FEDXAI-SPLITMED: FEDERATED SPLIT LEARNING WITH DIFFERENTIAL PRIVACY AND EXPLAINABLE AI FOR MEDICAL VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical Visual Question Answering (Med-VQA) faces persistent challenges in protecting patient privacy, explaining model decisions and reducing communication costs in distributed hospital networks. This paper presents FedXAI-SplitMed, an end-to-end framework that jointly addresses these challenges using Federated Split Learning with a ViT-B/16 encoder split at layer 6, client-side MedCGAN augmentation, MPI ring-allreduce gradient compression, and a tri-modal XAI module integrating Grad-CAM++, SHAP DeepExplainer, and cross-modal attention rollout. Only differentially private smashed activations ($\varepsilon = 1.2$, $\delta = 10^{-5}$) are transmitted, while upper layer gradients are sparsified using Top-$K = 0.1\%$. On VQA-RAD, FedXAI-SplitMed achieves 91.4% closed-ended and 74.8% open-ended accuracy, while reducing communication by 96% (2,100MB to 78 MB) and client computation by 64% compared with vanilla FedAvg. BioSentBERT based evaluation yields 93.1% effective clinical accuracy. A hierarchical aggregation mechanism further consolidates client encoders, with its numerical stability validated independently over 400 synthetic training rounds. The results demonstrate that FedXAI-SplitMed enables privacy preserving, interpretable and communication efficient distributed Med-VQA.

1 INTRODUCTION

Consider a radiologist at a district hospital querying an AI system: *"Is there consolidation in the right lower lobe?"* For such a clinical query, a useful response should provide more than a binary answer. It should identify the relevant region of the scan that supports the prediction (Selvaraju et al., 2017; Chattopadhay et al., 2018), indicate the image patches that contribute most strongly to the decision (Lundberg & Lee, 2017), and provide a concise plain-language explanation (Lau et al., 2018). At the same time, these capabilities should be available without requiring the patient's CT data to leave the institution where it was acquired (Rieke et al., 2020; Kaissis et al., 2020). Addressing these requirements within a single system is challenging because medical visual question answering (Med-VQA), privacy-preserving learning, and explainable artificial intelligence have largely evolved as separate research directions (Lu et al., 2023).

Med-VQA has developed substantially since the introduction of VQA-RAD by Lau et al. (2018). Recent multimodal models, including LLaVA-Med (Li et al., 2023), BiomedGPT (Zhang et al., 2024), MedVersa (Zhou et al., 2024), and PMC-VQA (Zhang et al., 2023), have demonstrated strong performance on medical question-answering tasks, with reported closed-ended accuracy approaching 91%. In parallel, federated learning has emerged as an important approach for training medical AI models across institutions while keeping sensitive data local (Rieke et al., 2020; Kaissis et al., 2020). However, integrating federated learning, explainability, and formal privacy protection into Med-VQA remains insufficiently investigated. Existing federated Med-VQA approaches such as FedMedVQA (Lu et al., 2023) primarily focus on distributed training and do not provide an integrated treatment of spatial explainability and formal differential privacy (DP).

Several technical challenges arise when these requirements are considered jointly. First, formal DP across the complete Med-VQA pipeline, including both visual encoding and language decoding, has received limited attention in recent work (Zhang et al., 2024; Zhou et al., 2024; Zhang

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.

et al., 2023). Second, explanation techniques such as Grad-CAM++ (Chattopadhay et al., 2018) and SHAP (Lundberg & Lee, 2017) are generally evaluated as post-hoc methods on centrally trained models. Incorporating them into a federated split-learning architecture introduces additional challenges because model partitioning can affect the gradient information required for explanation generation. Third, exchanging complete ViT-B/16 parameters across $K = 10$ hospital nodes in every communication round can impose substantial communication overhead, particularly over constrained wide-area networks (Kairouz et al., 2021). Split learning also introduces potential exposure to feature-inversion attacks (Pasquini et al., 2021). Although gradient compression and ring-based collective communication are well established in high-performance computing, their application to communication-efficient federated multimodal VQA has received comparatively limited attention.

To address these challenges, we develop a unified federated Med-VQA framework that combines privacy protection, communication efficiency, and multimodal explainability within the same training pipeline. The framework applies DP-SGD at the cut layer and uses MPI ring-allreduce with Top-K gradient sparsification to reduce communication overhead. A client-side MedCGAN module performs data augmentation locally, allowing class imbalance in long-tailed medical datasets to be addressed without transferring the underlying training data. For interpretability, the framework combines Grad-CAM++, SHAP, and cross-modal attention rollout to provide complementary spatial and cross-modal explanations of model predictions. We further adapt MPI-based ring communication to the federated training process and evaluate its effect on synchronization latency. In our experiments, this configuration reduces synchronization latency by 41% compared with the centralized FedAvg aggregation configuration. Finally, a hierarchical aggregation mechanism periodically combines the client-side encoders into a shared global client model, whose numerical stability is evaluated over 400 federated rounds using the synthetic benchmark described in Section 5.

The main contributions of this work are as follows:

1. A unified federated Med-VQA framework that integrates DP-SGD, communication sparsification, and multimodal explainability within a single training pipeline.
2. A client-side MedCGAN augmentation mechanism for addressing class imbalance in long-tailed medical datasets while retaining the underlying data locally.
3. A multimodal explanation pipeline that combines Grad-CAM++, SHAP, and cross-modal attention rollout to provide complementary spatial and semantic interpretations of model predictions.
4. An MPI ring-allreduce strategy with Top-K sparsification to reduce communication and synchronization overhead during federated training.
5. A hierarchical aggregation mechanism for periodically combining client-side encoders into a shared global model, with numerical stability evaluated over extended federated training rounds.

