10.1109/TNNLS.2026.3658522. URL https://doi.org/10.1109/TNNLS.2026.3658522.

Haoxuan Che, Haibo Jin, Zhengrui Guo, Yi Lin, Cheng Jin, and Hao Chen. LLM-driven medical report generation via communication-efficient heterogeneous federated learning. *IEEE Transactions on Medical Imaging*, 45(1):28–39, 2026. doi: 10.1109/TMI.2025.3591185.

Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23(15):6986, 2023. doi: 10.3390/s23156986.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11285–11293, 2024. doi: 10.1609/aaai.v38i10.29007.

Jieying Chen and Rong Pan. Medical report generation based on multimodal federated learning. *Computerized Medical Imaging and Graphics*, 113:102342, 2024. doi: 10.1016/j.compmedimag.2024.102342.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1439–1449. Association for Computational Linguistics, 2020.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2607–2615, 2024.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

Huy Q. Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh N. H. Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, 122:103219, 2025. doi: 10.1016/j.inffus.2025.103219.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics, 2021.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. KARGEN: Knowledge-enhanced automated radiology report generation using large language models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 382–392. Springer, 2024. doi: 10.1007/978-3-031-72086-4_36.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020.

Yuyang Sha, Hongxin Pan, Weiyu Meng, and Kefeng Li. Contrastive knowledge-guided large language models for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 111–120. Springer, 2025. URL https://papers.miccai.org/miccai-2025/0178-Paper1713.html.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8432–8440, 2022. doi: 10.1609/aaai.v36i8.20819. URL https://ojs.aaai.org/index.php/AAAI/article/view/20819.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Hnk1WRMAYqg.