Learning to Trim: End-to-End Causal Graph Pruning with Dynamic Anatomical Feature Banks for Medical VQA

Zibo Xu[1], Qiang Li[1], Weizhi Nie*[2], and Yuting Su[2]

[1] School of Microelectronics, Tianjin University, Tianjin 300072, China
[2] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
weizhinie@tju.edu.cn

Abstract. Medical Visual Question Answering (MedVQA) models often exhibit limited generalization due to reliance on dataset-specific correlations, such as recurring anatomical patterns or question-type regularities, rather than genuine diagnostic evidence. Existing causal approaches are typically implemented as static adjustments or post-hoc corrections. To address this issue, we propose a Learnable Causal Trimming (LCT) framework that integrates causal pruning into end-to-end optimization. We introduce a Dynamic Anatomical Feature Bank (DAFB), updated via a momentum mechanism, to capture global prototypes of frequent anatomical and linguistic patterns, serving as an approximation of dataset-level regularities. We further design a differentiable trimming module that estimates the dependency between instance-level representations and the global feature bank. Features highly correlated with global prototypes are softly suppressed, while instance-specific evidence is emphasized. This learnable mechanism encourages the model to prioritize causal signals over spurious correlations adaptively. Experiments on VQA-RAD, SLAKE, SLAKE-CP and PathVQA demonstrate that LCT consistently improves robustness and generalization over existing debiasing strategies.

Keywords: Medical VQA · Causal Inference · Causal Trimming · Dynamic Feature Bank.

1 Introduction

Medical Visual Question Answering (MedVQA) connects medical imaging with natural language to support clinical decision-making [1, 12, 25, 16]. Despite recent progress, existing models often suffer significant performance drops under Out-of-Distribution (OOD) settings. This issue largely arises from reliance on spurious correlations (e.g., associating question types with background textures) instead of true causal mechanisms [23].

* Corresponding author.

Recent approaches incorporate causal inference by modeling spurious correlations as confounders and applying backdoor adjustment with static dictionaries [10, 12, 18]. However, spurious correlations arising from variable anatomical features are inherently dynamic, rendering static dictionaries inadequate. Meanwhile, Test-Time Adaptation (TTA) methods attempt to prune biased features during inference [15], but they incur additional computational cost and treat the trained model as a black box, without improving intrinsic robustness. To address these limitations, we propose Learnable Causal Trimming (LCT), an end-to-end framework that integrates causal intervention directly into training. Instead of static confounder sets, we construct a Dynamic Anatomical Feature Bank (DAFB) updated via momentum to capture evolving anatomical priors and dataset biases. A Causal Trimming (CT) module operates as a soft gate, estimating the dependency between instance features and DAFB prototypes, and adaptively reweighting confounder-correlated components while preserving informative evidence. This learnable intervention encourages the model to disentangle causal signals from confounding noise during optimization.

To our knowledge, we are the first to embed dynamic causal trimming as a trainable module within MedVQA. Our contributions are threefold:

- We propose the LCT framework, transforming causal adjustment in MedVQA from static or test-time strategies into an end-to-end trainable paradigm.
- We design a Dynamic Anatomical Feature Bank with a Causal Trimming module to track and mitigate confounders during training without external annotations.
- Experiments on four datasets demonstrate that LCT achieves state-of-the-art performance and improves OOD generalization.

2 Methodology

2.1 Causal Formulation: The Confounding Problem

We analyze bias in MedVQA using a Structural Causal Model (SCM) [20], as shown in Fig. 1(c). I, Q, and A denote the medical image, question, and answer. Ideally, the model should estimate $P(A|I, Q)$ through the causal paths $I \rightarrow A$ and $Q \rightarrow A$ [24]. However, medical datasets contain strong priors. We introduce a latent confounder C to represent dataset-level regularities. The backdoor paths $I \leftarrow C \rightarrow A$ and $Q \leftarrow C \rightarrow A$ induce spurious correlations. For example, if C encodes lung-related priors, the model may predict "Pneumonia" from lung appearance alone, ignoring lesion-specific evidence in I. Standard supervised learning on the observational distribution may therefore capture shortcuts induced by C. To mitigate this issue, we simulate a causal intervention to reduce the confounding effect of C on answer prediction. As C is unobserved, we approximate it with a *Dynamic Anatomical Feature Bank* and apply *Causal Trimming* to reweight features associated with C during representation learning.