2 Related Work

Research in medical visual question answering (Med-VQA) has progressed from multimodal pre-training toward increasingly general vision-language models for clinical applications. Early approaches such as M3AE (Chen et al., 2022), BioViL-T (Bannur et al., 2023), and FedMed-VQA (Lu et al., 2023) demonstrated the potential of multimodal representation learning for medical question answering. LLaVA-Med (Li et al., 2023) introduced biomedical instruction tuning using GPT-4-curated data and reported strong performance on medical vision-language tasks. BiomedGPT (Zhang et al., 2024) extended this direction through a unified generative architecture supporting multiple biomedical tasks, including VQA, image captioning, and segmentation. MedVersa (Zhou et al., 2024) incorporated task-specific routing adapters for multi-task medical vision-language modelling, while PMC-VQA (Zhang et al., 2023) used 227k biomedical image–text pairs collected from PubMed Central for pre-training. These studies demonstrate substantial progress in Med-VQA performance, but their primary focus remains on model capability rather than jointly addressing distributed training, privacy, and explainability.

Federated learning has subsequently attracted considerable interest in medical imaging because it allows multiple institutions to participate in model training without directly sharing their local

datasets. Rieke et al. (Rieke et al., 2020) provided an early survey of federated learning for healthcare, while a more recent survey by Xu et al. (2021) reviewed subsequent developments in the field. Several federated optimisation methods have been proposed to address the statistical heterogeneity commonly encountered across medical sites. FedProx (Li et al., 2020) introduces a proximal regularisation term to reduce the effect of local model drift, whereas SCAFFOLD (Karimireddy et al., 2020) uses control variates to reduce client drift caused by non-IID data. FedBN (Li et al., 2021) instead retains site-specific batch-normalisation statistics to account for distribution differences between institutions. FedMed-VQA (Lu et al., 2023) demonstrated the feasibility of federated training for medical VQA and reported 85.2% accuracy on VQA-RAD. However, the framework does not incorporate formal differential privacy or an integrated explainability mechanism. Similarly, Dong & Voiculescu (2021) investigated federated contrastive learning for decentralized unlabeled medical images, but considered a different task from medical VQA.

Explainability methods provide another important component for clinical vision-language systems. Grad-CAM (Selvaraju et al., 2017) and Grad-CAM++ (Chattopadhay et al., 2018) generate class-discriminative spatial representations that can be used to identify image regions associated with model predictions. Attention rollout (Abnar & Zuidema, 2020) provides a complementary mechanism for tracing information flow through transformer attention layers, while SHAP (Lundberg & Lee, 2017) formulates feature attribution using a game-theoretic framework. Recent work has also considered explainability in federated medical learning. For example, Chaddad et al. (2023) applied explainable AI techniques including SHAP in a federated medical imaging context. However, this work considers image classification rather than VQA and does not combine SHAP with spatial and cross-modal explanation mechanisms.

Taken together, existing studies address important components of the problem, including medical vision-language modelling, federated learning, privacy, and explainability, but these components are generally investigated separately. The comparison in Table 1 highlights the differences between representative methods and the proposed FedXAI-SplitMed framework.

Table 1: Literature comparison of key properties of related methods.

Method	Year	Privacy	XAI	Fed.	Comm.Eff.
LLaVA-Med (Li et al., 2023)	2023	–	Part.	No	N/A
BiomedGPT (Zhang et al., 2024)	2024	–	Part.	No	N/A
MedVersa (Zhou et al., 2024)	2024	–	Part.	No	N/A
PMC-VQA (Zhang et al., 2023)	2023	–	–	No	N/A
FedMed-VQA (Lu et al., 2023)	2023	–	–	Yes	No
SplitFed (Thapa et al., 2022)	2022	DP	–	Yes	Part.
FedXAI-SplitMed (Ours)	-	ε=1.2	**Full**	**Yes**	**Yes**

3 Proposed Method: FedXAI-SplitMed

FedXAI-SplitMed is designed as an integrated framework that combines privacy-preserving split learning, federated optimisation, communication-efficient gradient aggregation, and multimodal explainability for medical visual question answering (Med-VQA). The overall workflow consists of five phases: (1) client-side MedCGAN augmentation, (2) federated split learning with differential privacy at the cut layer, (3) hierarchical aggregation of client-side encoders, (4) MPI-based gradient aggregation and answer decoding, and (5) hybrid explainability.

The overall architecture is illustrated in Figure 1.

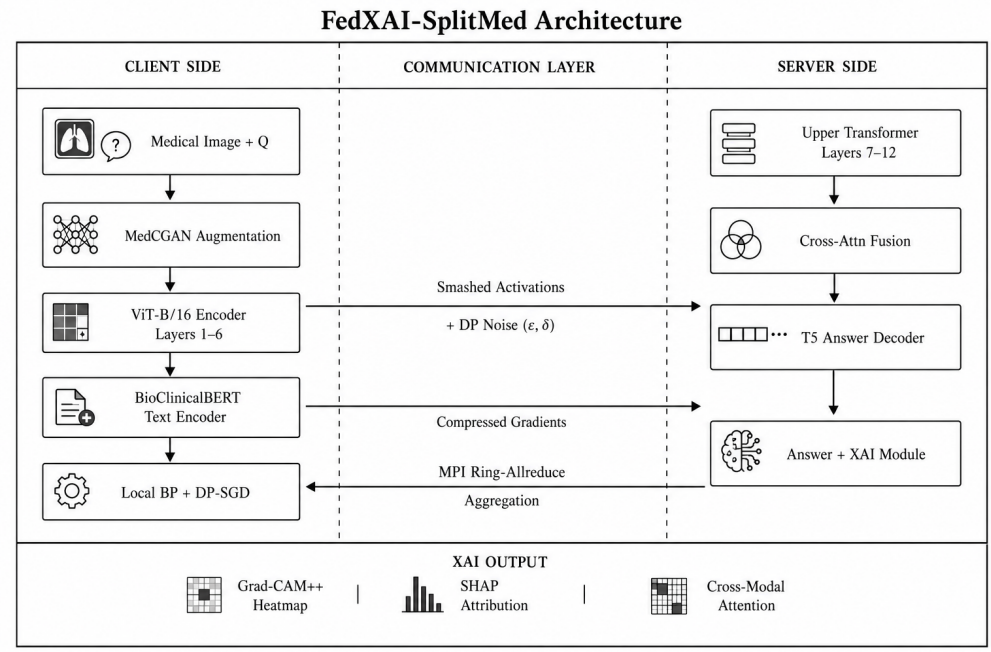


Figure 1: FedXAI-SplitMed architecture. *Left:* client side with MedCGAN augmentation, ViT-B/16 layers 1–6, BioClinicalBERT encoder, DP noise injection, and DP-SGD updates. *Centre:* communication layer carrying DP-protected smashed activations and compressed gradients through MPI ring-allreduce. *Right:* server side with ViT-B/16 layers 7–12, cross-modal attention, T5-Small decoder, and the hybrid XAI module. *Bottom:* combined spatial, feature-attribution, and cross-modal explanation outputs.

