Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Localizing Before Answering: A Benchmark for Grounded Medical Visual Question Answering

Dung Nguyen[1], **Minh Khoi Ho**[1], **Huy Ta**[2], **Thanh Tam Nguyen**[3], **Qi Chen**[2], **Kumar Rav**[4], **Quy Duong Dang**[2], **Satwik Ramchandre**[2], **Son Lam Phung**[6], **Zhibin Liao**[2], **Minh-Son To**[5], **Johan Verjans**[2], **Phi Le Nguyen**[1], **Vu Minh Hieu Phan**[2]

[1] Hanoi University of Science and Technology
[2] Australian Institute for Machine Learning, The University of Adelaide
[3] Griffith University
[4] SA Health
[5] College of Medicine and Public Health, Flinders University
[6] University of Wollongong

Abstract

Medical Large Multi-modal Models (LMMs) have demonstrated remarkable capabilities in medical data interpretation. However, these models frequently generate hallucinations contradicting source evidence, particularly due to inadequate localization reasoning. This work reveals a critical limitation in current medical LMMs: instead of analyzing relevant pathological regions, they often rely on linguistic patterns or attend to irrelevant image areas when responding to disease-related queries. To address this, we introduce HEAL-MedVQA (Hallucination Evaluation via Localization MedVQA), a comprehensive benchmark designed to evaluate LMMs' localization abilities and hallucination robustness. HEAL-MedVQA features (i) two innovative evaluation protocols to assess visual and textual shortcut learning, and (ii) a dataset of 67K VQA pairs, with doctor-annotated anatomical segmentation masks for pathological regions. To improve visual reasoning, we propose the Localize-before-Answer (LobA) framework, which trains LMMs to localize target regions of interest and self-prompt to emphasize segmented pathological areas, generating grounded and reliable answers. Experimental results demonstrate that our approach significantly outperforms state-of-the-art biomedical LMMs on the challenging HEAL-MedVQA benchmark, advancing robustness in medical VQA.

1 Introduction

Large Multimodal Models (LMMs), or Multi-modal Large Language Models (M-LLMs) [Alayrac *et al.*, 2022; Bai *et al.*, 2023; Liu *et al.*, 2024b] such as GPT-4V [OpenAI *et al.*, 2023] achieve superb performance in understanding data from multiple modalities, such as vision and language, and generating human-like texts. Leveraging the strong capabilities of those foundational models, recent works [Li *et al.*, 2024; Chen *et al.*, 2024; Zhang *et al.*, 2023a; Chen *et al.*, 2023; Moor *et al.*, 2023; Wu *et al.*, 2023b; Wu *et al.*, 2023a; Yan and Pei, 2022] have developed M-LLMs for medical imaging applications such as medical Visual Question Answering (VQA). Recently, many multimodal foundation models (LLaVA-Med [Li *et al.*, 2024], CheXagent [Chen *et al.*, 2024], GPT-4 [OpenAI *et al.*, 2023], Gemini [Team *et al.*, 2023]) have emerged and demonstrated impressive reasoning and comprehension capabilities on biomedical queries and related domains. These models can assist clinical professionals in interpreting and gaining insights on medical images, helping the diagnosis and treatment processes be more efficient.

Due to the sensitive nature of the biomedical domain, the accuracy and trustworthiness of these foundational models are important. Though impressive, Large Language Models (LLMs) and multimodal LLMs are prone to hallucinating contents. Specifically, the models produce false answers, which are not grounded in the visual evidence. Given a question. *"Does the patient have pneumonia in his right lung?"*, many LLMs are prone to answer *"Yes"* without reasoning the image content. This behavior arises because the observation of *pneumonia* frequently co-occurs with parts of the lung, leading to spurious correlations instead of rigorous reasoning. Fig. 1 illustrates the answer of some open-source LMMs, and the cross-attention map of the models during generation. These models tend to hallucinate and attend to irrelevant image regions when answering these questions.

