Frequency-Domain Dual-Branch Fusion for Medical Visual Question Answering

Yusra Tariq, Rakesh Chandra Joshi

Department of Artificial Intelligence, Amity University, Noida, India

arXiv:2608.08307v1 [cs.CV] 8 Aug 2026

Abstract

Medical Visual Question Answering (VQA) requires aligning subtle visual evidence, including lesion texture, boundary sharpness, and diffuse density changes, with clinical language. Existing multimodal fusion approaches operating in the spatial domain may not fully exploit complementary frequency information present in visual and textual representations. We introduce a dual-branch frequency-domain fusion module that conditions spectral filtering on the input question, enabling adaptive selection of global low-frequency structure and fine-grained high-frequency detail before reconstructing the spatial representation for answer generation. To provide a richer spectrum for filtering, we extract complementary features from early texture-sensitive and final semantic layers of a frozen BiomedCLIP encoder and align both with the question representation using a symmetric InfoNCE objective prior to staged joint training with a BioBART decoder. We pretrain the proposed model on PMC-VQA and fine-tune it on the VQA-RAD and SLAKE benchmarks, demonstrating that frequency-aware multimodal fusion improves medical VQA performance while maintaining a lightweight and efficient architecture.

Introduction

Medical Visual Question Answering (Med-VQA) requires a model to answer clinical questions by grounding them in fine grained visual evidence from medical images. Unlike general visual question answering, where object recognition is often sufficient, Med-VQA demands the interpretation of subtle anatomical structures, lesion boundaries, tissue texture, and diffuse pathological changes. Consequently, the mechanism that fuses visual and textual representations plays a central role in determining the quality of multimodal reasoning. A fusion strategy that treats all spatial regions and feature scales uniformly may fail to distinguish subtle textural abnormalities from larger structural patterns that are equally important for clinical decision making.

Recent Med-VQA approaches primarily improve performance by combining biomedical vision-language pretraining with spatial-domain cross-attention fusion, where question tokens attend over image patch embeddings. While effective at aligning semantic content, this paradigm does not separate global structure from fine-grained texture before fusion, leaving both entangled in the same representation, and incurs quadratic computational cost as token sequences grow.

Frequency-domain learning offers an alternative to spatial fusion: transforming representations via the Fourier transform decomposes them into globally-aggregated low-frequency structure and high-frequency texture. Global Filter Networks showed that learnable filtering in the frequency domain is an efficient alternative to spatial token mixing for visual representation learning (Rao et al. 2021), and FSRU extended this to multimodal rumor detection via unimodal spectrum compression and cross-modal spectrum co-selection (Lao et al. 2024). However, frequency-domain fusion remains unexplored for Medical VQA, where fine-grained image features must be integrated with clinically meaningful language representations.

Motivated by these observations, we propose a dual branch frequency domain fusion framework for Medical Visual Question Answering. Rather than directly fusing spatial representations, our framework transforms visual and textual features into the frequency domain, performs question guided spectral filtering and cross modal interaction, and reconstructs enhanced spatial representations for answer generation. The image branch operates on two dimensional Fourier representations extracted from a frozen BiomedCLIP vision encoder, while the text branch performs one dimensional Fourier analysis over contextual question embeddings. A bank of learnable spectral filters performs unimodal spectrum compression, followed by a cross modal gating mechanism that selectively emphasizes or suppresses frequency components according to complementary information from the other modality. This design enables the model to adaptively balance global structural information and fine grained texture based on the clinical question being answered.

We pretrain the proposed model on PMC-VQA (Zhang et al. 2023b) and evaluate it on the VQA-RAD and SLAKE benchmarks using Exact Match and Token F1 for both open-ended and closed-ended questions. The experimental results demonstrate the effectiveness of the proposed frequency-domain fusion strategy for multimodal medical reasoning.

Our contributions are threefold.

- We propose a dual branch question conditioned frequency domain fusion module for Medical Visual Question Answering that performs spectral filtering and cross modal gating before reconstructing enhanced spatial representa-

Copyright © 2027, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tions.

- We introduce a dual depth visual representation strategy that combines early texture sensitive features with late semantic features from a frozen BiomedCLIP encoder and aligns them with question representations through a contrastive pretraining objective.
- We demonstrate the effectiveness of the proposed frequency-domain fusion framework through extensive experiments on the VQA-RAD (Lau et al. 2018) and SLAKE (Liu et al. 2021) benchmarks after pretraining on PMC-VQA.