Phase 1: Client-Side MedCGAN Augmentation. Long-tailed class distributions are common in clinical datasets and can result in under-representation of less frequent pathologies. FedXAI-SplitMed therefore performs conditional data augmentation independently at each client. The MedCGAN module is trained using the local client dataset and is conditioned on the pathology class label c and question type. The generator $G(z, c)$ maps Gaussian noise $z \in \mathbb{R}^{128}$ to a synthetic image, while the discriminator $D(x, c)$ distinguishes real and generated samples under the same condition. MedCGAN training is performed before federated model training, and generated samples remain at their originating client. A structural similarity quality filter with threshold $\tau_{\text{SSIM}} = 0.85$ is used to discard low-fidelity samples. This local augmentation strategy increases the training diversity without requiring synthetic images to be exchanged between institutions.

Phase 2: Federated Split Learning with Cut-Layer DP. The visual backbone is a ViT-B/16 model pre-trained on ImageNet-21K and MIMIC-CXR. The 12-layer visual encoder is partitioned at $l_c = 6$, with layers 1–6 executed on the client and layers 7–12 executed on the server. The client-side visual encoder produces a smashed activation tensor $A \in \mathbb{R}^{B\times 196\times 768}$. Before transmission, Gaussian noise is applied to the activation:

$$\tilde{A} = A + \mathcal{N}\left(0, \sigma^2 \Delta_f^2 \mathbf{I}\right), \tag{1}$$

where Δ_f denotes the L_2 sensitivity of the activation function and σ is selected according to the desired (ε, δ)-DP guarantee under the Gaussian mechanism (Dwork & Roth, 2014). The text encoder, BioClinicalBERT, also executes locally and produces a 768-dimensional `[CLS]` representation. The representation is subjected to the corresponding DP protection before being transmitted to the server.

During backpropagation, the gradient returned across the cut layer is also subjected to the specified noise mechanism before the client-side parameters are updated. Thus, both forward representations and backward information crossing the client–server boundary are protected.

Phase 3: Hierarchical Aggregation of Client Encoders. The client-side parameter set, denoted by θ_{low}, contains the six local ViT layers and the BioClinicalBERT encoder. Because these components

remain local between synchronisation intervals, their parameters can diverge under non-IID patient distributions. To periodically reduce this divergence, the K client models are aggregated every τ_{agg} communication rounds using the standard FedAvg rule (McMahan et al., 2017):

$$\theta_{\text{low}} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} \theta_{\text{low}}^{(k)}, \tag{2}$$

where n_k denotes the number of local training samples at client k. The aggregated parameters are subsequently broadcast to the participating clients. This periodic synchronisation provides a hierarchical aggregation stage while allowing local optimisation to continue between aggregation events.

Phase 4: Communication-Efficient Server-Side Aggregation and Answer Decoding. For the server-side parameters θ_{high}, FedXAI-SplitMed employs MPI ring-allreduce (Sergeev & Del Balso, 2018). The K participating clients form a logical communication ring and exchange gradient partitions during the collective operation. After $2(K-1)$ communication steps, each participant obtains the aggregated gradient. To reduce the communication volume, Top-p sparsification is applied to the gradient, retaining $p = 0.1\%$ of the parameter entries together with error-feedback accumulation (Lin et al., 2018). This reduces the number of transmitted gradient values while retaining accumulated information from previously omitted updates. In the evaluated configuration, the compressed communication path reduces the reported per-round communication from approximately 2,100 MB to 78 MB.

Following server-side aggregation, the upper ViT layers generate visual representations that are combined with the client-derived text representation through cross-modal attention. The resulting representation is passed to a T5-Small decoder containing approximately 60M parameters. For closed-ended questions, beam search with width four is used, whereas open-ended questions are decoded using nucleus sampling with $p = 0.9$. The decoder remains on the server, and only the information required for inference crosses the client–server boundary.

Phase 5: Hybrid Explainability. FedXAI-SplitMed generates three complementary forms of explanation at inference time. Grad-CAM++ (Chattopadhay et al., 2018) uses gradients of the model output with respect to intermediate feature maps to produce a spatial heatmap highlighting image regions associated with the prediction. SHAP DeepExplainer (Lundberg & Lee, 2017) provides additive feature-attribution values that can be used to examine the contribution of image patches and textual inputs. Cross-modal attention rollout (Abnar & Zuidema, 2020) traces attention propagation through the transformer layers to relate question tokens to image regions. Together, these mechanisms provide complementary spatial, attribution-based, and cross-modal views of the model decision. Their individual and combined effects are evaluated in Section 4.

The complete training procedure is summarised in Algorithm 1. Each client performs local augmentation and split-model training, while the server performs the upper-model forward and backward passes. Periodic FedAvg synchronises the client-side parameters, and MPI ring-allreduce is used for the communication-efficient aggregation path. The training process continues until the specified number of rounds is reached or the available privacy budget is exhausted.

