Causal Dual-Interventional MedVQA via Textual Perturbation and Counterfactual Visual Verification

Jiuxiang You , Yi Yu * , Zhenguo Yang

Posted Date: 27 May 2026

doi: 10.20944/preprints202605.1802.v1

Keywords: medical visual question answering; concept-agnostic perturbations; clinical term grounding; counterfactual visual verifier

 doi:10.20944/preprints202605.1802.v1

Article

Causal Dual-Interventional MedVQA via Textual Perturbation and Counterfactual Visual Verification

Jiuxiang You [1], Yi Yu [1,*] and Zhenguo Yang[2]

[1] Hiroshima University, Japan
[2] Guangdong University of Technology, China
* Correspondence: yiyu@hiroshima-u.ac.jp

Abstract

Medical Visual Question Answering (MedVQA) aims to answer medical questions from clinical images. However, current models often rely on spurious language shortcuts rather than visual evidence, compromising clinical reliability. To this end, we propose a causal dual-interventional framework to mitigate language shortcuts in MedVQA. Our method incorporates two components: a textual de-confounding module and a counterfactual visual verifier. The textual de-confounding module disrupts linguistic shortcut biases via concept-agnostic perturbations to block backdoor pathways. Meanwhile, it aligns clinical terms with anatomical regions, compelling the model to establish genuine visual dependencies. In addition, the counterfactual visual verifier evaluates visual reliance by masking key regions and measuring prediction confidence drops under occlusion, thereby reducing language-driven artifacts. Extensive experiments on two public datasets demonstrate that our method significantly outperforms existing baselines.

Keywords: medical visual question answering; concept-agnostic perturbations; clinical term grounding; counterfactual visual verifier

1. Introduction

Medical Visual Question Answering (MedVQA)[1–3] aims to answer clinical questions about medical images, serving as a critical component of intelligent diagnosis systems. Early approaches employed attention mechanisms[4,5] and joint embeddings but struggled with limited medical knowledge and labeled data. With the emergence of Large Vision Language Models (LVLMs), recent methods have leveraged large-scale biomedical corpora to improve domain alignment. For example, Wu et al.[6] and Wang et al.[7] pre-train vision–language encoders to enhance medical awareness. Nguyen et al.[8] explore conditional reasoning and generative strategies to improve robustness under limited data, and Cai et al.[9] employ layer-wise relevance propagation to generate counterfactual visual explanations for interpretability.

Despite these advances, existing methods overlook a fundamental problem: models often exploit spurious linguistic shortcuts rather than grounding predictions in visual evidence. We identify two key sources of shortcut learning: (1) question patterns (e.g., "What view of the chest...") frequently correlate with specific answers like "Frontal" and (2) clinical terms (e.g., mention of "lung consolidation") commonly associate with "Yes" answers. Medical datasets are typically small with biased distributions, making models vulnerable to exploiting these linguistic cues instead of performing rigorous visual reasoning.

We analyze this phenomenon through the lens of causal inference, as illustrated in the right panel of Figure 1. Ideally, predictions should depend on visual and textual evidence (Q and $I \rightarrow K \rightarrow A$). However, a confounder (C) representing linguistic biases creates a spurious pathway directly to the answer, bypassing visual reasoning. Specifically, the confounder directly influences both the question-answer distribution in training data and the model's prediction mechanism, enabling shortcuts that

 doi:10.20944/preprints202605.1802.v1

circumvent the image. This "backdoor" path allows models to predict answers by exploiting language patterns alone, fundamentally compromising clinical reliability.