Related Work

Early Med-VQA methods combined CNN visual encoders with recurrent question encoders, fusing modalities via concatenation or bilinear pooling; these established task feasibility but were limited by local receptive fields and scarce annotated medical data. Recent approaches instead pair biomedical vision-language pretraining, such as BiomedCLIP (Zhang et al. 2023a), with spatial-domain cross-attention fusion, improving semantic alignment but entangling global structure and local texture within the same fused representation, at quadratic cost in token count.

Frequency-domain learning offers an efficient alternative to spatial or sequential token mixing: GFNet (Rao et al. 2021) and FNet (Lee-Thorp et al. 2022) replace spatial self-attention with learnable filtering in the 2D and 1D Fourier domains, respectively, for unimodal vision and language representation. FSRU (Lao et al. 2024) extends this to multimodal fusion via unimodal spectrum compression and cross-modal spectrum co-selection, though for rumor detection rather than question-conditioned reasoning. Frequency-domain fusion for Med-VQA remains unexplored: existing methods rely on spatial-domain cross-attention (Zhang et al. 2023a), motivating the question-conditioned frequency-domain fusion approach developed here.

Problem Definition

We formulate Medical Visual Question Answering (Med-VQA) as a multimodal question answering task. Let $\mathcal{D} = \{(v_i, q_i, a_i, t_i)\}_{i=1}^{N}$ denote a Med-VQA dataset of N samples, where v_i is a medical image, q_i is the associated clinical question represented as a token sequence, a_i is the corresponding ground-truth answer, also represented as a token sequence, drawn from a shared vocabulary $\mathcal{V}$ with q_i, and $t_i \in \{\text{yes/no}, \text{open}, \text{other}\}$ is the answer-type label associated with the sample. Given an image-question pair (v, q), our goal is to learn a parametric mapping

$$f_\theta : (v, q) \longmapsto \hat{a}, \tag{1}$$

where θ denotes the learnable parameters of the model and $\hat{a}$ is the predicted answer sequence generated by jointly reasoning over the visual content of v and the linguistic content of q. The objective is to learn the parameters θ such that the predicted answer $\hat{a} = f_\theta(v, q)$ closely matches the ground-truth answer a, enabling clinically accurate and contextually relevant responses to medical image-question pairs.

During training, the model additionally predicts the answer type $\hat{t}$ from a shared encoder representation, so that Equation (1) is optimized jointly with an auxiliary classification objective over t. This auxiliary objective encourages the encoder to represent coarse answer-type structure alongside fine-grained answer content, and is used only as a training-time regularizer. At inference time, only $\hat{a} = f_\theta(v, q)$ is used for evaluation against the ground-truth answer a.

Methodology

Overall Framework

Given an image-question pair (v, q), our model predicts the answer $\hat{a}$ through four stages: (1) a frozen vision transformer extracts patch-level visual features at two depths, capturing low-level spatial detail and high-level semantic content respectively; (2) a pretrained sequence-to-sequence text encoder produces contextual question representations; (3) a dual-branch frequency-domain fusion module filters the low-level visual and question representations in the spectral domain and exchanges information between them via learned cross-modal gates; and (4) the frequency-enhanced representations are additively fused with the late-semantic visual stream and the original textual stream, and passed to a sequence decoder that autoregressively generates the answer. An auxiliary classification head jointly predicts the coarse answer category (yes/no, open, other), providing an additional training signal without altering the generation pathway. Figure 1 illustrates the overall pipeline.

Visual and Question Feature Extraction

Dual-depth visual encoding. A frozen Vision Transformer Φ (BiomedCLIP (Zhang et al. 2023a)), pretrained on biomedical image-text pairs, decomposes an image v into $N = 196$ patch tokens (14×14 grid). We extract hidden states at two depths: an *early* intermediate block, $\mathbf{T}_e \in \mathbb{R}^{N \times d}$, retaining local texture and edge information relevant to radiological findings, and the *final* block, $\mathbf{T}_s \in \mathbb{R}^{N \times d}$, carrying globally contextualized semantics. Both are linearly projected into the shared model dimension D:

$$\mathbf{V}_e = \mathbf{T}_e \mathbf{W}_e + \mathbf{b}_e, \qquad \mathbf{V}_s = \mathbf{T}_s \mathbf{W}_s + \mathbf{b}_s. \tag{2}$$

$\mathbf{V}_e$ feeds the frequency-domain fusion module (Section 3.3); $\mathbf{V}_s$ is later fused additively (Section 3.4). Φ remains frozen throughout training; only $\mathbf{W}_e$ and $\mathbf{W}_s$ are learned.