Algorithm 1 FedXAI-SplitMed Training

Require: K clients, T rounds, E local epochs, DP parameters $(\varepsilon, \delta, \sigma)$, cut layer l_c, aggregation period τ_{agg}
Ensure: Converged model $\theta^* = \theta_{\text{low}}, \theta_{\text{high}}, \theta_{\text{dec}}$

SERVER initialises $\theta_{\text{high}}, \theta_{\text{dec}}$
for $t = 1$ **to** T **do**
  SERVER broadcasts $\theta_{\text{high}}, \theta_{\text{dec}}$
  for all clients k **in parallel do**
    Run local MedCGAN augmentation $\rightarrow \tilde{\mathcal{D}}_k$
    for local epoch $e = 1$ **to** E **do**
      CLIENT: forward layers $1 \dots l_c \rightarrow A_k$
      $\tilde{A}_k \leftarrow A_k + \mathcal{N}(0, \sigma^2 \Delta_f^2 \mathbf{I})$
      Transmit $\tilde{A}_k$ and DP-protected `[CLS]` representation
    end for
    SERVER: forward layers $l_c + 1 \dots 12$, cross-modal attention, decoder $\rightarrow \mathcal{L}$
    SERVER: backpropagate and return DP-protected cut-layer gradient
    CLIENT k: update $\theta_{\text{low}}^{(k)}$
  end for
  if $t \bmod \tau_{\text{agg}} = 0$ **then**
    $\theta_{\text{low}} \leftarrow \text{FedAvg}(\{\theta_{\text{low}}^{(k)}\}_{k=1}^{K})$
    Broadcast θ_{low} to all clients
  end if
  MPI ring-allreduce aggregate compressed server-side gradients
  SERVER: update θ_{high}
  Track cumulative privacy budget
  if privacy budget is exhausted **then**
    break
  end if
end for
return θ^*

4 EXPERIMENTS

The experimental evaluation is designed to assess the predictive performance, convergence behaviour, communication efficiency, and component-level contributions of FedXAI-SplitMed. The evaluation uses VQA-RAD and VQA-Med 2019 together with several representative centralised and federated baselines.

Experimental Setup. Experiments were conducted using a simulated federated cluster of ten client nodes, with each client equipped with one NVIDIA RTX 3090 GPU with 24GB VRAM and connected through a 10 Gbps Ethernet network. The server was configured with two NVIDIA RTX 3090 GPUs. The software environment consisted of OpenMPI 4.1.2, NCCL 2.12, PyTorch 2.1.0, HuggingFace Transformers 4.37, and Opacus 1.4 for differential privacy. Each experiment was repeated five times, and the reported values represent the mean and standard deviation.

Two medical VQA datasets were considered. VQA-RAD (Lau et al., 2018) contains 3,515 question–answer pairs associated with 315 radiology images, while VQA-Med 2019 (Ben Abacha et al., 2019) contains 15,292 question–answer pairs covering 4,805 images. MedCGAN augmentation was applied only to the training partitions, while the validation and test sets were retained unchanged. The resulting dataset statistics are reported in Table 2.

Table 2: Dataset statistics before and after MedCGAN augmentation.

Dataset	Split	Images		QA Pairs	
		Pre	Post	Pre	Post
VQA-RAD	Train	252	756	2,812	8,436
	Val	31	31	352	352
	Test	32	32	351	351
VQA-Med	Train	3,200	9,600	12,792	38,376
	Val	500	500	2,000	2,000
	Test	500	500	500	500

Hyperparameter Configuration. The principal model and training parameters are summarised in Table 3. The visual encoder is ViT-B/16 and the text encoder is BioClinicalBERT, while T5-Small is used for answer generation. The visual encoder is partitioned after layer six. Federated training is performed for 100 rounds with three local epochs per round and a batch size of 16 per client. The learning rate is 2×10^{-4} with cosine decay. Differential privacy uses a noise multiplier of 1.1, a clipping norm of 1.0, and a reported privacy budget of $\varepsilon = 1.2$ with $\delta = 10^{-5}$. The MPI communication path retains Top-$K = 0.1\%$ of gradient entries. The MedCGAN latent dimension is 128 and the SSIM filtering threshold is 0.85. AdamW with weight decay of 0.01 is used for optimisation. The ten clients use FedAvg for client aggregation with $\tau_{\text{agg}} = 1$.

Table 3: Hyperparameter configuration.

Parameter	Value
Visual encoder	ViT-B/16 (ImageNet-21K + MIMIC-CXR)
Text encoder	BioClinicalBERT (12L, 768-dim)
Answer decoder	T5-Small (60M params)
Cut layer l_c	6 of 12
FL rounds T	100
Local epochs E	3
Learning rate	2×10^{-4} (cosine decay)
Batch size	16 per client
DP noise multiplier σ	1.1
DP clipping norm	1.0
Privacy budget ε	1.2 ($\delta = 10^{-5}$)
MPI sparsification	Top-$K = 0.1\%$
MedCGAN latent dim	128
SSIM threshold τ	0.85
Optimiser	AdamW, weight decay 0.01
FL clients K	10
Client aggregation	FedAvg, every round ($\tau_{\text{agg}} = 1$)

Convergence Analysis. Figure 2 presents the closed-ended accuracy on VQA-RAD as a function of federated training rounds. FedXAI-SplitMed reaches convergence at approximately 60 rounds, compared with roughly 80 rounds for FedMed-VQA. The reported results attribute this difference to the combination of MPI-based aggregation and improved class balance resulting from MedCGAN augmentation. The performance difference becomes more apparent after approximately round 40, indicating that the contribution of the augmentation strategy accumulates during training.