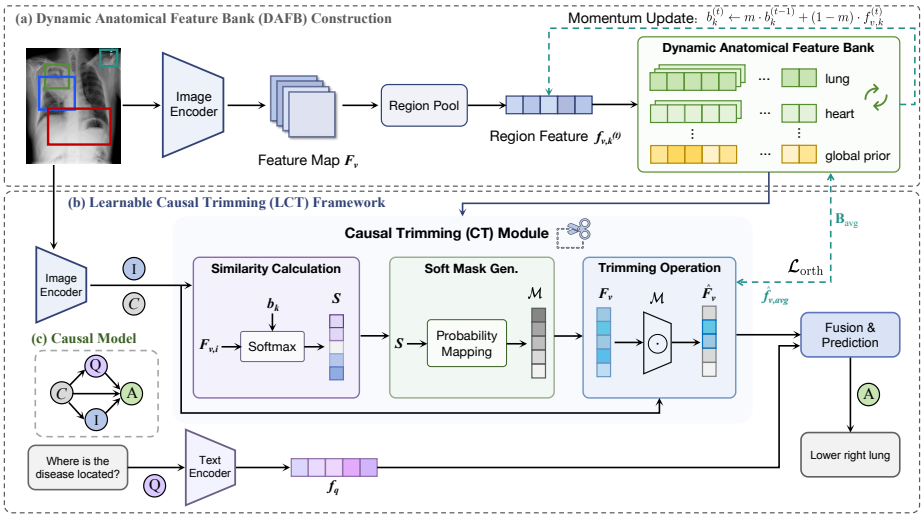

Fig. 1: The proposed Learnable Causal Trimming (LCT) Framework with Dynamic Anatomical Feature Bank.

2.2 Learnable Causal Trimming (LCT) Framework

Based on Sec. 2.1, the key to achieving robust MedVQA is executing the causal intervention by eliminating the confounding effect of C. To this end, we propose the Learnable Causal Trimming (LCT) framework. Unlike previous Test-Time Adaptation methods [15] that perform hard pruning on frozen models during inference, LCT internalizes the causal intervention mechanism into an end-to-end training pipeline. The overall architecture is depicted in Fig. 1.

The framework processes the inputs through three main stages:

Dual-Stream Encoding: The input medical image I and question Q are processed by an image encoder and a text encoder to extract the regional visual feature map $F_v \in \mathbb{R}^{N \times D}$ and the question embedding f_q, respectively.

Confounder Proxy Construction: A confounder bank is maintained via momentum updates to aggregate global, high-frequency anatomical priors across the dataset, serving as a proxy of the unobserved confounder C.

Causal Intervention via Trimming: The Causal Trimming (CT) Module acts as a soft gate. It evaluates the similarity between the instance-specific visual features F_v and the confounder prototypes in the DAFB, dynamically fading out spurious correlations to output the purified visual representation $\hat{F}_v$.

Finally, the trimmed visual features are fused with f_q for answer prediction, supervised by a joint loss function that enforces causal disentanglement.

2.3 Dynamic Anatomical Feature Bank (DAFB)

As defined in our SCM, the confounder C represents dataset-specific priors, such as the frequent co-occurrence of specific anatomical backgrounds and diagnostic

answers. Since C is latent and constantly evolving during training, we propose to explicitly model it using the **DAFB**.

