X-FLoRA: Cross-modal Federated Learning with Modality-expert LoRA for Medical VQA

Min Hyuk Kim, Chang Heon Kim, Seok Bong Yoo*
Department of Artificial Intelligence Convergence, Chonnam National University,
Gwangju, Korea sbyoo@jnu.ac.kr
*Corresponding author

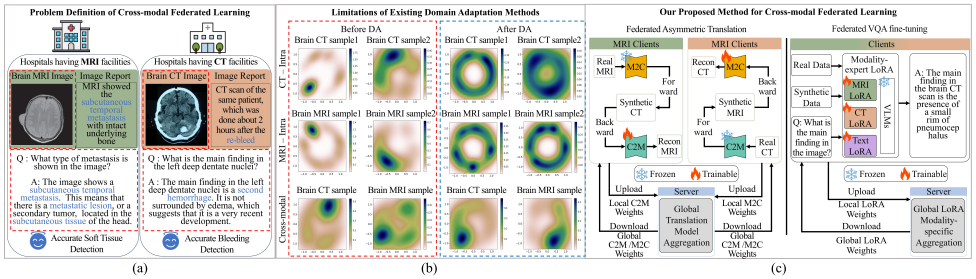

Figure 1: (a) Illustration for problem definition of cross-modal federated learning, highlighting the complementary strengths of each modality in visual question answering tasks. (b) Visualization of limitations of existing domain adaptation methods using Gaussian kernel density estimation, emphasizing the challenges of adapting MRI and CT modalities. (c) Our proposed method for cross-modal federated learning.

Abstract

Medical visual question answering (VQA) and federated learning (FL) have emerged as vital approaches for enabling privacy-preserving, collaborative learning across clinical institutions. However, both these approaches face significant challenges in cross-modal FL scenarios, where each client possesses unpaired images from only one modality. To address this limitation, we propose X-FLoRA, a cross-modal FL framework that uses modality-expert low-rank adaptation (LoRA) for medical VQA. Specifically, X-FLoRA enables the synthesis of images from one modality to another without requiring data sharing between clients. This is achieved by training a backward translation model within a federated asymmetric translation scheme that integrates clinical semantics from textual data. Additionally, X-FLoRA introduces modality-expert LoRA, which fine-tunes separate LoRA modules to strengthen modality-specific representations in the VQA task. The server aggregates the trained backward translation models and fine-tuned LoRA modules using discriminator quality scores and expert-aware weighting, which regulate the relative contributions from different clients. Experiments were conducted on VQA datasets encompassing different medical modalities, and the results demonstrate that X-FLoRA outperforms existing FL methods in terms of VQA performance.

1 Introduction

Medical visual question answering (VQA) (Lin et al., 2023; Khare et al., 2021) has emerged as a promising tool in computer-aided diagnosis, supporting clinical decision-making by generating answers to diagnostic questions based on medical images. However, the broader application of VQA methods is often constrained by data privacy concerns. Federated learning (FL) has gained significant attention for enabling privacy-preserving, decentralized model training across clinical institutions. In response to this, several federated VQA frameworks (Lao et al., 2023; Zhu et al., 2024; Tobaben et al., 2024) have been proposed to address medical VQA tasks without requiring patient data sharing. Despite this progress, federated VQA remains limited in cross-modal FL settings (Qayyum et al., 2022; Dai et al., 2024), where each client only possesses data from a single imaging modality and lacks paired samples from other modalities.

In typical clinical cross-modal FL scenarios, individual clients may have access to only magnetic resonance imaging (MRI) or computed tomography (CT) data, but not both. These modalities differ

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8391–8409
November 4-9, 2025 ©2025 Association for Computational Linguistics

substantially due to their distinct imaging mechanisms and diagnostic purposes—MRI uses magnetic fields and radio waves, whereas CT relies on ionizing radiation. As shown in Fig. 1(a), MRI and CT reports from electronic medical records (EMRs) for the same brain region often highlight different pathological features. For instance, an MRI report may describe a "subcutaneous temporal metastasis," while a CT report for the same region may note a "re-bleed," reflecting their respective strengths in soft tissue characterization and hemorrhage detection. Similarly, in the GPT-4-based medical VQA dataset (Li et al., 2023), modality-aligned responses such as "metastatic lesion" for MRI and "second hemorrhage" for CT further demonstrate the need for modality-aware understanding (Achiam et al., 2023).

The challenges of cross-modal FL arise not only from inter-modality gaps but also from intra-modality variations due to differences in imaging devices and patient characteristics. To address these issues, domain adaptation (DA) techniques have been explored (Zhao et al., 2022). As a preliminary study, we employ ResNet50 as a feature extractor to visualize feature distributions of MRI and CT datasets using the approach proposed by Chen et al.. The results, shown in Fig. 1(b), compare feature distributions before and after applying DA. The first and second rows represent intra-modal DA effects for CT and MRI, respectively, while the third row illustrates the cross-modal DA impact. These results indicate that DA alone is insufficient for addressing cross-modal heterogeneity (Chen et al., 2020) and can even lead to performance degradation (Yang et al., 2024).

To overcome these challenges, we propose X-FLoRA, a cross-modal FL framework that incorporates modality-expert low-rank adaptation (LoRA) for medical VQA, as illustrated in Fig. 1(c). X-FLoRA consists of two primary phases: federated asymmetric translation and federated VQA fine-tuning. In the first phase, each client independently trains a text-driven backward translation model—either CT-to-MRI (C2M) or MRI-to-CT (M2C)—using its data. During training, only the backward model is updated, while the forward model remains frozen. These translation models integrate images with their corresponding EMR reports to capture clinically significant textual features that may not be visually evident. Clients then upload their trained backward translation weights to a central server, which aggregates them.

In the second phase, modality-expert LoRA modules fine-tune representations for each modality—MRI, CT, and text—independently. Given the distinct characteristics of each modality, these specialized modules improve the quality of modality-specific representation. The server aggregates the fine-tuned LoRA modules using modality-specific aggregation, balancing the contributions from real and synthetic data across clients. This design allows X-FLoRA to effectively address the limitations of clinical cross-modal FL environments, enhancing both modality diversity and modality-specific representation without requiring any data sharing.

We summarize our main contributions as follows:

- To the best of our knowledge, we propose the first unified VQA framework to mitigate cross-modal heterogeneity by combining a cross-modal translation strategy with modality-specific expert fine-tuning. This approach improves both modality diversity and representation quality in federated VQA.

- We propose an FL framework for asymmetric translation, where each client trains only the backward text-driven model to complement visual features with clinical insights derived from EMRs. Furthermore, aggregation based on discriminator quality scores increases the influence of clients with higher-quality translation models.

- We introduce modality-expert LoRA, a lightweight and modality-specific adaptation mechanism. Separate LoRA modules are applied to each modality, and a modality-specific aggregation strategy ensures a balanced integration of real and synthesized data from diverse clients.

2 Related Work

2.1 Visual Question Answering

VQA (Ji et al., 2024; Naik et al., 2024; Xing et al., 2024; Song et al., 2024; Li et al., 2024a, 2023; Liu et al., 2023; Wang et al., 2024a; Yan et al., 2024) is an interdisciplinary task that integrates computer vision and natural language processing to generate answers to natural language questions about visual content. Building on the progress in general-domain VQA, there has been a surge of interest in

adapting VQA for medical applications. Recent studies show that autoregressive decoder-based large language models (LLMs) and visual language models (VLMs), when fine-tuned on medical datasets, demonstrate strong performance on clinical tasks. For example, BioMistral (Labrak et al., 2024), adapted from Mistral-7B (Jiang et al., 2023), has shown impressive results on complex medical question answering benchmarks, such as medical licensing exams and PubMed-based queries (Jin et al., 2019). Similarly, specialized medical VLMs like LLaVA-Med (Li et al., 2023), derived from LLaVA (Liu et al., 2023), and MedFlamingo (Moor et al., 2023), based on Open-Flamingo (Awadalla et al., 2023), have demonstrated effectiveness in radiological (Lau et al., 2018) and pathological (He et al., 2020) VQA tasks. Despite these advancements, current methods are limited in FL settings involving cross-modal medical data, where they struggle to model the inherent differences in visual and textual modality characteristics.

2.2 Vertical Multimodal Federated Learning

Protecting patient privacy has become a critical concern in the digital healthcare era, especially given the risks associated with misuse or unauthorized commercialization of sensitive data (Chiruvella et al., 2021). To address these issues, several vertical multimodal FL approaches have been developed (Zhang et al., 2023; Qayyum et al., 2022; Yang et al., 2022). Zhang et al. introduced UTMP, a federated learning framework in which unimodal clients collaboratively train a multimodal model through hierarchical encoder-decoder aggregation. Qayyum et al. proposed a collaborative FL framework for multimodal COVID-19 diagnosis on edge devices, enabling clients with either X-ray or ultrasound data to train a shared model without exchanging raw data. Yang et al. presented a cross-modal federated human activity recognition framework that uses a feature-disentangled network with both modality-agnostic and modality-specific encoders, enabling collaborative learning from clients with heterogeneous sensor and video modalities. However, these prior works do not consider the intrinsic characteristics of medical imaging, such as differences in imaging physics (e.g., CT vs. MRI) and semantic focus in clinical reports. To bridge this gap, we propose a new FL strategy combining federated asymmetric translation and federated VQA fine-tuning, which explicitly considers modality-specific features through the use of asymmetric forward/backward models and modality-expert LoRA modules.