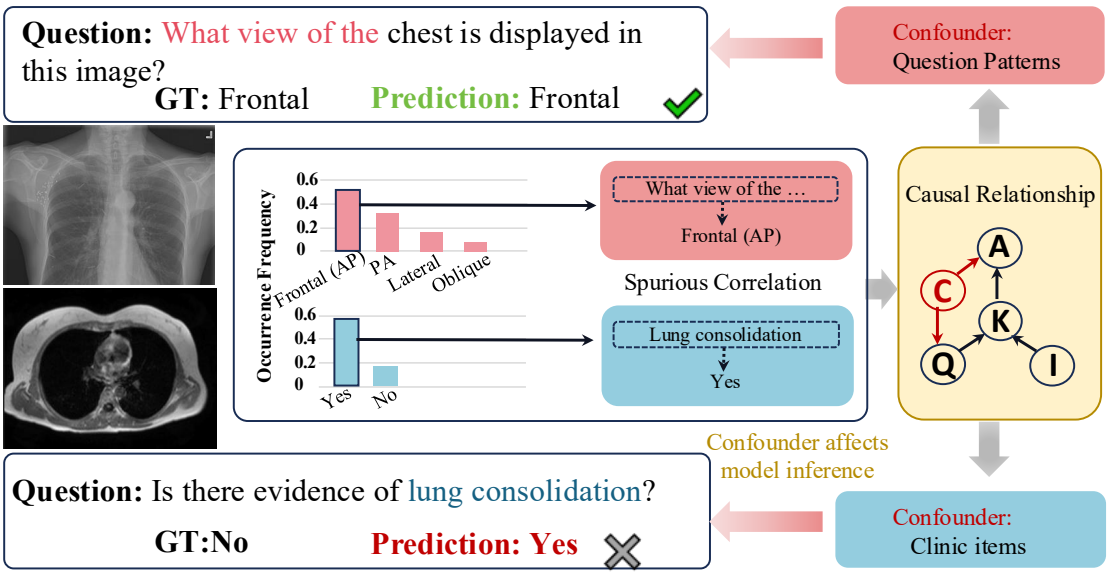

Figure 1. Illustration of how these biases introduce confounders during inference, creating misleading causal paths. As a result, the model relies on spurious correlations rather than true causal relationships, leading to incorrect predictions.

To this end, we propose a causal dual-interventional framework (DiMed) to mitigate language shortcuts via textual perturbation and counterfactual visual verification. Specifically, our method consists of two key stages: (1) During training, we introduce a textual de-confounding module combined with a Clinical Term Grounding Module (CTGM). This branch disrupts linguistic bias via question rewriting and entity masking, while simultaneously forcing the model to recover information from specific visual anatomical regions, thereby strengthening the true causal path (Q and $I \rightarrow K \rightarrow A$). (2) During inference, we employ a counterfactual visual verifier which serves as a causal filter by measuring the decrease in prediction confidence. In summary, our contributions are as follows:

- We propose a dual-interventional MedVQA framework that systematically mitigates language shortcuts by unifying textual perturbation strategies with counterfactual visual verification.
- We devise a clinical term grounding module to enforce visual dependency via anatomical alignment along with a counterfactual visual verifier that refines inference by isolating pure visual evidence from spurious correlations.
- Extensive experiments on two public benchmarks demonstrate that our method significantly outperforms the existing methods.

2. Related Work

2.1. Medical Visual Question Answering

Traditional MedVQA approaches employ joint embeddings and attention mechanisms[4,5] to align visual and textual features. To address data scarcity, Nguyen et al.[8] leverage meta-learning and conditional reasoning mechanisms, to improve generalization under limited supervision. However, these methods assume balanced and unbiased training data. With LLM emergence, PMC-LLaMA[6] fine-tunes LLaMA[6] on PubMed Central biomedical corpus to enhance medical reasoning through domain knowledge. PMC-CLIP[7] pre-trains vision encoders on medical image-text pairs to improve alignment, where domain-aware pre-training still optimizes likelihood without explicit confounding treatment. Recent methods like UnICLAM[10] and VG-CALF[11] introduce contrastive learning and vision-guided cross-attention mechanisms, where architectural advances improve fusion but remain vulnerable to linguistic shortcuts.

 doi:10.20944/preprints202605.1802.v1

2.2. Shortcut Bias and Causal Intervention

Shortcut bias arises when models rely on spurious correlations rather than true causal relationships. In Natural Language Processing (NLP) tasks, Gururangan et al.[12] reveal that Natural Language Inference models exploit annotation artifacts by relying on lexical patterns. Kaushik et al.[13] introduce counterfactual data augmentation to weaken lexical shortcuts. These interventions, however, remain limited to text-only settings and cannot address cross-modal confounders. In computer vision, Tang et al.[14] apply counterfactual feature editing to reduce visual spurious cues, where methods focus on visual confounders but overlook linguistic biases in multimodal tasks. Within MedVQA, this problem is exacerbated by small-scale datasets with biased distributions, where linguistic patterns become particularly strong confounders. Cai et al.[9] generate counterfactual visual explanations to expose shortcut behaviors, where the approach remains inference-time focused and does not systematically disrupt linguistic shortcuts during training.

3. Methods

In this section, we present the framework of DiMed, including the construction of de-confounding data, the textual de-confounding module with clinical term grounding, and the counterfactual visual verifier.

3.1. Overview of the Framework

The framework of the proposed DiMed is shown in Figure 2, which operates in two complementary phases. Training Phase: We augment training data with constructed negative examples to break spurious entity-answer correlations while preserving medical knowledge. The textual de-confounding module disrupts linguistic shortcuts by rewording questions and masking clinical terms, while the clinical term grounding module (CTGM) enforces visual dependency by aligning clinical terms to anatomical regions. Inference Phase: The counterfactual visual verifier filters non-causal predictions by measuring confidence drops when key visual evidence is masked.