Question encoding. The question q is encoded by the encoder of BioBART (Yuan et al. 2022), a pretrained biomedical sequence-to-sequence language model,

$$\mathbf{H}_q = \text{Encoder}(q) \in \mathbb{R}^{L_q \times D}, \tag{3}$$

and linearly projected to $\mathbf{X}_t = \mathbf{H}_q \mathbf{W}_t + \mathbf{b}_t \in \mathbb{R}^{L_q \times D}$, the representation consumed by the fusion module. Unlike Φ, the question encoder is fine-tuned jointly in later training stages.

Contrastive feature alignment (pretraining). Prior to joint training, the early- and late-depth visual projections are aligned with the question representation using a symmetric InfoNCE objective (van den Oord, Li, and Vinyals 2018),

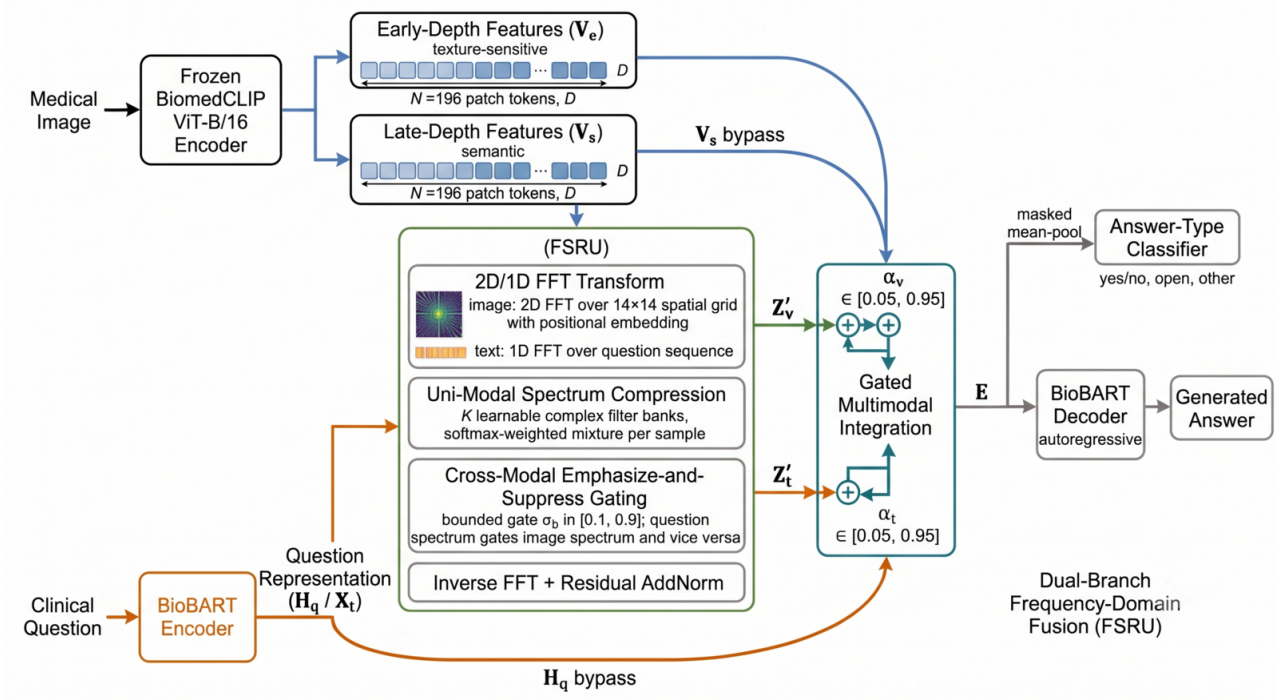

Figure 1: Overall architecture. A frozen BiomedCLIP encoder extracts early- and late-depth visual features ($\mathbf{V}_e$, $\mathbf{V}_s$); a BioBART encoder produces the question representation ($\mathbf{H}_q$). $\mathbf{V}_e$ and the question representation are processed by the dual-branch frequency-domain fusion module (FSRU), while $\mathbf{V}_s$ and $\mathbf{H}_q$ bypass FSRU directly. Gated multimodal integration combines both paths into the fused representation $\mathbf{E}$, which feeds an auxiliary answer-type classifier and the BioBART decoder for answer generation.

