SynTeX-FL: Cross-Modal Text Transfer in Federated Learning for Medical Visual Question Answering

Min Hyuk Kim[*], **Do Hyun Ki**[*], **Seok Bong Yoo**[†]
Department of Artificial Intelligence Convergence
Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

Abstract

Medical visual question answering (VQA) and federated learning (FL) have gained prominence as critical methods for clinical artificial intelligence, with VQA supporting image-based diagnostic reasoning and FL allowing institutions to train models jointly without sharing sensitive patient data. However, existing approaches face substantial challenges in cross-modal vertical FL settings, where each participating institution possesses unimodal, modality-specific textual data without access to paired multimodal data. To address this limitation, we introduce SynTeX-FL, a federated framework designed to enable cross-modal text transfer for medical VQA without sharing raw data. Moreover, SynTeX-FL applies a reconstruction-based text synthesis model that extracts clinically relevant semantics from medical images to generate cross-modal textual representations. Furthermore, the proposed method incorporates modality-specialized low-rank adaptation (LoRA) modules to enhance modality-aware representations for each imaging domain. During aggregation, the central server employs discriminator-based quality scores and source-aware weighting strategies to integrate heterogeneous client contributions. Extensive experiments on multimodal medical VQA benchmarks demonstrate that SynTeX-FL significantly outperforms current FL baselines, providing a robust and efficient solution for cross-modal reasoning in decentralized clinical environments.

1 Introduction

Medical visual question answering (VQA) [44, 3, 64] has gained attention as a practical approach in computer-aided diagnosis, assisting clinicians by providing answers to diagnostic queries grounded in medical images. However, the widespread adoption of VQA methods is often limited by concerns regarding data privacy. Federated learning (FL) enables collaborative model training across clinical institutions while preserving patient data privacy [14]. Federated VQA frameworks [28, 67, 51] have been introduced to address this problem, enabling medical VQA without sharing patient data. Despite recent advances, federated VQA encounters limitations in cross-modal vertical FL scenarios [9, 62], where each client has data from only one modality and lacks access to same-patient paired samples across imaging modalities. In typical cross-modal vertical FL scenarios across hospitals, individual clients typically possess either X-ray, computed tomography (CT), or magnetic resonance imaging (MRI) data, but not more than one of these. This situation commonly arises in real clinical environments where hospitals maintain different imaging equipment and clinical specializations. These imaging modalities differ substantially in both acquisition mechanisms and diagnostic roles. For example, X-rays employ low-dose radiation primarily for chest imaging, whereas CT employs ionizing radiation, and MRI uses magnetic fields and radio waves. As illustrated in

[*]These authors contributed equally.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

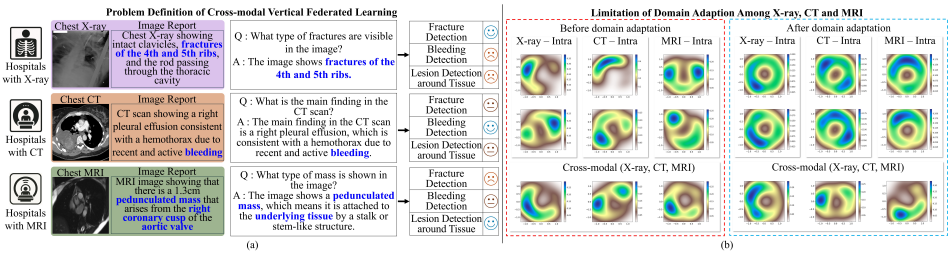

Figure 1: (a) Illustration revealing the need for cross-modal vertical federated learning (FL) across hospitals, highlighting the complementary strengths of each modality in visual question answering (VQA) tasks. (b) Visualization of intra- and cross-modal domain adaptation effects using a contrastive learning-based method [57], emphasizing the challenges of adapting X-ray, CT, MRI modalities.

Figure 1(a), electronic medical record reports from X-ray, CT, and MRI for the same chest region often emphasize different pathological characteristics. For example, an X-ray report might describe a fracture of the ribs, whereas a CT report for the same region may indicate bleeding, and an MRI report may highlight a pedunculated mass. These differences reflect the respective strengths of each modality in detecting fractures, hemorrhages, and soft tissue abnormalities. Likewise, in the GPT-4-based medical VQA dataset [32], the mentioned responses for each modality highlight the necessity of modality-specialized reasoning.

In response to these differences, researchers have explored domain adaptation techniques. As a preliminary experiment, Figure 1(b) illustrates the feature distributions before and after applying SEA [57], a contrastive learning-based domain adaptation method. To visualize the feature distributions of electronic medical records, we adopt the approach proposed by [6], using ResNet50 as the feature extractor. After domain adaptation, intra-modal feature distributions become aligned, while cross-modal distributions remain largely separated. It highlights the insufficiency of domain adaptation in addressing cross-modal heterogeneity [5] and its potential to degrade performance [63].

To overcome these challenges, we propose SynTeX-FL, a cross-modal vertical FL framework that uses cross-modal text transfer as a training-time semantic augmentation for medical VQA. The proposed SynTeX-FL framework comprises two stages: cross-modal text generation and modality-specialized fine-tuning for VQA. In the first stage, each client independently performs cross-modal text transfer training using its data. The reconstruction transfer module is updated during training, whereas the synthesis transfer module remains fixed. By integrating electronic medical records and their corresponding images, these transfer modules extract clinically important visual cues that are not clearly stated in the textual data. Clients can upload their trained reconstruction transfer weights to the central server for aggregation. In the second stage, modality-specialized low-rank adaptation (LoRA) modules independently fine-tune representations for each modality, including X-ray, CT, and MRI. Due to the unique properties of each modality, these modules improve the quality of modality-specific representations. The server aggregates the fine-tuned LoRA modules using a modality-specific strategy, balancing the influence of real and synthetic data from different clients. We summarize the primary contributions as follows:

- We propose a cross-modal vertical FL framework that enables unpaired modality-to-modality semantic transfer through image-guided text synthesis and modality-specialized tuning, without requiring paired cross-modal data, image-space translation, or feature-level alignment. This framework improves the robustness of modality-level diversity and representation.

- We propose an FL framework for cross-modal text transfer, in which each client trains only the reconstruction-based text transfer module to enrich textual features with clinically relevant information extracted from the images. In addition, a quality-aware aggregation strategy enhances the influence of clients with higher-quality transfer models.

- We introduce modality-specialized LoRA, a lightweight adaptation mechanism tailored to each modality. Distinct LoRA modules are assigned per modality, and a source-aware weighting scheme facilitates a balanced aggregation of real and synthetic data across clients.

2 Related Work

2.1 Visual Question Answering

VQA [3, 64, 31, 33, 32, 37, 54, 61, 40, 22] combines natural language processing with computer vision, a field encompassing robust recognition and multimodal fusion [12, 20, 23, 30], to answer image-grounded questions. Based on advancements in general-domain VQA, considerable attention has been focused on extending its utility to the medical field. Recent research has indicated that large language models and vision language models achieve notable effectiveness on clinical tasks when optimized using medical datasets. Specifically, BioMistral [26], a variant of Mistral-7B [17], performs well on challenging clinical question answering tasks, including medical licensing examinations and queries based on PubMed [19]. Likewise, medical-specific VLMs such as LLaVA-Med [32], adapted from the LLaVA [37], and MedFlamingo [42], which build on OpenFlamingo [1], produce impressive results on VQA tasks in radiology [29] and pathology [11]. Hallucination-mitigation method for medical LVLMs [40], including VCD- and DoLa-based decoding strategies, improves reliability through sparse visual-contrastive decoding. In contrast to prior VQA approaches, X-FLoRA [22] focuses on image-level translation, enabling asymmetric translation between MRI and CT images. While effective for visual modality alignment, such image translation–based approaches inherently confine cross-modal interaction to the visual representation space and remain restricted to a limited set of modalities, namely MRI and CT, where visual correspondence can be reasonably assumed. To overcome these limitations, we introduce SynTeX-FL, a vertical federated framework that performs cross-modal transfer at the textual level, enabling unified semantic integration of heterogeneous imaging modalities while avoiding unreliable image-space synthesis.

2.2 Cross-modal Vertical Federated Learning

Ensuring patient privacy has emerged as a critical concern in contemporary digital healthcare due to threats related to unauthorized disclosure or the misuse of sensitive medical information [7, 48]. Vertical multimodal FL methods have been proposed to mitigate these challenges [66, 46, 62]. For instance, [66] developed UTMP, a federated approach where clients with unimodal data cooperatively train a multimodal network by hierarchically aggregating encoder-decoder architectures. Additionally, [46] introduced an FL method aimed at multimodal coronavirus disease 2019 diagnostics, allowing clients with distinct imaging modalities, such as X-rays or ultrasounds, to train the models collaboratively without sharing raw data. Furthermore, [62] proposed a cross-modal vertical FL framework, achieving a robust MRI reconstruction performance without the domain shift problem. However, existing approaches do not consider critical properties of medical imaging, including the distinction of imaging mechanisms (e.g., X-ray, CT, and MRI) and the clinical semantic emphasis of the reports. To overcome this limitation, we propose a novel FL method integrating cross-synthesis and modality-specialized tuning stages, using a synthesis and reconstruction model and modality-specific encoder to capture the features of each modality.

2.3 Text Style Transfer

Text style transfer (TST) aims to convert the style of an input sentence into a target style while preserving the style-independent content. Several studies [16, 4, 56] have trained the sequence-to-sequence model on parallel corpora for TST. Specifically, [16] included a pointer network with the sequence-to-sequence model to transfer modern English into Shakespearean English. However, these studies rely on parallel datasets, which are limited in real-world settings due to the scarcity of parallel corpora. Methods for TST using non-parallel datasets have been proposed to address this limitation [27, 60, 15, 25]. Particularly, adversarial learning [15, 25] has demonstrated robust performance. For example, [25] applied self-supervision to the input sequence based on cycle-consistency, making the model robust to mixed-style inputs. Nevertheless, these prior studies have focused on certain tasks, such as sentiment transfer and formality transfer, and have not addressed TST for electronic medical records across modalities. In addition, they do not consider the constraints of vertical FL, where direct access to different modality data across institutions is restricted. To overcome this problem, we introduce a federated method for non-parallel cross-modal TST, using modality-specific visual input as semantic support. This approach enhances TST by incorporating clinical information that is often missing in text alone.

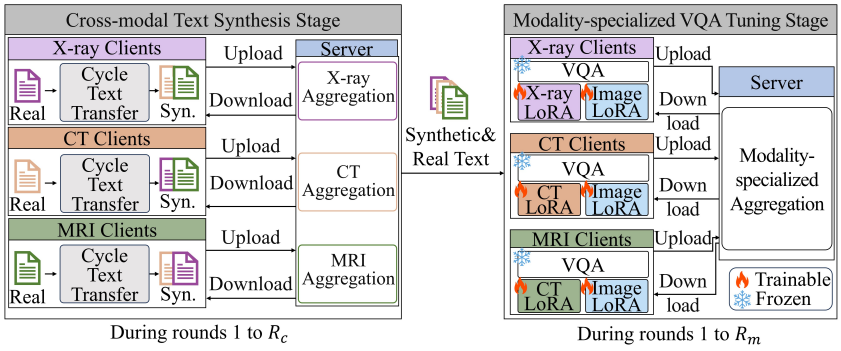

Figure 2: Architecture of the SynTeX-FL framework.