This study shows two types of shortcut learning that current multimodal LLMs suffer from: textual and visual shortcut learning. *First*, the models attend to non-important text tokens with higher relevancy score than any image tokens. As shown in Fig. 1b, the model produces higher relevancy scores on the token *lung* than on any image token. *Second*, on the image, the model attends to non-queried image regions (as shown in Fig. 1a). This shows the model learns shortcuts on irrelevant text and image tokens, instead of basing on the visual evidence when answering the questions.

To address these concerns, we introduce a new bench-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

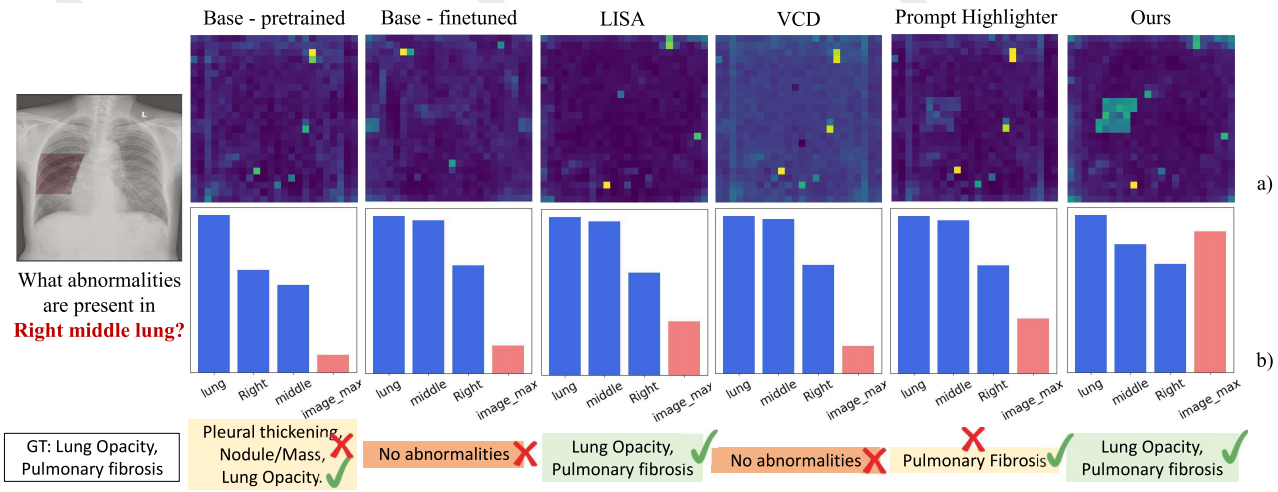

Figure 1: An example of Med-VQA from our dataset. We follow LVLM-Interpret [Stan *et al.*, 2024] to visualize a) the attention map of the image with respect to the model response, and b) the relevancy score between the answer and input words. LLaVA-Med [Li *et al.*, 2024], as reference model, fails to attend to the region of interest, and its answer has a low relevancy score to the image. Several hallucination reduction methods can generate responses more relevant to the image, but may not localize the target region.

mark, dubbed HEAL-MedVQA (Hallucination EvAluation via Localization in Medical VQA), to dissect the LLM's ability to localize at visual evidence when answering. Specifically, this paper proposes two evaluation protocols, called Textual Perturbation Test (TPT) and Visual Perturbation Test (VPT), which probe LLMs' sensitivity towards textual and visual shortcuts. To support this evaluation, we curate adversarial VQAs and procure pixel-level annotations of affected anatomy regions from three board-certified radiologists from two large-scale public datasets, MIMIC-CXR [Johnson *et al.*, 2019] and VinDr-CXR [Nguyen *et al.*, 2020]. In total, we construct over 60,000 question-answer pairs probing LLM's textual and visual shortcut learning. Our curated medical benchmark provides a structured codebase, facilitating comprehensive comparisons of advanced LLMs.

To this end, we propose Localize-before-Answer (LobA), a framework that forces LMM to generate visual-evidence tokens, which is then input to the segmentation decoder to localize the area of interest before answering. Based on the localization map, LobA self-prompts to enhance the model's attention, forcing it to leverage visual evidence relevant to the question before answering. In summary, our main contributions are as follows.