2.3 Image-to-Image Translation

A wide range of image-to-image translation techniques have been developed in recent years (Zhu et al., 2017; Huang et al., 2018; Isola et al., 2017; Cheng et al., 2023; Xia et al., 2024; Li et al., 2024b; Xu et al., 2024). Zhu et al. introduced CycleGAN, which enables unpaired image-to-image translation using a cycle-consistency loss. Huang et al. proposed MUNIT, which separates images into shared content and domain-specific style codes to generate diverse outputs. Isola et al. developed pix2pix for supervised image-to-image translation using paired data, directly learning mappings from input to output images. Although effective, most of these methods assume access to paired multimodal datasets—an assumption that does not hold in vertical multimodal FL scenarios, where data are distributed across institutions. To address this limitation, we propose a federated approach to unpaired cross-modal image translation that leverages modality-specific clinical text reports as semantic guidance. This strategy enriches the translation process with medically relevant details that may be implicit or missing in visual data alone.

3 Methodology

3.1 Overall

As shown in Fig. 2, X-FLoRA consists of two key phases: federated asymmetric translation and federated VQA fine-tuning. In the federated asymmetric translation phase, there are N_m clients with MRI data and N_c clients with CT data. Each group of clients trains backward translation models specific to their modality, while the forward translation model is provided by other modality clients and remains frozen. The central server aggregates the backward translation models from all clients. Subsequently, clients download the aggregated backward weights, enabling both MRI and CT clients to perform federated asymmetric translation without sharing data directly. This phase is repeated over R_t rounds. After these rounds, each client generates synthetic images of the other modality.

In the next phase, modality-expert LoRA modules are applied to the respective modality encoders using the synthetic images. The weights from these LoRA modules are then uploaded to the server for global aggregation, specific to each modality. Addi-

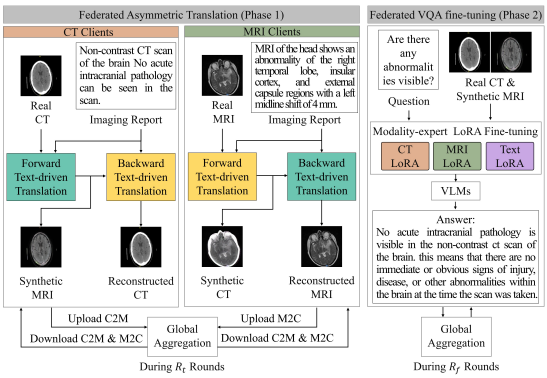

Figure 2: Overall architecture of the X-FLoRA framework.

tionally, expert-aware weighting is used to balance the contributions of real and synthetic data. This fine-tuning phase is repeated over R_f rounds, promoting increased modality diversity and enhancing the robustness of modality-aware representations. After all rounds are completed, the final global VQA model is obtained.

3.2 Federated Asymmetric Translation

Inspired by cycle consistency (Radford et al., 2021), we propose a federated asymmetric translation that enables each client to train cross-modal translation even if it possesses only a single type of modality data as shown in Fig. 3. In this phase, each client possesses real data x and corresponding imaging report t. In addition, clients perform forward translator F, which receives x and t as inputs, to generate a synthetic image.

3.2.1 Forward and Backward Text-driven Translation

In the forward process, the text encoder extracts text features, while the image encoder and residual blocks extract image features. The extracted image and text features are then fused via text-driven attention, enabling the translator to generate modality-consistent synthetic images enriched with clinically relevant textual cues.

Each client generates synthetic images using the frozen forward translator F, and subsequently applies a backward translator B, with the same architecture as F, to reconstruct the original image. This reconstruction is used to train B and a discriminator D, which distinguishes between real and reconstructed images. Specifically D and B are trained as follows:

$$\hat{D}, \hat{B} = \underset{D}{\operatorname{argmax}} \underset{B}{\min} \, \mathcal{L}_{total}, \tag{1}$$

where, $\mathcal{L}_{total}$ denotes the total loss function, defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \eta \mathcal{L}_{id}, \tag{2}$$

where η balances two objectives: adversarial loss ($\mathcal{L}_{adv}$), which ensures realism of the reconstructed image, and identity loss ($\mathcal{L}_{id}$), which ensures fidelity to the original input. The adversarial loss is formulated as follows:

$$\mathcal{L}_{adv} = \mathbb{E}_x \sim p_{\text{data}}(x)[||1 - D(B(F(x,t),t))||_2], \tag{3}$$

where $x \sim p_{\text{data}}(x)$ denotes the data distribution of real data and $\|\cdot\|_2$ denotes the L2 norm. The identity loss minimizes the pixel-wise difference between the reconstructed image and the original input. This loss is formulated as follows:

$$\mathcal{L}_{id} = ||B(F(x,t),t) - x||_1, \tag{4}$$

where $\|\cdot\|_1$ denotes the L1 norm. After local training, the central server aggregates the backward translation weights $\theta^{r,i}_{c2m}$ and $\theta^{r,j}_{m2c}$ received from the i-th MRI and j-th CT clients in the r-th communication round, respectively.

3.2.2 Discriminator Score-based Aggregation

To enhance reliability and stability across clients, we introduce a discriminator score-based aggregation. Each MRI client transmits three components to the server: (1) backward translation weights $\theta^{r,i}_{c2m}$, (2) the backward model-based gradient $g^{r,i}_m$, and (3) a discriminator-based reliability scores $s^{r,i}_m$ (from 0 to 1) . The server aggregates MRI client updates using:

$$\theta^{r+1}_{c2m} = \frac{1}{N_m} \sum_{i=1}^{N_m} (\omega^{r,i}_{m,s} + \omega^{r,i}_{m,g}), \tag{5}$$

$$\omega^{r,i}_{m,s} = \frac{s^{r,i}_m \cdot \theta^{r,i}_{c2m}}{\sum_{i=1}^{N_m}(s^{r,i}_m) + \sqrt{G^r_m}}, \tag{6}$$

$$\omega^{r,i}_{m,g} = \frac{g^{r,i}_m \cdot \theta^{r,i}_{c2m}}{\sum_{i=1}^{N_m}(s^{r,i}_m) + \sqrt{G^r_m}}. \tag{7}$$

Here, $\omega^{r,i}_{m,s}$ and $\omega^{r,i}_{m,g}$ denote the reliability-based normalized model weight and the gradient-based normalized model weight from i-th client, respectively. Moreover, θ^{r+1}_{c2m} denotes the aggregated weights in the $(r+1)$-th round. In addition, the discriminator score $s^{r,i}_m$ is defined as follows:

$$s^{r,i}_m = \mathbb{E}_{x^i_m} \sim p_{\text{data}}(x^i_m)[D^{r,i}_m(x^i_m)], \tag{8}$$

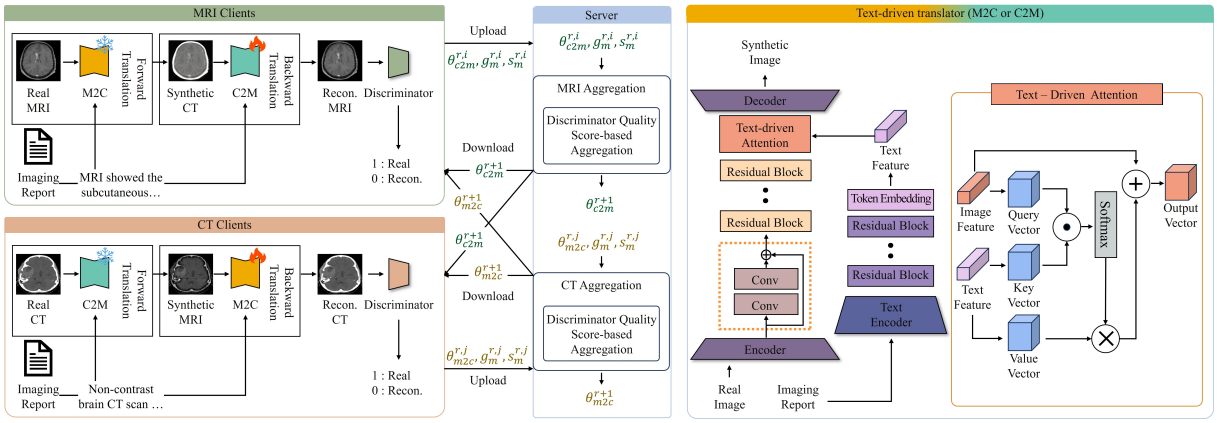

Figure 3: Architecture of federated asymmetric translation.

where x_m^i denotes the real data x_m^i held by the i-th MRI client and $D_m^{r,i}$ denotes the local discriminator of the i-th MRI client in r-th round. Moreover, G_m^r represents the accumulated squared sum of the gradients (momentum) in r-th round, formulated as follows:

$$G_m^r = G_m^{r-1} + \sum_{i=1}^{N_m} (g_m^{r,i})^2. \tag{9}$$

By using G_m^r and $s_m^{r,i}$, this aggregation approach prioritizes contributions from clients whose discriminators better distinguish real from generated images, thereby enhancing model robustness.

For CT clients, the server aggregates the backward translation weights $\theta_{m2c}^{r,j}$ in a similar strategy, following the definition provided in Eq. (7). Specifically, the j-th CT client's momentum G_c^r and discriminator score $s_c^{r,j}$ are calculated using the gradient of the backward translator $g_c^{r,j}$, x_c^j, and discriminator $D_c^{r,j}$ based on Eqs. (8) and (9).