3 Method

3.1 Overview

The SynTeX-FL method comprises two stages, cross-modal text synthesis and modality-specialized VQA tuning, (Figure 2). During the cross-modal text synthesis stage, N_x X-ray clients, N_c CT clients, and N_m MRI clients participate using modality-specific data. Clients train reconstruction transfer models tailored to their respective modalities, whereas synthesis models from other modalities remain fixed. These locally trained models are uploaded to the central server, which performs aggregation across clients. The aggregated reconstruction transfer weights are redistributed, enabling all clients to continue the synthesis process without sharing raw data. After R_c rounds of this stage, each client generates synthetic text corresponding to the other modalities. The next stage applies modality-specialized LoRA modules to the corresponding text encoders. The LoRA weights are uploaded to the server for modality-specialized aggregation. A source-aware weighting scheme is employed during aggregation to balance the contributions from real and synthetic data. This tuning stage is conducted over R_m rounds, enhancing the modality diversity and robustness of modality-aware representations. The final global VQA model is obtained upon completion. *Appendix A* provides pseudo-code algorithms of the cross-modal text synthesis, quality-aware aggregation, modality-specialized VQA tuning, and modality-specialized aggregation.

3.2 Cross-modal Text Synthesis Stage

In vertical FL environments, each institution holds only unimodal data (see Figure 1(a)), limiting access to complementary cross-modal information. Although prior approaches, such as domain adaptation (Figure 1(b)), attempt feature alignment, they require access to multimodal datasets. Unlike feature-level alignment, text-level transfer operates in a shared clinical semantic space, allowing complementary findings across X-ray, CT, and MRI to be integrated without suppressing modality-specific diagnostic cues. To address these limitations, we propose a cross-modal text synthesis stage that enriches electronic medical records and enables explicit semantic alignment without requiring same-patient paired samples across medical imaging modalities. This design allows each client to benefit from modality diversity while preserving data privacy.

Inspired by adversarial learning approaches [15, 25], we introduce a cross-modal text synthesis stage that makes it possible for each client to learn cross-modal text transfer even with unimodal text data (Figure 3). Each client in this stage possesses real text t and the corresponding image I, and uses a synthesis transfer S to produce synthetic text based on these inputs.

Cycle Text Transfer. The proposed cycle text transfer module applies an attention layer [52] to learn the textual representations. This layer comprises a multihead self-attention mechanism followed by a positionwise feedforward network. In parallel, this approach implements the image embedding module as a convolutional layer with a kernel size of three. Average pooling, followed by a linear projection, transforms the output of this module into image queries. The cross-attention mechanism uses these image queries with the encoded text features. The resulting features are processed through six stacked Transformer blocks, each employing multi-head attention with 12

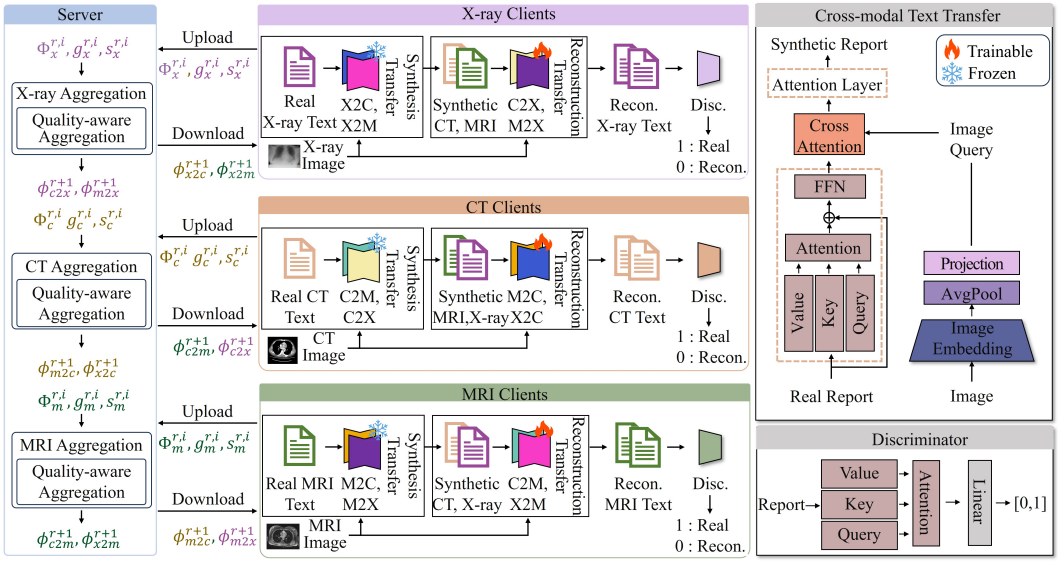

Figure 3: Cross-modal text synthesis stage architecture, where $\Phi_a^{r,i}$ denotes the set of reconstruction transfer weights $\phi_{b2a}^{r,i}$ trained by the ith client in round r for modality pairs (a, b).

attention heads, and projected to vocabulary logits for decoding into the target-modality report. Additionally, the discriminator comprises an attention layer followed by a linear projection. The input sequence is encoded and passed through a linear layer to produce a sequence-level discriminator score.

Each client employs these architectures to perform cycle text transfer during local training. Each client generates synthetic text using the frozen synthesis transfer module S, and applies a reconstruction transfer module R, sharing the same architecture as S, to recover the original text. Each client has access to only a single modality; thus, S from the other modality client cannot be trained locally and remains fixed during training. The reconstructed text is employed to train R and a discriminator D, which learns to distinguish between real and reconstructed samples. Specifically, D and R are optimized as follows:

$$\hat{D}, \hat{R} = \arg\min_{R} \max_{D} \mathcal{L}_{cross}, \tag{1}$$

where, $\mathcal{L}_{cross}$ denotes the total loss function in the cross-modal text synthesis stage, defined as follows:

$$\mathcal{L}_{cross} = \mathcal{L}_{disc} + \lambda \mathcal{L}_{rec}, \tag{2}$$

where λ controls the trade-off between two objectives: 1) the discriminator loss ($\mathcal{L}_{disc}$), which encourages the reconstructed text to appear realistic, and 2) the reconstruction loss ($\mathcal{L}_{rec}$), promoting consistency with the original input. The discriminator loss is defined as follows:

$$\mathcal{L}_{disc} = ||1 - D(R(S(t, I), I))||_2, \tag{3}$$

where $\|\cdot\|_2$ denotes the L2 norm. Additionally, the reconstruction loss reduces the sequence-level discrepancy between the reconstructed text and the original text, defined as follows:

$$\mathcal{L}_{rec} = ||R(S(t, I), I) - t||_1, \tag{4}$$

where $\|\cdot\|_1$ denotes the L1 norm. After local training, the central server aggregates the reconstruction transfer weights uploaded by each client.

Quality-aware Aggregation. We introduce quality-aware aggregation to enhance reliability and stability across clients. Each client with modality a uploads the reconstruction transfer weights $\phi_{b2a}^{r,i}$ for every other modality b, the corresponding gradient $g_a^{r,i}$, and the discriminator score $s_a^{r,i}$ to the server. The discriminator score reflects the realism and clinical plausibility of the generated text, while

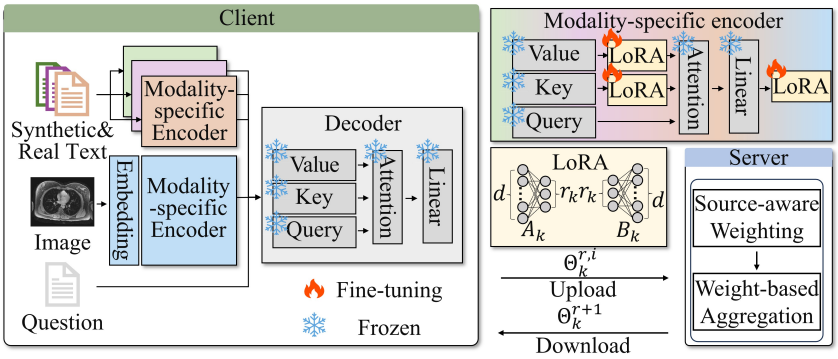

Figure 4: Architecture of the modality-specialized VQA tuning stage.

the gradient-based term captures the stability of local training. By jointly considering these scores, the aggregation process balances output quality and optimization reliability, preventing unreliable updates from the global model. The server aggregates $\phi_{b2a}^{r,i}$ as follows:

$$\phi_{b2a}^{r+1} = \frac{1}{N_a}\sum_{i=1}^{N_a}(\omega_{b2a,s}^{r,i} + \omega_{b2a,g}^{r,i}), \forall\, a, b \in \{x, c, m\},\ a \neq b, \tag{5}$$

where ϕ_{b2a}^{r+1} denotes the aggregated weights in the $(r+1)$th round and N_a represents the number of clients with modality a. Additionally, $\omega_{b2a,s}^{i}$ and $\omega_{b2a,g}^{i}$ represent the score- and gradient-based normalized model weights, respectively, from the ith client for modality pairs $(a, b) \in m, c, x$ with $a \neq b$, where x, c, and m denote X-ray, CT, and MRI texts. These normalized model weights are formulated as follows:

$$\omega_{b2a,s}^{r,i} = \frac{s_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} s_a^{r,i} + \sqrt{G_a^r}}, \quad \forall\, a, b \in \{x, c, m\},\ a \neq b, \tag{6}$$

$$\omega_{b2a,g}^{r,i} = \frac{g_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} s_a^{r,i} + \sqrt{G_a^r}}, \quad \forall\, a, b \in \{x, c, m\},\ a \neq b. \tag{7}$$

Based on $\omega_{b2a,s}^{i}$ and $\omega_{b2a,g}^{i}$, the quality-aware aggregation prioritizes clients with a higher discriminator reliability and training stability. Moreover, G_a^r represents the accumulated squared sum of the gradients (momentum) in the rth round, formulated as follows:

$$G_a^r = G_a^{r-1} + \sum_{i=1}^{N_a}(g_a^{r,i})^2. \tag{8}$$

By incorporating G_a^r and $s_a^{r,i}$, the aggregation process stresses updates from clients with more reliable discriminators, enhancing the robustness of the model.

3.3 Modality-specialized VQA Tuning Stage

Effective fine-tuning methods are necessary due to the high memory and computational demands of vision language models in VQA tasks. In federated medical environments, each modality of data possesses distinct characteristics and fundamental differences in clinical focus (Figure 1(a)). Building on the concept of LoRA [13], this work introduces modality-specialized VQA tuning that independently learns discriminative features for each imaging modality, enabling the effective capture of modality-aware representations in cross-modal vertical FL.