following the temperature-scaled contrastive formulation of CLIP (Radford et al. 2021). Mean-pooled, ℓ_2-normalized projections of $\mathbf{V}_e$ and $\mathbf{V}_s$ are each contrasted against the pooled question representation $\mathbf{X}_t$ across the batch, with a learnable temperature (logit scale) shared across both depths. During this stage, only the projection layers ($\mathbf{W}_e$, $\mathbf{W}_s$, $\mathbf{W}_t$) and the temperature parameter are updated; all other parameters, including the frozen vision backbone and the question encoder itself, remain fixed. This warmup stage ensures both visual depths enter the frequency-domain fusion module (Section 3.3) already loosely aligned with the question's semantic space, before any generation-based supervision is introduced.

Dual-Branch Frequency-Domain Fusion

The core of our approach filters and cross-modulates $\mathbf{V}_e$ and $\mathbf{X}_t$ in the frequency domain, treating the image and text streams as independent spectral branches connected only through learned cross-modal gates.

Spectral transform. The early-visual tokens are reshaped to their native 14×14 grid and combined with a learned positional embedding $\mathbf{P} \in \mathbb{R}^{14\times14\times D}$, compensating for the translation equivariance of Fourier-domain filtering, which would otherwise treat the spatial grid as circular (toroidal).

$$\mathbf{G} = \mathrm{reshape}(\mathbf{V}_e) + \mathbf{P}. \tag{4}$$

A two-dimensional real FFT is applied over the spatial dimensions and a one-dimensional real FFT over the question sequence:

$$\mathbf{F}_v = \mathcal{F}_{2D}(\mathbf{G}) \in \mathbb{C}^{14\times8\times D}, \tag{5}$$

$$\mathbf{F}_t = \mathcal{F}_{1D}(\mathbf{X}_t) \in \mathbb{C}^{L'_t\times D}, \tag{6}$$

where $L'_t = \lfloor L_q/2 \rfloor + 1$.

Uni-modal spectrum compression. Each branch is filtered by a bank of K learnable complex filters, combined with sample-dependent mixture weights derived from the pooled magnitude spectrum:

$$w^v = \mathrm{softmax}\big(g_v\big(\tfrac{1}{HW}\textstyle\sum|\mathbf{F}_v|\big)\big)\,, \tag{7}$$

$$\tilde{\mathbf{F}}_v = \sum_{k=1}^{K} w_k^v\,(\mathbf{F}_v \odot \mathbf{\Phi}_k^v)\,, \tag{8}$$

with $\mathbf{\Phi}_k^v$ learnable complex kernels and $\odot$ elementwise complex multiplication. The text branch is compressed analogously with its own filter bank and scoring head, yielding $\tilde{\mathbf{F}}_t$. Conditioning the mixture on the input spectrum, rather than fixing it, allows the effective filter to adapt per sample.

Cross-modal emphasize-and-suppress. Each branch's compressed spectrum modulates the other. A source spectrum is pooled via learned-query attention, normalized, and

mapped through channel- and band/position-specific linear heads to a bounded gate,

$$\sigma_b(x) = 0.1 + 0.8 \cdot \mathrm{sigmoid}(x), \tag{9}$$

which keeps gate values in $[0.1, 0.9]$ so that a branch's frequency content can be strongly attenuated but never fully discarded or fully passed through. The question spectrum produces gate $\boldsymbol{\gamma}^v$, applied to the image spectrum; the image spectrum produces gate $\boldsymbol{\gamma}^t$, applied to the text spectrum:

$$\hat{\mathbf{F}}_v = \tilde{\mathbf{F}}_v \odot \boldsymbol{\gamma}^v(\tilde{\mathbf{F}}_t), \qquad \hat{\mathbf{F}}_t = \tilde{\mathbf{F}}_t \odot \boldsymbol{\gamma}^t(\tilde{\mathbf{F}}_v). \tag{10}$$

