

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

Min Hyuk Kim^{1, 2}, Taeyeong Kim¹, Seok Bong Yoo^{1,*}

¹Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Republic of Korea

²Korea Electronics Technology Institute, Seongnam-si, Gyeonggi-do, Republic of Korea

alsgur0720@keti.re.kr, sbyoo@jnu.ac.kr

Appendix

The appendix provides supplementary details supporting the proposed framework and experiments. Appendix A outlines the core algorithms of SynTeX-FL. 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. 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.

A Method Algorithms

A.1 Cross-modal Text Synthesis

Algorithm 1 describes the training process of text-driven translation. Each client produces synthetic images and reconstructs the original image.

Quality-aware Aggregation Algorithm 2 specifies the aggregation process using discriminator quality scores and gradient information.

A.2 Modality-specialized VQA tuning

Algorithm 3 presents the fine-tuning phase for modality-specialized low-rank adaptation (LoRA) modules. Each client updates only the lightweight LoRA parameters for the X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and image modalities.

Modality-specialized Aggregation Algorithm 4 defines the aggregation process for the modality-specialized LoRA weights.

B Ratio of Clients

Tables 1 to 3 compare the performance of SynTeX-FL with several existing FL methods under various client settings on the LLaVA-Med dataset. Table 1 evaluates the case with four

Algorithm 1: Cross-modal Text Synthesis

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

- 1: $S(t, I) \triangleright$ Generate transferred text using synthesis transfer module
- 2: $R(S(t, I), I) \triangleright$ Reconstruct using reconstruction transfer module
- 3: $D(R(S(t, I), I)) \triangleright$ Distinguish between real and reconstruction text
- 4: $\mathcal{L}_{disc} \leftarrow \|1 - D(R(S(t, I), I))\|_2$
- 5: $\mathcal{L}_{rec} \leftarrow \|R(S(t, I), I) - t\|_1$
- 6: $\mathcal{L}_{total} \leftarrow \mathcal{L}_{disc} + \lambda \mathcal{L}_{rec}$
- 7: Optimize R and D to minimize and maximize \mathcal{L}_{total}
- 8: **return** weight of R , discriminator score $s^{r,i}$ and gradient $g^{r,i}$ to server =0

Algorithm 2: Quality-aware Aggregation

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

- 1: $G_k^r \leftarrow G_k^{r-1} + \sum_{i=1}^{N_k} (g_k^{r,i})^2 \triangleright$ Update cumulative gradient for each clients
- 2: **for** each client $i \in N_k$ **do**
- 3: $\omega_{b2a,s}^{r,i} = \frac{s_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} s_a^{r,i} + \sqrt{G_a^r}}, \quad \forall a, b \in \{x, c, m\}, a \neq b \triangleright$ Score-based normalized model weight
- 4: $\omega_{b2a,s}^{r,i} = \frac{g_a^{r,i} \cdot \phi_{b2a}^{r,i}}{\sum_{i=1}^{N_a} g_a^{r,i} + \sqrt{G_a^r}}, \quad \forall a, b \in \{x, c, m\}, a \neq b \triangleright$ Gradient-based normalized model weight
- 5: $\phi_{b2a}^{r+1} = \frac{1}{N_a} \sum_{i=1}^{N_a} (\omega_{b2a,s}^{r,i} + \omega_{b2a,g}^{r,i}), \quad \forall a, b \in \{x, c, m\}, a \neq b \triangleright$ Normalized model weight-based aggregation
- 6: **end for**
- 7: **return** Aggregated weights $\phi_{b2a}^{r+1} = 0$

X-ray clients, one CT client, and one MRI client. Table 2 examines the scenario with one X-ray client, four CT clients, and one MRI client, and Table 3 presents the results with one X-ray client, one CT client, and four MRI clients. Additionally, Table 4 represents the results with nine clients, consisting of three X-ray, CT and MRI clients.

Across the settings, SynTeX-FL consistently outperforms

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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

- 1: $\hat{z}_k = E_k z_k + B_k A_k z_k \triangleright$ Refine modality-specific representation
- 2: Fine-tune only LoRA parameters A_k, B_k
- 3: **return** modality-specific LoRA weights $\Theta_k^{r,i} = \{A_k^{r,i}, B_k^{r,i}\} = 0$