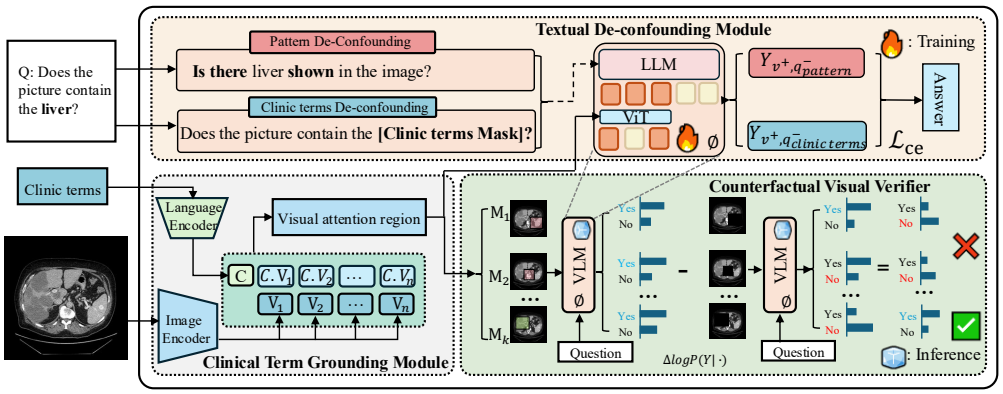

Figure 2. Overview of the proposed dual-interventional framework. During training, we apply textual de-confounding with clinical term grounding to disrupt linguistic shortcuts and enforce visual dependency. During inference, we utilize a counterfactual visual verifier to filter out non-causal predictions by measuring confidence drops when key visual evidence is masked.

3.2. Construction of De-confounding Data

Linguistic shortcuts in MedVQA stem from two primary confounders: (1) Pattern Bias from fixed question templates, and (2) Clinical Terms Bias from entity-answer co-occurrences. To mitigate these, we develop an automated pipeline with three steps: (i) extract clinical terms using biomedical NLP tools, (ii) filter confounders via statistical scoring, and (iii) construct augmented data by perturbing identified confounders.

 doi:10.20944/preprints202605.1802.v1

3.2.1. Identification of Clinical Terms

We employ a two-stage approach to identify strong confounders:

- **Entity Extraction & Normalization:** We use ScispaCy [15], a biomedical NLP tool, to extract clinical term from questions. We normalize these terms by unifying case formats and merging morphological variants (e.g., singular/plural forms).
- **Confounder Filtering via PMI:** Not all extracted terms act as confounders. To identify spurious correlations, we measure the statistical dependence between each clinical term C and answer A using Pointwise Mutual Information:

$$\text{PMI}(C, A) = \log \frac{P(C, A)}{P(C)P(A)} \tag{1}$$

High PMI indicates that term C and answer A co-occur more frequently than chance, suggesting a spurious association. We retain terms with PMI $> \theta$ (threshold $\theta = 0.5$) as clinical confounders, and discard generic terms (PMI ≈ 0) like "patient" that appear with all answers uniformly.

3.2.2. Negative Example Construction

To further break entity-answer statistical associations, we augment the dataset with negative examples. For each original sample with clinical term C, we create a negative variant by:

- Replacing C with an alternative term C' (using mixed strategy: 60% semantically distant, 20% similar, 20% low-frequency)
- Flipping the answer to negative.

This reduces the PMI score toward the threshold, ensuring the model cannot rely on entity-answer co-occurrence statistics.

3.2.3. Standardization of Question Patterns

Models exploit question templates as shortcuts: different syntactic variations of semantically equivalent questions often correlate with different answer distributions due to dataset biases. For example, questions like "Does the picture contain a liver lesion?", "Is there a liver lesion?", and "Can you identify a liver lesion?" are semantically equivalent but may be associated with different training label distributions.

To decouple semantic content from syntactic patterns, we rewrite all questions into a canonical form using Lingshu [16] with a designed prompt. The standardized template is: "Is there [entity/condition] in the image?". The prompt instructs the model to preserve semantic meaning while removing syntactic variations. As a practical proxy, we retain rewrites whose semantics remain consistent with the original gold answer. During training, we employ a pattern de-confounding strategy: for each training sample, we randomly select between the original and rewritten question. This forces the model to recognize that semantically equivalent questions should yield identical answers regardless of phrasing, thereby disrupting shortcut learning from question templates.

3.3. Textual De-Confounding with Clinical Term Grounding

To mitigate reliance on language shortcuts, we propose a dual-branch training strategy that perturbs textual inputs while explicitly grounding reasoning in visual evidence.