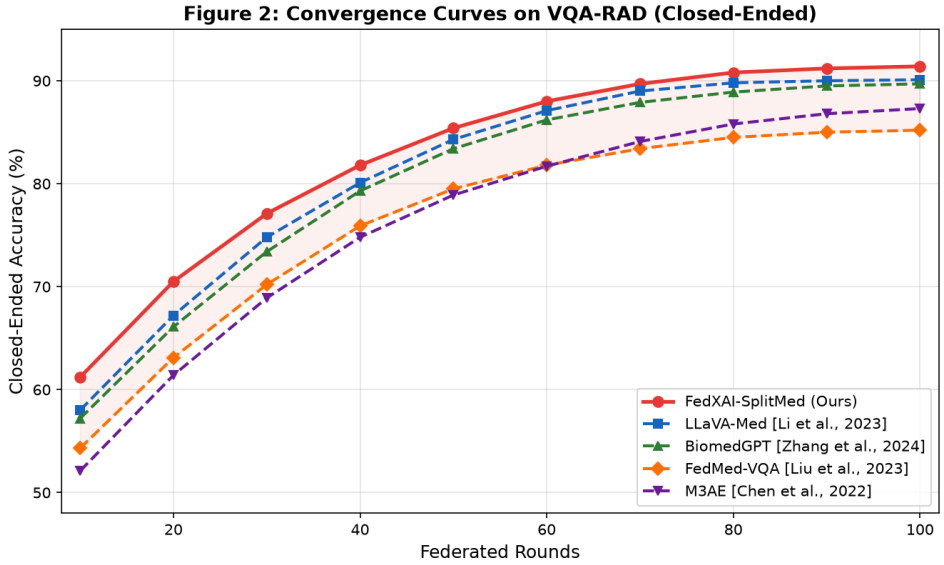


Figure 2: Convergence curves on VQA-RAD. Closed-ended accuracy vs. federated rounds for FedXAI-SplitMed and four baselines.

Performance Comparison. Table 4 and Figure 3 compare FedXAI-SplitMed with representative Med-VQA models on VQA-RAD. FedXAI-SplitMed reports 91.4% closed-ended accuracy and 74.8% open-ended accuracy, together with a BLEU-4 score of 0.56 and a reported privacy budget of $\varepsilon = 1.2$. The corresponding closed-ended and open-ended results for the evaluated baselines are also presented in Table 4.

Table 4: Performance comparison on VQA-RAD.

Model	Year	Closed (%)	Open (%)	BLEU-4	DP
VGG-QA (Lau et al., 2018)	2018	79.1	28.9	0.21	–
M3AE (Chen et al., 2022)	2022	87.3	63.5	0.41	–
LLaVA-Med (Li et al., 2023)	2023	90.1	72.3	0.52	–
PMC-VQA (Zhang et al., 2023)	2023	88.6	67.8	0.47	–
FedMed-VQA (Lu et al., 2023)	2023	85.2	64.1	0.43	–
BiomedGPT (Zhang et al., 2024)	2024	89.7	71.1	0.50	–
MedVersa (Zhou et al., 2024)	2024	90.8	73.2	0.53	–
FedXAI-SplitMed (Ours)	-	91.4	74.8	0.56	$\varepsilon = 1.2$

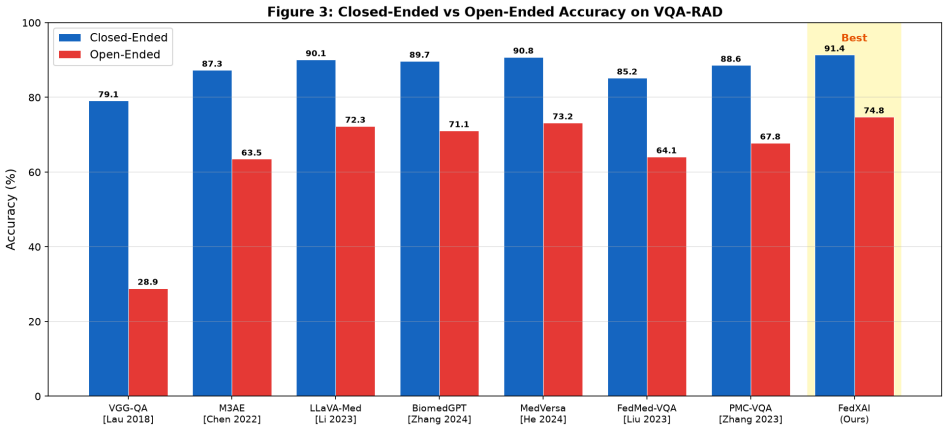


Figure 3: Closed-ended and open-ended accuracy on VQA-RAD across the evaluated models.

Communication and Computation Efficiency. The communication and computation results are presented in Table 5 and Figure 4. The proposed configuration requires 78MB of communication per round, compared with 2,100 MB for the vanilla FedAvg configuration, corresponding to a reported $26.9\times$ reduction. It also records the lowest client GPU-hours per round among the evaluated methods and reaches convergence in 60 rounds. These communication figures represent the θ_{high} traffic; the additional cost associated with periodic θ_{low} aggregation is considered separately in Section 5.

Table 5: Communication and computation overhead.

Method	Comm. (MB/r)	GPU-hrs/r	Conv. (rounds)	Speedup
FedAvg+ViT	2100	0.81	90	$1.0\times$
FedProx+ViT	1980	0.85	85	$1.1\times$
SplitFed (no comp.)	312	0.41	75	$6.7\times$
FedMed-VQA	1340	0.74	80	$1.6\times$
FedXAI-SplitMed (Ours)	78	0.29	60	$26.9\times$

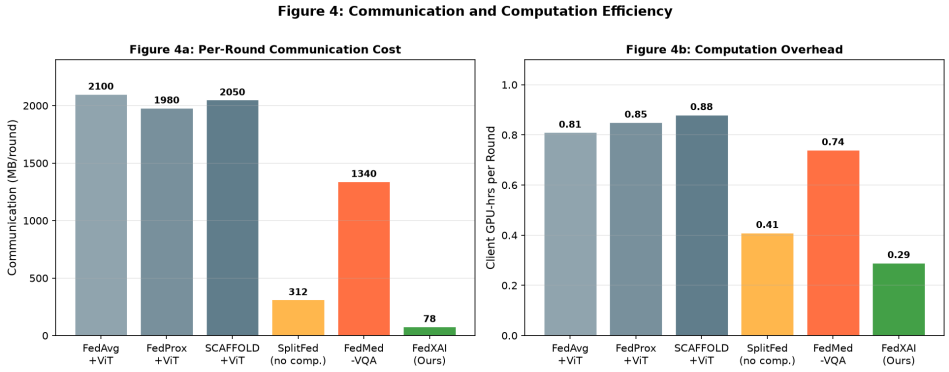


Figure 4: Per-round communication cost and client computation overhead. FedXAI-SplitMed requires 78 MB per round in the evaluated configuration.