Inverse transform and fusion. The gated spectra are mapped back via inverse FFTs and combined with their pre-transform inputs through a residual feed-forward sub-layer (pre-norm, with an internal residual connection, denoted $\mathrm{AddNorm}$):

$$\mathbf{O}_v = \mathrm{AddNorm}\Big(\mathcal{F}_{2D}^{-1}(\hat{\mathbf{F}}_v) + \mathbf{V}_e\Big), \tag{11}$$

$$\mathbf{O}_t = \mathrm{AddNorm}\Big(\mathcal{F}_{1D}^{-1}(\hat{\mathbf{F}}_t) + \mathbf{X}_t\Big). \tag{12}$$

The branches are concatenated, projected, and normalized,

$$\mathbf{Z} = \mathrm{LayerNorm}(\mathrm{concat}(\mathbf{O}_v, \mathbf{O}_t)\,\mathbf{W}_o + \mathbf{b}_o), \tag{13}$$

then mapped back into the language model's representation space, $\mathbf{Z}' = \mathrm{LayerNorm}(\mathbf{Z}\mathbf{W}_z + \mathbf{b}_z)$, and split into $\mathbf{Z}'_v \in \mathbb{R}^{N\times D}$ and $\mathbf{Z}'_t \in \mathbb{R}^{L_q\times D}$.

Gated Multimodal Integration

The frequency-processed streams are additively blended with the original visual and textual representations rather than replacing them, using two learnable scalar gates bounded to $[0.05, 0.95]$ by the same σ_b function-style function:

$$\bar{\mathbf{V}} = \mathrm{LayerNorm}(\mathbf{V}_s + \alpha_v\,\mathbf{Z}'_v), \tag{14}$$

$$\bar{\mathbf{H}} = \mathrm{LayerNorm}(\mathbf{H}_q + \alpha_t\,\mathbf{Z}'_t), \tag{15}$$

where α_v, α_t are learned scalars, ensuring the spectral contribution is neither negligible nor dominant. A learnable source-type embedding is added to each stream so the decoder can distinguish visual from textual tokens:

$$\mathbf{V}^* = \bar{\mathbf{V}} + \mathbf{e}_v, \qquad \mathbf{H}^* = \bar{\mathbf{H}} + \mathbf{e}_t. \tag{16}$$

The two streams are concatenated into a single fused encoder representation, with the attention mask formed by concatenating an all-ones visual mask with the question's padding mask:

$$\mathbf{E} = \mathrm{concat}(\mathbf{V}^*, \mathbf{H}^*) \in \mathbb{R}^{(N+L_q)\times D}. \tag{17}$$

Answer Generation and Auxiliary Objectives

Generation. $\mathbf{E}$ substitutes for the encoder output of a pre-trained sequence-to-sequence decoder, which attends over $\mathbf{E}$ and autoregressively generates the answer:

$$\hat{a} = \arg\max_a \prod_{j=1}^{L_a} p_\theta\big(a^j \mid a^{<j}, \mathbf{E}\big). \tag{18}$$

The decoder is supervised with teacher forcing using a label-smoothed cross-entropy loss $\mathcal{L}_{\mathrm{gen}}$ over non-padding positions.

Answer-type classification. A masked mean pooling of $\mathbf{E}$ is passed through a linear head to predict the coarse answer type $y \in \{\text{yes/no}, \text{open}, \text{other}\}$:

$$\hat{y} = \mathrm{softmax}(\bar{\mathbf{e}}\,\mathbf{W}_y + \mathbf{b}_y), \qquad \bar{\mathbf{e}} = \frac{\sum_i m_i \mathbf{E}_i}{\sum_i m_i}, \tag{19}$$

with m_i the attention mask at position i. This head is trained with a class-balanced focal loss (Lin et al. 2017) $\mathcal{L}_{\mathrm{type}}$ to counteract answer-category imbalance.

Representation-level auxiliary losses. Two auxiliary terms encourage the fused branches to carry answer-relevant, sample-specific information. A contrastive term $\mathcal{L}_{\mathrm{contrast}}$ pulls the correct (mean-pooled, ℓ_2-normalized) image and text branch representations closer to a contextual encoding of the ground-truth answer than mismatched in-batch counterparts via a margin-based hinge, and additionally penalizes correct-pair similarity that falls below a fixed target, averaged over both branches. A diversity term $\mathcal{L}_{\mathrm{div}}$ penalizes high average pairwise cosine similarity between different samples' visual branch outputs within a batch, discouraging representational collapse.