3.3.1. Textual Perturbation Strategy

We construct two complementary counterfactual question variants:

- **Pattern De-confounding ($Q_{Pattern}$):** Question templates themselves bias predictions (e.g., "Is there...?" questions tend to yield "Yes" in training data). We rewrite Q into a canonical form, removing stylistic biases. This forces the model to recognize that different phrasings should yield equivalent answers based on visual content alone.

 doi:10.20944/preprints202605.1802.v1

- **Clinical Terms De-confounding ($Q^-_{\text{Clinical items}}$):** Clinical items often shortcut predictions, correlating with answers regardless of visual evidence. We mask these tokens to create $Q^-_{\text{Clinical items}}$. Crucially, rather than simply removing information, we compensate by grounding the masked entity in visual evidence via CTGM, forcing the model to recover semantic information from the image.

3.3.2. Clinical Term Grounding Module (CTGM)

Simply masking clinical terms causes information loss, preventing the model from accessing semantic information about the clinical concept. CTGM addresses this by providing an alternative information channel: instead of the textual entity token, the model learns to access semantic information through relevant visual regions. This achieves dual goals: (1) disrupts text-shortcut pathways, (2) establishes robust visual reasoning for the same concept.

Visual Relevance Calculation. We identify relevant anatomical regions by measuring semantic alignment between the clinical concept and visual patches. First, we extract patch-level features using a frozen vision encoder: $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. Second, we encode the clinical entity C using the language encoder: $\mathbf{C} \in \mathbb{R}^d$. We compute cosine similarity between $\mathbf{C}$ and each patch:

$$\text{sim}(\mathbf{C}, v_j) = \frac{\mathbf{C} \cdot v_j}{\|\mathbf{C}\|\|v_j\|}. \tag{2}$$

Region Selection and Injection. We select the top-k patches with highest relevance scores:

$$\mathcal{I}_{grounded} = \{I_m \mid v_m \in \text{TopK}(\{\text{sim}(\mathbf{C}, v_j)\}_{j=1}^N, k)\}. \tag{3}$$

Training Objective. We train on three variants with combined loss:

$$\mathcal{L}_{total} = \mathcal{L}_{original} + \lambda(\mathcal{L}_{pattern} + \mathcal{L}_{grounding}), \tag{4}$$

where

$$\mathcal{L}_{original} = -\log P(Y \mid Q, I), \tag{5}$$
$$\mathcal{L}_{pattern} = -\log P(Y \mid Q_{pattern}, I), \tag{6}$$
$$\mathcal{L}_{grounding} = -\log P(Y \mid Q^-_{clinic}, \mathcal{I}_{grounded}). \tag{7}$$

The weight $\lambda = 1.0$ (in experiments) balances perturbation objectives. Joint training on three variants forces the model to produce consistent answers across: original questions, pattern-normalized questions, and visually-grounded masked questions. This prevents shortcuts and ensures causal reasoning.

3.4. Counterfactual Visual Verifier

During inference, we employ a region-level verification mechanism inspired by contrastive guidance[17]. We measure the model's output distribution shift when key visual regions are masked, ensuring predictions are grounded in relevant anatomical evidence. Rather than simply comparing single predictions, we leverage the probability distribution over the model's output tokens. For each generated answer $\hat{y}$, we compute the probability distribution under two conditions: with the full image and with masked regions.

Probability Contrast Measurement. For each identified clinical region, we measure the contrast between full and masked conditions:

$$\tilde{p}(y_t|\cdot) \propto p_\theta(y_t \mid I, X, y_{<t}) \left(\frac{p_\theta(y_t \mid I, X, y_{<t})}{p_\theta(y_t \mid I_{mask}, X, y_{<t})} \right)^\alpha. \tag{8}$$

Taking logarithms for numerical stability:

 doi:10.20944/preprints202605.1802.v1

$$\log \tilde{p}(y_t) \approx \log p_\theta(y_t \mid I, X, y_{<t}) - \alpha \log p_\theta(y_t \mid I_{mask}, X, y_{<t}), \tag{9}$$

where α is the region guidance strength (set to $\alpha = 0.5$ in our experiments). This formulation directly contrasts the model's token probabilities between unmasked and masked images.