- We introduce HEAL-MedVQA (Hallucination Evaluation via Localization in Medical VQA), a benchmark that consists of 67K QA pairs, and doctor-annotated segmentation masks with two new evaluation protocols, assessing LLMs' robustness to textual and visual shortcut learning. Our benchmark includes comprehensive comparisons of 8 state-of-the-art LLMs and hallucination techniques, thus standardizing hallucination and accuracy evaluations of future multi-modal LLMs.
- We propose the Localize-before-Answer (LobA) framework, which forces the model to localize the affected anatomical regions and self-prompt to enhance attention on the target regions before answering.
- Our proposed LobA framework significantly surpasses recent LLMs by up to 5.44%, while showing the highest robustness to textual and visual shortcut learning.

2 Related Works

Medical Visual Question Answering. Several benchmarks in Med-VQA propose specialized datasets [Hasan *et al.*, 2018; Abacha *et al.*, 2019; Abacha *et al.*, 2020; Abacha *et al.*, 2021]. VQA-RAD [Lau *et al.*, 2018] offers both open-ended and closed-ended questions. SLAKE [Liu *et al.*, 2021] develops a bilingual dataset with a knowledge graph. Overall, existing benchmarks use accuracy as the main metric, which overlooks the hallucination evaluation of multi-modal LLMs.

A majority of Med-VQA models are finetuned from other VLMs, namely LLaVA-Med [Li *et al.*, 2024] from LLaVA [Liu *et al.*, 2024b], CheXagent [Chen *et al.*, 2024] from BLIP-2 [Li *et al.*, 2023a], or [Zhang *et al.*, 2023a; Chen *et al.*, 2023; Moor *et al.*, 2023; Wu *et al.*, 2023b; Wu *et al.*, 2023a; Yan and Pei, 2022; Huy *et al.*, 2025]. Recent approaches, which localize relevant regions in Med-VQA, have gained attention in recent times [Tascon-Morales *et al.*, 2023; Tascon-Morales *et al.*, 2024; Zhang *et al.*, 2024a]. However, during inference time, these VQA models require human-annotated masks to locate regions of interest. In contrast, our method explicitly learns to localize during training, bypassing the need to acquire human annotations when inferring.

Hallucination in VQA. To overcome hallucination, there are two main approaches. *First*, data-centric methods focus on augmenting existing VQA training datasets and propose new benchmarks for evaluation. In [Li *et al.*, 2023b] and [Liu *et al.*, 2023], they show that frequently appearing objects tend to guide models' answers during inference time more than absent or less frequent ones. As such, POPE [Li *et al.*, 2023b] adds adversarial questions asking for the presence of objects

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

that are absent from the image. LRV-Instruction [Liu *et al.*, 2023] introduces negative instructions, which are irrelevant to the visual content.

Second, model-centric methods focus on modifying the model architecture to support the visual component. InternVL [Chen and others, 2024], and LLaVA-1.5-HD [Liu *et al.*, 2024a] scale up the vision foundation model to allow better visual understanding. Others, such as [He *et al.*, 2024; Jain *et al.*, 2024] append several levels of information (segmentation mask, object list, task-based knowledge) to enrich the visual context. Lastly, several methods propose inference-based methods, which augment the inference of the pre-trained LLMs. VCD [Leng *et al.*, 2024] and M3ID [Favero *et al.*, 2024] aim to ground inference by negating a contrastive version of the input where the image is distorted or dropped. OPERA [Huang *et al.*, 2024] discourages overly confident but erroneous information accumulated in self-attention layers by applying a penalty. Overall, existing hallucination techniques are applied on the image level, while neglecting region-level details. In contrast, we propose to explicitly reason on the pixel level before answering.

3 Hallucination Evaluation Benchmark

This section introduces HEAL-MedVQA (Hallucination Evaluation via Localization in Medical VQA) benchmark consists of 67K QA pairs, and doctor-annotated segmentation masks with two new evaluation protocols.