Modality-specialized LoRA. As illustrated in Figure 4, the real and synthetic text obtained from the cross-modal text synthesis stage are fed into each corresponding modality-specific encoder [32]. The proposed modality-specialized LoRA architecture is applied to the key projection, value projection, and linear layer in this encoder. Moreover, the decoder shares the same architecture as the encoder,

comprising the fixed value, key, and query projections and a linear layer, but LoRA is not applied to the decoder. Each client employs fixed encoders that are specific to each modality, denoted as E_k for modality $k \in \{x, c, m, I\}$. Each encoder extracts feature representations $E_k z_k$, where z_k denotes the input vector for modality k. This approach adopts LoRA-based tuning to enrich these representations while keeping the pretrained encoders frozen, as follows:

$$\hat{z}_k = E_k z_k + B_k A_k z_k, \quad k \in \{x, c, m, I\}, \tag{9}$$

where A_k and B_k represent low-rank matrices in the LoRA architecture. Specifically, A_k reduces the input dimension by projecting features into $\mathbb{R}^{d \times r_k}$. The matrix B_k projects the features back to the original feature space. This decomposition and reconstruction strategy enables efficient fine-tuning of cross-modal textual feature representations. These fine-tuned features are integrated via a projector that facilitates intermodal representation learning. Following VQA tuning, the modality-specialized LoRA weights are sent to the central server for aggregation.

Modality-specialized Aggregation. The server aggregates the fine-tuned LoRA weights received from clients across the X-ray, CT, and MRI modalities. In the proposed framework, the modality-specialized LoRA weights are organized based on the modality on which they were trained, with separate sets maintained for X-ray, CT, MRI, and image data. In contrast to conventional FL methods [35, 41, 34, 21] that aggregate weights from all modalities together, we introduce modality-specialized aggregation, where weights are addressed separately for each modality. This separation allows each modality-specialized LoRA module to learn and represent the distinct characteristics of the data for each modality more effectively. We enhance the aggregation quality by introducing a source-aware weighting scheme that distinguishes between weights derived from real and synthetic sources. This approach allows the dynamic adjustment of client contributions during aggregation. The source-aware weighting is computed as follows:

$$\alpha_k^i = \begin{cases} \dfrac{\gamma}{\gamma \cdot N_k^r + N_k^s} & \text{if } i \in \mathcal{R}_k \\ \dfrac{1}{\gamma \cdot N_k^r + N_k^s} & \text{otherwise} \end{cases}, \quad k \in \{x, c, m\}, \tag{10}$$

where α_k^i represents the aggregation weight assigned to the ith client, and $\mathcal{R}_k$ denotes the set of clients holding real data for modality k. The balance ratio γ determines the relative weight assigned to real versus synthetic data. For image modality of all clients, α_I^i is fixed at 1. Based on α_k^i, the server aggregates each modality-specialized LoRA weight $\Theta_k^{r,i} = \{A_k^{r,i}, B_k^{r,i}\}$ in the rth round. The weight adjusts the contribution of the real and synthetic data to modality-aware representations. The aggregation process based on modality-aware weights is defined as follows:

$$\Theta_k^{r+1} = \frac{1}{N_k} \sum_{i=1}^{N_k} \alpha_k^i \Theta_k^{r,i}, \quad k \in \{x, c, m, I\}, \tag{11}$$

where N_I denotes the total number of clients ($N_x + N_c + N_m$). The final global VQA model is obtained after the modality-specialized VQA tuning stage is completed.

4 Experiments

4.1 Dataset and Evaluation Metric

We evaluated the proposed framework on a unified dataset that integrates data from the LLaVA-Med [32] and VQA-RAD [29] datasets. The LLaVA-Med dataset is constructed to facilitate instruction-tuning in multi-institutional settings. This dataset comprises image-text pairs derived from PubMed, along with an instruction-tuning subset generated by GPT-4, consisting of 10K examples across modalities, including X-ray, CT, and MRI.

The VQA-RAD dataset includes 3515 question-answer pairs authored by medical professionals and 315 corresponding radiological images. Moreover, GPT-4 was employed to generate imaging reports based on these question-answer pairs, covering yes or no questions and short-form open-ended responses. In the FL setup, SynTeX-FL was trained across six clients. Two clients used X-ray data (N_x = 2), and the other four used the CT and MRI data ($N_c = 2$ and $N_m = 2$), on the LLaVA-Med

Table 1: Comparison with prior FL methods for BLEU, METEOR, ROUGE-L and CIDEr on the LLaVA-Med dataset.

Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
FedAvg [41]	0.3682	0.3467	0.5003	0.2892	0.1486
FedProx [35]	0.3702	0.3450	0.5064	0.2859	0.1512
MOON [34]	0.3604	0.3492	0.5152	0.2935	0.1561
FedProto [50]	0.3541	0.3486	0.5176	0.2943	0.1568
IOS [58]	0.3587	0.3508	0.5190	0.2913	0.1510
FedTGP [65]	0.3672	0.3561	0.5237	0.3012	0.1572
FedMedVLP [38]	0.3597	0.3533	0.5196	0.2955	0.1540
FedKIM [55]	0.3701	0.3588	0.5279	0.3015	0.1581
SynTeX-FL	**0.4057**	**0.3727**	**0.5543**	**0.3309**	**0.1740**

Table 2: Comparison of accuracy with FL methods and SynTeX-FL on the VQA-RAD dataset.

Metric	Accuracy (%)		
	Overall	Open	Closed
FedAvg [41]	64.76	51.39	73.56
FedProx [35]	65.38	52.09	74.13
MOON [34]	67.21	53.31	76.37
FedProto [50]	68.19	54.04	77.51
IOS [58]	69.04	55.37	78.05
FedTGP [65]	70.00	57.15	78.46
FedMedVLP [38]	69.53	55.81	78.30
FedKIM [55]	70.14	56.12	78.49
SynTeX-FL	**73.89**	**61.46**	**82.10**

Table 3: VQA performance of FL methods with domain adaptation models on the LLaVA-Med and VQA-RAD datasets.

Method		LLaVA-Med		VQA-RAD		
		Metric				
		METEOR	CIDEr	Accuracy (%)		
Domain Adaptation	FL			Open	Closed	Overall
SEA [57]	FedTGP	0.3569	0.5231	57.31	78.50	70.07
CAF [59]	IOS	0.3514	0.5205	55.42	78.37	69.12
SEA [57]	IOS	0.3510	0.5181	55.50	78.53	70.19
CAF [59]	FedTGP	0.3558	0.5207	57.40	78.55	69.59
SynTeX-FL		**0.3727**	**0.5543**	**61.46**	**82.10**	**73.89**

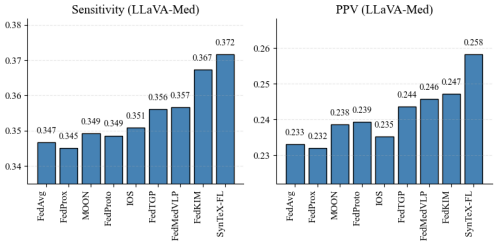

Figure 5: Comparison with prior FL methods for sensitivity and positive predictive value (PPV) on the LLaVA-Med datasets.

and VQA-RAD datasets. To reflect diverse hospital environments, we combined these datasets and distributed the unified data across clients using a Dirichlet distribution (α = 10), which corresponds to a moderately heterogeneous setting commonly adopted in medical FL scenarios [10, 49]. *Appendix B* provides experiments that explore different client configurations, including extreme cases and a nine-client setting.

We evaluated the response quality using ROUGE-L [36], METEOR [2], CIDEr [53], and BLEU [43] for the LLaVA-Med benchmark, and used accuracy as the evaluation criterion for the VQA-RAD dataset. These evaluation metrics assess the lexical and semantic quality of the generated responses. In addition, CIDEr emphasizes the informative content via TF-IDF-weighted similarity, and ROUGE-L measures the n-gram overlap. Further, METEOR balances precision and recall, and BLEU quantifies the lexical precision. Following prior studies [39, 8], we evaluated the style transfer quality by measuring the style accuracy with a pretrained classifier [24] and computed the source-BLEU (sBLEU) [45] to compare the generated texts against their sources. We also evaluated the transferred sentences in terms of perplexity, using the GPT2-based model [47].

4.2 Implementation Details

We adopted a federated-by-dataset setting [41], in which each client independently creates a local dataset and engages in FL with a server. All experiments were performed on NVIDIA L40S GPU. We employed the stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.001. SynTeX-FL was trained for 150 global rounds: R_c = 50 rounds for the text synthesis stage and R_m = 100 rounds for VQA tuning stage. We also set λ and γ to 0.5 and 1.5, respectively, and r_x, r_c, r_m, and r_I to 16, 32, 8, and 16, respectively. *Appendix C* details the experiments to optimize these hyperparameters.

4.3 Results and Analysis

We evaluated SynTeX-FL against multiple baseline FL frameworks, including FedAvg [41], FedProx [35], MOON [34], FedProto [50], IOS [58], FedTGP [65], FedMedVLP [38], and FedKIM [55], on the LLaVA-Med and VQA-RAD datasets. We adopted the VQA model introduced by Liu et al. as the backbone, given its widespread application in the medical domain. Each baseline model was trained and assessed, following the authors' hyperparameter settings and open-source codes.

Table 4: Comparison of FL methods with medical VQA backbones on the VQA-RAD dataset.

VQA Backbone	FL Method	Overall	Open	Closed
LLaVA-Med [32]	FedTGP	70.52	57.21	78.56
	IOS	70.66	57.82	78.93
Med-MoE [18]	FedTGP	70.17	56.95	79.25
	IOS	70.95	57.19	80.17
AMiF [68]	FedTGP	71.51	58.35	80.34
	IOS	72.08	58.87	80.82
LLaVA [37]	SynTeX-FL	**73.89**	**61.46**	**82.10**

Table 5: Comparison with real-only LoRA on LLaVA-Med and VQA-RAD.

LLaVA-Med					
Method	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
Real-only LoRA	0.3690	0.3398	0.5112	0.2884	0.1427
SynTeX-FL	**0.4057**	**0.3727**	**0.5543**	**0.3309**	**0.1740**
VQA-RAD					
Method	Overall		Open		Closed
Real-only LoRA	65.76		52.04		73.26
SynTeX-FL	**73.89**		**61.46**		**82.10**

Table 6: Computational complexity of FL methods on the LLaVA-Med dataset.

Models	Trainable Params	Convergence Round	Training Time (hours)
FedProto [50]	13G	202	33.6
IOS [58]	13G	188	30.6
FedTGP [65]	13G	184	30.6
SynTeX-FL	**434M**	**148**	**26.3**

Table 7: Ablation study on SynTeX-FL on the LLaVA-Med dataset for CIDEr.