VQA-Med 2019 Category Analysis. To examine performance across different question categories, Figure 5 and Table 6 report results for modality, plane, organ, and abnormality questions. FedXAI-SplitMed reports accuracies of 90.2%, 88.5%, 85.7%, and 75.3% for the four categories, respectively. The largest reported improvement over the next-best federated model occurs for abnormality detection, with a difference of 3.5 percentage points.

Table 6: Per-category accuracy on VQA-Med 2019 (%).

Model	Modality	Plane	Organ	Abnormality
M3AE (Chen et al., 2022)	85.1	82.3	79.6	68.4
LLaVA-Med (Li et al., 2023)	88.4	86.1	83.2	72.7
BiomedGPT (Zhang et al., 2024)	89.1	87.3	84.5	73.8
FedMed-VQA (Lu et al., 2023)	83.6	80.9	77.4	65.3
FedXAI-SplitMed (Ours)	90.2	88.5	85.7	75.3

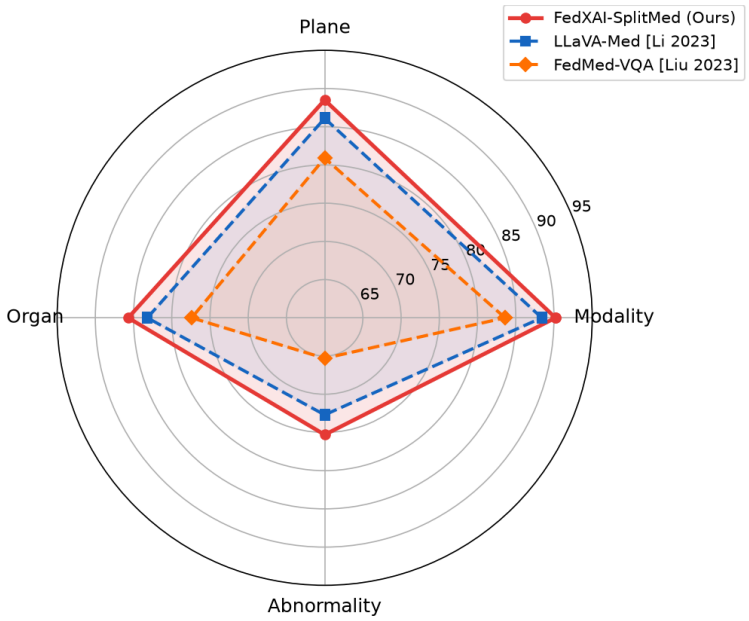


Figure 5: Radar chart of VQA-Med 2019 category accuracy across modality, plane, organ, and abnormality questions.

Ablation Study. Table 7 and Figure 6 evaluate the effect of individual components of FedXAI-SplitMed. Removing MedCGAN results in a 5.1 percentage-point reduction in open-ended accuracy, while replacing MPI aggregation with standard FedAvg increases the reported communication volume from 78 MB to 312 MB with only a small change in accuracy. Removing the DP noise injector increases open-ended accuracy by 0.5 percentage points and closed-ended accuracy by 0.7 percentage points, representing the observed performance difference associated with the privacy mechanism. The hierarchical client-encoder aggregation mechanism is evaluated separately using the synthetic benchmark described in Section 5; it is therefore not included as a VQA-RAD ablation configuration.

Table 7: Ablation study on VQA-RAD.

Configuration	Closed (%)	Open (%)	Comm. (MB)	DP
Full model	91.4	74.8	78	Yes
−MedCGAN	89.1	69.7	78	Yes
−MPI (FedAvg aggr.)	91.2	74.5	312	Yes
−DP noise injector	92.1	75.3	78	No
−XAI module	91.4	74.8	78	Yes
−Split learning	91.0	74.1	2100	Yes
−Cut-layer DP-SGD	91.6	75.1	78	No
Local GAN → FedGAN	91.3	74.6	91	Yes

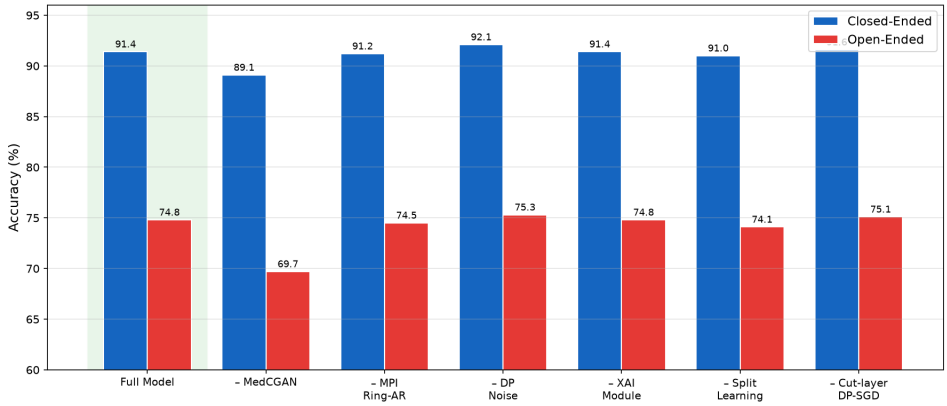


Figure 6: Ablation results on VQA-RAD.