3.1 New Evaluation Protocol — Sensitivity Test

Conventional accuracy metrics [Abacha *et al.*, 2019; Liu *et al.*, 2021; Lau *et al.*, 2018] overlook bias factors influencing the model's responses. This paper assesses two key phenomena in medical LMMs:

- **Language Shortcut:** Despite generating correct answers, the model relies on language shortcuts learnt in the training dataset, rather than the actual visual content. For example, given a question *"Does the right lower lung have pneumonia?"*, models can answer *"Yes"* due to the co-occurrence of *lung* and *pneumonia* in the training set.
- **Vision Shortcut:** The model can concentrate on non-relevant regions (as in Fig. 1a). This behavior results in the unreliability of LMMs, even when providing correct answers, as they ignore visual evidence.

To this end, we propose two new evaluation protocols, called Textual Perturbation Test (TPT) and Visual Perturbation Test (VPT), as shown in Fig. 2.

Textual Perturbation Test. To assess how sensitive a model is to language biases, we propose to swap the key entities, *i.e.*, anatomy or disease, and evaluate the model's sensitivity to the perturbed questions. The evaluation protocol is as follows.

- From the test set of binary questions - "Does *anatomy* have *disease*?", we collect True Positive samples having correct "Yes" predictions.
- We randomly replace either the anatomy or disease with another valid co-occurring term. For example, in Fig. 2 (Above), the question *"Does left lower lung suffer from pneumonia?"* is replaced by either switching the anatomy to *right upper lung*, or to another disease.
- The percentage of "Yes" → "No" flips defines the *Textual Perturbation Test* (TPT) score.

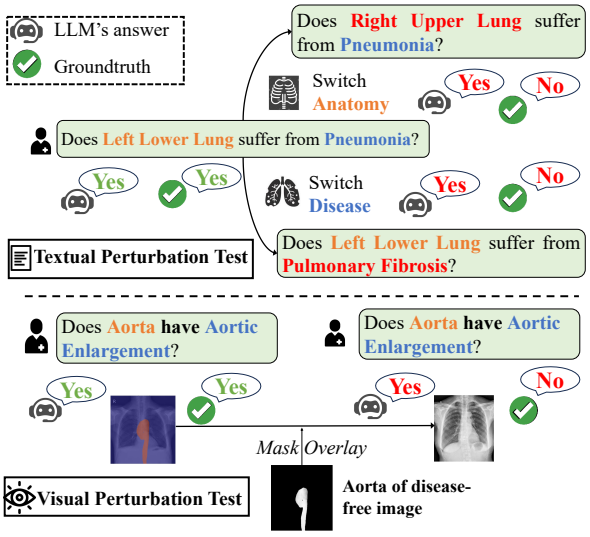

Figure 2: Textual and Visual Perturbation Tests. TPT: Replace the anatomy or disease term in the question. VPT: Swap the region of interest with the same region from an image of a different class. We report the percentage of "Yes" → "No" flips to assess model sensitivity to shortcuts.

Visual Perturbation Test. We propose to assess whether the model grounds the answer on the anatomical region of interest. Specifically, when querying the presence of a disease in a particular area, *e.g.*, the left lower lung, we take a segmented region of the left lower lung from another image that does not share the same diagnosis as the current one. This segmented region is resized and blended onto the original image. By doing so, we expect the model to focus its attention on the replaced region of interest to produce the correct diagnosis, altering the response from "Yes" to "No". We define the visual perturbation test (VPT) score as the *percentage of answers that change* from "Yes" to "No" under this protocol.

3.2 Dataset Curation

To assess the LLM's robustness against both textual and visual shortcuts, we introduce a new benchmark called HEAL-MedVQA (Hallucination Evaluation via Localization in Medical VQA). Built upon VinDr-CXR [Nguyen *et al.*, 2020] and MIMIC-CXR [Johnson *et al.*, 2019], we construct HEAL-MedVQA by mapping spatial relationships between diseases and anatomical locations, and probing the model's robustness towards disease and location perturbation on both texts and images. Fig. 4 illustrates our overall pipeline of data construction. We extract the disease bounding boxes and anatomical masks from the image. Based on the overlap between disease and anatomy localization, we determine their spatial relationships. Then, several QA templates are used to generate VQA samples. This process can be broken down into four main steps:

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

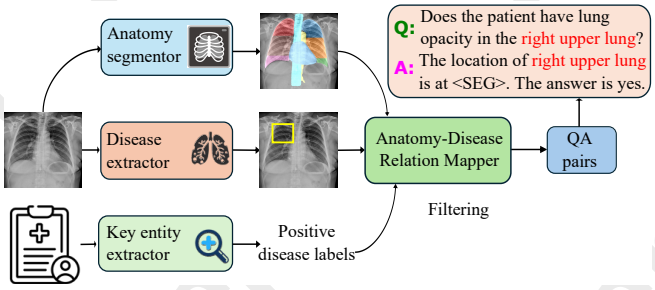

Figure 3: Our data processing pipeline.

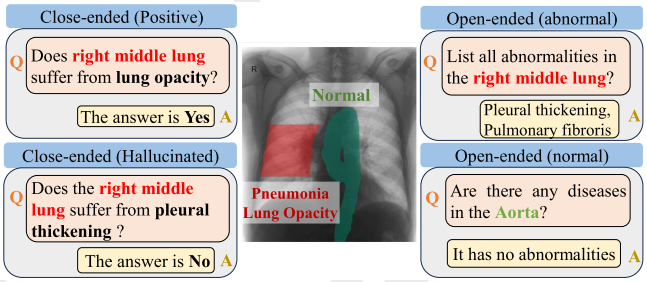

Figure 4: Four question types in our benchmark.

Anatomy segmentation. We define 8 main anatomies, e.g. *Right lower lung, Heart, etc.*, where chest diseases commonly occur. On the training set, we obtain the segmentation mask using a model pre-trained on PAXRay++ dataset [Seibold and others, 2022]. On the test set, we consult three board-certified radiologists to annotate the segmentation. By providing doctor-annotated pixel-level segmentation, this research offers a new benchmark to evaluate visual reasoning and robustness to hallucination of current M-LLMs.

Disease extraction. For the disease bounding boxes, the VinDr-CXR dataset provides bounding box annotations for 22 medical findings. We train a standard detection model, YOLOv5 [Jocher *et al.*, 2020], on the dataset and obtain the bounding boxes on MIMIC-CXR. To reduce false positives, we extract the disease labels from the associated medical reports and filter out absent diseases.

Anatomy-disease relation mapping. To get the anatomy-disease relation, we measure the overlap between the disease bounding box and the anatomical segmentation mask. Since the disease area is often small compared to the anatomy's, we compute the intersection over the disease area, instead of a union of both. If this IoU_{dis} score between a disease-anatomy pair is over $\delta = 0.5$, then that disease occurs on that anatomy.

Question Generation. After obtaining a list of disease–anatomy mappings, we randomly generate 2–5 QA pairs per image, focusing on two main formats: close-ended and open-ended, as shown in Fig. 4. Close-ended questions are binary: "Does *anatomy* have *disease*?", "Is there *disease* present in *anatomy*?". Each is either a Positive question, which queries a disease present in the anatomy, or a Hallucinated question, which queries a disease that is **not** present. Open-ended questions have two types: normal and abnormality-query questions, which respectively probe regions without and with diseases.

In total, our HEAL-MedVQA dataset comprises 45,331 closed-ended questions and 22,573 open-ended ones. We compare our benchmark with those previously reported in Table 1. To the best of our knowledge, we are the first benchmark providing a large-scale dataset with comprehensive hallucination metrics and doctor-annotated masks to evaluate the hallucination of emerging LLMs.

4 Proposed Method

This section presents our proposed simple-yet-effective framework, **Lo**calize-**b**efore-**A**nswer (**LobA**) to (i) localizes the queried regions, and then (ii) reweigh the attention to ground on correct visual evidence before answering. An overview of the framework is shown in Figure 5.

4.1 Grounded Text Generation

Given an image x_{img} and a question about it x_{text}. We train LLMs to generate an answer $\hat{y}_{text}$, while grounding on the queried region, and producing a segmentation mask $\hat{M}$. The mask $\hat{M}$ corresponds to both x_{text} and $\hat{y}_{text}$, explaining the response of the model.