Cross-modal Text Synthesis Stage			Modality-specialized VQA Tuning Stage		
Image	Transfer	Quality-based Aggregation	Modality-specialized LoRA	Modality-specialized Aggregation	CIDEr
✓	✓	✓	✓	✓	**0.5543**
✓	✓		✓	✓	0.5492
✓	✓	✓	✓		0.5485
✓	✓	✓			0.5410
	✓	✓	✓	✓	0.5509
					0.5003

The results reflect the mean performance across three runs initialized with distinct random seeds, showing low standard deviations (e.g. 5.4×10^{-4} for SynTeX-FL) across runs and demonstrating the consistency of the findings. Bold indicates the best results in each table. Tables 1 and 2 reveal that SynTeX-FL outperforms the baseline FL methods across all evaluation metrics. This performance gain results from incorporating synthetic data, facilitating joint model optimization despite the lack of same-patient cross-modal pairs. Moreover, modality-specialized VQA tuning improved the representation quality, tailoring the model to the unique properties of each imaging domain. *Appendix I* presents qualitative examples of synthetic cross-modal texts and VQA responses, including clinical expert-guided analysis of whether the generated texts preserve modality-specific diagnostic findings. *Appendix G* also discusses the extensibility of the proposed framework to additional modalities, highlighting the challenges of aligning heterogeneous modality-specific representations. Additionally, *Appendix F* quantifies the performance gap between FL and centralized training, where the centralized model serves as an upper bound. For example, on LLaVA-Med, BLEU-1 decreases from 0.3512 to 0.3309 and CIDEr from 0.5734 to 0.5543, while on VQA-RAD, accuracy drops from 77.25 to 73.89. This gap reflects inherent FL challenges such as modality heterogeneity and restricted data sharing.

Table 3 presents the evaluation of SynTeX-FL when combined with domain adaptation techniques, namely SEA [57] and CAF [59]. We also assessed the integration of these methods with state-of-the-art FL frameworks, including FedTGP and IOS. Moreover, SynTeX-FL demonstrates superior performance in all combinations. These results demonstrate that text transfer outperforms feature-level domain adaptation [57, 59], as it explicitly describes modality-specific information in text rather than aligning feature spaces. Additionally, Table 4 compares SynTeX-FL with FedTGP and IOS using medical VQA backbones, including LLaVA-Med [32], Med-MoE [18], and AMiF [68], on VQA-RAD. SynTeX-FL with the original LLaVA backbone achieves the best performance across all metrics, demonstrating its competitiveness in cross-modal vertical FL without relying on a specialized medical VQA architecture. Furthermore, Table 5 compares SynTeX-FL with a real-only LoRA baseline in which cross-modal synthetic text is not used for LoRA tuning. SynTeX-FL consistently outperforms this baseline on both datasets, demonstrating the benefit of cross-modal synthetic supervision beyond real-data-only LoRA tuning.

Five clinical experts from Nuclear Medicine and Cardiology conducted blinded evaluations of 996 LLaVA-Med cases across five clinical-label categories. Detailed protocols and results are provided in *Appendix D*. SynTeX-FL consistently outperforms the baseline methods in both linguistic quality and clinical reliability. Figure 5 reveals that SynTeX-FL outperformed all compared models concerning sensitivity and positive predictive value. This result indicates that SynTeX-FL generates fewer incorrect responses, which is essential in healthcare applications.

Table 6 presents the comparison of the training efficiency of multiple FL methods. SynTeX-FL achieves faster convergence in fewer rounds (148 rounds) and requires significantly fewer trainable parameters (434M) than other methods (13G) using lightweight LoRA modules. This approach enables SynTeX-FL to operate effectively in clinical environments with limited resources.

4.4 Ablation Study

This section analyzes the contribution of each SynTeX-FL component. In Table 7, a checkmark (✓) indicates the module activation. The first and last rows present the performance of SynTeX-FL and the backbone (FedAvg), respectively. The second and third rows report the results when quality-aware and modality-specialized aggregation, respectively, are excluded. The fourth row provides the results when synthetic text is used without the modality-specialized VQA tuning stage. The fifth row presents the results when only text is used in the transfer. Comparing each module with SynTeX-FL confirms that each component contributes to performance improvement. *Appendix E* presents the experiments on layer selection for applying LoRA in modality-specific encoders and fine-tuning without LoRA.

5 Conclusion

This study addresses the critical challenge of cross-modal heterogeneity in federated VQA. We introduce SynTeX-FL, a cross-modal vertical FL framework that moves beyond feature-level alignment by enabling unpaired modality-to-modality semantic transfer through image-guided text synthesis and modality-specialized VQA tuning. SynTeX-FL operates under the FL framework by selectively training reconstruction transfers, sharing synthesis transfers, applying modality-specialized VQA tuning, and aggregating a global model to improve VQA accuracy. We empirically validated that SynTeX-FL achieves state-of-the-art performance in federated VQA on the LLaVA-Med and VQA-RAD datasets while maintaining high computational efficiency. The results demonstrate that the proposed approach effectively addresses the challenges arising from the absence of same-patient paired samples across imaging modalities in federated clinical scenarios.

Acknowledgments and Disclosure of Funding

This work was supported by the IITP grant funded by the Korea Government (MSIT) (No. RS-2026-25619158, RS-2024-00437718, RS-2023-00256629, RS-2022-00156287). The authors declare no competing interests.

References

[1] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

[2] Banerjee, S. and Lavie, A. (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.

[3] Bawazir, A., Alshanqiti, H., Wu, K., and Albreiki, F. (2025). Falcon medical visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29622–29624.

[4] Carlson, K., Riddell, A., and Rockmore, D. (2018). Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

[5] Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P. A. (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.

[6] Chen, W., Yu, Z., De Mello, S., Liu, S., Alvarez, J. M., Wang, Z., and Anandkumar, A. (2021). Contrastive syn-to-real generalization. *arXiv preprint arXiv:2104.02290*.

[7] Chiruvella, V., Guddati, A. K., et al. (2021). Ethical issues in patient data ownership. *Interactive journal of medical research*, 10(2):e22269.

[8] Dai, N., Liang, J., Qiu, X., and Huang, X. (2019). Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

[9] Dai, Q., Wei, D., Liu, H., Sun, J., Wang, L., and Zheng, Y. (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.

[10] Gohari, R. J., Aliahmadipour, L., and Valipour, E. (2024). Fedbrain-distill: Communication-efficient federated brain tumor classification using ensemble knowledge distillation on non-iid data. In *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 49–54. IEEE.

[11] He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

[12] Hong, Y., Kim, M. J., Lee, I., and Yoo, S. B. (2023). Fluxformer: Flow-guided duplex attention transformer via spatio-temporal clustering for action recognition. *IEEE Robotics and Automation Letters*, 8(10):6411–6418.

[13] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

[14] Huang, D., Ye, X., and Sakurai, T. (2024). Multi-party collaborative drug discovery via federated learning. *Computers in Biology and Medicine*, 171:108181.

[15] Huang, F., Chen, Z., Wu, C. H., Guo, Q., Zhu, X., and Huang, M. (2021). Nast: A non-autoregressive generator with word alignment for unsupervised text style transfer. *arXiv preprint arXiv:2106.02210*.

[16] Jhamtani, H., Gangal, V., Hovy, E., and Nyberg, E. (2017). Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

[17] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., evendra Singh Chaplot, las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and El, W. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

[18] Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., and Liu, Z. (2024). Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 3843–3860.

[19] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

[20] Jung, J. W., Yoon, J. H., and Yoo, S. B. (2024). Densesphere: Multimodal 3d object detection under a sparse point cloud based on spherical coordinate. *Expert Systems with Applications*, 251:124053.

[21] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

[22] Kim, M. H., Kim, C., and Yoo, S. B. (2025). X-flora: Cross-modal federated learning with modality-expert lora for medical vqa. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8390–8408.

[23] Kim, M. H., Kim, M. J., and Yoo, S. B. (2024). Occluded part-aware graph convolutional networks for skeleton-based action recognition. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7310–7317. IEEE.

[24] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[25] La Quatra, M., Gallipoli, G., and Cagliero, L. (2025). Self-supervised text style transfer using cycle-consistent adversarial networks. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–38.

[26] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R. (2024). Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

[27] Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M., and Boureau, Y.-L. (2019). Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

[28] Lao, M., Pu, N., Zhong, Z., Sebe, N., and Lew, M. S. (2023). Fedvqa: Personalized federated visual question answering over heterogeneous scenes. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7796–7807.

[29] Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

[30] Lee, S.-J. and Yoo, S. B. (2021). Super-resolved recognition of license plate characters. *Mathematics*, 9(19):2494.

[31] Lei, S. W., Gao, D., Wu, J. Z., Wang, Y., Liu, W., Zhang, M., and Shou, M. Z. (2023). Symbolic replay: Scene graph as prompt for continual learning on vqa task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1250–1259.

[32] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

[33] Li, P., Si, Q., Fu, P., Lin, Z., and Wang, Y. (2024). Object attribute matters in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18545–18553.

[34] Li, Q., He, B., and Song, D. (2021). Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722.

[35] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

[36] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

[37] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

[38] Lu, S., Liu, Z., Liu, T., and Zhou, W. (2023). Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, 126:107037.

[39] Luo, F., Li, P., Zhou, J., Yang, P., Chang, B., Sui, Z., and Sun, X. (2019). A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060*.

[40] Mahdavi, Z., Khodakaramimaghsoud, Z., Khaloo, H., Taleshani, S. B., Hashemi, E., Kaleybar, J. M., and Manzari, O. N. (2026). Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347.

[41] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

[42] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. (2023). Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

[43] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

[44] Peng, P., Fan, W., Shen, Y., Yang, X., and Zhou, D. (2025). A global visual information intervention model for medical visual question answering. *Computers in Biology and Medicine*, 192:110195.

[45] Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

[46] Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A., and Qadir, J. (2022). Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, 3:172–184.

[47] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[48] Raggio, C. B., Zabaleta, M. K., Skupien, N., Blanck, O., Cicone, F., Cascini, G. L., Zaffino, P., Migliorelli, L., and Spadea, M. F. (2025). Fedsynthct-brain: A federated learning framework for multi-institutional brain mri-to-ct synthesis. *Computers in Biology and Medicine*, 192:110160.

[49] Riyadi, M. A. Q., Dewi, A. M., Mukhlishin, Z. A. N., and Arep, Z. R. A. (2026). Comparative evaluation of federated learning algorithms in dirichlet non-iid medical imaging. *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, 5(1):32–44.

[50] Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. (2022). Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8432–8440.

[51] Tobaben, M., Souibgui, M. A., Tito, R., Nguyen, K., Kerkouche, R., Jung, K., Jälkö, J., Kang, L., Barsky, A., d'Andecy, V. P., et al. (2024). Neurips 2023 competition: Privacy preserving federated learning document vqa. *arXiv preprint arXiv:2411.03730*.

[52] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[53] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

[54] Wang, Q., Ji, R., Peng, T., Wu, W., Li, Z., and Liu, J. (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.

[55] Wang, X., Wang, J., Xiao, H., Chen, J., and Ma, F. (2024b). Fedkim: Adaptive federated knowledge injection into medical foundation models. *arXiv preprint arXiv:2408.10276*.

[56] Wang, Y., Wu, Y., Mou, L., Li, Z., and Chao, W. (2019). Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578.

[57] Wang, Y., Xu, Y., Yang, J., Chen, Z., Wu, M., Li, X., and Xie, L. (2023). Sensor alignment for multivariate time-series unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10253–10261.

[58] Wu, Z., Chen, T., and Ling, Q. (2023). Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*.

[59] Xie, B., Li, S., Lv, F., Liu, C. H., Wang, G., and Wu, D. (2022). A collaborative alignment framework of transferable knowledge extraction for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6518–6533.

[60] Xu, P., Cheung, J. C. K., and Cao, Y. (2020). On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR.

[61] Yan, Q., Duan, J., and Wang, J. (2024a). 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.