Overall objective. The full training loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{type}}\,\mathcal{L}_{\mathrm{type}} + \lambda_{\mathrm{contrast}}\,\mathcal{L}_{\mathrm{contrast}} + \lambda_{\mathrm{div}}\,\mathcal{L}_{\mathrm{div}}, \tag{20}$$

with $\lambda_{\mathrm{type}}, \lambda_{\mathrm{contrast}}, \lambda_{\mathrm{div}} \geq 0$ non-negative weights, held fixed within a stage but reweighted at each stage transition to reflect that stage's training priorities; λ_{div} is held constant throughout.

Experiments

Experimental Setup

Datasets. The proposed model is pretrained on PMC-VQA (Zhang et al. 2023b) and subsequently fine-tuned on two publicly available medical VQA benchmarks: VQA-RAD and SLAKE.VQA-RAD contains 1,793 official training QA pairs, from which we carve a 269-sample validation set via an image-level 85/15 split (ensuring no image appears in both splits), leaving 1,524 samples for training; the official 451-sample test set is used unmodified for evaluationSLAKE consists of 4,165 training, 1,053 validation, and 1,061 testing samples after restricting to the English-language subset, which we use throughout for consistency with the evaluation protocol. Following standard practice, all input images are resized to 224×224.

Implementation Details. Our framework employs BiomedCLIP (Zhang et al. 2023a) as the frozen visual encoder and BioBART-v2 (Yuan et al. 2022) as the text encoder-decoder. The model is optimized using AdamW (Loshchilov and Hutter 2019) with mixed-precision training and a linear-warmup, cosine-decay learning-rate schedule. Pretraining is performed on PMC-VQA for 20 epochs, followed by fine-tuning on VQA-RAD for 10 epochs and SLAKE for 15 epochs, using a base learning rate of 5×10^{-6} (discriminatively reduced to $0.1\times$ for the pretrained text encoder and, once unfrozen, the vision backbone; annealed to 1×10^{-6} in the final fine-tuning

stage). All experiments were conducted on a Google Colab G4 GPU (NVIDIA RTX PRO 6000 Blackwell Server Edition) using PyTorch 2.0 with AMP.

Evaluation Metrics. Following prior medical VQA studies, performance is evaluated using Exact Match (EM) and Token-level F1. Results are reported on the official test sets, with separate evaluation on open-ended and closed-ended question categories whenever applicable.

Baselines

We compare against established Med-VQA methods evaluated under the same Exact Match protocol on VQA-RAD and SLAKE: M3AE (Chen et al. 2022), M2I2 (Li et al. 2023b) (262.15M params, closest in scale to our 260M model), MUMC (Li et al. 2023a) (211.06M params), and PeFoMed (He et al. 2024). All baseline scores are taken directly from their original publications. Our results are averaged over 3 seeds (42, 123, 2024) on the official test splits.

Results and Analysis

Table 1 presents the quantitative comparison of the proposed method with representative medical VQA approaches on the VQA-RAD and SLAKE benchmarks. Our model achieves an overall Exact Match (EM) score of 40.21% on VQA-RAD and 69.21% on SLAKE, with standard deviations of 0.56% and 1.10%, respectively, across three random seeds. The relatively small variance indicates that the proposed training strategy produces stable and reproducible results. In addition, the proposed model achieves Token F1 scores of 44.01% on VQA-RAD (Open 28.4, Closed 56.4) and 72.90% on SLAKE (Open 74.7, Closed 70.7), suggesting that many generated answers are semantically relevant even when they do not exactly match the reference answer.

Compared with recent Med-VQA methods, including M3AE, M2I2, MUMC, and PeFoMed, the proposed framework achieves lower overall benchmark performance. These methods benefit from stronger multimodal pretraining strategies and specialized architectures designed specifically for medical visual question answering. Nevertheless, our objective is not to replace these systems directly, but rather to investigate whether frequency-domain multimodal fusion can improve generative medical VQA. Despite its relatively lightweight architecture based on BiomedCLIP and BioBART, the proposed model demonstrates competitive performance on the larger SLAKE benchmark and provides a practical foundation for evaluating the effectiveness of spectral feature fusion.