We use a *<SEG>* token, denoting the token that will be used in later steps to obtain the segmentation of the relevant anatomies. Using this token, we instruct LLMs to localize on visual evidence before answering. As such, we format the response as *"The location of {relevant areas} is at <SEG>"*. For example, if the question is *"Are there any abnormalities at the right lower lung?"*, the output of the LMM will be *"The location of the right lower lung is at <SEG>"*. Note that, in our format, the model has to produce the visual evidence first, before giving the final answers. The LLM is trained to generate the above response:

$$\hat{y}_{\text{text}} = LLM(x_{\text{text}}, x_{\text{img}}). \tag{1}$$

We use the projected hidden state h_{seg} from the *<SEG>* token as the query to the segmentation model. This can be formulated as $\hat{M} = Seg(x_{\text{img}}, h_{\text{seg}})$. Our LLM is trained via the text objectives. Similar to other LLMs [Touvron *et al.*, 2023; OpenAI *et al.*, 2023], we formulate the text generation loss as the cross-entropy loss between the generated text and the ground truth:

$$\mathcal{L}_{text} = \text{CE}(\hat{y}_{\text{text}}, y_{\text{text}}). \tag{2}$$

We combine BCE loss and DICE loss to train the segmentation module:

$$\mathcal{L}_{seg} = \lambda_{\text{BCE}}\text{BCE}(\hat{M}, M) + \lambda_{\text{DICE}}\text{DICE}(\hat{M}, M). \tag{3}$$

Finally, the overall objective can be seen as the weighted sum of the two losses:

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}. \tag{4}$$

4.2 Attention Highlighter via Self-Prompting

While the proposed training paradigm allows the model to localize the region of interest when answering, hallucination can still occur when the model pays little attention to

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Datasets	# Images	# QA pairs	Adversarial questions	Closed-ended	Open-ended	Visual grounding	Annotation
VQA-Rad [Lau *et al.*, 2018]	315	3.5K	✗	✓	✓	✗	✗
SLAKE [Liu *et al.*, 2021]	642	14K	✗	✓	✓	✗†	Mask/Bbox
VQA-Med [Abacha *et al.*, 2021]	5.5K	5.5K	✗	✗	✓	✗	✗
PMC-VQA [Zhang *et al.*, 2023b]	149K	227K	✗	✓	✓	✗	✗
OmniMedVQA [Hu *et al.*, 2024]	118k	128K	✗	✓	✓	✗	✗
CARES [Xia *et al.*, 2024]	18K	41K	✗	✓	✓	✗	✗
Halt-MedVQA [Wu *et al.*, 2024]	1736	2359	✓	✓	✗	✗	✗
ProbMed [Yan *et al.*, 2024]	6.3K	57K	✓	✓	✗	✗	✗
HEAL-MedVQA (Ours)	34K	67K	✓	✓	✓	✓	Mask

Table 1: Comparison of Med-VQA datasets. † indicates visual annotations for all anatomic regions in the image instead of the one questioned.