Let $\mathbf{B} \in \mathbb{R}^{K \times D}$ denote the feature bank, where K is the bank size and D is the feature dimension. The bank $\mathbf{B} = \{b_1, b_2, \ldots, b_K\}$ stores global prototypes representing common visual patterns (i.e., potential confounders). Instead of using a static dictionary, we employ a **Momentum Update** mechanism to maintain the stability of $\mathbf{B}$, ensuring it captures the long-term distribution of the dataset rather than the noisy fluctuations of a single mini-batch. Specifically, at training iteration t, we update the specific prototypes in the bank using the region-level features from the current batch. Let $f_{v,k}^{(t)}$ denote the aggregated visual feature of the current batch associated with the k-th prototype. We update the individual prototype $b_k \in \mathbf{B}$ via a momentum mechanism:

$$b_k^{(t)} \leftarrow m \cdot b_k^{(t-1)} + (1 - m) \cdot f_{v,k}^{(t)}, \tag{1}$$

where $m \in [0, 1)$ is a momentum coefficient. A large m (e.g., 0.99) ensures that the bank evolves smoothly, effectively filtering out instance-specific details and retaining only the robust, high-frequency anatomical priors (e.g., generic lung shapes or background artifacts) that constitute the confounder C.

2.4 Causal Trimming (CT) Module

With the confounder proxy $\mathbf{B}$ constructed, the next critical step is to reduce the confounding effect of C on the final prediction. To achieve this, we design the **Causal Trimming (CT) Module**, which adaptively reweights regional features according to their correlations with the dataset priors $\mathbf{B}$. As shown in Fig. 1, this module operates in three sequential steps:

1. Similarity Calculation: Given the input visual feature map $F_v \in \mathbb{R}^{N \times D}$ (where N corresponds to the number of patch-level feature tokens extracted by the visual encoder), we first compute the Confounder Relevance Score matrix $\mathbf{S} \in \mathbb{R}^{N \times K}$ by measuring the similarity between each image region and the bank prototypes:

$$\mathbf{S}_{i,k} = \frac{\exp(F_{v,i} \cdot b_k^T / \tau)}{\sum_{j=1}^{K} \exp(F_{v,i} \cdot b_j^T / \tau)}, \tag{2}$$

where τ is a temperature hyperparameter. A high score $\mathbf{S}_{i,k}$ indicates that the i-th region strongly resembles a common prototype in the bank, implying it is likely associated with dataset-level priors.

2. Soft Mask Generation: To make the trimming process differentiable, we map the relevance scores into a Soft Trimming Mask $\mathcal{M} \in \mathbb{R}^{N \times 1}$. We aggregate the scores across all prototypes to determine the total confounding degree of each region:

$$\mathcal{M}_i = 1 - \sigma\left(\psi\left(\sum_{k=1}^{K} \mathbf{S}_{i,k} \cdot w_k\right)\right), \tag{3}$$

where $\sigma(\cdot)$ is the Sigmoid function, ψ is a learnable projection layer, and w_k are learnable weights for the prototypes. The resulting mask $\mathcal{M}_i$ controls the

residual gain of each region: it approaches 0 for regions highly correlated with confounder prototypes, reducing their additional amplification, and approaches 1 for unique, instance-specific diagnostic regions, enhancing their contribution.

3. Trimming Operation: Finally, the reweighted Trimmed Feature $\hat{F}_v$ is obtained via element-wise modulation combined with a residual connection:

$$\hat{F}_v = F_v \odot \mathcal{M} + F_v. \tag{4}$$

The residual connection is crucial here; it preserves gradient flow and retains the fundamental anatomical context, while the mask selectively emphasizes diagnostic evidence over confounder-correlated patterns.