A clear difference is observed between the two datasets. The proposed method performs considerably better on SLAKE than on VQA-RAD, particularly for open-ended questions. A possible contributing factor is the substantially larger training set provided by SLAKE, which offers more supervision for learning robust multimodal representations. In contrast, VQA-RAD contains fewer training examples and a wider variety of answer expressions, making exact-match evaluation substantially more challenging for generative models. This trend is also reflected in the Token F1 scores, which consistently exceed Exact Match, indicating that the generated responses often contain partially correct medical information despite failing strict string-level matching.

Ablation Study

To evaluate the contribution of the proposed Frequency Spectrum Representation and Fusion Unit (FSRU), we compare the complete Q-FSRU framework with a variant in which the FSRU module is removed while keeping all other components and the training protocol unchanged. The results are summarized in Table 2.

The proposed model consistently outperforms the variant without the FSRU module on both VQA-RAD and SLAKE. Specifically, incorporating the proposed frequency-domain fusion module improves Exact Match (EM) from 36.66% to 40.21% on VQA-RAD and from 66.26% to 69.21% on SLAKE. These improvements indicate that the FSRU module enhances multimodal feature fusion and contributes to more accurate answer generation.

The consistent performance gains across both benchmarks demonstrate the effectiveness of the proposed frequency-domain fusion strategy. Although the proposed model does not achieve state-of-the-art performance, the ablation study shows that the FSRU module makes a positive contribution to the overall framework. These findings suggest that frequency-domain multimodal fusion is a promising direction for generative medical visual question answering and could be further improved through larger vision-language models and more extensive medical pretraining.

Qualitative Analysis

Figure 2 presents a t-SNE visualization of the image token embeddings before and after the proposed FSRU module on the SLAKE validation set. Before frequency-domain fusion, the embeddings are widely scattered with noticeable overlap between different question categories. After applying FSRU, the embeddings become more compact and better separated, indicating that the proposed frequency-spectrum fusion learns a more structured and discriminative feature representation. This improved feature organization is consistent with the quantitative performance gains achieved by the proposed model.

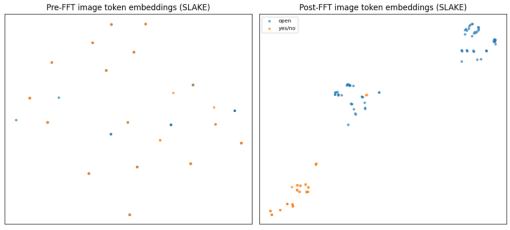

Figure 2: t-SNE (van der Maaten and Hinton 2008) visualization of the image token embeddings before and after the proposed FSRU module on the SLAKE validation set.

Table 1: Accuracy (Exact Match, %) on VQA-RAD and SLAKE test sets.

Method	VQA-RAD			SLAKE		
	Open	Closed	Overall	Open	Closed	Overall
M3AE	67.2	83.5	77.0	80.3	87.8	83.2
M2I2	66.5	83.5	76.8	74.7	91.1	81.2
MUMC	71.5	84.2	79.2	81.5	91.1	84.9
PeFoMed	62.6	87.1	77.4	77.8	88.7	82.1
Ours	19.8	56.4	40.2 ± 0.6	68.3	70.7	69.2 ± 1.1

Table 2: Ablation study on the FSRU module. Results are mean $\pm$ std over three seeds.

Configuration	VQA-RAD EM (%)	SLAKE EM (%)
w/o FSRU	36.66 ± 0.78	66.26 ± 0.09
Ours (Full Model)	$\mathbf{40.21 \pm 0.56}$	$\mathbf{69.21 \pm 1.10}$

Case study

Figure 3 shows representative predictions with and without FSRU. Incorporating FSRU correctly identifies anatomical structures, diseases, and abnormality locations that the ablated model misses, illustrating the benefit of frequency-domain fusion for clinical reasoning.

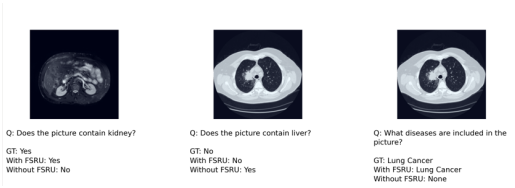

Figure 3: Representative case studies on the SLAKE validation set comparing the proposed model with and w/o FSRU. The proposed model correctly answers several medical VQA questions that are incorrectly or incompletely predicted by the ablated model, demonstrating the effectiveness of frequency-domain multimodal fusion.