3.3 Federated Modality-Expert Fine-tuning

Training VLMs for VQA typically demands extensive VRAM and significant computational resources, necessitating efficient fine-tuning strategies. Moreover, in federated medical environments, data across modalities (e.g., MRI, CT) exhibit inherently distinct characteristics. To effectively capture modality-specific features in cross-modal FL for medical VQA, we propose modality-expert LoRA fine-tuning, which independently learns discriminative features from each imaging modality.

3.3.1 Modality-expert LoRA

As illustrated in Fig. 4, the proposed modality-expert LoRA architecture is designed separately for each modality and enables efficient training with substantially reduced computational over-head compared to full-scale VLM fine-tuning. Each client uses fixed, modality-specific encoders denoted as W_m for MRI, W_c for CT, and W_t for text. These encoders extract feature representations $W_m v_m$, $W_c v_c$, and $W_t v_t$ where v_m, v_c, and v_t are the modality-specific input vectors. To enhance these representations without updating the pre-trained encoders, we apply LoRA fine-tuning as follows:

$$\hat{v}_k = W_k v_k + \beta_k \alpha_k v_k, \quad k \in \{m, c, t\}, \tag{10}$$

where α and β are low-rank weight matrices in the LoRA layers. Specifically, α_m, α_c, and α_t project input features into low-rank subspaces of dimensions $\mathbb{R}^{d \times r_m}$, $\mathbb{R}^{d \times r_c}$, and $\mathbb{R}^{d \times r_t}$, respectively. Corresponding matrices β_m, β_c, and β_t project them back into the original feature space. This decomposition-reconstruction approach enables efficient fine-tuning of MRI and CT modality-specific feature. The LoRA modules are applied to the linear projection matrices of the modality-specific encoders, including the key and value projection layers in the attention blocks as well as the linear layers in the feedforward blocks. Each LoRA module is integrated alongside each linear matrix in the model (Hu et al., 2022). Moreover, these fine-tuned features are fused by using a projector to consider inter-modal representation. After local training, the fine-tuned modality-expert LoRA weights are transmitted to the central server for aggregation.

3.3.2 Modality-specific Aggregation

The central server receives the fine-tuned LoRA weights from MRI and CT clients and performs

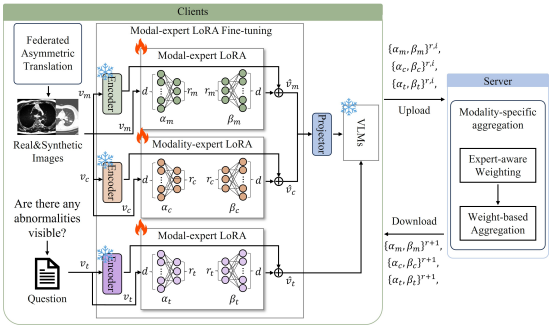

Figure 4: Architecture of the federated visual question answering fine-tuning.

aggregation. In our framework, these modality-expert LoRA weights are categorized according to the modality on which they were trained, maintaining separate sets of weights for MRI, CT, and text data. Unlike conventional FL approaches (Li et al., 2020; McMahan et al., 2017; Li et al., 2021; Kairouz et al., 2021) that aggregate all modality weights jointly, we propose modality-specific aggregation, which processes the weights independently for each modality. This separation enables each modality-expert LoRA module to better capture and represent the unique characteristics of CT, MRI, and text inputs.

To further enhance aggregation quality, we introduce an expert-aware weighting scheme that differentiates the contributions of weights based on whether they were trained on real or synthetic data. This allows the system to adjust the influence of each client's update during aggregation. The expert-aware weight for the i-th MRI client is defined as:

$$\lambda_m^i = \begin{cases} \dfrac{\epsilon}{\epsilon \cdot N_m^r + N_m^s} & \text{if } i \in \mathcal{R}_m \\ \dfrac{1}{\epsilon \cdot N_m^r + N_m^s} & \text{otherwise} \end{cases}, \quad (11)$$

where λ_m^i denotes the MRI aggregation weight for the i-th client, and $\mathcal{R}_m$ is the set of indices corresponding to clients with real MRI data. Additionally, N_m^r and N_m^s represent the number of clients with real and synthetic MRI data, respectively. The hyperparameter ϵ controls the relative scaling between from real and synthetic data. A similar expert-aware weighting strategy is applied to CT clients, producing CT aggregation weights λ_c^i using the same formulation as in Eq. (11). For all clients contributing text data, the aggregation weight λ_t^i is set to 1.

Based on λ_k^i, the server aggregates the MRI LoRA weights $\{\alpha_m, \beta_m\}^{r,i}$, CT LoRA weights $\{\alpha_c, \beta_c\}^{r,i}$, and text LoRA weights $\{\alpha_t, \beta_t\}^{r,i}$ in the r-th round. It balances the influence of real and synthetic data on modality-specific representations for MRI and CT. The weight-based aggregation process is defined as follows:

$$\{\alpha_k, \beta_k\}^{r+1} = \frac{1}{N_k} \sum_{i=1}^{N_k} \lambda_k^i \{\alpha_k, \beta_k\}^{r,i}, \quad k \in \{m, c, t\}, \quad (12)$$

where N_t denotes a total number of clients ($N_m + N_c$). After federated VQA fine-tuning phase is completed, the final global VQA model with the modality-specific LoRA module is obtained.

4 Experiments

4.1 Dataset and Evaluation Metric

The experiments utilize a combined dataset drawn from the LLaVA-Med dataset (Li et al., 2023) and the VQA-RAD dataset (Lau et al., 2018). LLaVA-Med is designed to support instruction-following multimodal learning across multiple institutions. It is built using image-text pairs sourced from PubMed Central and includes a GPT-4-generated instruction-tuning set, comprising 10K samples across modalities such as CT and MRI. The VQA-RAD dataset comprises 3,515 clinician-authored QA pairs and 315 radiology images, with imaging reports generated using GPT-4 based on the QA pairs, which include closed-ended answers (i.e., yes/no) and open-ended answers with a short phrase. In our federated learning setup, X-FLoRA is trained across eight clients. Four clients use MRI data (N_m = 4), and the other four use the CT data (N_c = 4), both the LLaVA-Med and VQA-RAD datasets. Appendix C provides additional experiments varying the number of MRI and CT clients, with comparisons against baseline FL methods.

To evaluate the quality of the generated responses, we use four standard automatic metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) for the LLaVA-med dataset and accuracy for the VQA-RAD dataset. These metrics assess both surface-level and semantic aspects of generation. BLEU captures lexical precision; METEOR balances precision and recall; ROUGE evaluates n-gram overlap; and

Dataset	LLaVA-Med					VQA-RAD		
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr	Accuracy (%)		
						Open	Closed	Overall
FedAvg (McMahan et al., 2017)	0.2892	0.1486	0.3467	0.3682	0.5003	51.39	73.56	64.76
FedProx (Li et al., 2020)	0.2859	0.1512	0.3450	0.3702	0.5064	52.09	74.13	65.38
MOON (Li et al., 2021)	0.2935	0.1561	0.3492	0.3604	0.5152	53.31	76.37	67.21
FedProto (Tan et al., 2022)	0.2943	0.1568	0.3486	0.3541	0.5176	54.04	77.51	68.19
IOS (Wu et al., 2023)	0.2913	0.1510	0.3508	0.3587	0.5190	55.37	78.05	69.04
FedTGP (Zhang et al., 2024)	0.3012	0.1572	0.3561	0.3672	0.5237	57.15	78.46	70.00
FedMedVLP (Lu et al., 2023)	0.2955	0.1540	0.3533	0.3597	0.5196	55.81	78.30	69.53
FedKIM (Wang et al., 2024b)	0.3015	0.1581	0.3588	0.3701	0.5279	56.12	78.49	70.14
X-FLoRA	**0.3191**	**0.1630**	**0.3704**	**0.3954**	**0.5430**	**60.42**	**81.10**	**72.89**

Table 1: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, CIDEr and accuracy on the LLaVA-Med and VQA-RAD dataset.

CIDEr measures TF-IDF-weighted similarity, placing higher importance on informative content in vision-language tasks. In addition, We evaluate translators with four metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), learned perceptual image patch similarity (LPIPS), and frechet inception distance (FID).

4.2 Implementation Details

The experiments follow a federated-by-dataset scenario (McMahan et al., 2017), where each client constructs its own local dataset and collaborates with a central server through FL. All experiments were conducted using a single NVIDIA L40S GPU.

The text encoder (Radford et al., 2021) consists of 12 transformer blocks, each comprising layer normalization, multi-head self-attention (heads of eight, input length of 77, hidden size of 512), a residual connection, and a feed-forward network with GELU activation. This structure is repeated across all layers. After the transformer, the embedding token is passed through a linear projection to obtain the final text representation.

The image encoder (Zhu et al., 2017) begins with a 7×7 convolutional layer using reflection padding and ReLU activation. This is followed by two downsampling blocks, each with a 3×3 convolution and ReLU, reducing spatial resolution by a factor of 4. Next, nine residual blocks are applied, each composed of two 3×3 convolutional layers, normalization, and ReLU. The discriminator extends this encoder with a final 1-channel convolutional layer followed by a sigmoid activation.

We employ stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.001. X-FLoRA is trained for a total of 150 global rounds, consisting R_t = 50 rounds for translational pretraining and R_f = 100 rounds for federated fine-tuning. Additionally, we set both η and ϵ to 1.5, and r_m, r_c, and r_t to 16, 32, and 8. Appendix C provides experiments to optimize these parameters.