[62] Yan, Y., Wang, H., Huang, Y., He, N., Zhu, L., Xu, Y., Li, Y., and Zheng, Y. (2024b). Cross-modal vertical federated learning for mri reconstruction. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6384–6394.

[63] Yang, M., Wu, Z., Zheng, H., Huang, L., Ding, W., Pan, L., and Yin, L. (2024). Cross-modality medical image segmentation via enhanced feature alignment and cross pseudo supervision learning. *Diagnostics*, 14(16):1751.

[64] Yu, T., Tong, Z., Yu, J., and Zhang, K. (2025). Fine-grained adaptive visual prompt for generative medical visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9662–9670.

[65] Zhang, J., Liu, Y., Hua, Y., and Cao, J. (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.

[66] Zhang, R., Chi, X., Liu, G., Zhang, W., Du, Y., and Wang, F. (2023). Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation. *arXiv preprint arXiv:2303.15486*.

[67] Zhu, H., Togo, R., Ogawa, T., and Haseyama, M. (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.

[68] Zou, Y. and Yin, Z. (2025). Alignment, mining and fusion: Representation alignment with hard negative mining and selective knowledge fusion for medical visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29623–29633. IEEE.

Appendix

The appendix provides supplementary methodological details and experimental analyses supporting the proposed framework. Appendix A outlines the core algorithms of SynTeX-FL, covering local training and server-side aggregation. Appendix B examines the stability of quality-aware aggregation across client settings. Appendix C presents ablation studies on hyperparameters and LoRA ranks. Appendix D reports cross-modal text synthesis results with clinical evaluation. Appendix E investigates optimal LoRA injection layers. Appendix F quantifies the contribution of each module to overall VQA performance. Appendix G discusses the extensibility to other modalities. Appendix H provides further discussion on aggregation stability, cross-modal performance, and clinical applicability. Finally, Appendix I shows qualitative VQA examples across different imaging modalities.

A Method Algorithms

A.1 Cross-modal Text Synthesis

Algorithm 1 describes the local training process for cross-modal text synthesis. Each client generates synthetic text from its report and corresponding image using a frozen synthesis transfer module, and then reconstructs the original text using a trainable reconstruction transfer module. The reconstruction module is trained with adversarial and reconstruction objectives, while the discriminator learns to distinguish real text from reconstructed text.

A.2 Quality-aware Aggregation

Algorithm 2 specifies the server-side aggregation process using discriminator quality scores and gradient information. The server accumulates squared gradients across communication rounds and uses this history together with current discriminator scores and gradients to normalize client contributions. The resulting score- and gradient-based weighted parameters are combined to update the reconstruction transfer models for each modality pair.

A.3 Modality-specialized VQA Tuning

Algorithm 3 presents the fine-tuning procedure for modality-specialized low-rank adaptation (LoRA) modules. Each client updates separate LoRA parameters for the X-ray, computed tomography (CT), and magnetic resonance imaging (MRI) text encoders and the image encoder, while keeping the pretrained encoder weights fixed. The low-rank updates adapt the corresponding representations to the VQA task, and the trained LoRA parameters are uploaded to the server for aggregation.

A.4 Modality-specialized Aggregation

Algorithm 4 defines the aggregation process for the modality-specialized LoRA weights, maintaining separate parameter sets for each modality. For each text modality, source-aware weighting distinguishes contributions learned from real and synthetic data, with the balance ratio controlling their relative influence. Image-encoder LoRA weights are aggregated uniformly across clients. The server combines these contributions to obtain the updated global LoRA parameters for each modality.

Algorithm 1 Cross-modal Text Synthesis

Require: text report t, image data I, synthesis transfer module S, reconstruction transfer module R

$S(t, I)$ ▷ Generate transferred text using synthesis transfer module
$R(S(t, I), I)$ ▷ Reconstruct using reconstruction transfer module
$D(R(S(t, I), I))$ ▷ Distinguish between real and reconstruction text
$\mathcal{L}_{disc} \leftarrow \|1 - D(R(S(t, I), I))\|_2$
$\mathcal{L}_{rec} \leftarrow \|R(S(t, I), I) - t\|_1$
$\mathcal{L}_{total} \leftarrow \mathcal{L}_{disc} + \lambda \mathcal{L}_{rec}$
Optimize R and D to minimize and maximize $\mathcal{L}_{total}$
return weight of R, discriminator score $s^{r,i}$ and gradient $g^{r,i}$ to server

Algorithm 2 Quality-aware Aggregation

Require: modality-specific clients $N_k, k \in \{x, c, m\}$, discriminator gradient $g^{r,i}$ and score $s^{r,i}$

$G_k^r \leftarrow G_k^{r-1} + \sum_{i=1}^{N_k} (g_k^{r,i})^2$ ▷ Update cumulative gradient for each clients
for each client $i \in N_k$ **do**
  $\omega_{b2a,s}^{r,i} = \frac{s_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} s_a^{r,i} + \sqrt{G_a^r}}, \quad \forall a, b \in \{x, c, m\}, \ a \neq b$ ▷ Score-based normalized model weight
  $\omega_{b2a,g}^{r,i} = \frac{g_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} s_a^{r,i} + \sqrt{G_a^r}}, \quad \forall a, b \in \{x, c, m\}, \ a \neq b$ ▷ Gradient-based normalized model weight
  $\phi_{b2a}^{r+1} = \frac{1}{N_a} \sum_{i=1}^{N_a} (\omega_{b2a,s}^{r,i} + \omega_{b2a,g}^{r,i}), \quad \forall a, b \in \{x, c, m\}, \ a \neq b$ ▷ Normalized model weight-based aggregation
end for
return Aggregated weights ϕ_{b2a}^{r+1}

Algorithm 3 Modality-specialized VQA Tuning

Require: For each modality $k \in \{x, c, m, I\}$: input z_k, encoder weights E_k, and LoRA weights A_k, B_k

$\hat{z}_k = E_k z_k + B_k A_k z_k$ ▷ Refine modality-specific representation
Fine-tune only LoRA parameters A_k, B_k
return modality-specific LoRA weights $\Theta_k^{r,i} = \{A_k^{r,i}, B_k^{r,i}\}$

Algorithm 4 Modality-specialized Aggregation

Require: i-th clients N_k^r (real), N_k^s (synthetic), LoRA weights $\Theta_k^{r,i}$, and normalization ratio γ

for each client i **do**
  if i is real **then**
    $\alpha_k^i \leftarrow \frac{\gamma}{\gamma \cdot N_k^r + N_k^s}$ Compute real-client weight
  else
    $\alpha_k^i \leftarrow \frac{1}{\gamma \cdot N_k^r + N_k^s}$ ▷ Compute synthetic-client weight
  end if
end for
$\Theta_k^{r+1} = \frac{1}{N_k} \sum_{i=1}^{N_k} \alpha_k^i \Theta_k^{r,i}, k \in \{x, c, m, I\}$ ▷ Aggregate modality-specialized LoRA weights
return Aggregated weights Θ_k^{r+1}

Table 8: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE-L, and CIDEr on LLaVA-Med dataset **with 4 X-ray clients, 1 CT client and 1 MRI client.**

Dataset	LLaVA-Med				
Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
FedAvg	0.3641	0.3408	0.4968	0.2857	0.1446
FedProx	0.3652	0.3414	0.5003	0.2804	0.1478
MOON	0.3576	0.3455	0.5108	0.2908	0.1512
FedProto	0.3557	0.3447	0.5117	0.2915	0.1530
IOS	0.3602	0.3486	0.5178	0.2884	0.1497
FedTGP	0.3629	0.3515	0.5201	0.2990	0.1546
SynTeX-FL	**0.3902**	**0.3677**	**0.5451**	**0.3165**	**0.1615**

Table 9: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE-L, and CIDEr on LLaVA-Med dataset **with 1 X-ray client, 4 CT clients and 1 MRI client.**

Dataset	LLaVA-Med				
Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
FedAvg	0.3656	0.3421	0.4960	0.2864	0.1459
FedProx	0.3630	0.3405	0.4968	0.2797	0.1437
MOON	0.3650	0.3437	0.5011	0.2813	0.1467
FedProto	0.3639	0.3433	0.5027	0.2856	0.1497
IOS	0.3671	0.3453	0.5113	0.2901	0.1523
FedTGP	0.3601	0.3478	0.5188	0.2965	0.1523
SynTeX-FL	**0.3982**	**0.3709**	**0.5497**	**0.3168**	**0.1635**

Table 10: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE-L, and CIDEr on LLaVA-Med dataset **with 1 X-ray client, 1 CT client and 4 MRI clients.**

Dataset	LLaVA-Med				
Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
FedAvg	0.3658	0.3424	0.4990	0.2869	0.1461
FedProx	0.3628	0.3472	0.4946	0.2716	0.1476
MOON	0.3602	0.3454	0.5000	0.2820	0.1477
FedProto	0.3652	0.3443	0.5036	0.2864	0.1403
IOS	0.3678	0.3455	0.5119	0.2925	0.1525
FedTGP	0.3635	0.3459	0.5242	0.2987	0.1544
SynTeX-FL	**0.3992**	**0.3711**	**0.5465**	**0.3165**	**0.1639**

Table 11: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE-L, and CIDEr on LLaVA-Med dataset **with 3 X-ray client, 3 CT client and 3 MRI clients.**

Dataset	LLaVA-Med				
Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
FedAvg	0.3976	0.3536	0.5394	0.3185	0.1700
FedProx	0.4017	0.3706	0.5368	0.3185	0.1595
MOON	0.4043	0.3533	0.5375	0.3259	0.1695
FedProto	0.4012	0.3659	0.5433	0.3217	0.1675
IOS	0.3931	0.3690	0.5477	0.3229	0.1643
FedTGP	0.3898	0.3679	0.5435	0.3186	0.1721
SynTeX-FL	**0.4055**	**0.3720**	**0.5540**	**0.3300**	**0.1732**

B Ratio of Clients

Tables 8 to 10 compare the performance of SynTeX-FL with several existing FL methods under various client settings on the LLaVA-Med dataset. Table 8 evaluates the case with four X-ray clients, one CT client, and one MRI client. Table 9 examines the scenario with one X-ray client, four CT clients, and one MRI client, and Table 10 presents the results with one X-ray client, one CT client, and four MRI clients. Additionally, Table 11 represents the results with nine clients, consisting of three X-ray, CT and MRI clients.

Across the settings, SynTeX-FL consistently outperforms all existing methods in terms of the BLEU, METE-OR, ROUGE, and CIDEr metrics. These results underscore the robustness and effectiveness of SynTeX-FL, even with varying distributions of modality-specific clients. The superior performance demonstrates that SynTeX-FL effectively manages cross-modal heterogeneity and maintains high-quality VQA generation, regardless of the client composition.

C Hyperparameter

C.1 Weight of Total Loss

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

C.2 Modality-specialized LoRA

Table 13 summarizes the results of an ablation study evaluating the effect of various combinations of ranks (r_m, r_c, r_t, and r_I) assigned to modality-expert LoRA modules. Although the configuration of (16, 32, 8, 16) previously demonstrated promising results, we further validated its effectiveness by experimenting with alternative rank combinations. Table 13 reveals that the (16, 32, 8, 16) setting consistently outperforms other configurations across all evaluation metrics, including BLEU, METEOR, ROUGE, and CIDEr. This result confirms that assigning a moderate capacity to the X-ray,

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

λ	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	**0.3309**	**0.1740**	**0.3727**	**0.4057**	**0.5543**
0.6	0.3170	0.1610	0.3542	0.3911	0.5413

Table 13: 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.