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

- 1: **for** each client i **do**
- 2: **if** i is real **then**
- 3: $\alpha_k^i \leftarrow \frac{\gamma}{\gamma \cdot N_k^r + N_k^s}$ Compute real-client weight
- 4: **else**
- 5: $\alpha_k^i \leftarrow \frac{1}{\gamma \cdot N_k^r + N_k^s}$ \triangleright Compute synthetic-client weight
- 6: **end if**
- 7: **end for**
- 8: $\Theta_k^{r+1} = \frac{1}{N_k} \sum_{i=1}^{N_k} \alpha_k^i \Theta_k^{r,i}, k \in \{x, c, m, I\} \triangleright$ Aggregate modality-specialized LoRA weights
- 9: **return** Aggregated weights $\Theta_k^{r+1} = 0$

all existing methods in terms of the BLEU, METEOR, 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 5 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 6 presents an ablation study analyzing the contribution of rank (r_m, r_c, r_t , and r_I) using varying its rank, where only one modality-specialized LoRA is fine-tuned. The evaluation was conducted on the LLaVA-Med dataset using the BLEU, METEOR, ROUGE, and CIDEr metrics. Moreover, Table 6 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

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

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

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

Table 2: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 1 X-ray client, 4 CT clients and 1 MRI client**.

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

Table 3: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 1 X-ray client, 1 CT client and 4 MRI clients**.

Dataset	LLaVA-Med				
Metric	BLEU-1	BLUE-5	METEOR	ROUGE	CIDEr
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

Table 4: Comparison with prior federated learning methods in terms of BLEU, METEOR, ROUGE, and CIDEr on LLaVA-Med dataset **with 3 X-ray client, 3 CT client and 3 MRI clients**.

text and image modalities of X-ray, CT, and MRI is most effective for the VQA task.

Table 7 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.

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

Rank	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
r_x					
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	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
r_c					
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	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
r_m					
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	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
r_I					
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 6: 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.

Although the configuration of (16, 32, 8, 16) previously demonstrated promising results, we further validated its effectiveness by experimenting with alternative rank combinations. Table 7 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, CT, and MRI encoders and a smaller capacity to the image encoder leads to the most balanced and effective performance.

Moreover, Tables 8 and 9 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 (favoring synthetic data). This finding highlights the importance of prioritizing real data for robust VQA model training.

D Cross-modal Text Synthesis

Table 10 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

Rank	BLEU-1	BLEU-5	METEOR	ROUGE	CIDEr
r_x, r_c, r_m, r_I					
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 7: 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.

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

Table 9: Effect of the normalization ratio (γ) on accuracy in expert-aware weighting on the VQA-RAD dataset.

SynTeX-FL consistently outperforms all baseline models 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 11, 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.

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 12, we systematically assess all combinations of these layers using standard generation metrics on the LLaVA-Med dataset. The results indicate that simultaneously applying

Synthesis and Reconstruction Transfer	Method	Metric		
		ACC(%) (\uparrow)	sBLEU (\downarrow)	PPL (\downarrow)
CT \rightarrow MRI	NAST	52.28	0.6698	501.17
	TST-CycleGAN	84.47	0.6623	405.07
	Ours	99.73	0.5906	271.50
CT \rightarrow 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 \rightarrow CT	NAST	70.88	0.7038	342.84
	TST-CycleGAN	82.35	0.6567	309.55
	Ours	99.78	0.6116	278.52
MRI \rightarrow 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 \rightarrow 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 \rightarrow MRI	NAST	88.45	0.5867	1237.30
	TST-CycleGAN	81.37	0.6105	655.98
	Ours	99.93	0.1807	276.57

Table 10: Performance of cross-modal text synthesis compared with baseline 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 11: Comparison with prior text style transfer methods in terms of Sensitivity and PPV on the LLaVA-Med dataset.

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 13 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

Table 14 presents an additional ablation study of the individual contributions of each module in the SynTeX-FL framework. The results reveal that each module considerably 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 15 reports the results of centralized learning under identical settings, serving as an upper bound for comparison with the SynTeX-FL framework.

W_q	W_k	W_v	W_f	BLEU-1	BLEU-5	METEOR	ROUGE-L	CIDEr
✓	✓	✓	✓	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 12: 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.

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

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

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

Table 14: Ablation study for SynTeX-FL on the VQA-RAD dataset in terms of the accuracy.

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

Table 15: Ablation study on federated learning in terms of BLEU, METEOR, ROUGE-L, CIDEr and overall accuracy on the LLaVA-Med and VQA-RAD dataset.

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 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 1–12 present qualitative examples of VQA results using the LLaVA-Med dataset. Specifically, figures 1–4 illustrate cases based on CT data, while figures 5–8 focus on MRI-based VQA scenarios. Figures 9–12 provides the cases based on X-ray data.

Moreover, Figures 13–15 present qualitative CT, MRI and X-ray examples of VQA results using the VQA-RAD dataset, respectively. In the figure, the red arrow highlights a lesion or anatomical structure described in the imaging report, red text indicates incorrect or inconsistent responses, while blue text represents accurate and contextually appropriate answers.