Despite these gains, FSRU is not uniformly beneficial. Figure 4 shows representative cases where it degrades predictions relative to the w/o-FSRU baseline. On closed-ended organ-presence questions, FSRU occasionally flips a correct binary answer, and on abdominal CT slices containing several adjacent organs, it produces over-inclusive or malformed multi-token spans, such as naming two organs where only one was asked about or generating an incoherent locational phrase, whereas the baseline produces a single well-formed answer. These failures are concentrated in denser, multi-structure CT images, suggesting that cross-modal gating can occasionally over-emphasize competing high-frequency content when several anatomically similar structures share a spectral neighborhood, an effect the question-conditioning does not fully suppress.

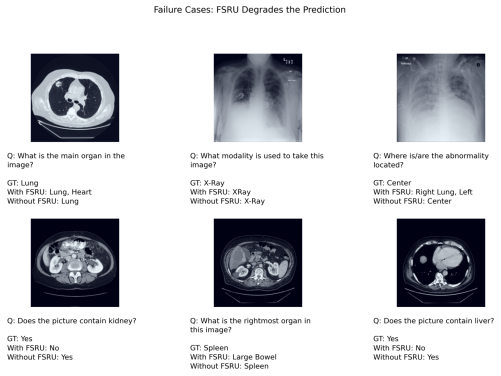

Figure 4: Representative failure cases where FSRU degrades predictions relative to the w/o-FSRU baseline on the SLAKE validation set. FSRU occasionally flips correct closed-ended (yes/no) organ-presence answers, and on abdominal CT slices with several adjacent organs, produces over-inclusive or malformed multi-token spans (e.g., predicting two organs, or an incoherent location phrase) where the baseline gives a single correct answer.

Conclusion

We presented a dual-branch frequency-domain fusion framework for medical VQA, combining BiomedCLIP and BioBART with a question-guided Frequency Spectral Representation Unit (FSRU). The ablation study confirms FSRU consistently improves answer prediction over the non-frequency baseline, though results do not surpass current state-of-the-art methods. Our failure case analysis further indicates that these gains are not uniform: FSRU is most beneficial on open-ended, texture-sensitive questions, but can occasionally degrade closed-ended organ-presence answers and produce over-inclusive multi-token spans on visually dense, multi-organ CT slices. Future work will explore larger vision-language backbones, additional imaging datasets, and gating mechanisms with tighter per-token control to mitigate these failure modes.

References

Chen, Z.; Du, Y.; Hu, J.; Liu, Y.; Li, G.; Wan, X.; and Chang, T.-H. 2022. Multi-Modal Masked Autoencoders for Medical Vision-and-Language Pre-Training. In *Medical Image Com-*

puting and Computer Assisted Intervention – MICCAI 2022, 679–689. Cham: Springer Nature Switzerland.

He, J.; Liu, G.; Li, P.; He, G.; Chen, Z.; and Zhong, S. 2024. PeFoMed: Parameter Efficient Fine-tuning of Multimodal Large Language Models for Medical Imaging. arXiv:2401.02797.

Lao, A.; Zhang, Q.; Shi, C.; Cao, L.; Yi, K.; Hu, L.; and Miao, D. 2024. Frequency Spectrum Is More Effective for Multimodal Representation and Fusion: A Multimodal Spectrum Rumor Detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18426–18434. AAAI Press.

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: 180251.

Lee-Thorp, J.; Ainslie, J.; Eckstein, I.; and Ontañón, S. 2022. FNet: Mixing Tokens with Fourier Transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4296–4313. Seattle, United States: Association for Computational Linguistics.

Li, P.; Liu, G.; He, J.; Zhao, Z.; and Zhong, S. 2023a. Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses for Medical Visual Question Answering. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 374–383. Cham: Springer Nature Switzerland.

Li, P.; Liu, G.; Tan, L.; Liao, J.; and Zhong, S. 2023b. Self-Supervised Vision-Language Pretraining for Medical Visual Question Answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global Filter Networks for Image Classification. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 980–993.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Yuan, H.; Yuan, Z.; Gan, R.; Zhang, J.; Xie, Y.; and Yu, S. 2022. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, 97–109. Dublin, Ireland: Association for Computational Linguistics.

Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Lungren, M.; Naumann, T.; and Poon, H. 2023a. BiomedCLIP: a Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. arXiv:2303.00915.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023b. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv:2305.10415.