Rank r_x,r_c,r_m,r_I	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
16,32,8,16	**0.3309**	**0.1740**	**0.3727**	**0.4057**	**0.5543**
32,32,8,16	0.3126	0.1543	0.3625	0.4002	0.5432
16,16,8,8	0.3129	0.1537	0.3695	0.4021	0.5459
16,32,16,8	0.3124	0.1533	0.3690	0.4012	0.5433

Table 14: 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_x	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_m	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
Rank r_I	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
8	0.3132	0.1550	0.3603	0.3802	0.5389
16	**0.3139**	**0.1565**	**0.3610**	**0.3868**	**0.5399**
32	0.3121	0.1538	0.3605	0.3855	0.5360

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

γ	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	**0.3309**	**0.1740**	**0.3727**	**0.4057**	**0.5543**
2	0.3105	0.1684	0.3589	0.3823	0.5317
2.5	0.3052	0.1647	0.3604	0.3760	0.5208

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

γ	Accuracy (%)		
	Open	Closed	Overall
0.5	56.48	76.98	68.84
1	58.45	77.95	70.21
1.5	**61.46**	**82.10**	**73.89**
2	59.10	77.58	70.24
2.5	58.38	76.99	69.60

CT, and MRI encoders and a smaller capacity to the image encoder leads to the most balanced and effective performance.

Table 14 presents an ablation study analyzing the contribution of each rank by varying one modality-specialized LoRA rank at a time. The evaluation was conducted on the LLaVA-Med dataset using the BLEU, METEOR, ROUGE, and CIDEr metrics. Moreover, Table 14 reveals that setting all modality-specialized LoRA ranks (r_m, r_c, r_t, and r_I) to 16, 32, 8, and 16 yields the best overall performance across the BLEU, METEOR, ROUGE, and CIDEr metrics on the LLaVA-Med dataset. This result implies that a balanced representation capacity across the text and image modalities of X-ray, CT, and MRI is most effective for the VQA task.

Moreover, Tables 15 and 16 explore the effect of the aggregation weight hyperparameter γ, controlling the balance between real and synthetic data contributions during modality-specific aggregation. As γ increases, real client data are assigned higher weights. The model performs best at $\gamma = 1.5$; however, performance degrades when $\gamma = 1$ (equal weighting) or $\gamma < 1$ (favoring synthetic data). This finding highlights the importance of prioritizing real data for robust VQA model training.

D Cross-modal Text Synthesis

Table 17 reports the performance of cross-modal text style transfer (TST) on the LLaVA-Med dataset, evaluated using accuracy, sBLEU, and perplexity. The results reveal that SynTeX-FL consistently outperforms all baseline models across different modalities, suggesting that clinical cues from visual information can steer TST.

We collaborated with clinicians from the Departments of Nuclear Medicine and Cardiology to assess the clinical validity of cross-modal text transfer. Upon their recommendations, we incorporated evaluation metrics critical for medical text generation: sensitivity, reflecting diagnostic completeness, and positive predictive value, quantifying the accuracy of positive outputs. As presented in Table 18,

Table 17: Performance of cross-modal text synthesis compared with baseline on the LLaVA-Med dataset.

Synthesis and Reconstruction Transfer	Method	Metric		
		ACC(%) (↑)	sBLEU (↓)	PPL (↓)
CT→MRI	NAST	52.28	0.6698	501.17
	TST-CycleGAN	84.47	0.6623	405.07
	Ours	**99.73**	**0.5906**	**271.50**
CT→X-ray	NAST	76.81	0.5521	346.11
	TST-CycleGAN	80.31	0.6395	547.21
	Ours	**99.85**	**0.3739**	**297.56**
MRI→CT	NAST	70.88	0.7038	342.84
	TST-CycleGAN	82.35	0.6567	309.55
	Ours	**99.78**	**0.6116**	**278.52**
MRI→X-ray	NAST	71.82	0.5927	644.31
	TST-CycleGAN	84.13	0.5790	622.15
	Ours	**99.86**	**0.1855**	**473.44**
X-ray→CT	NAST	89.91	0.6876	1324.63
	TST-CycleGAN	90.06	0.6526	591.30
	Ours	**99.89**	**0.3416**	**525.93**
X-ray→MRI	NAST	88.45	0.5867	1237.30
	TST-CycleGAN	81.37	0.6105	655.98
	Ours	**99.93**	**0.1807**	**276.57**

Table 18: Comparison with prior text style transfer methods in terms of Sensitivity and PPV on the LLaVA-Med dataset.

Metric	Sensitivity	PPV
NAST	79.90	78.58
TST-CycleGAN	82.64	81.09
SynTeX-FL	**98.52**	**98.20**

Table 19: Clinical-label consistency and visibility violations on the LLaVA-Med dataset.

Method	Accuracy	Precision	Recall	F1 score	Visibility Violation Rate
NAST	72.2	70.5	70.8	70.6	18.4
TST-CycleGAN	79.1	80.1	73.9	76.9	14.5
SynTeX-FL	**97.5**	**94.3**	**98.9**	**96.5**	**3.4**

Table 20: Ablation study on the selection of each layer (W_q, W_k, W_v, W_f) in terms of BLEU, METEOR, ROUGE-L, and CIDEr metrics on the LLaVA-Med dataset.

W_q	W_k	W_v	W_f	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
✓	✓	✓	✓	0.3225	0.1608	0.3525	0.3852	0.5398
	✓	✓	✓	**0.4057**	**0.3727**	**0.5543**	**0.3309**	**0.1740**
✓		✓	✓	0.3580	0.2819	0.4104	0.2229	0.0963
✓	✓		✓	0.3701	0.1998	0.4135	0.2579	0.1436
✓	✓	✓		0.3555	0.2597	0.4412	0.2381	0.1338
✓			✓	0.3169	0.2745	0.4621	0.2440	0.1190
	✓		✓	0.3013	0.2691	0.4421	0.2237	0.1558
		✓	✓	0.3720	0.2871	0.4884	0.2770	0.1600

Table 21: Comparison with prior FL methods in terms of finding-level correctness on the LLaVA-Med dataset.

Model	FedAvg	FedProx	MOON	FedProto	IOS	FedTGP	FedMedVLP	FedKIM	SynTeX-FL
Finding-level Correctness	42.1	43.0	44.2	44.8	45.3	46.5	45.7	47.2	**52.8**

Table 22: Ablation study on LoRA fine-tuning in terms of BLEU, METEOR, ROUGE-L, and CIDEr metrics on the LLaVA-Med dataset.

Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
Full Fine-tuning	0.3225	0.1608	0.3525	0.3852	0.5398
SynTeX-FL	**0.3309**	**0.1740**	**0.3727**	**0.4057**	**0.5543**

Table 23: Ablation study on synthetic text quality in terms of BLEU, METEOR, ROUGE-L, and CIDEr metrics on the LLaVA-Med dataset.

Dataset	LLaVA-Med				
Metric	ROUGE-L	METEOR	CIDEr	BLEU-1	BLEU-5
SynTeX-FL (Low-quality Synthetic Text)	0.4042	0.3698	0.5503	0.3289	0.1730
SynTeX-FL	**0.4055**	**0.3720**	**0.5540**	**0.3300**	**0.1732**

SynTeX-FL demonstrates superior performance in sensitivity and positive predictive value compared to prior TST baselines, such as NAST and TST-CycleGAN. These results suggest that SynTeX-FL produces more clinically reliable responses by reducing false positives and enhancing the capture of relevant findings, which is crucial for downstream tasks, such as medical VQA, where textual precision directly affects diagnostic decision-making. To further evaluate hallucination risk and clinical consistency on LLaVA-Med, we assessed 996 source cases across five prespecified clinical label categories: primary finding, anatomical localization, imaging appearance, severity, and additional findings. The 996 source cases were approximately evenly divided among five clinicians. Blinded to the generation method, each clinician evaluated the original report and all corresponding target-modality reports generated by NAST, TST-CycleGAN, and SynTeX-FL for their assigned cases under the same protocol. The clinicians annotated their assigned cases by labeling each finding as present or absent in the source report and in each corresponding generated target-modality report. We also report on the visibility-violation rate as a visibility-specific hallucination indicator, defined as the proportion of generated positive statements that refer to findings not assessable in the target modality. Clinicians additionally labeled whether each generated positive finding was assessable in the specified target modality. Metrics were micro-averaged over the five clinical-label categories across all reviewed cases. As shown in Table 19, SynTeX-FL consistently achieves strong performance across all metrics, indicating reliable preservation of clinician-annotated findings with limited omissions and unsupported additions.

We also evaluate finding-level correctness, defined as a binary metric that measures whether the predicted clinical findings fully match the ground-truth findings at the case level as shown in Table 21. Specifically, a prediction is considered correct only if all clinically relevant findings are accurately identified without omission or addition; otherwise, it is counted as incorrect. This metric provides a coarse-grained assessment of clinical accuracy, reflecting whether the generated report is entirely consistent with the ground truth in terms of diagnostic findings.

Figure 6 compares the average performance of S and R in cross-modal text transfer. We evaluated S and R regarding accuracy, sBLEU, and perplexity. Cross-modal text transfer outperforms the baseline

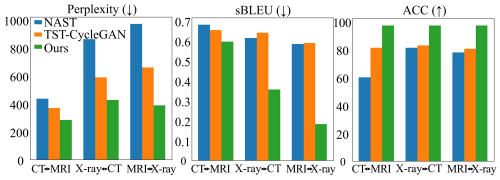

Figure 6: Performance of cross-modal text synthesis compared with the baseline for perplexity, sBLEU and accuracy (ACC) on the LLaVA-Med dataset.

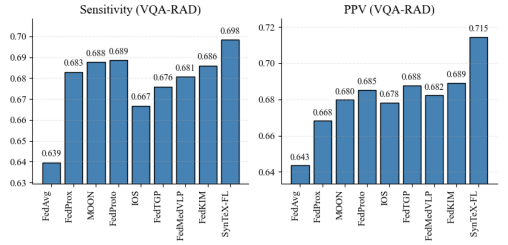

Figure 7: Comparison with prior FL methods for sensitivity and positive predictive value (PPV) on the VQA-RAD datasets.

Table 24: Ablation study for SynTeX-FL on the VQA-RAD dataset in terms of the accuracy.

Federated Asymmetric Translation			Federated VQA Finetuning		Overall Accuracy (%)
Text	Translation	Discriminator-based Aggregation	Modality-expert LoRA	Modality-specific Aggregation	
✓	✓	✓	✓	✓	**73.89**
✓	✓		✓	✓	73.11
✓	✓	✓	✓		73.10
✓	✓	✓			72.55
	✓	✓	✓	✓	72.98
					64.76

Table 25: Ablation study on federated learning in terms of BLEU, METEOR, ROUGE-L, CIDEr and overall accuracy on the LLaVA-Med and VQA-RAD dataset.