2.5 Optimization Objectives

To ensure accurate answers and causally independent representations, LCT optimizes a joint objective comprising a BCE classification loss ($\mathcal{L}_{\text{vqa}}$) and an orthogonality loss ($\mathcal{L}_{\text{orth}}$):

$$\mathcal{L}_{\text{vqa}} = -\sum_{j=1}^{|\mathcal{A}|} y_j \log(P(a_j|\hat{F}_v, f_q)), \quad \mathcal{L}_{\text{orth}} = \left\| \frac{\hat{f}_{v,\text{avg}} \cdot \mathbf{B}_{\text{avg}}^T}{\|\hat{f}_{v,\text{avg}}\| \|\mathbf{B}_{\text{avg}}\|} \right\|^2, \tag{5}$$

where y is the ground truth. To explicitly enforce causal intervention, $\mathcal{L}_{\text{orth}}$ minimizes the cosine similarity between the mean vectors of the trimmed features ($\hat{f}_{v,\text{avg}}$) and the confounder bank ($\mathbf{B}_{\text{avg}}$), penalizing any residual correlation. The total objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vqa}} + \lambda \mathcal{L}_{\text{orth}}$, where λ balances accuracy and causal disentanglement.

3 Experiments

3.1 Datasets and Implementation Details

Datasets. VQA-RAD [6] contains 3,148 QA pairs on 315 radiology images. **SLAKE** [14] is a semantically-labeled bilingual dataset with 7,033 QA pairs on 642 radiology images. **PathVQA** [5] comprises 32,799 QA pairs on 4,998 pathology images, evaluating the model's generalizability across drastically different medical domains. Furthermore, to explicitly evaluate out-of-distribution (OOD) robustness, we utilize **SLAKE-CP** [25].

Implementation Details. Our framework is implemented in PyTorch. We leverage the pre-trained PMC-CLIP [11] for dual-stream encoding to extract initial text embeddings and patch-level visual features. For the causal intervention, the DAFB is initialized with $K = 256$ prototypes and updated with a momentum coefficient of m=0.99 to capture stable dataset priors. For the trimming module, the temperature scaling factor τ is set to 0.07. The entire network is trained end-to-end using the AdamW optimizer with a learning rate of $1e-4$ for 50 epochs. The balancing weight for the orthogonality loss, λ, is empirically set to 0.1.

Table 1: Comparison of accuracy (%) on VQA-RAD and SLAKE datasets. Best results are highlighted in **bold**.

Methods	VQA-RAD			SLAKE		
	Open	Closed	Overall	Open	Closed	Overall
MEVF+SAN [17]	49.2	73.9	64.1	75.3	78.4	76.5
MEVF+BAN [17]	49.2	77.2	66.1	77.8	79.8	78.6
CPRD+BAN [13]	52.5	77.9	67.8	79.5	83.4	81.1
M2I2 [9]	61.8	81.6	73.7	74.7	**91.1**	81.2
MISS [2]	**71.8**	80.4	76.1	81.5	82.9	82.0
M3AE [3]	67.2	83.5	77.0	80.3	87.8	83.3
CCIS-MVQA [1]	68.8	79.2	75.1	80.1	86.7	84.1
PMC-CLIP [11]	67.0	84.0	77.6	81.9	88.0	84.3
CLIPQCR [4]	58.0± 1.4	79.6± 1.1	71.1± 1.2	78.2± 1.3	82.6± 1.5	80.1± 1.3
DeBCF [25]	58.6± 1.1	80.9± 0.8	71.6± 1.0	80.8± 0.9	84.9± 0.7	82.6± 0.9
DeCoCT [23]	67.1± 0.5	85.7± 0.4	78.3± 0.5	82.5± 0.3	87.0± 0.6	84.9± 0.5
CIMB-MVQA [12]	69.3± 0.2	86.2± 0.2	79.4± 0.2	82.1± 0.1	89.4± 0.1	85.1± 0.2
LCT (Ours)	70.9± 0.6	**87.9**± 0.4	**81.2**± 0.4	**83.4**± 0.3	89.9± 0.5	**86.0**± 0.4

3.2 Comparison with State-of-the-Arts

Table 1 compares LCT with SOTA methods on VQA-RAD and SLAKE. LCT achieves the best Overall accuracy on both datasets, surpassing CIMB-MVQA by 1.8% and 0.9%, respectively. For **Closed-ended questions**, LCT obtains the highest accuracy on VQA-RAD (87.9%). Although M2I2 achieves 91.1% on SLAKE Closed, it drops notably on Open questions (74.7%). In contrast, LCT maintains balanced performance, achieving 89.9% on Closed and a leading 83.4% on Open questions. For **Open questions**, LCT reaches 70.9% on VQA-RAD and achieves SOTA on SLAKE Open. While slightly below MISS on VQA-RAD Open, LCT exceeds MISS by 5.1% in Overall accuracy, indicating stronger holistic reasoning performance.