4.3 Results and Analysis

We compare X-FLoRA with several baseline FL methods, including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), MOON (Li et al., 2021), FedProto (Tan et al., 2022), IOS (Wu et al., 2023), FedTGP (Zhang et al., 2024), FedMedVLP (Lu et al., 2023) and FedKIM (Wang et al., 2024b), using both the LLaVA-Med and VQA-RAD datasets. The VQA model architecture proposed by Liu et al. (2023) is used as the backbone because it has been broadly utilized in the medical domain. All the baseline models are trained and evaluated from scratch using the respective authors' experimental settings and open-source code. We report the average performance over three runs using different random seeds, with a standard deviation of 6.3×10^{-3}, confirming the consistency of our results. The best scores are highlighted in bold across all tables. As shown in Table 1, X-FLoRA achieves superior VQA performance across all five metrics compared to baseline FL methods. This improvement stems from the integration of cross-modal synthetic data, which enables collaborative training even under unpaired modality settings. Additionally, modality-specific fine-tuning via LoRA modules enhances representation quality by adapting to the distinct characteristics of each imaging domain.

Figure 5 provides a qualitative comparison of responses generated by X-FLoRA, IOS, and FedTGP for a given CT scan. In the figure, the red arrow indicates a mass-arising lesion near the rib, red text indicates incorrect or inconsistent responses, while BLEU text represents accurate and contextually appropriate answers. X-FLoRA successfully identifies key clinical features—such as the intact

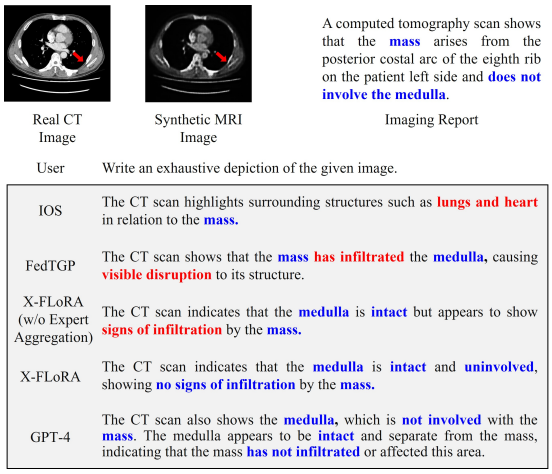

Figure 5: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

Method		LLaVA-Med		VQA-RAD		
		Metric				
		METEOR	CIDEr	Accuracy (%)		
DA	FL			Open	Closed	Overall
SEA (Wang et al., 2023)	FedTGP	0.3569	0.5231	57.31	78.50	70.07
CAF (Xie et al., 2022)	IOS	0.3514	0.5205	55.42	78.37	69.12
CAF (Xie et al., 2022)	IOS	0.3510	0.5181	55.50	78.53	70.19
SEA (Wang et al., 2023)	FedTGP	0.3558	0.5207	57.40	78.55	69.59
X-FLoRA		**0.3704**	**0.5430**	**60.42**	**81.10**	**72.89**

Table 2: Performance of FL methods with DA models for VQA performance on the LLaVA-Med and VQA-RAD dataset.

and uninvolved state of the medulla and the absence of mass infiltration—matching the GPT-4-generated reference from the imaging report. Moreover, in this synthetic MRI image, it appears that the medulla remains intact. This demonstrates X-FLoRA's strong grounding capability in clinically relevant visual content. Appendix C also provides additional example comparisons of X-FLoRA and other FL methods.

Table 2 evaluates X-FLoRA when integrated with DA techniques, specifically SEA (Wang et al., 2023) and CAF (Xie et al., 2022). We also examine combinations of DA methods with state-of-the-art FL models such as FedTGP and IOS. X-FLoRA consistently outperforms these combinations, highlighting the benefit of federated asymmetric translation in improving VQA performance.

Moreover, Table 3 presents a comparison of the performance of F and B of asymmetric translation with CycleGAN. We evaluate F with LPIPS and FID, and assess both F and B using PSNR, SSIM. Asymmetric translation surpasses CycleGAN through higher PSNR and SSIM and lower LPIPS and FID. This result is attributed to com-

Forward	Architecture	Metric		Forward + Backward	Architecture	Metric	
		LPIPS(↓)	FID (↓)			PSNR (↑)	SSIM (↑)
CT→MRI	CycleGAN (Only Image)	0.25	119.83	CT→MRI→CT	CycleGAN (Only Image)	25.51	0.78
	Ours (Image + Text)	**0.22**	**90.22**		Ours (Image + Text)	**27.23**	**0.87**
MRI→CT	CycleGAN (Only Image)	0.24	109.66	MRI→CT→MRI	CycleGAN (Only Image)	27.24	0.81
	Ours (Image + Text)	**0.23**	**105.05**		Ours (Image + Text)	**28.57**	**0.88**

Table 3: Performance of asymmetric translation compared with CycleGAN on the LLaVA-Med dataset.

Models	Trainable Params	Convergence Round	Training Time (hours)
FedProto (Tan et al., 2022)	13G	202	33.6
IOS (Wu et al., 2023)	13G	188	30.6
FedTGP (Zhang et al., 2024)	13G	184	30.6
X-FLoRA	**58M**	**149**	**25.1**

Table 4: Computational complexity of FL methods on the LLaVA-Med dataset.

Federated Asymmetric Translation			Federated VQA Finetuning		CIDEr
Text	Translation	Discriminator-based Aggregation	Modality-expert LoRA	Modality-specific Aggregation	
✓	✓	✓	✓	✓	**0.5430**
✓	✓		✓	✓	0.5407
✓	✓	✓	✓		0.5401
✓	✓	✓			0.5357
	✓	✓	✓	✓	0.5304
					0.5003

Table 5: Ablation study for X-FLoRA on the LLaVA-Med dataset in terms of the CIDEr.

plementing visual features with clinical insights through text corresponding to the images.

Table 4 compares the training efficiency in several FL methods. X-FLoRA not only converges faster in fewer training rounds (149 rounds) but also requires much fewer trainable parameters (58 mega) compared to other methods (13 giga), owing to the use of lightweight LoRA modules. Specifically, we adopted the ViT-L/14 model for visual encoder and transformer layers within Vicuna-7B model for text encoder, as introduced in the original LLaVA architecture. Since the total parameter size of visual and text encoders is 4 giga parameters out of the 13 giga parameters of the entire model, the use of LoRA is reasonable for efficient fine-tuning. This makes X-FLoRA particularly suitable for resource-constrained clinical environments.

4.4 Ablation Study

This section analyzes the contribution of each X-FLoRA component. Table 5 presents ablation results, where a checkmark (✓) indicates module activation. The first and last rows show X-FLoRA and backbone performance, respectively. The second and third rows report the results when discriminator-based aggregation and modality-specific aggregation are excluded, respec-

tively. The fourth row reports the results when synthetic images are used without federated VQA finetuning, which reflects the performance of full fine-tuning. It demonstrates that fine-tuning with LoRA yields better performance than full fine-tuning. This is supported by experimental evidence showing that selectively fine-tuning leads to better performance compared to full fine-tuning (Hu et al., 2022). The fifth row reports the results when only images are used in translator. Comparing each module with the X-FLoRA confirms that each component contributes to performance improvements.

5 Conclusion

This study tackles the critical challenge of cross-modal heterogeneity in federated VQA. We propose X-FLoRA, a comprehensive framework that integrates asymmetric text-driven translation, modality-expert LoRA modules, and global aggregation strategies to effectively address this issue. X-FLoRA selectively trains backward translation models, shares forward translations, applies modality-specific fine-tuning, and aggregates a global model, all within the FL paradigm to enhance VQA accuracy. Our experimental results demonstrate that X-FLoRA outperforms existing FL baselines, achieving state-of-the-art VQA performance on both the LLaVA-Med and VQA-RAD datasets, while maintaining computational efficiency. These results underscore the effectiveness of the proposed design in managing unpaired multimodal data in decentralized clinical settings.

6 Limitations

In addition, although X-FLoRA demonstrates improved quantitative performance on benchmark datasets such as LLaVA-Med and VQA-RAD, the clinical interpretability and reliability of the generated responses have not yet been directly assessed through expert review. Medical decision-making often involves context-specific and nuanced reasoning, which cannot be fully captured by automated metrics alone. Therefore, it would be valuable to examine whether the model's outputs align with clinical expectations in real-world scenarios. Future work should include qualitative evaluations by domain experts, such as structured assessments conducted by radiologists or physicians, to better understand how the model's responses are perceived and trusted in clinical environments. Such evaluations would help bridge the gap between algorithmic performance and practical usability, ultimately contributing to the safe and effective deployment of federated VQA systems in healthcare.

This study focuses on MRI and CT, widely used and clinically complementary imaging modalities, providing a robust foundation for evaluating the proposed framework. Although these modalities are robust, other modalities such as ultrasound, PET, and digital pathology remain unexplored. In future work, we will extend X-FLoRA by using specialized forward and backward translators adapted to other modalities. Expanding the number of forward and backward translators enables the framework to accommodate a wider range of modalities. However, this poses the challenge of mapping modality-specific representations to a common feature space due to the substantial heterogeneity in imaging and semantic characteristics across modalities. This challenge becomes even more pronounced when incorporating modalities beyond MRI and CT, such as ultrasound, PET, and digital pathology. To address this, we propose advanced feature alignment techniques, including modality-invariant representation learning and contrastive alignment with clinical text embeddings. These methods aim to enhance cross-modal knowledge transfer despite significant inter-modality gaps.