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

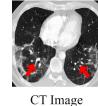
Synthetic MRI Text

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

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 .



CT Image

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

Real Text

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

Synthetic MRI Text

Chest X-ray shows bilateral patchy opacities in the lower lung zones, with possible peribronchial thickening. Bronchiectasis 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 bronchiectasis .

Figure 1: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

Figure 3: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.



CT Image

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

Real Text

Synthetic MRI Text

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 .



CT Image

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

Real Text

Synthetic MRI Text

Synthetic X-ray Text

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

Real Text

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**.

Synthetic MRI Text

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

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 2: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

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

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.	Postoperative head CT reveals post-surgical changes in the right frontotemporoparietal region with persistent irregular density near the right lateral ventricular wall.	Skull radiograph is limited, but postoperative changes may be visible over the right frontotemporoparietal region; ventricular involvement cannot be assessed.
Real Text	Synthetic CT Text	Synthetic X-ray Text	
User	What region of the brain is affected?		
IOS	The affected region in the brain is the left centrottemporal 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 5: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

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.	Non-contrast head CT shows subtle hypodensities in the occipital and posterior parietal subcortical white matter.	Skull X-ray is not suitable for evaluating white matter changes; findings suggestive of PRES are not visualizable .
Real Text	Synthetic CT Text	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 6: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

MRI Image	Enhanced T1 axial MRI demonstrates substantial interval enlargement of the left middle cerebral artery aneurysm, currently measuring MCA to 9.3 \times 0.60mm. MCA, middle cerebral artery; MRI, magnetic resonance imaging.	Contrast-enhanced cranial CT shows interval increase in the size of the left middle cerebral artery aneurysm, currently measuring approximately 9.3 mm in diameter.	Conventional skull radiograph is limited in evaluating intracranial vasculature, but may suggest calcified contours in the region of the left MCA , possibly corresponding to an aneurysmal dilation.
Real Text	Synthetic CT Text	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 left frontal arachnoid cyst .		
FedTGP	The image reveals calcification in the left middle cerebral artery instead of aneurysmal changes .		
SynTeX-FL (w/o Image)	The main finding in the image is the chronic stability of an aneurysm in the left MCA without interval changes		
SynTeX-FL	The main finding is a marked interval growth of an aneurysm in the left MCA .		
GPT-4	The main finding in the image is the substantial interval enlargement of an aneurysm in the left middle cerebral artery (MCA) .		

Figure 7: Example comparison of SynTeX-FL and other FL methods on LLaVA-Med dataset. The GPT-4 is considered as the ground truth.

MRI Image	Axial T2-weighted brain MRI scan shows a solitary metastatic lesion in the right temporal lobe with surrounding hypodense areas consistent with vasogenic edema, suggestive of metastatic disease.	Non-contrast head CT demonstrates a solitary hyperdense lesion in the right temporal lobe with surrounding hypodense areas consistent with vasogenic edema, suggestive of metastatic disease.	Skull radiograph is not diagnostic for intracranial metastasis; no definitive abnormality is visualized in the right temporal region .
Real Text	Synthetic CT Text	Synthetic X-ray Text	
User	Is there any sign of edema?		
IOS	According to the image, there is no sign of edema .		
FedTGP	The MRI scan indicates brain edema surrounding a primary tumor in the right temporal lobe .		
SynTeX-FL (w/o Image)	The main finding in the image is the chronic stability of an aneurysm in the left MCA without interval changes		
SynTeX-FL	The MRI scan shows brain edema surrounding a metastatic lesion located in the right temporal lobe .		
GPT-4	Yes, the MRI scan indicates the presence of brain edema surrounding the metastatic lesion in the right temporal lobe .		

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



X-ray Image

Chest x-ray showing prominent **left pulmonary artery** with congested hyperinflated left lung and shifting of the mediastinum to the right side.

Real Text

MRI of the thorax demonstrates an enlarged **left pulmonary artery** with signs of vascular engorgement and overinflation in the left lung parenchyma.

Synthetic MRI Text

CT of the chest reveals a prominent **left pulmonary artery**, associated with hyperinflation and vascular congestion of the left lung.

Synthetic CT Text

User What does the chest x-ray show about the left pulmonary artery?

IOS	There is no clear visualization of the left pulmonary artery in the chest x-ray.
FedTGP	The chest x-ray indicates a collapsed left pulmonary artery , suggesting decreased blood flow .
SynTeX-FL (w/o Image)	The chest x-ray shows a normal-sized left pulmonary artery with no abnormalities .
SynTeX-FL	The chest x-ray reveals that the left pulmonary artery is more pronounced than normal, indicating increased visibility or enlargement .
GPT-4	The chest x-ray shows a prominent left pulmonary artery , which means that it appears larger or more visible than usual .