To further assess cross-modality generalization, we evaluate on PathVQA (Table 2). LCT achieves the highest Overall accuracy of 65.6%, outperforming MUMC (65.1%). Notably, LCT establishes a new SOTA on Open-ended questions (41.1%), while achieving comparable Closed accuracy (89.9%). These results suggest that explicitly modeling dynamic confounders via causal trimming improves robustness, particularly for fine-grained pathological reasoning.

3.3 Out-of-Distribution Generalization on SLAKE-CP

Table 3 presents results on SLAKE-CP, a challenging OOD benchmark for evaluating diagnostic robustness. LCT achieves the best performance across all metrics, reaching 51.0% Overall, 28.9% on Open questions, and 71.6% on Closed questions. These results indicate that the proposed method improves the model's ability to mitigate dataset-specific biases under distribution shifts.

Under the OOD setting, conventional representation learning models exhibit substantial performance degradation. For example, MISS and M2I2 drop to

Table 2: Comparison with SOTA methods on PathVQA.

Methods	PathVQA		
	Open	Closed	Overall
MEVF-BAN [17]	8.1	81.4	44.8
MedGemma [21]	20.4	65.1	47.7
AMAM [19]	18.2	84.4	50.4
LLaVA-Med [7]	32.7	89.8	61.3
M2I2 [9]	36.3	88.0	62.2
CLIP-ViT [22]	40.0	87.0	63.6
MUMC [8]	39.0	**90.4**	65.1
LCT (Ours)	**41.1**	89.9	**65.6**

Table 3: Comparison with SOTA methods on SLAKE-CP.

Methods	SLAKE-CP		
	Open	Closed	Overall
MEVF+BAN [17]	$13.0_{\pm 1.4}$	$29.8_{\pm 1.2}$	$29.1_{\pm 1.3}$
CLIPQCR [4]	$13.4_{\pm 1.2}$	$30.5_{\pm 1.1}$	$30.0_{\pm 1.2}$
CPRD+BAN [13]	$13.9_{\pm 1.3}$	$31.2_{\pm 1.5}$	$30.4_{\pm 1.5}$
DeBCF [25]	$18.6_{\pm 1.1}$	$35.4_{\pm 1.0}$	$34.2_{\pm 1.2}$
MISS [2]	$17.3_{\pm 1.1}$	$49.2_{\pm 1.0}$	$33.8_{\pm 1.1}$
M2I2 [9]	$17.3_{\pm 1.1}$	$51.9_{\pm 0.9}$	$35.2_{\pm 1.0}$
M3AE [3]	$24.4_{\pm 1.0}$	$65.1_{\pm 0.9}$	$45.4_{\pm 1.0}$
DeCoCT [23]	$27.3_{\pm 0.5}$	$69.6_{\pm 0.5}$	$49.2_{\pm 0.6}$
LCT (Ours)	$\mathbf{28.9}_{\pm 0.6}$	$\mathbf{71.6}_{\pm 0.4}$	$\mathbf{51.0}_{\pm 0.5}$

33.8% and 35.2%, respectively. In comparison, LCT outperforms M2I2 by 15.8% in Overall accuracy, demonstrating stronger robustness to distribution shifts. Compared with recent debiasing and causal approaches, LCT also shows consistent improvements. Relative to DeCoCT, LCT improves Overall accuracy by 1.8%, Open accuracy by 1.6%, and Closed accuracy by 2.0%. In addition, LCT maintains a low standard deviation (±0.5% Overall), suggesting stable performance across runs. These findings support that the proposed learnable trimming strategy enhances both robustness and training stability in OOD scenarios.