Acknowledgments

This work was supported by the IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, RS-2023-00256629, RS-2022-00156287, RS-2024-00437718).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. 2020. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505.

Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M Alvarez, Zhangyang Wang, and Anima Anandkumar. 2021. Contrastive syn-to-real generalization. *arXiv preprint arXiv:2104.02290*.

Bin Cheng, Zuhao Liu, Yunbo Peng, and Yue Lin. 2023. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22736–22746.

Varsha Chiruvella, Achuta Kumar Guddati, and 1 others. 2021. Ethical issues in patient data ownership. *Interactive journal of medical research*, 10(2):e22269.

Qian Dai, Dong Wei, Hong Liu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. 2024. Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1445–1453.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134.

Huishan Ji, Qingyi Si, Zheng Lin, Yanan Cao, and Weiping Wang. 2024. Towards one-to-many visual question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16931–16943.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, evendra Singh Chaplot, Diegode las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, and 1 others. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1033–1036. IEEE.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Mingrui Lao, Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S Lew. 2023. Fedvqa: Personalized federated visual question answering over heterogeneous scenes. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7796–7807.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. MMedAgent: Learning to use medical tools with multi-modal agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8745–8760.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

Zhenglin Li, Bo Guan, Yuanzhou Wei, Yiming Zhou, Jingyu Zhang, and Jinxin Xu. 2024b. Mapping new realities: Ground truth image creation with pix2pix image-to-image translation. *arXiv preprint arXiv:2404.19265*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

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

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Nandita Shankar Naik, Christopher Potts, and Elisa Kreiss. 2024. CommVQA: Situating visual question answering in communicative contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13362–13377.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir. 2022. Collaborative federated learning for healthcare: Multimodal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, 3:172–184.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Lingyun Song, Chengkun Yang, Xuanyu Li, and Xuequn Shang. 2024. A robust dual-debiasing VQA model based on counterfactual causal effect. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4242–4252.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. 2022. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8432–8440.

Marlon Tobaben, Mohamed Ali Souibgui, Rubèn Tito, Khanh Nguyen, Raouf Kerkouche, Kangsoo Jung, Joonas Jälkö, Lei Kang, Andrey Barsky, Vincent Poulain d'Andecy, and 1 others. 2024. Neurips 2023 competition: Privacy preserving federated learning document vqa. *arXiv preprint arXiv:2411.03730*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Qunbo Wang, Ruyi Ji, Tianhao Peng, Wenjun Wu, Zechao Li, and Jing Liu. 2024a. Soft knowledge prompt: Help external knowledge become a better teacher to instruct llm in knowledge-based vqa. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6132–6143.

Xiaochen Wang, Jiaqi Wang, Houping Xiao, Jinghui Chen, and Fenglong Ma. 2024b. Fedkim: Adaptive federated knowledge injection into medical foundation models. *arXiv preprint arXiv:2408.10276*.

Yucheng Wang, Yuecong Xu, Jianfei Yang, Zhenghua Chen, Min Wu, Xiaoli Li, and Lihua Xie. 2023. Sensor alignment for multivariate time-series unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10253–10261.

Zhaoxian Wu, Tianyi Chen, and Qing Ling. 2023. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*.

Mengfei Xia, Yu Zhou, Ran Yi, Yong-Jin Liu, and Wenping Wang. 2024. A diffusion model translator for efficient image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Binhui Xie, Shuang Li, Fangrui Lv, Chi Harold Liu, Guoren Wang, and Dapeng Wu. 2022. A collaborative alignment framework of transferable knowledge extraction for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6518–6533.

Xiaoying Xing, Peixi Xiong, Lei Fan, Yunxuan Li, and Ying Wu. 2024. Learning to ask denotative and connotative questions for knowledge-based VQA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8301–8315.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and 1 others. 2024. Mlevlm: Improve multi-level progressive capabilities based

on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.

Quan Yan, Junwen Duan, and Jianxin Wang. 2024. Multi-modal concept alignment pre-training for generative medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5378–5389.

Mingjing Yang, Zhicheng Wu, Hanyu Zheng, Liqin Huang, Wangbin Ding, Lin Pan, and Lei Yin. 2024. Cross-modality medical image segmentation via enhanced feature alignment and cross pseudo supervision learning. *Diagnostics*, 14(16):1751.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. 2022. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3063–3071.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. 2024. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16768–16776.

Rongyu Zhang, Xiaowei Chi, Guiliang Liu, Wenyi Zhang, Yuan Du, and Fangxin Wang. 2023. Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation. *arXiv preprint arXiv:2303.15486*.

Ziyuan Zhao, Fangcheng Zhou, Kaixin Xu, Zeng Zeng, Cuntai Guan, and S Kevin Zhou. 2022. Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE transactions on medical imaging*, 42(3):633–646.

He Zhu, Ren Togo, Takahiro Ogawa, and Miki Haseyama. 2024. Prompt-based personalized federated learning for medical visual question answering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1821–1825. IEEE.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.

Appendix

The appendix of this study provides comprehensive details that support the main framework, methodology, and experimental results presented in the paper. Below is a summary of each section: Section A provides a detailed explanation of the core algorithms of X-FLoRA. Section B presents a discussion of the stability of discriminator-based aggregation and experiments on cross-modality. Section C presents additional quantitative and qualitative results. Section D presents qualitative VQA results of X-FLoRA and other methods.

A Method Algorithms

A.1 Federated Asymmetric Translation

This algorithm 1 describes the training process of text-driven translation. Each client generates synthetic images and reconstructs the original image.

A.1.1 Discriminator Quality Score-based Aggregation

This algorithm 2 details the aggregation process using discriminator quality scores and gradient information.

A.2 Federated VQA FInetuning

This algorithm 3 presents the finetuning phase for modality-expert LoRA modules. Each client updates only the lightweight LoRA parameters for MRI, CT, and text modalities.

A.2.1 Modality-specific Aggregation

This algorithm 4 defines the aggregation process for modality-expert LoRA weights.

Algorithm 1 Federated Asymmetric Translation

Require: Real data x, text report t, frozen forward translation F

1: $F(x, t)$ ▷ Generate synthetic image using forward translation
2: $B(F(x, t), t)$ ▷ Reconstruct using backward translation
3: $D(B(F(x, t), t))$ ▷ Distinguish between real and reconstruction image
4: $\mathcal{L}_{adv} \leftarrow \|1 - D(B(F(x, t), t))\|_2$
5: $\mathcal{L}_{id} \leftarrow \|B(F(x, t), t) - x\|_1$
6: $\mathcal{L}_{total} \leftarrow \mathcal{L}_{adv} + \eta\mathcal{L}_{id}$
7: Optimize B and D to minimize and maximize $\mathcal{L}_{total}$
8: **return** weight of B, discriminator score $s^{r,i}$ and gradient $g^{r,i}$ to server

Algorithm 2 Discriminator Quality Score-based Aggregation

Require: MRI clients N_m, CT clients N_c, discriminator gradient $g_m^{r,i}$ and score $s_m^{r,i}$

$G_m^r \leftarrow G_m^{r-1} + \sum_{i=1}^{N_m} (g_m^{r,i})^2$ ▷ Update cumulative gradient for MRI clients

$G_c^r \leftarrow G_c^{r-1} + \sum_{j=1}^{N_c} (g_c^{r,j})^2$ ▷ Update cumulative gradient for CT clients

for each MRI client $i \in N_m$ **do**

$\theta_{c2m}^{r+1} = \sum_{i=1}^{N_m} \frac{(s_m^{r,i} + g_m^{r,i}) \cdot \theta_{c2m}^{r,i}}{\sum_{k=1}^{N_m} s_m^{r,k} + \sqrt{G_m^r}}$ ▷ Perform weight-based aggregation for MRI clients

end for

for each CT client $j \in N_c$ **do**

$\theta_{m2c}^{r+1} = \sum_{i=1}^{N_c} \frac{(s_c^{r,j} + g_c^{r,j}) \cdot \theta_{m2c}^{r,j}}{\sum_{k=1}^{N_c} s_c^{r,k} + \sqrt{G_c^r}}$ ▷ Perform weight-based aggregation for CT clients

end for

return Aggregated weights θ_{c2m}^{r+1} and θ_{m2c}^{r+1}

Algorithm 3 Federated VQA Finetuning

Require: For each modality $k \in \{m, c, t\}$: input v_k, encoder weights W_k, and LoRA weights α_k, β_k.

$\hat{v}_k = W_k v_k + \beta_k \alpha_k v_k$ ▷ Refines the modality-specific representation.