Dataset	LLaVA-Med					VQA-RAD
Metric	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr	Overall Accuracy
Centralized Learning (LLaVA)	**0.3512**	**0.1924**	**0.3928**	**0.4112**	**0.5734**	**77.25**
SynTeX-FL	0.3309	0.1740	0.3727	0.4057	0.5543	73.89

TST models, including NAST and TST-CycleGAN, by achieving higher accuracy and a lower sBLEU and perplexity. A lower sBLEU score is desirable because it indicates an effective style transfer with reduced lexical overlap and better incorporation of modality-specific clinical language. This result is attributed to the enrichment of the textual features with clinical cues derived from the associated medical images. Moreover, Figure 7 reveals that SynTeX-FL outperformed all compared models concerning sensitivity and positive predictive value. This result indicates that SynTeX-FL generates fewer incorrect responses, which is essential in healthcare applications.

E Modality-specialized LoRA

We conducted an ablation study on four candidate projection matrices in the transformer architecture: the query (W_q), key (W_k), value (W_v), and feedforward (W_f) layers, to determine the optimal configuration of LoRA injection layers. In Table 20, we systematically assess all combinations of these layers using standard generation metrics on the LLaVA-Med dataset. The results indicate that simultaneously applying LoRA to all four components yields the best performance across all metrics. Notably, configurations omitting either the attention (e.g., W_k, W_v) or feedforward (W_f) layers substantially degrade performance, highlighting the importance of capturing contextual interactions and feedforward transformations for effective multimodal representation learning.

To verify the efficiency of the proposed LoRA-based fine-tuning strategy, we compared SynTeX-FL against conventional full fine-tuning in terms of generation quality on the LLaVA-Med dataset. Table 22 reveals that SynTeX-FL achieves comparable or superior performance across all evaluation metrics while requiring considerably fewer trainable parameters. This result emphasizes the effectiveness of lightweight LoRA modules in preserving expressiveness without incurring the computational over-head of full model updates. These findings validate SynTeX-FL as a more scalable and resource-efficient alternative for fine-tuning large vision language models in federated clinical environments.

F Ablation Study

To analyze the effect of low-quality synthetic data on the proposed quality-aware aggregation, Table 23 presents the results obtained when the generated cross-modal texts are deliberately degraded before being used in training. Specifically, low-quality synthetic texts are constructed by randomly masking clinically relevant tokens and replacing a subset of medical entities with mismatched findings from other samples, simulating incomplete and clinically inconsistent descriptions.

Although performance slightly decreases compared with the standard setting using clean synthetic texts, SynTeX-FL avoids severe degradation under low-quality synthetic inputs. This behavior indicates that the proposed aggregation strategy effectively reduces the influence of unreliable synthetic information by down-weighting corrupted updates during training.

Table 24 presents an additional ablation study of the individual contributions of each module in the SynTeX-FL framework. The results reveal that each module consider-ably enhances the overall performance of SynTeX-FL. The combination of these modules operates synergistically to maximize VQA performance, addressing challenges posed by cross-modal FL heterogeneity.

Moreover, Table 25 reports the results of centralized learning under identical settings, serving as an upper bound for comparison with the SynTeX-FL framework.

G Limitation

This study focuses on X-ray, CT and MRI, 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 SynTex-FL by using specialized synthesis and reconstruction transfers adapted to other modalities. Expanding the number of synthesis and reconstruction transfers 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 semantic characteristics across modalities. This challenge becomes even more pronounced when incorporating modalities beyond X-ray, CT and MRI, 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.

H Discussion

H.1 Quality-aware Aggregation

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

Moreover, the proposed framework does not solely rely on discriminator scores for determining the aggregation weight. Instead, the framework incorporates additional signals, including the gradient of the discriminator loss and 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 support more robust aggregation decisions.

H.2 Experiments on Cross-modality

This study validates the superiority of the proposed approach by effectively addressing cross-modal heterogeneity via a combination of domain adaptation and federated learning (FL) strategies. While combinations of domain adaptation 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 in medical imaging modalities, such as X-ray, CT, and MRI.

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, these methods struggle to fully capture modality-specific semantic cues and achieve effective cross-modal representation learning. In contrast, the proposed SynTeX-FL framework mitigates these challenges by employing a federated asymmetric translation and federated VQA

Table 26: Representative examples of source-supported findings and their synthesized expressions across target imaging modalities.

Figure	Original Modality	Original-text Key Phrase	Target Modality	Synthesized-text Key Phrase
8	CT	"narrowing of the left main bronchus to 5.7 mm"	MRI	"**focal narrowing** of the left main bronchus ... approximately 5.7 mm"
			X-ray	"left main bronchus ... **cannot be reliably assessed on plain radiographs**"
10	CT	"ground glass ... peribronchial consolidation ... mild bronchiolectasis"	MRI	"**increased T2 signal intensity** ... **peribronchial signal alterations**"
			X-ray	"bilateral **patchy opacities**; **peribronchial thickening**; bronchiolectasis is **not clearly visualized**"
13	MRI	"high T2 and FLAIR signal in the occipital and posterior parietal subcortical white matter ... PRES"	CT	"**subtle hypodensities** in the occipital and posterior parietal white matter"
			X-ray	"white matter changes ... **are not visualizable**"
15	MRI	"metastatic lesion in the right temporal lobe with brain edema"	CT	"**hyperdense lesion** ... surrounding **hypodense areas** consistent with vasogenic edema"
			X-ray	"**not diagnostic** for intracranial metastasis"
19	X-ray	"bilateral fluffy infiltrates ... normal cardiac shadow ... non-cardiogenic pulmonary edema"	MRI	"bilateral areas of **increased signal intensity**; **no evidence of cardiomegaly**"
			CT	"bilateral **ground-glass opacities**; **perihilar distribution**"

fine-tuning. This design allows each modality to retain its unique characteristics while enabling effective cross-modal representation learning.

The experimental results demonstrate that the proposed approach consistently outperforms a combination of domain adaptation and FL strategies, particularly in handling complex modality-specific reasoning tasks. This result underscores the effectiveness of explicitly modeling cross-modal heterogeneity via structured translation and fine-tuning mechanisms, rather than by relying solely on shared representations.

H.3 Application of SynTeX-FL

The SynTeX-FL approach has strong potential for integration into real-world clinical decision-support systems. Similar to LLaVA-Med, which has utility in answering instruction-style diagnostic queries, the proposed model can assist clinicians by providing consistent answers across imaging modalities even without paired data. For example, SynTeX-FL can be applied in scenarios where CT or MRI may be unavailable, offering enriched responses using modality-informed representations. Additionally, SynTeX-FL could support radiology training by presenting cross-modal reasoning paths, akin to Med-Flamingo, which has demonstrated clinical question answering across diverse modalities. These cases highlight the capacity of SynTeX-FL for performance gains and practical deployment in federated, resource-constrained healthcare settings.

I VQA Results

Figures 8−19 present qualitative examples of VQA results using the LLaVA-Med dataset. Specifically, figures 8−11 illustrate cases based on CT data, while figures 12−15 focus on MRI-based VQA scenarios. Figures 16−19 provide the cases based on X-ray data.

Moreover, Figures 20−22 present qualitative CT, MRI and X-ray examples of VQA results using the VQA-RAD dataset, respectively. The annotations in these figures were prepared based on clinician assessment: red arrows indicate the lesion or anatomical structure referenced in the report, red text marks clinically incorrect or inconsistent responses, and blue text marks clinically accurate and

contextually appropriate responses. As shown in figures, SynTeX-FL preserves source-supported pathology while adapting it to the target modality.

Table 26 provides representative examples of how source-supported clinical findings are expressed according to the diagnostic characteristics of each target modality. For example, in Fig. 19, X-ray "bilateral fluffy infiltrates" are expressed on MRI as "bilateral areas of increased signal intensity," which is physically plausible because pulmonary edema increases lung water content and may produce high signal on fluid-sensitive MRI sequences. The absence of cardiomegaly is also preserved, supporting the non-cardiogenic interpretation. In Fig. 13, the source MRI finding of "high T2/FLAIR signal" is represented on CT as "subtle hypodensities" in the same occipital and posterior parietal white-matter regions. MRI sensitively depicts white-matter abnormalities through changes in tissue signal intensity, whereas CT reflects differences in X-ray attenuation, causing the same process to appear as more subtle low-density changes. In Fig. 10, the source CT findings of "ground-glass" and "peribronchial consolidation" are represented on X-ray as "bilateral patchy opacities" and "peribronchial thickening." CT provides detailed cross-sectional images of pulmonary abnormalities, whereas X-ray projects three-dimensional thoracic structures onto a two-dimensional image, causing the same process to appear as broader changes in lung opacity and bronchial contours.

These examples provide qualitative evidence that SynTeX-FL preserves source-supported findings while expressing them in forms consistent with the physical and diagnostic observability of each target modality. We further qualitatively examine whether such modality-consistent preservation of clinical findings contributes to downstream VQA performance. In each example, 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 SynTeX-FL 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.



CT Image

Computed tomography of the chest with intravenous contrast in the axial plane demonstrating **narrowing** of the **left main bronchus** to 5.7 mm.

Real Text

Contrast-enhanced axial thoracic MRI reveals focal **narrowing** of the **left main bronchus,** measuring approximately 5.7 mm in diameter.

Synthetic MRI Text

Chest X-ray does not clearly visualize the **left main bronchus**; bronchial narrowing cannot be reliably assessed on plain radiographs.

Synthetic X-ray Text

User	What is the main finding in the image?
IOS	The main finding in the image is the presence of a large **right pleural effusion.**
FedTGP	The main finding in the image is the **narrowing** of the **right main bronchus.**
SynTeX-FL (w/o Image)	The main finding is a **narrowing** of the **left pulmonary artery.**
SynTeX-FL	The main finding in the image is the **left main bronchial narrowing**
GPT-4	The main finding in the image is the **narrowing of the left main bronchus**.

Figure 8: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.



CT Image

Axial view of a contrast-enhanced CT of the **abdomen** demonstrating a **decompressed gallbladder and biliary leakage**. CT: computed tomography.

Real Text

Axial abdominal MRI with contrast demonstrates a collapsed **gallbladder and hyperintense signal along the biliary tract,** suggestive of bile leakage.

Synthetic MRI Text

Abdominal radiograph does not visualize the **gallbladder or biliary tract.**

Synthetic X-ray Text

User	What is the main finding in the image?
IOS	The main finding in the image is the presence of a **large amount of ascites**, which is an **abnormal accumulation of fluid** in the **abdominal** cavity.
FedTGP	The main finding in the axial view of the contrast-enhanced CT image of the **abdomen** is a **thickened gallbladder wall suggestive of cholecystitis**.
X-FLoRA (w/o Expert Aggregation)	The main finding in the axial view of the contrast-enhanced CT image of the **abdomen** is a **normal gallbladder**.
X-FLoRA	The main finding in the image is the **decompressed gallbladder and biliary leakage.**
GPT-4	The main finding in the axial view of the contrast-enhanced CT image of the **abdomen** is a **decompressed gallbladder and biliary leakage.**

Figure 9: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.