Token-level Confidence Drop Calculation. For the answer sequence $\hat{y} = [y_1, y_2, \ldots, y_T]$, we compute the log-likelihood drop for each token when the region is masked:

$$\Delta\mathcal{P}_t = \log p_\theta(y_t \mid I, X, y_{<t}) - \log p_\theta(y_t \mid I_{mask}, X, y_{<t}). \tag{10}$$

The overall causal contribution of the masked region is:

$$\Delta\mathcal{P}_{answer} = \sum_{t=1}^{T} \Delta\mathcal{P}_t. \tag{11}$$

Verification Decision. After masking each identified clinical region and computing $\Delta\mathcal{P}$, we identify the region with maximum contribution:

$$region^* = \arg\max_{region} \Delta\mathcal{P}(region). \tag{12}$$

An answer is considered visually grounded if the maximum regional importance score exceeds the threshold: $\Delta\mathcal{P}(region^*) > \tau$ (empirically set to $\tau = 0.5$ based on validation performance). This ensures that the model's answer depends on analyzing specific anatomical regions rather than relying on language shortcuts or background patterns.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. To comprehensively evaluate the effectiveness of proposed DiMed, we evaluate DiMed on two publicly available benchmarks for medical visual question answering:

- **VQA-RAD** [18]: A manually curated dataset containing 315 radiology images and 3,515 QA pairs. It serves as a rigorous test for few-shot reasoning capabilities due to its small scale and high quality.
- **SLAKE** [19]: A large-scale bilingual dataset comprising 642 images and approximately 14,000 QA pairs with rich semantic annotations. We utilized the English subset to evaluate performance on diverse body parts and modalities.

Implementation Details. We adopted Lingshu-7B [16] as our backbone model. During training, we applied LoRA to fine-tune the language encoder while keeping the visual encoder frozen. The model was trained for 10 epochs using AdamW with a learning rate of 2e-4. The hyperparameter k for region selection in CTGM was set to 3. For the counterfactual visual verifier, features were extracted using biomedclip [20], and the guidance scale α was set to 0.5.

4.2. Baseline and Performance on the Two Datasets

We compared DiMed against methods spanning different paradigms: traditional attention-based approaches (MEVE-BAN[8], M3AE), knowledge-distillation methods (PubMedCLIP[21], CPCR[4]), and recent vision-language models (LaPA[22], CCIS-MVQA[9], VG-CALF[11], UniCLAM[10], CIMB-MVQA[23]). Results are presented in Table 1.

 doi:10.20944/preprints202605.1802.v1

Table 1. Accuracy (%) comparison of different methods on VQA-RAD and SLAKE datasets. Boldface indicates the best performance in each column.

Methods	Reference	VQA-RAD			SLAKE		
		Open-ended	Close-ended	Overall	Open-ended	Close-ended	Overall
MEVE-BAN [8]	MICCAI'19	40.33	73.90	59.20	75.19	81.49	77.66
M3AE [24]	MICCAI'22	63.10	83.31	75.40	79.83	86.30	82.37
PubMedCLIP [21]	EACL'23	60.10	80.00	72.10	78.40	82.50	80.10
CPCR [4]	TMI'23	60.50	80.40	72.50	80.50	84.10	81.90
LaPA [22]	CVPR'24	66.48	85.29	77.82	79.84	86.53	82.46
CCIS-MVQA [9]	TMI'24	68.78	79.24	75.06	80.12	86.72	84.08
VG-CALF [11]	Neurocomputing'25	67.00	85.50	76.10	81.40	83.80	83.30
UniCLAM [10]	MedIA'25	59.80	82.60	73.20	81.10	85.70	83.10
CIMB-MVQA [23]	MIA'25	69.33	**86.19**	79.42	82.08	89.42	85.09
DiMed (Ours)	**Proposed**	**74.00**	86.05	**80.60**	**84.10**	**90.90**	**86.83**

On the VQA-RAD dataset, DiMed achieves an overall accuracy of 80.60%, outperforming the previous best method CIMB-MVQA by 1.18 percentage points (80.60% vs. 79.42%). This gain indicates that our dual-interventional framework effectively mitigates linguistic shortcuts and enhances visual grounding, especially for open-ended medical questions requiring detailed visual understanding. On the SLAKE dataset covering diverse anatomical regions and question types, DiMed achieves an overall accuracy of 86.83%, surpassing CIMB-MVQA by 1.74 percentage points (86.83% vs. 85.09%). The improvement is most pronounced for close-ended questions (90.90% vs. 89.42%), highlighting the effectiveness of clinical term grounding in anchoring medical concepts to visual evidence. We attribute these improvements to two key factors.

First, textual de-confounding disrupts linguistic shortcuts by enforcing robustness across question variants, preventing exploitation of spurious pattern–answer correlations. Second, clinical term grounding strengthens visual reasoning by anchoring masked clinical terms to anatomically relevant regions, encouraging meaningful associations between clinical concepts and visual evidence.