Finetune only LoRA parameters α_k, β_k.

return Modality-specific LoRA weights $\{\alpha_m, \beta_m\}^{r,i}$, $\{\alpha_c, \beta_c\}^{r,i}$ and $\{\alpha_t, \beta_t\}^{r,i}$

Algorithm 4 Modality-specific Aggregation

Require: i-th clients N_m^r (real), N_m^s (synthetic), CT clients N_c^r, N_c^s, LoRA weights $\{\alpha_m, \beta_m\}^{r,i}$, $\{\alpha_c, \beta_c\}^{r,i}$, $\{\alpha_t, \beta_t\}^{r,i}$ and normalization ratio ϵ

for each client i **do**

if i is real **then**

$\lambda^i \leftarrow \frac{\epsilon}{\epsilon \cdot N^r + N^s}$ ▷ Compute real-client weight

else

$\lambda^i \leftarrow \frac{1}{\epsilon \cdot N^r + N^s}$ ▷ Compute synthetic-client weight

end if

end for

$\{\alpha_k, \beta_k\}^{r+1} \leftarrow \sum_{i=1}^{N} \lambda^i \cdot \{\alpha_k, \beta_k\}^{r,i}, \ k \in \{m, c, t\}$ ▷ Aggregate modality-specific LoRA weights

return Aggregated weights $\{\alpha_k, \beta_k\}^{r+1}$

B Discussion

B.1 Discriminator Score-based Aggregation

Discriminator-based aggregation may raise concerns about stability, especially in cross-modal scenario. However, the proposed framework addresses this issue through a carefully designed weighting mechanism. Specifically, our method does not directly rely on the discriminator's confusion between real and synthetic data. Instead, aggregation weights are determined based on the discriminator's confidence and accuracy exclusively on real images, reflecting its reliability in recognizing genuine data rather than its susceptibility to well-generated synthetic examples.

Moreover, the proposed framework does not rely solely on discriminator scores for aggregation weight determination. Instead, it incorporates additional signals, including the gradient of the discriminator loss and the cumulative gradient sum (e.g., G^r), to ensure more stable and reliable weighting. These complementary factors help mitigate potential biases caused by temporary discriminator confusion and contribute to more robust aggregation decisions.

B.2 Experiments on Cross-modality

In this study, we validate the superiority of our approach by effectively addressing cross-modal heterogeneity through a combination of DA and FL strategies. While combination of DA and FL strategies primarily rely on aggregating modality-specific features into a shared representation, they often fail to bridge the substantial semantic and visual gaps inherent in medical imaging modalities, such as MRI and CT.

To ensure a fair and comprehensive comparison, we selected state-of-the-art FL baselines that explicitly incorporate domain adaptation mechanisms (e.g., FedTGP with SEA and IOS with CAF). These methods represent the approaches for mitigating domain shifts. However, even with these enhancements, they struggle to fully capture modality-specific semantic cues and achieve effective cross-modal representation learning. In contrast, our proposed X-FLoRA framework mitigates these challenges by employing a federated asymmetric translation and federated VQA finetuning. This design allows each modality to retain its unique characteristics while still enabling effective cross-modal representation learning.

Experimental results demonstrate that our ap-

Dataset	LLaVA-Med		VQA-RAD	
Metric	PPV	Sensitivity	PPV	Sensitivity
FedAvg	34.67	23.30	63.95	64.35
FedProx	34.50	23.20	68.30	66.82
MOON	34.92	23.85	68.77	67.98
FedProto	34.86	23.93	68.86	68.52
IOS	35.08	23.52	66.67	67.83
FedTGP	35.61	24.36	67.59	68.77
X-FLoRA	**37.04**	**25.67**	**69.67**	**71.24**

Table 6: Comparison with prior federated learning methods in terms of ppv and sensitivity on LLaVA-Med and VQA-RAD datasets.

Dataset	LLaVA-Med				
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
FedAvg	0.2857	0.1446	0.3408	0.3641	0.4968
FedProx	0.2804	0.1478	0.3414	0.3652	0.5003
MOON	0.2908	0.1512	0.3455	0.3576	0.5108
FedProto	0.2915	0.1530	0.3447	0.3557	0.5117
IOS	0.2884	0.1497	0.3486	0.3602	0.5178
FedTGP	0.2990	0.1546	0.3515	0.3629	0.5201
X-FLoRA	**0.3158**	**0.1614**	**0.3667**	**0.3899**	**0.5403**

Table 7: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 6 CT clients and 2 MRI clients.**

proach consistently outperforms combination of DA and FL strategies, particularly in handling complex modality-specific reasoning tasks. This underscores the effectiveness of explicitly modeling cross-modal heterogeneity through structured translation and fine-tuning mechanisms, rather than relying solely on shared representations.

B.3 RAG with X-FLoRA

Integrating RAG into our FL framework poses significant challenges. In FL setting, clients are constrained from sharing raw data due to privacy regulations. Moreover, RAG requires access to a large, centralized, and searchable corpus at inference time. Unfortunately, this assumption conflicts with the privacy-preserving nature of FL, particularly in medical domains. Hence, RAG can potentially improve QA performance but integrating it into X-FLoRA requires a privacy-preserving retrieval method. It is because client queries may contain sensitive medical information that must not be exposed during external document retrieval.

C Additional Experiments

C.1 Clinical Validation

This work was conducted in collaboration with clinical experts in the Department of Nuclear Medicine and the Department of Cardiology. Specifically, our qualitative evaluations (Figs 5 and 8−18) are annotated lesion areas (marked with red arrows) by clinical experts. To further validate the clinical usefulness, we consulted clinical experts, and incorporated additional recommended evaluation metrics such as sensitivity, which relates to diagnostic accuracy, and positive predictive value (PPV), which reflects the rate of false positives. As shown in Table 6, X-FLoRA outperforms all compared models in both sensitivity and PPV. This indicates that X-FLoRA generates fewer incorrect responses, which is vital in healthcare applications.

Dataset	LLaVA-Med				
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
FedAvg	0.2864	0.1459	0.3421	0.3656	0.4960
FedProx	0.2797	0.1437	0.3405	0.3630	0.4968
MOON	0.2813	0.1467	0.3437	0.3650	0.5011
FedProto	0.2856	0.1497	0.3433	0.3639	0.5027
IOS	0.2901	0.1523	0.3453	0.3671	0.5113
FedTGP	0.2965	0.1523	0.3478	0.3601	0.5188
X-FLoRA	**0.3160**	**0.1611**	**0.3685**	**0.3902**	**0.5398**

Table 8: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 2 CT clients and 6 MRI clients.**

Dataset	LLaVA-Med					VQA-RAD		
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr	Accuracy (%)		
						Open	Closed	Overall
IOS	0.2957	0.1556	0.3514	0.3552	0.5176	55.38	78.02	69.13
FedTGP	0.2997	0.1560	0.3567	0.3644	0.5236	57.52	79.43	69.69
X-FLoRA	**0.3287**	**0.1642**	**0.3731**	**0.3900**	**0.5415**	**59.27**	**81.14**	**72.50**

Table 9: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 4 X-ray clients and 4 CT clients.**

C.2 Ratio of Clients

Tables 7 and 8 compare the performance of X-FLoRA with several existing FL methods under different client settings on the LLaVA-Med dataset. Specifically, Table 7 evaluates the case with 6 CT clients and 2 MRI clients, while Table 8 examines the scenario with 2 CT clients and 6 MRI clients.

Across both settings, X-FLoRA consistently outperforms all existing methods in terms of BLEU, METEOR, ROUGE, and CIDEr metrics. These results highlight the robustness and effectiveness of X-FLoRA, even under varying distributions of modality-specific clients. The superior performance demonstrates that X-FLoRA effectively handles cross-modal heterogeneity and maintains high-quality VQA generation, regardless of the client composition.

Dataset	LLaVA-Med					VQA-RAD		
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr	Accuracy (%)		
						Open	Closed	Overall
LLaVA	0.2937	0.1519	0.3508	0.3558	0.5167	55.33	78.06	69.30
X-FLoRA	**0.3191**	**0.1630**	**0.3704**	**0.3954**	**0.5430**	**60.42**	**81.10**	**72.89**

Table 10: Comparison with LLaVA in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset.

Federated Asymmetric Translation			Federated VQA Finetuning		Overall Accuracy (%)
Text	Translation	Discriminator-based Aggregation	Modality-expert LoRA	Modality-specific Aggregation	
✓	✓	✓	✓	✓	**72.89**
✓	✓		✓	✓	71.08
✓	✓	✓	✓		71.12
✓	✓	✓			69.83
	✓	✓	✓	✓	70.89
					64.76

Table 11: Ablation study for X-FLoRA on the VQA-RAD dataset in terms of the accuracy.

η	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
0.3	0.3095	0.1578	0.3473	0.3846	0.5349
0.4	0.3158	0.1604	0.3516	0.3911	0.5410
0.5	**0.3191**	**0.1630**	**0.3604**	**0.3954**	**0.5430**
0.6	0.3170	0.1610	0.3542	0.3911	0.5413

Table 12: Effect of the adjusting hyperparameter (η) in terms of BLEU, METEOR, ROUGE and CIDEr in federated learning of shared asymmetric translation on the LLaVA-Med dataset.

C.3 Additional Modality

Our proposed architecture is inherently extensible, as it does not assume fixed modality pairs and supports potential extensions, as mentioned by Limitation section. To present empirical evidence for potential extensions, we conducted the experiment with X-ray (additional modality) and CT clients. As presented in Table 9, X-FLoRA outperforms recent compared models, demonstrating generalization across more diverse settings. In particular, the superior results on both CT and newly introduced X-ray clients provide strong empirical evidence that our framework is not confined to specific modality pairs, but can be effectively extended to additional modalities. This highlights that X-FLoRA consistently maintains performance advantages across heterogeneous modalities, thereby reinforcing its potential as a general federated learning solution for real-world multi-modal medical environments.