3.4 Ablation Study on Key Components

We conduct ablation studies on VQA-RAD, SLAKE, and PathVQA to evaluate the individual contributions of each component. Results are in Table 4.

Baseline + DAFB (w/o CT): We remove the causal trimming mechanism and directly concatenate DAFB features with the joint vision-language embeddings, followed by an MLP classifier. Using DAFB alone brings moderate improvements (+1.5% on VQA-RAD and +1.3% on PathVQA), suggesting that enriched anatomical priors are beneficial but insufficient to fully mitigate dataset biases when applied through standard feature fusion.

Baseline + CT (w/o DAFB): We remove the dynamic feature bank and compute trimming masks using global visual pooled features. Applying CT alone leads to larger gains(+3.5% on VQA-RAD and +2.2% on PathVQA). This indicates that explicitly suppressing spurious correlations contributes significantly to robustness, even without precise anatomical prototypes.

Full LCT: Combining DAFB and CT achieves the best performance across all datasets. The consistent improvements over individual variants demonstrate that the two modules are complementary: DAFB enhances anatomical representation, while CT performs causal trimming. Their integration yields stronger and more stable diagnostic reasoning performance.

Table 4: Ablation study on the key components of the proposed LCT framework. **CT**: Causal Trimming module. **DAFB**: Dynamic Anatomical Feature Bank.

Baseline	DAFB	CT	VQA-RAD			SLAKE			PathVQA
			Open	Closed	Overall	Open	Closed	Overall	Overall
✓			66.5	83.1	76.5	81.6	87.7	84.0	61.4
✓	✓		68.2	84.6	78.0	82.2	88.2	84.5	62.7
✓		✓	69.3	87.1	80.0	82.2	88.9	84.8	63.6
✓	✓	✓	**70.9**	**87.9**	**81.2**	**83.4**	**89.9**	**86.0**	**65.6**

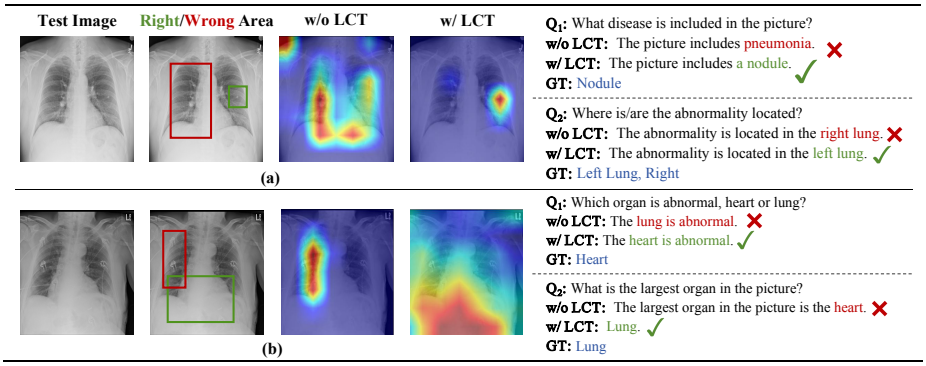

Fig. 2: Qualitative comparison of Grad-CAM visualizations and QA results.

3.5 Qualitative Analysis and Visualization

To illustrate how LCT mitigates spurious correlations, we present Grad-CAM visualizations and QA outputs in Figure 2. Without LCT, the baseline often attends to irrelevant or biased regions. In (a), attention is diffusely distributed over the right lung (red box), failing to localize the actual lesion. In Example (b), the baseline focuses on an upper lung region instead of the target organs.

With LCT, the attention maps become more anatomically focused. In the "w/ LCT" column, activations concentrate on the relevant structures, such as the small nodule in the left lung (Example a) and the enlarged cardiac silhouette (Example b), indicating improved visual grounding.

The refined attention is reflected in the predicted answers. In Example (a), the baseline predicts "pneumonia" and incorrectly identifies the "right lung", whereas LCT correctly recognizes the "nodule" and localizes it in the "left lung". In Example (b), LCT accurately identifies the heart as the abnormal organ and the lung as the largest organ, while the baseline produces inconsistent responses. These results suggest that causal trimming improves anatomical grounding and reduces reliance on dataset-specific shortcuts.