4.3. Ablation Study

We incrementally add components to evaluate their individual and combined effects. As shown in Table 2, pattern de-confounding improves accuracy by +1.54%, while CTGM yields a larger gain of +1.98%, indicating the stronger impact of visual grounding. Their combination further improves performance to 77.80% (+2.40%). The verifier alone provides limited benefit (+1.54%), but when combined with training-time interventions, the full model achieves 80.60%, corresponding to a total gain of +5.20%, demonstrating strong synergy between training and inference.

Table 2. Ablation study on VQA-RAD. TRAINING: TEXTUAL DE-CONFOUNDING + CTGM; INFERENCE: COUNTERFACTUAL VISUAL VERIFIER.

Model	Pattern	CTGM	Verifier	Overall
Baseline (Lingshu-7B)	-	-	-	75.40
+ Pattern De-conf.	✓	-	-	76.94
+ CTGM	-	✓	-	77.38
+ Pattern + CTGM	✓	✓	-	77.80
+ Verifier Only	-	-	✓	76.94
DiMed (Full)	✓	✓	✓	**80.60**

4.4. Counterfactual Visual Verifier Analysis

Figure 3 shows the verifier's effectiveness. Before training, masking any region produces minimal confidence drop ($\Delta P < 0.5$), indicating shortcut-based prediction. After training, masking the critical region causes substantial drop ($\Delta P > 0.5$), confirming visual grounding. PMI reduction validates statistical debiasing.

 doi:10.20944/preprints202605.1802.v1

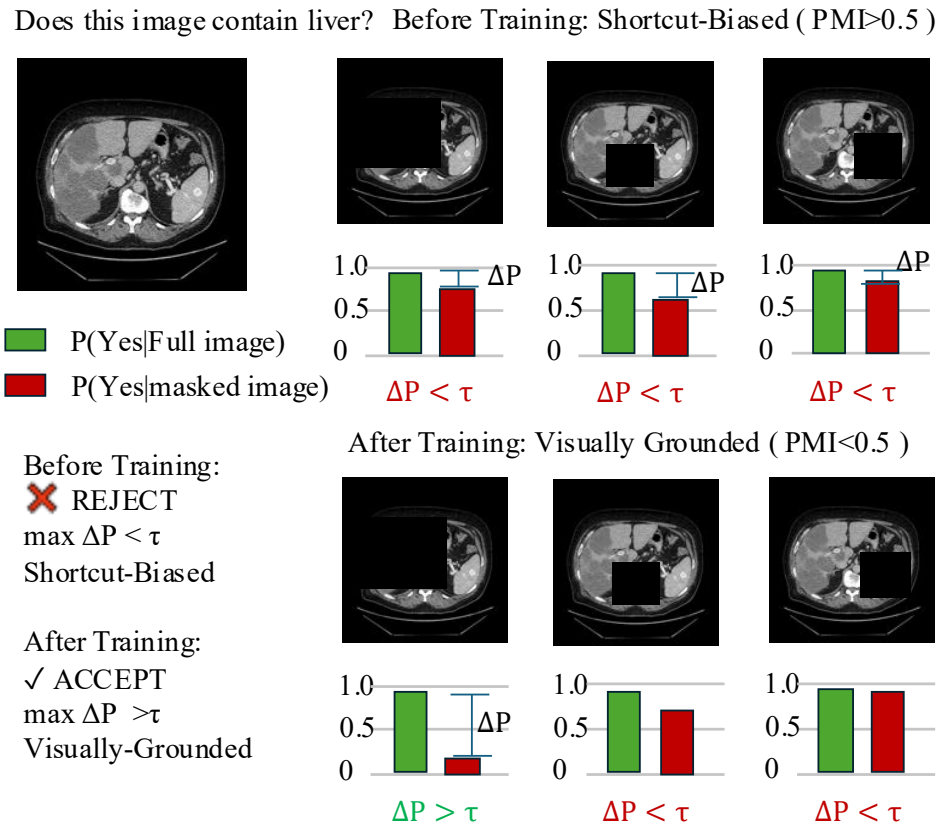

Figure 3. Counterfactual Verification: Before training (top), predictions rely on shortcut cues ($\Delta P < \tau$, REJECT). After training (bottom), predictions become visually grounded ($\Delta P > \tau$, ACCEPT).

4.5. Impact of Top-k Region Selection

To understand how the number of grounded regions affects performance, we vary the hyperparameter k in CTGM. Figure 4 shows accuracy as a function of k.