C.4 Comparison with LLaVA

As shown in Table 10, X-FLoRA outperforms the LLaVA (Liu et al., 2023). This indicates that our framework enhances performance without degrading the frozen LLM's capabilities, validating the effectiveness of our design. The improvement primarily stems from the modality-expert LoRA finetuning, which injects modality-specific knowledge into the encoders while preserving the general reasoning ability of the backbone LLM. By selectively adapting key and value projections in the attention layers and linear transformations in the feedforward layers, our LoRA modules achieve fine-grained alignment with medical imaging modalities at minimal computational cost. This confirms that lightweight, targeted adaptation not only avoids catastrophic forgetting but also leads to consistent gains across all evaluation metrics.

C.5 Ablation Study

Table 11 presents an additional ablation study of the individual contributions of each module in the X-FLoRA framework. The results demonstrate that each module significantly enhances the overall performance of X-FLoRA. The combination of these modules operates synergistically to maximize VQA performance, effectively addressing challenges posed by cross-modal FL heterogeneity.

C.6 Weight of Total Loss

Table 12 presents the impact of the hyperparameter η on the performance of federated learning with shared asymmetric translation, evaluated using BLEU, METEOR, ROUGE, and CIDEr metrics on the LLaVA-Med dataset. The results indicate that setting η to 0.5 yields the best overall performance across all metrics, suggesting that this value provides an effective balance between adversarial and identity losses in training.

C.7 Modality-expert LoRA

Table 13 presents an ablation study analyzing the contribution of rank (r_m, r_c, and r_t) by varying its rank, where only one modality-expert LoRA is fine-tuned. The evaluation was conducted on the LLaVA-Med dataset using BLEU, METEOR, ROUGE, and CIDEr metrics. Moreover, Table 13 shows that setting all modality-specific LoRA ranks (r_m, r_c, and r_t) to 16, 32 and 8 yields the best overall performance across BLEU, METEOR, ROUGE, and CIDEr metrics on the LLaVA-Med dataset. This result suggests that a balanced representation capacity across MRI, CT, and text modalities is most effective for the VQA task.

Table 14 summarizes the results of an ablation study evaluating the impact of different combinations of ranks (r_m, r_c, and r_t) assigned to modality-expert LoRA modules. While the configuration of (16, 32, 8) had previously shown promising

Rank r_m	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
8	0.3121	0.1582	0.3601	0.3876	0.5367
16	**0.3150**	**0.1598**	**0.3645**	**0.3907**	**0.5406**
32	0.3122	0.1577	0.3605	0.3869	0.5371
Rank r_c	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
8	0.3133	0.1569	0.3605	0.3858	0.5376
16	0.3128	0.1575	0.3611	0.3861	0.5364
32	**0.3149**	**0.1580**	**0.3651**	**0.3911**	**0.5402**
Rank r_t	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
8	**0.3138**	**0.1555**	**0.3637**	**0.3882**	**0.5390**
16	0.3125	0.1534	0.3610	0.3868	0.5351
32	0.3120	0.1533	0.3600	0.3851	0.5355

Table 13: Ablation study on the contribution of each rank (r_m, r_c, and r_t) in terms of BLEU, METEOR, ROUGE, and CIDEr metrics on the LLaVA-Med dataset.

Rank r_m,r_c,r_t	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
16,32,8	**0.3191**	**0.1630**	**0.3704**	**0.3954**	**0.5430**
32,32,8	0.3175	0.1608	0.3685	0.3925	0.5417
16,16,8	0.3177	0.1611	0.3672	0.3923	0.5401
16,32,16	0.3156	0.1596	0.3655	0.3896	0.5399

Table 14: Effect of the combination of rank (r_m, r_c, and r_t) in terms of BLEU, METEOR, ROUGE, and CIDEr metrics on the LLaVA-Med dataset.

results, we further validated its effectiveness by experimenting with alternative rank combinations. As shown in Table 14, the (16, 32, 8) setting consistently outperforms other configurations across all evaluation metrics, including BLEU, METEOR, ROUGE, and CIDEr. This confirms that assigning moderate capacity to the MRI and CT experts and a smaller capacity to the text expert leads to the most balanced and effective performance.

Moreover, Tables 15 and 16 explore the impact of the aggregation weight hyperparameter ϵ, which controls the balance between real and synthetic data contributions during modality-specific aggregation. As ϵ increases, real client data receives higher weight. The best performance is achieved at $\epsilon = 1.5$, while performance degrades when $\epsilon = 1$ (equal weighting) or $\epsilon = 0.5$ (favoring synthetic data). This highlights the importance of prioritizing real data for robust VQA model training.

C.8 Effect of Text

Figure 6 demonstrates the effectiveness of the textual cues associated with images in the LLaVA-Med dataset. As shown, CT→MRI and MRI→CT translations performed without textual cues significantly degrade visual quality, introducing severe noise and distorting anatomical regions. Compared with translations without textual cues, the proposed text-driven translations leverage image-associated textual information to preserve clinical insights. Specifically, the second and third rows of the CT→MRI results show that translations without textual cues introduce severe noise. Furthermore, in the MRI→CT results, the second row highlights Posterior Reversible Encephalopathy Syndrome—emphasizing this finding during the CT conversion process. Notably, these results demonstrate that text-driven translation effectively preserves and emphasizes clinically relevant regions.

ϵ	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
0.5	0.2959	0.1514	0.3547	0.3778	0.5201
1	0.3091	0.1598	0.3602	0.3845	0.5289
1.5	**0.3291**	**0.1730**	**0.3704**	**0.4054**	**0.5530**
2	0.3105	0.1684	0.3589	0.3823	0.5317
2.5	0.3052	0.1647	0.3604	0.3760	0.5208

Table 15: Effect of the normalization ratio (ϵ) on BLEU, METEOR, ROUGE, and CIDEr scores in expert-aware weighting on the LLaVA-Med dataset.

ϵ	Accuracy (%)		
	Open	Closed	Overall
0.5	56.48	76.98	68.84
1	58.45	77.95	70.21
1.5	**60.42**	**81.10**	**72.89**
2	59.10	77.58	70.24
2.5	58.38	76.99	69.60

Table 16: Effect of the normalization ratio (ϵ) on accuracy in expert-aware weighting on the VQA-RAD dataset.

C.9 Visual analysis of federated asymmetric translation.

Figure 7 exhibits the visualized results of the forward and backward processes of federated asymmetric translation for each modality across global training rounds. Initially, both forward and backward translations exhibit significant noise. However, as training progresses, the proposed federated asymmetric translation—which focuses on enhancing the backward translator—progressively improves its ability to capture the features of the input images. These results demonstrate that our training methodology enables efficient model learning even in cross-modal FL scenarios where each client holds data from only a single modality.

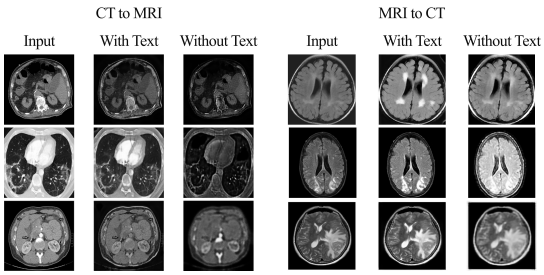

Figure 6: Effect of textual cues on clinical feature augmentation in the forward translator of asymmetric translation on the LLaVA-Med dataset.

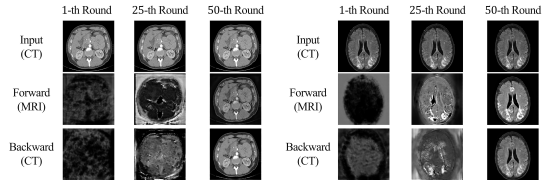

Figure 7: Visual results of federated asymmetric translation across global rounds. In the first round, the forward translator produces severe noise. By the 25th round, noise is substantially reduced and the overall structure begins to take shape. By the 50th round, the output closely resembles the input form with noise almost entirely eliminated.

D VQA Results

Figures 8−16 present qualitative examples of VQA results using the LLaVA-Med dataset. Specifically, figures 8−12 illustrate cases based on CT data, while figures 13−16 focus on MRI-based VQA scenarios. Moreover, Figures 17 and 18 present qualitative CT and MRI examples of VQA results using the VQA-RAD dataset, respectively. In the figure, the red arrow highlights a lesion or anatomical structure described in the imaging report, red text indicates incorrect or inconsistent responses, while BLEU text represents accurate and contextually appropriate answers.

Figure 12 presents a failure case analysis of our model. Although the imaging report indicates multiple abnormalities in the lower lobes—including ground glass opacities, arcade-like bands of parenchymal consolidation, peribronchial consolidation, and mild bronchiolectasis—our model successfully identified one of the true abnormalities but additionally predicted unrelated findings such as multiple cavitary lesions. However, it is important to note that other baseline models performed even more poorly. This suggests that despite the imperfect prediction, our model demonstrates a comparatively stronger ability to recognize at least some clinically relevant abnormalities.

In each example without Fig. 12, various models are evaluated by their ability to correctly identify the main imaging findings when presented with corresponding medical images and diagnostic queries. The figures demonstrate that X-FLoRA consistently provides more accurate and clinically relevant responses. This highlights the importance of diverse modality data and modality-specific expert representation for achieving reliable VQA performance in medical imaging contexts.