4 Conclusion

In this paper, we introduced the Learnable Causal Trimming (LCT) framework to overcome the spurious correlations and dataset biases in MedVQA. By explicitly purifying visual representations at the feature level, LCT integrates a Dynamic Anatomical Feature Bank (DAFB) to continuously capture domain-specific medical prototypes and a Causal Trimming (CT) module to dynamically filter out confounding visual noise while preserving authentic pathological signals. Extensive experiments demonstrate that our approach not only establishes new state-of-the-art performance across multiple in-distribution benchmarks—including VQA-RAD, SLAKE, and PathVQA—but also exhibits exceptional robustness on the out-of-distribution SLAKE-CP dataset. Ultimately, LCT forces the network to ground its predictions in genuine anatomical evidence rather than exploiting statistical shortcuts, representing a significant step toward developing highly reliable multi-modal AI systems for clinical applications.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62272337.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, L., Fang, H., Xu, N., Ren, B.: Counterfactual causal-effect intervention for interpretable medical visual question answering. IEEE Transactions on Medical Imaging **43**(12), 4430–4441 (2024)
2. Chen, J., Yang, D., Jiang, Y., Lei, Y., Zhang, L.: Miss: A generative pre-training and fine-tuning approach for med-vqa. In: International Conference on Artificial Neural Networks. pp. 299–313 (2024)
3. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022)
4. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163 (2023)
5. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
6. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

8. Li, P., Liu, G., He, J., Zhao, Z., Zhong, S.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 374–383 (2023)
9. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medical visual question answering. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)
10. Li, Q., Liu, M., Chang, R., Nie, W., Bai, S., Liu, A.: Multilabel chest x-ray image classification via category disentangled causal learning. IEEE Transactions on Artificial Intelligence **7**(2), 1048–1061 (2026)
11. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536 (2023)
12. Liu, B., Liu, L., Ding, J., Yang, X., Peng, W., Liu, L.: Cimb-mvqa: Causal intervention on modality-specific biases for medical visual question answering. Medical Image Analysis p. 103850 (2025)
13. Liu, B., Zhan, L.M., Wu, X.M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: International conference on medical image computing and computer-assisted intervention. pp. 210–220 (2021)
14. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654 (2021)
15. Liu, Y., Qiao, R., Lee, M.L., Hsu, W.: Test-time adaptation by causal trimming. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
16. Lu, Z., Zeng, Q., Lu, M., Chen, G., Xia, Y.: Bridging the semantic gap in medical visual question answering with prompt learning. IEEE Transactions on Medical Imaging **44**(11), 4605–4616 (2025)
17. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: International conference on medical image computing and computer-assisted intervention. pp. 522–530 (2019)
18. Nie, W., Zhang, C., Song, D., Bai, Y., Xie, K., Liu, A.A.: Chest x-ray image classification: A causal perspective. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 25–35 (2023)
19. Pan, H., He, S., Zhang, K., Qu, B., Chen, C., Shi, K.: Amam: an attention-based multimodal alignment model for medical visual question answering. Knowledge-Based Systems **255**, 109763 (2022)
20. Pearl, J.: Causal inference in statistics: An overview (2009)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 726–736 (2023)

23. Wan, X., Teng, Q., Chen, J., Lu, Y., Yuan, D., Liu, Z.: Eliminating language bias for medical visual question answering with counterfactual contrastive training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 194–204 (2025)
24. Xu, Z., Li, Q., Nie, W., Wang, W., Liu, A.: Structure causal models and llms integration in medical visual question answering. IEEE Transactions on Medical Imaging **44**(8), 3476–3489 (2025)
25. Zhan, C., Peng, P., Zhang, H., Sun, H., Shang, C., Chen, T., Wang, H., Wang, G., Wang, H.: Debiasing medical visual question answering via counterfactual training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–393 (2023)