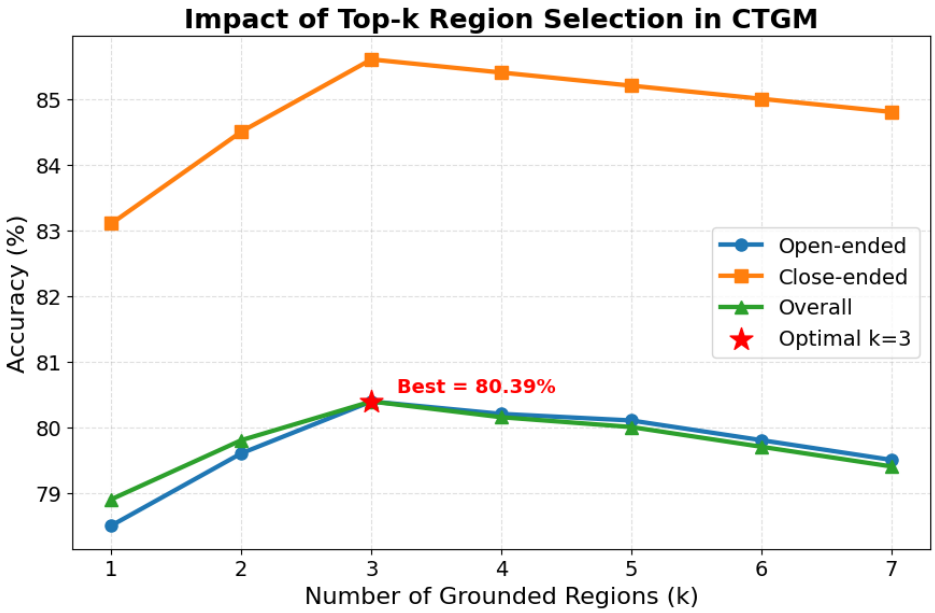

Figure 4. Impact of top-k region selection in CTGM on VQA-RAD (Open, Close, Overall).

With $k = 1$, performance is limited (78.50% overall), indicating that a single region is insufficient to capture the full semantic content of a masked clinical term. Performance improves with $k = 2$

 doi:10.20944/preprints202605.1802.v1

(79.60%) and $k = 3$ (80.39%), showing that 2-3 regions are optimal for grounding clinical concepts. Beyond $k = 5$, performance plateaus or slightly decreases (80.10%), suggesting that overly diffuse region selection introduces noise. We set $k = 3$ as the default based on this analysis.

4.6. Qualitative Examples and Failure Cases

Successful Examples. Figure 5 shows two cases (BP means baseline prediction). Case 1 demonstrates how CTGM disrupts clinical terms bias: the baseline predicts "No" for a normal liver due to the spurious correlation between "liver" and abnormality in the training set. DiMed correctly grounds the concept to the actual liver region and predicts "Yes". **Failed Cases.** Despite improvements, DiMed has limitations. Case 2 highlights a limitation on multi-class modality questions, where the model still fails to distinguish fine-grained acquisition types.

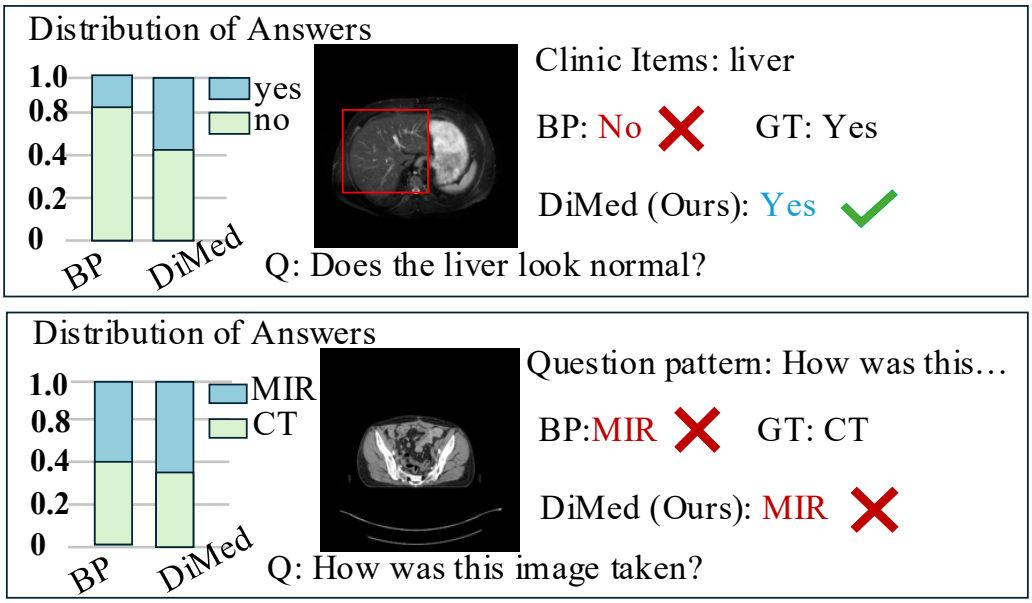

Figure 5. Examples of successful and failed cases of the proposed causal dual-interventional MedVQA.