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



X-ray Image

Chest x-ray notable for new findings of **spontaneous pneumomediastinum, subcutaneous emphysema, and bilateral patchy airspace disease** compared to chest x-ray on initial presentation.

Real Text

Thoracic MRI demonstrates evidence of newly developed **pneumomediastinum and soft tissue emphysema**, which were not observed in the prior imaging.

Synthetic MRI Text

Chest CT reveals new-onset **spontaneous pneumomediastinum and subcutaneous emphysema**, which were not present on the initial scan.

Synthetic CT Text

User What are the new findings in the chest x-ray?

IOS	Only bilateral patchy airspace disease is observed; there are no signs of spontaneous pneumomediastinum or subcutaneous emphysema .
FedTGP	The imaging indicates unilateral consolidation confined to the right lung .
SynTeX-FL (w/o Image)	The chest x-ray demonstrates subcutaneous emphysema and bilateral patchy airspace disease due to trauma, rather than spontaneous pneumomediastinum .
SynTeX-FL	New findings on the chest x-ray include spontaneous pneumomediastinum, subcutaneous emphysema, and bilateral patchy airspace disease , suggesting multifocal lung involvement and soft tissue air leakage.
GPT-4	The new findings in the chest x-ray include spontaneous pneumomediastinum, subcutaneous emphysema, and bilateral patchy airspace disease .

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



X-ray Image

Posterior anterior chest X-ray showing the right upper lobe **fibroavitory lesion and left upper lobe fibrosis**.

Real Text

Synthetic MRI Text

Synthetic CT Text

User What does the chest x-ray show about the left pulmonary artery?

IOS	There is extensive fibrosis in the lower lobes of both lungs without any cavitory lesion .
FedTGP	The chest X-ray indicates a fibroavitory lesion in the left upper lobe and fibrosis in the right lower lobe .
SynTeX-FL	The imaging shows a fibroavitory lesion in both lungs but no evidence of fibrosis .
GPT-4	The chest X-ray reveals a fibroavitory lesion located in the right upper lobe and fibrotic changes in the left upper lobe of the lungs .

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



X-ray Image

Thoracic MRI demonstrates bilateral areas of increased signal intensity **within the lung parenchyma**, with no evidence of cardiomegaly, suggestive of **non-cardiogenic pulmonary edema**.

Real Text

Synthetic MRI Text

Synthetic CT Text

User What do the infiltrates look like?

IOS	The image reveals clear lung fields without any signs of bilateral fluffy opacities or pulmonary edema .
FedTGP	The chest X-ray shows unilateral fluffy infiltrates , likely indicating localized infection rather than non-cardiogenic pulmonary edema .
SynTeX-FL	The chest X-ray demonstrates bilateral fluffy opacities due to cardiogenic pulmonary edema .
SynTeX-FL	The chest X-ray shows bilateral fluffy opacities, representing infiltrates consistent with non-cardiogenic pulmonary edema affecting both lungs .
GPT-4	The infiltrates in the chest X-ray appear as bilateral fluffy opacities . These fluffy infiltrates are visible in both lungs and suggest the presence of non-cardiogenic pulmonary edema .

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



Real CT Image

This is a **noncontrast CT**. This image is taken in axial. The finding is located at **right convexity**.

Real Text

Synthetic MRI Text

Synthetic X-ray Text

User Is this a noncontrast CT?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	Yes	Yes	Yes	Yes

User Where is the abnormality located?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
Right convexity	Left convexity	Right convexity	Right convexity	Right convexity

User Is a noncontrast CT the first imaging test for a suspected brain bleed?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	No	No	Yes	Yes

Figure 13: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.



The MRI image is the presence of blunting of the sulci and brain edema.

Real MRI Image Real Text Synthetic CT Text Synthetic X-ray Text

User	Is the brain swollen?				
IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth	
No	Yes	Yes	Yes	Yes	Yes

User Are the sulci blunted?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	Yes	Yes	Yes	Yes

User Is/Are there edema in the patient's brain?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	No	No	Yes	Yes

Figure 14: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.



In this case, the patient presents with a known diagnosis of small cell lung carcinoma.

Real X-ray Image Real Text Synthetic MRI Text Synthetic X-ray Text

User	Was the patient positioned appropriately without tilting?				
IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth	
No	Yes	Yes	Yes	Yes	Yes

User Are the sulci blunted?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	Yes	Yes	Yes	Yes

User Is/Are there edema in the patient's brain?

IOS	FedTGP	SynTeX-FL (w/o Image)	SynTeX-FL	Ground Truth
No	No	No	Yes	Yes

Figure 15: Example comparison of SynTeX-FL and other FL methods on VQA-RAD.