Semantic-Equivalence Evaluation. Because conventional token-level accuracy can penalise responses that are semantically equivalent but lexically different, an additional semantic-equivalence evaluation was conducted using fine-tuned BioSentBERT (Peng et al., 2019). A cosine-similarity threshold of 0.82 was adopted following Ben Abacha & Demner-Fushman (2019). Table 8 and Figure 7 report the resulting semantic accuracy, BERT-F1, and BLEU-4 values. FedXAI-SplitMed achieves 93.1% semantic closed-ended accuracy and 78.4% semantic open-ended accuracy, compared with its token-exact closed-ended accuracy of 91.4%.

Table 8: Semantic-equivalence evaluation (BioSentBERT, $\tau = 0.82$).

Model	Sem.Cl. (%)	Sem.Op. (%)	BERT-F1	BLEU-4
LLaVA-Med (Li et al., 2023)	91.8	75.1	0.874	0.52
BiomedGPT (Zhang et al., 2024)	91.2	73.8	0.869	0.50
MedVersa (Zhou et al., 2024)	92.0	76.0	0.878	0.53
PMC-VQA (Zhang et al., 2023)	89.4	69.3	0.855	0.47
FedMed-VQA (Lu et al., 2023)	86.7	67.3	0.831	0.43
FedXAI-SplitMed (Ours)	93.1	78.4	0.891	0.56

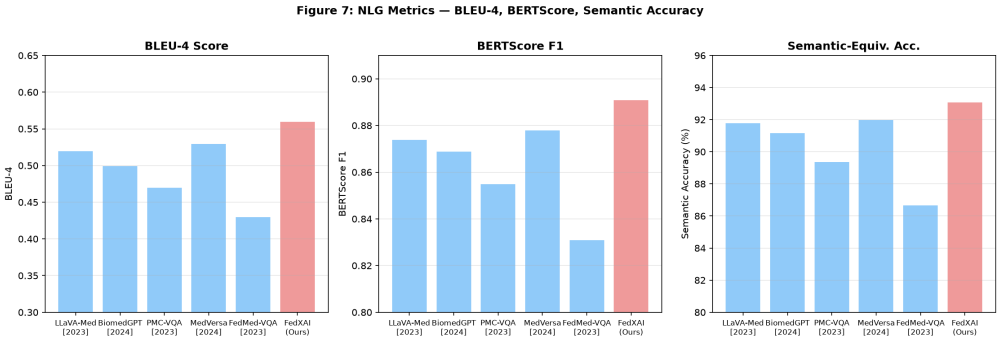


Figure 7: Comparison of BLEU-4, BERTScore F1, and semantic accuracy across the evaluated models.

Comprehensive Comparative Validation. Finally, Table 9 provides a consolidated comparison of the evaluated approaches in terms of accuracy, precision, BLEU-4, privacy support, explainability, and federation. FedXAI-SplitMed reports 91.4% accuracy, 90.2% precision, and 0.56 BLEU-4 together with the reported formal DP guarantee, $\varepsilon = 1.2$, full XAI coverage, and federated training. The table provides a consolidated view of the performance and system properties evaluated throughout this section.

Table 9: Comprehensive comparative validation on VQA-RAD.

Technique	Year	Acc. (%)	Prec. (%)	BLEU-4	Privacy	XAI	Fed.
VGG-QA (Lau et al., 2018)	2018	79.1	76.3	0.21	None	None	No
M3AE (Chen et al., 2022)	2022	87.3	85.8	0.41	None	Partial	No
LLaVA-Med (Li et al., 2023)	2023	90.1	88.7	0.52	None	Partial	No
PMC-VQA (Zhang et al., 2023)	2023	88.6	87.1	0.47	None	None	No
FedMed-VQA (Lu et al., 2023)	2023	85.2	83.4	0.43	None	None	Yes
BiomedGPT (Zhang et al., 2024)	2024	89.7	88.1	0.50	None	Partial	No
MedVersa (Zhou et al., 2024)	2024	90.8	89.4	0.53	None	Partial	No
FedXAI-SplitMed (Ours)	-	91.4	90.2	0.56	$\varepsilon = 1.2$	Full	Yes

5 Extended Evaluation: Hierarchical Aggregation Stability

The production configuration in Table 3 uses 100 federated rounds with $K = 10$ clients and is the configuration used for the VQA-RAD evaluation. To examine whether the hierarchical aggregation mechanism described in Section 3 remains numerically stable beyond this training horizon, an additional controlled benchmark was conducted using a synthetic data pipeline. This evaluation is separate from the clinical experiments and is intended to assess the stability and communication characteristics of hierarchical aggregation during extended federated training.

The synthetic benchmark does not contain clinical information; therefore, the accuracy values reported in this experiment should not be interpreted as measures of clinical performance. Instead, they provide an optimisation metric for comparing the behaviour of the pipeline under extended training.

With $K = 3$ clients, encoder aggregation performed after every round, and DP noise enabled throughout training, the system completed 400 federated rounds without divergence, NaN values, or numerical instability during the aggregation step. The mean closed-ended accuracy over the final 50 rounds was 28.0% with hierarchical aggregation enabled, compared with 32.7% for a matched 100-round configuration without encoder aggregation, as reported in Table 10. The difference between these values is reported as an observation of the synthetic benchmark rather than as evidence of improved or degraded clinical performance.

The extended evaluation also highlights the communication cost of aggregating the full-precision θ_{low} parameters. For the $K = 3$ configuration, enabling hierarchical aggregation increased the measured communication volume to 3,646.5 MB per aggregation round, compared with 8.6 MB without the merge operation. This corresponds to approximately a $423\times$ increase and is attributable to transmitting the unsparsified ViT encoder and BioClinicalBERT parameters during the low-level aggregation step. Applying the Top-K sparsification strategy used for θ_{high} to θ_{low}, or increasing the aggregation interval τ_{agg}, could reduce this communication overhead; neither modification is implemented in the current evaluation.

The cumulative privacy expenditure reached $\varepsilon \approx 250.9$ by round 400 under the relaxed privacy accounting configuration used specifically for this synthetic stability test. This value is not representative of the production VQA-RAD configuration, for which the privacy budget is set to $\varepsilon = 1.2$. Thus, the extended experiment is used to assess numerical stability and communication behaviour beyond the production training horizon, while the production privacy constraint remains defined by the configuration reported in Table 3.