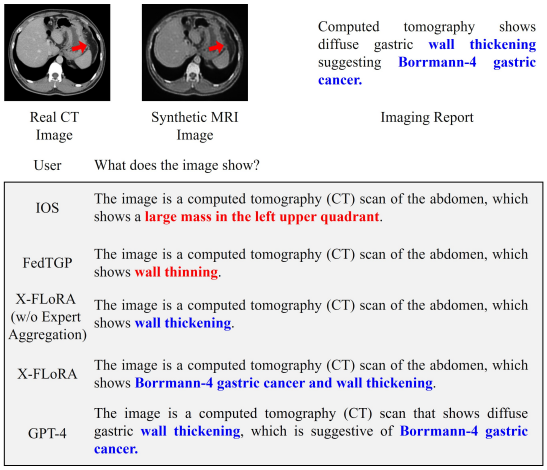

Figure 8: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

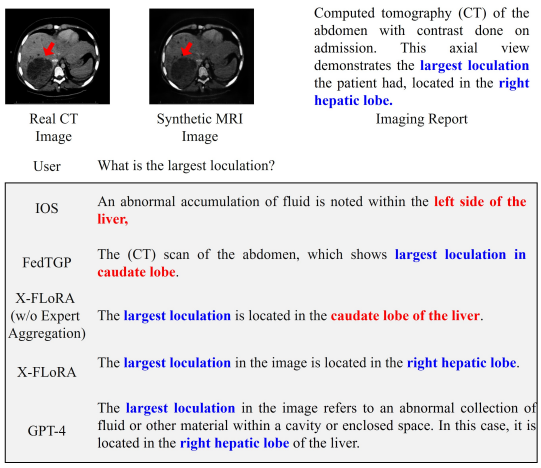

Figure 9: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

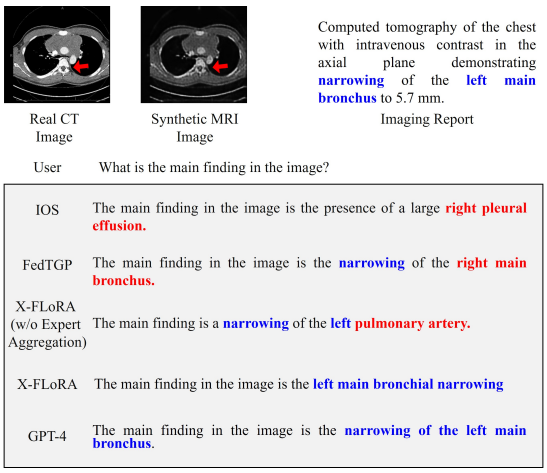

Figure 10: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

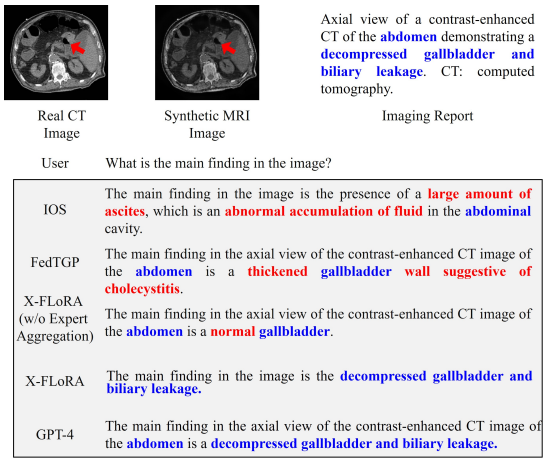

Figure 11: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

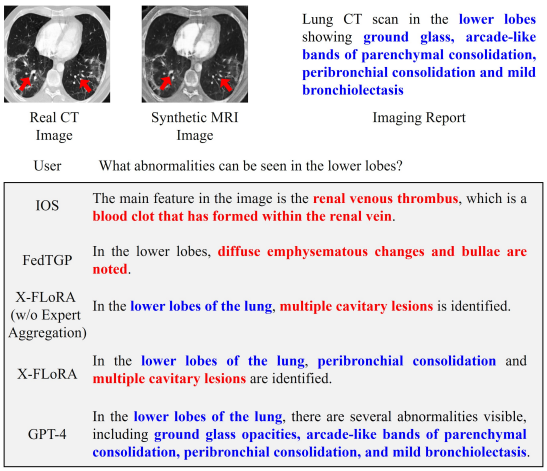

Figure 12: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

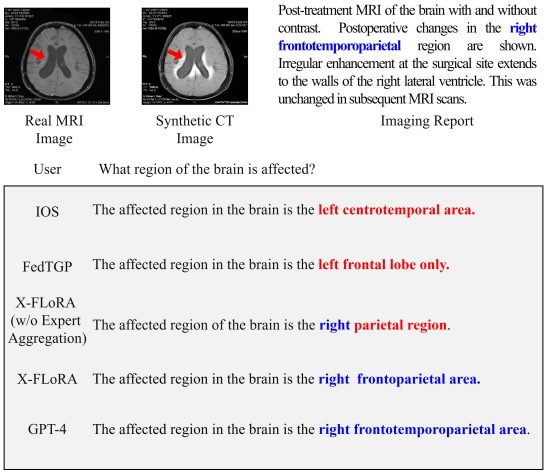

Figure 13: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

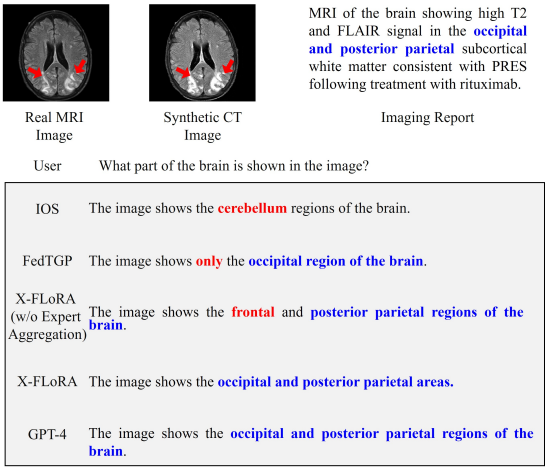

Figure 14: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

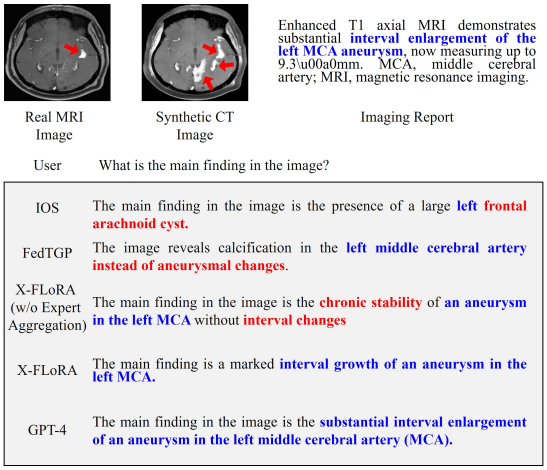

Figure 15: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

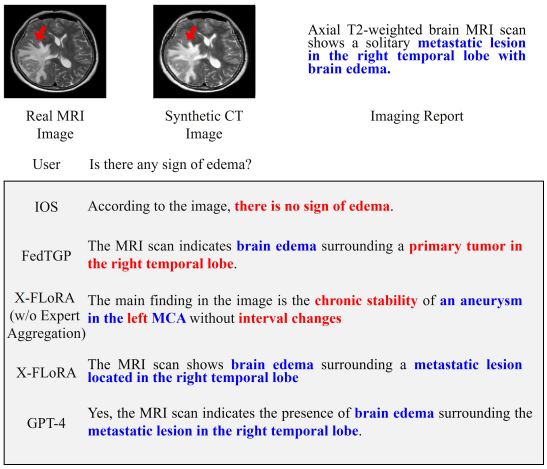

Figure 16: Example comparison of X-FLoRA and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

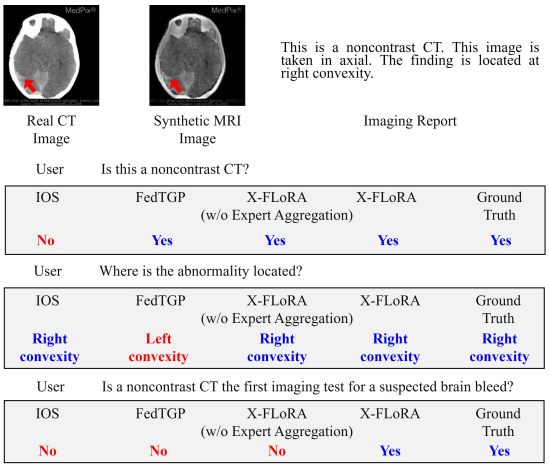

Figure 17: Example comparison of X-FLoRA and other FL methods on VQA-RAD.

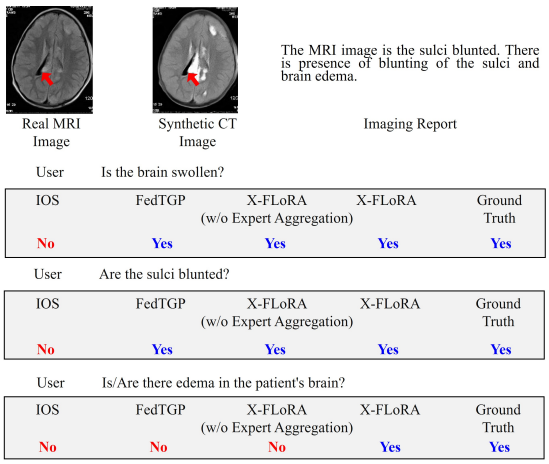

Figure 18: Example comparison of X-FLoRA and other FL methods on VQA-RAD.