CT Image

Lung CT scan in the **lower lobes** showing **ground glass, arcade-like bands of parenchymal consolidation, peribronchial consolidation and mild bronchiolectasis**

Real Text

Thoracic MRI reveals areas of increased T2 signal intensity in the **lower lobes, with curvilinear parenchymal changes and peribronchial signal alterations suggestive of mild bronchiolectasis.**

Synthetic MRI Text

Chest X-ray shows bilateral patchy opacities in the **lower lung zones, with possible peribronchial thickening. Bronchiolectasis is not clearly visualized on radiograph.**

Synthetic X-ray Text

User	What abnormalities can be seen in the lower lobes?
IOS	The main feature in the image is the **renal venous thrombus**, which is a **blood clot that has formed within the renal vein**.
FedTGP	In the lower lobes, **diffuse emphysematous changes and bullae are noted**.
SynTeX-FL (w/o Image)	In the **lower lobes of the lung**, **multiple cavitary lesions** is identified.
SynTeX-FL	In the **lower lobes of the lung**, **peribronchial consolidation** and **multiple cavitary lesions** are identified.
GPT-4	In the **lower lobes of the lung**, there are several abnormalities visible, including **ground glass opacities, arcade-like bands of parenchymal consolidation, peribronchial consolidation, and mild bronchiolectasis**.

Figure 10: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.



CT Image

A computed tomography scan shows that the mass arises from the posterior costal arc of the eighth rib on the patient left side and **does not involve the medulla**.

Real Text

Magnetic resonance imaging reveals a mass originating from the posterior arc of the eighth left rib, **with no evidence of medullary involvement**.

Synthetic MRI Text

Chest X-ray demonstrates a radiopaque mass along the posterior aspect of the left eighth rib, **with no signs of medullary extension.**

Synthetic X-ray Text

User	Write an exhaustive depiction of the given image.
IOS	The CT scan highlights surrounding structures such as **lungs and heart** in relation to the mass.
FedTGP	The CT scan shows that the mass **has infiltrated** the **medulla,** causing **visible disruption** to its structure.
SynTeX-FL (w/o Image)	The CT scan indicates that the **medulla** is intact but appears to show **signs of infiltration** by the mass.
SynTeX-FL	The CT scan indicates that the **medulla** is intact and **uninvolved**, showing **no signs of infiltration** by the mass.
GPT-4	The CT scan also shows the **medulla,** which is **not involved** with the mass. The medulla appears to be intact and separate from the mass, indicating that the mass **has not infiltrated** or affected this area.

Figure 11: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.



MRI Image

Post-treatment MRI of the brain with and without contrast. Postoperative changes in the **right frontotemporoparietal** region are shown. Irregular enhancement at the surgical site extends to the walls of the right lateral ventricle. This was unchanged in subsequent MRI scans.

Real Text

Postoperative head CT reveals post-surgical changes in the **right frontotemporoparietal** region with persistent irregular density near the right lateral ventricular wall.

Synthetic CT Text

Skull radiograph is limited, but postoperative changes may be visible over the **right frontotemporoparietal** region; ventricular involvement cannot be assessed.

Synthetic X-ray Text

User	What region of the brain is affected?
IOS	The affected region in the brain is the **left centrotemporal area.**
FedTGP	The affected region in the brain is the **left frontal lobe only.**
SynTeX-FL (w/o Image)	The affected region of the brain is the **right parietal region**.
SynTeX-FL	The affected region in the brain is the **right frontoparietal area.**
GPT-4	The affected region in the brain is the **right frontotemporoparietal area**.

Figure 12: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.



MRI Image

MRI of the brain showing high T2 and FLAIR signal in the **occipital and posterior parietal** subcortical white matter consistent with PRES following treatment with rituximab.

Real Text

Non-contrast head CT shows subtle hypodensities in the **occipital and posterior parietal subcortical white matter.**

Synthetic CT Text

Skull X-ray is not suitable for evaluating white matter changes; **findings suggestive of PRES are not visualizable.**

Synthetic X-ray Text

User	What part of the brain is shown in the image?
IOS	The image shows the **cerebellum** regions of the brain.
FedTGP	The image shows **only** the **occipital region of the brain**.
SynTeX-FL (w/o Image)	The image shows the **frontal** and **posterior parietal regions of the brain**.
SynTeX-FL	The image shows the **occipital and posterior parietal areas.**
GPT-4	The image shows the **occipital and posterior parietal regions of the brain**.

Figure 13: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

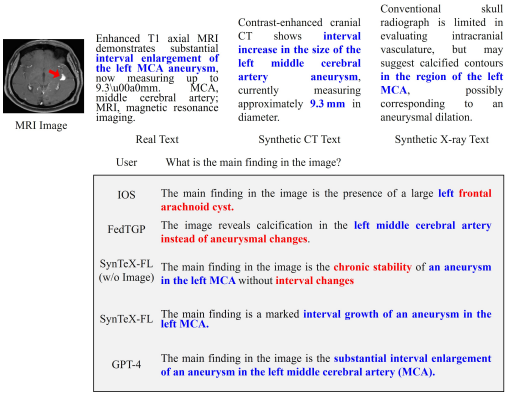

Figure 14: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

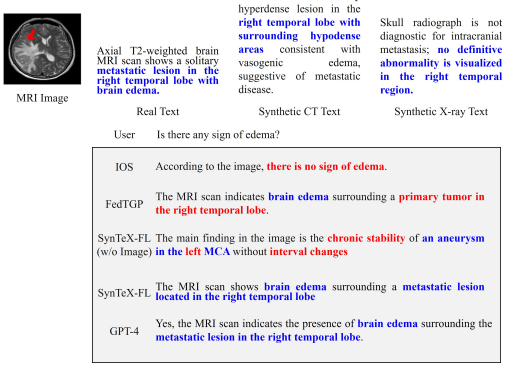

Figure 15: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

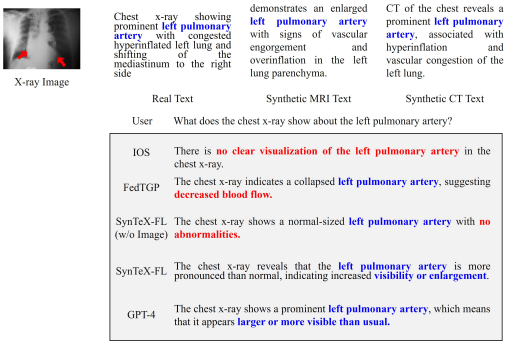

Figure 16: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

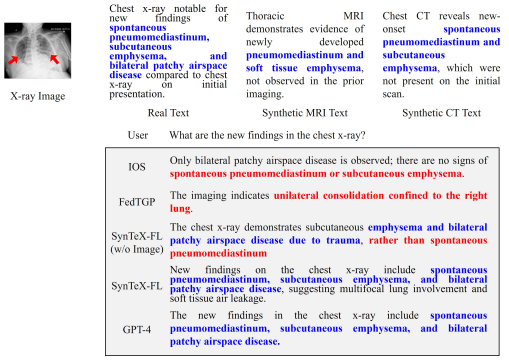

Figure 17: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

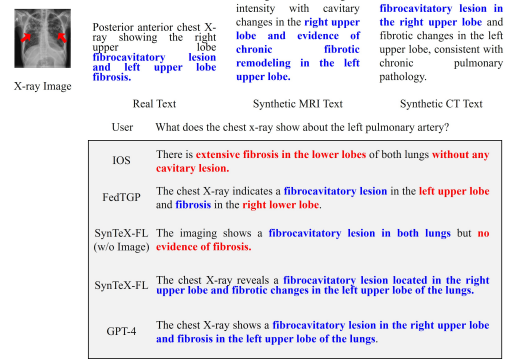

Figure 18: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

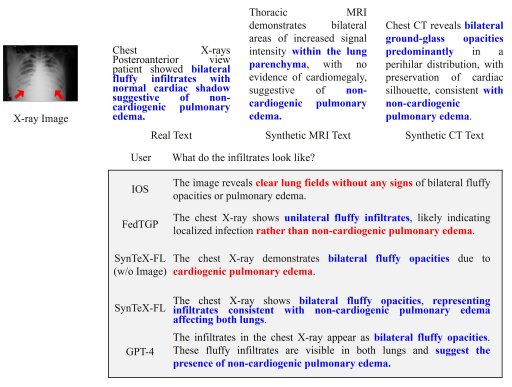

Figure 19: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. GPT-4-assisted reports are used as reference reports for qualitative comparison.

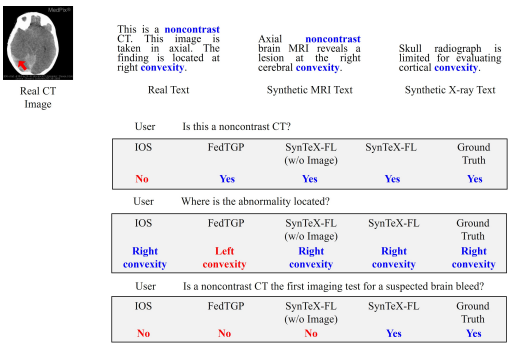

Figure 20: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.

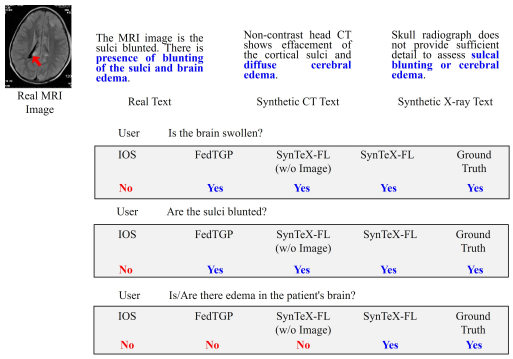

Figure 21: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.

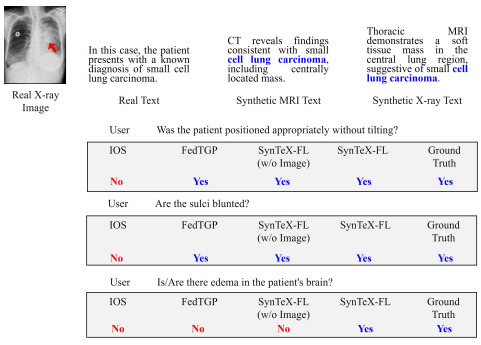

Figure 22: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our proposed framework is clearly stated in the abstract and introduction. Our claims are supported by the experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These discussions and limitations are addressed in Appendix G, along with considerations for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide mathematical formulations of the proposed method, including aggregation strategies.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the necessary information for reproducing the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable aPublication for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be fully organized and released once the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the details of the training procedure, including hyperparameters, optimizer type, and the datasets used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports the mean performance over three independent runs with different random seeds, along with the corresponding standard deviation, which captures the variability due to random initialization.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information about the required computer resources in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have thoroughly reviewed the NeurIPS Code of Ethics and strictly adhere to its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts of our work in Appendix H.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve the release of models or datasets that pose a high risk of misuse. The proposed method focuses on improving model performance in a federated learning setting, and does not introduce new generative capabilities or publicly release sensitive data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper fully cites all methods and datasets used in our approach.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our study does not include crowdsourcing experiments or research involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our study does not include crowdsourcing experiments or research involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: Our research does not include LLMs as part of the core methodology and only uses them for writing or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.