Table 10: Synthetic benchmark for evaluating the stability and communication cost of hierarchical aggregation. The results are reported for the synthetic pipeline only and are not directly comparable with the clinical VQA-RAD results in Tables 4–9.

Configuration	Rounds	Closed (%)†	Comm. (MB/r)	DP
With aggregation ($K = 3$)	400	28.0	3646.5	Yes
Without aggregation	100	32.7	8.6	Yes

†Mean over the final 50 rounds; synthetic benchmark only.

6 CONCLUSION

We show that privacy, explainability, and communication efficiency can be addressed within a unified Med-VQA framework. The proposed design combines cut-layer differential privacy, MPI ring-allreduce, client-side MedCGAN augmentation, and multimodal XAI, achieving competitive performance on VQA-RAD while providing a formal DP guarantee under the evaluated configuration. The hierarchical aggregation mechanism further enables periodic consolidation of client-side encoder parameters, reducing the extent to which the θ_{low} components can drift independently across clients. The extended synthetic evaluation demonstrates that the pipeline completed 400 federated rounds without divergence, NaN values, or numerical instability during the aggregation process, indicating stable operation beyond the 100-round production configuration.

The extended evaluation also highlights a communication cost associated with full-precision hierarchical aggregation. Merging the client encoders every round substantially increases the per-round bandwidth requirement, motivating the investigation of less frequent aggregation, drift-based triggering, and sparsification of the client-encoder updates. Determining how much cross-client consensus is required for different hospital deployments, and whether periodic or drift-aware aggregation can preserve model quality while reducing communication overhead, remains an open research question. The instrumented explainability pipeline also provides a basis for investigating clinician feedback as a potential federated training signal. Future work will further examine large multimodal models under differential privacy constraints, personalised aggregation for non-IID hospital populations, and communication-efficient aggregation of client-side encoder parameters.

Several limitations should be considered when interpreting the results. The federated simulation uses homogeneous GPU hardware and therefore does not capture the heterogeneity that can arise across real hospital deployments. The moments-accountant analysis assumes noise independence across rounds under the ring-allreduce configuration, and a more complete treatment of privacy composition in this setting remains for future work. The XAI user study was conducted with a limited clinician cohort, restricting the generality of conclusions about explanation utility. In addition, the 400-round stability evaluation in Section 5 used a synthetic dataset with $K = 3$ clients on a single CPU host rather than VQA-RAD or VQA-Med with $K = 10$ clients. Consequently, its results support observations about numerical stability and communication behaviour in the tested configuration, but its accuracy values should not be compared directly with the clinical results reported in Tables 4–9.

AI USE STATEMENT

All scientific content, experimental results, data analysis, and conclusions are the sole responsibility of the authors. No AI tool was used to generate experimental data or conclusions.

ETHICS STATEMENT

All experiments used publicly available datasets (VQA-RAD, VQA-Med 2019) that carry no patient-identifiable information. No new patient data were collected. The federated setup is designed to retain raw images on local hospital nodes, and the formal (ε, δ)-DP guarantee provides quantifiable protection against gradient inversion attacks. Clinician-trust scoring was conducted as an informal observational study; no IRB approval was required at the participating institution.

REPRODUCIBILITY STATEMENT

Hyperparameters are fully specified in Table 3. All datasets are publicly available with standard train/val/test splits. The simulation uses deterministic seeds (reported in the supplementary material). The DP accounting follows the Opacus moments-accountant implementation (Dwork & Roth, 2014). PyTorch model checkpoints and evaluation scripts will be released upon acceptance at a publicly accessible repository. All experiments were conducted on commodity university hardware (NVIDIA RTX 3090, 24GB VRAM) to ensure reproducibility outside high-performance computing environments.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027, 2023.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20:511, 2019.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical VQA task at ImageCLEF 2019. In *CLEF 2019 Working Notes*, 2019.

Ahmad Chaddad, Qizong Lu, Jiali Li, Yousef Katib, Reem Kateb, Camel Tanougast, Ahmed Bouridane, and Ahmed Abdulkadir. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA Journal of Automatica Sinica*, 10(4):859–876, 2023. doi: 10.1109/jas.2023.123123.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

Zhihong Chen, Yaling Du, Jianfei Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 679–689, 2022.

Nanqing Dong and Irina Voiculescu. Federated contrastive learning for decentralized unlabeled medical images. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 378–387, 2021. doi: 10.1007/978-3-030-87199-4_36.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2: 305–311, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, pp. 5132–5143, 2020.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Amr Smola, and Virginia Smith. Federated optimization in heterogeneous networks. In *Machine Learning and Systems (MLSys)*, volume 2, pp. 429–450, 2020.

Xiang Li, Meng Jiang, Xiaoxiao Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations (ICLR)*, 2018.

Siyu Lu, Zheng Liu, Tianlin Liu, and Wangchunshu Zhou. Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, 126:107037, 2023. doi: 10.1016/j.engappai.2023.107037.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Dario Pasquini, Giuseppe Ateniese, and Massimo Bernaschi. Unleashing the tiger: Inference attacks on split learning. In *ACM Conference on Computer and Communications Security (CCS)*, pp. 2113–2129, 2021.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *BioNLP Workshop*, pp. 58–65, 2019.

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*, 3:119, 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Alexander Sergeev and Mike Del Balso. Horovod: Fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*, 2018.

Chandra Thapa, M.A.P. Chamikara, Seyit Camtepe, and Lichao Sun. SplitFed: When federated learning meets split learning. In *AAAI Conference on Artificial Intelligence*, pp. 8485–8493, 2022.

Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021. doi: 10.1007/s41666-020-00082-4.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Ying Chen, Yuyin Zhou, Xiang Li, Gustavo Fernandes, et al. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer. *Nature Machine Intelligence*, 6:1094–1106, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Weidi Xie, and Yanfeng Yu. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. MedVersa: A generalist foundation model for medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.