5. Conclusions

In this paper, we propose a dual-interventional framework aimed at explicitly weakening linguistic shortcut pathways in medical VQA. By integrating pattern and clinical-level textual de-confounding with visual grounding of masked clinical terms, DiMed enforces a stronger dependency on image evidence during training. An inference-time counterfactual verifier further constrains predictions to anatomically meaningful regions. Extensive experiments on VQA-RAD and SLAKE confirm that our r approach achieves significant improvements. Current limitations in fine-grained lesion reasoning highlight the need for future extensions toward multi-scale evidence modeling and hierarchical grounding.

Acknowledgments: This work was supported by JST SPRING, Grant Number JPMJSP2132.

References

1. Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *Artif. Intell. Med.*, vol. 143, p. 102611, 2023.
2. K. Zhang, Y. Yang, J. Yu, H. Jiang, J. Fan, Q. Huang, and W. Han, "Multi-task paired masking with alignment modeling for medical vision-language pre-training," *IEEE TMM*, vol. 26, pp. 4706–4721, 2023.
3. C. Zhan, Y. Zhang, Y. Lin, G. Wang, and H. Wang, "Unidcp: unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts," *IEEE TMM*, vol. 26, pp. 9736–9748, 2024.
4. B. Liu, L.-M. Zhan, L. Xu, and X.-M. Wu, "Medical visual question answering via conditional reasoning and contrastive learning," *IEEE TMI*, vol. 42, no. 5, pp. 1532–1545, 2022.

 doi:10.20944/preprints202605.1802.v1

5. L. Qiao, R. Wang, Y. Shu, X. Xu, B. Li, W. Li, and X. Gao, "Re 3 adapter: Efficient parameter fing-tuning with triple reparameterization for adapter without inference latency," in *ICME*, 2024, pp. 1–6.
6. C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "Pmc-llama: toward building open-source language models for medicine," *J. Am. Med. Inform.*, vol. 31, no. 9, pp. 1833–1843, 2024.
7. W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, "Pmc-clip: Contrastive language-image pre-training using biomedical documents," in *MICCAI*, 2023, pp. 525–536.
8. B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *MICCAI*, 2019, pp. 522–530.
9. L. Cai, H. Fang, N. Xu, and B. Ren, "Counterfactual causal-effect intervention for interpretable medical visual question answering," *IEEE TMI*, 2024.
10. C. Zhan, P. Peng, H. Wang, G. Wang, Y. Lin, T. Chen, and H. Wang, "Uniclam: Contrastive representation learning with adversarial masking for unified and interpretable medical vision question answering," *Med. Image Anal.*, vol. 101, p. 103464, 2025.
11. A. Lameesa, C. Silpasuwanchai, and M. S. B. Alam, "Vg-calf: A vision-guided cross-attention and late-fusion network for radiology images in medical visual question answering," *Neurocomput.*, vol. 613, p. 128730, 2025.
12. S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and I. Dagan, "Annotation artifacts in natural language inference data," in *ACL*, 2018, pp. 107–112.
13. D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *EMNLP*, 2020, pp. 325–345.
14. Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *CVPR*, 2021, pp. 12 700–12 710.
15. M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispacy: fast and robust models for biomedical natural language processing," *arXiv:1902.07669*, 2019.
16. W. Xu, H. P. Chan, L. Li, M. Aljunied, R. Yuan, J. Wang, C. Xiao, G. Chen, C. Liu, Z. Li *et al.*, "Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning," *arXiv:2506.07044*, 2025.
17. D. Wan, J. Cho, E. Stengel-Eskin, and M. Bansal, "Contrastive region guidance: Improving grounding in vision-language models without training," in *ECCV*, 2024, pp. 198–215.
18. J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Sci. Data*, vol. 5, no. 1, pp. 1–10, 2018.
19. B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *IEEE ISBI*, 2021, pp. 1650–1654.
20. S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv:2303.00915*, 2023.
21. S. Eslami, G. De Melo, and C. Meinel, "Does clip benefit visual question answering in the medical domain as much as it does in the general domain?" *arXiv:2112.13906*, 2021.
22. T. Gu, K. Yang, D. Liu, and W. Cai, "Lapa: Latent prompt assist model for medical visual question answering," in *CVPR*, 2024, pp. 4971–4980.
23. B. Liu, L. Liu, J. Ding, X. Yang, W. Peng, and L. Liu, "Cimb-mvqa: Causal intervention on modality-specific biases for medical visual question answering," *Med. Image Anal.*, p. 103850, 2025.
24. Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *MICCAI*, 2022, pp. 679–689.