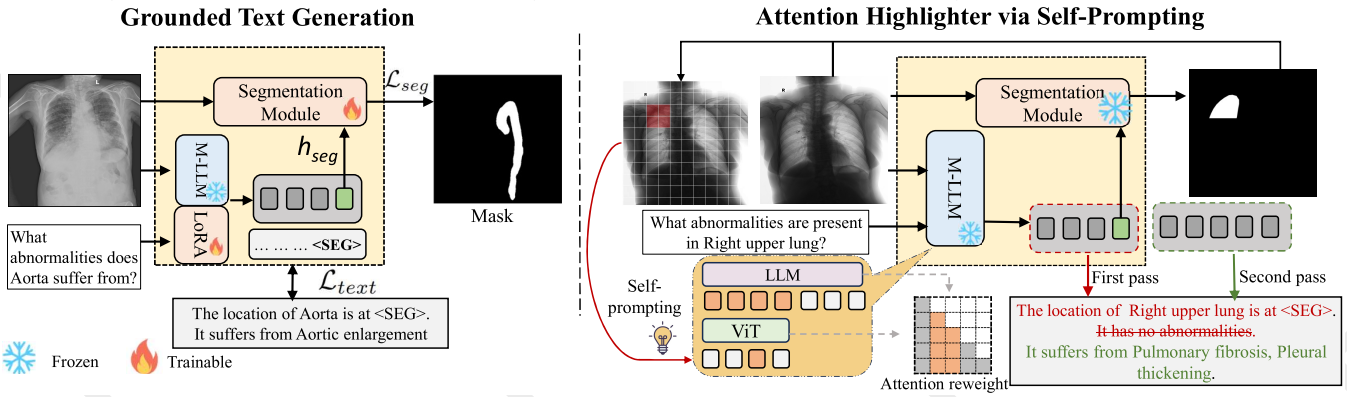

Figure 5: The proposed LoBA frameworks consists of two phases. **Grounded text generation**: during training, our model learns to localize and segment a region of interest mentioned in the question. **Self-prompting**: during inferencing, the model attention is calibrated to attend to the segmented region. With the attention, the model refines and generates visually grounded answers that are more robust to hallucinations.

the image compared to the text. To deal with this, we propose to highlight the attention on the queried regions by self-prompting the segmentation map. Our module consists of two functions: attention reweighting and contrastive decoding.

For *attention reweighting*, patches of interest x_{hl} are added extra weight before feeding through the softmax layer so that those regions get more focus during answer generation. Denote $h_{j,i}$ the original value of the attention logits of patch j to patch i, and β the hyperparameter of added weight, then the new attention weight after reweighting $\hat{h}_{j,i}$ becomes

$$\hat{h}_{j,i} = \begin{cases} h_{j,i} & \text{if } i \notin x_{hl}, \\ h_{j,i} + \log \beta, \beta > 0 & \text{if } i \in x_{hl}. \end{cases} \tag{5}$$

Then the attention probability after Softmax is

$$\tilde{a}_{j,i} = \frac{\beta^{m_i} \exp\left(a_{j,i}\right)}{\sum_k \beta^{m_k} \exp a_{j,k}}. \tag{6}$$

Here, $m_i = 1$, indicating token i is highlighted, and $m_i = 0$ otherwise. We apply this function to the visual backbone of M-LLM where the patches of interest are interpolated from the mask. Applying the reweighted attention $\tilde{A}$, we obtain the highlighted image tokens $\tilde{x}_{\text{img}}$:

$$\tilde{x}_{\text{img}} = \tilde{A}V. \tag{7}$$

To further alleviate shortcut learning, we perform *contrastive decoding* [Zhang *et al.*, 2024b], which contrasts the model decisions after and before highlighting regions of interest. Particularly, we compute the difference between the token probability after highlighting p_{hl} and probability before highlighting p_{bh} to obtain the decoded probability p:

$$\begin{aligned} p_{\text{bh}} &= p(\hat{y}_{\text{text}} | x_{\text{text}}, x_{\text{img}}) \\ p_{\text{hl}} &= p(\hat{y}_{\text{text}} | x_{\text{text}}, \tilde{x}_{\text{img}}) \\ p &= \text{softmax}((1+\alpha)\log p_{\text{hl}} - \alpha \log p_{\text{bh}}). \end{aligned} \tag{8}$$

where α is a hyperparameter controlling the degree of grounding. Applying contrastive decoding minimizes the effect of other tokens on the model output, preventing hallucination.

Discussion. Although Prompt Highlighter [Zhang *et al.*, 2024b] also emphasizes the "highlighting" of regions of interest to reduce hallucination, ours differs in two ways. First, Prompt Highlighter requires human prompting for attention highlighting, which is expensive in medical image analysis. In contrast, ours automatically localizes regions of interest, thus obviating the need for medical professionals to manually localize regions of interest. Second, while Prompt Highlighter only enhances attention on the language decoder, our methodology reweighs attention scores on the visual back-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Model	MIMIC						VinDr					
	Yes/No			Open-ended			Yes/No			Open-ended		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Proprietary Models												
Gemini-1.5-Flash-8B [Team *et al.*, 2023]	0.548	0.520	0.569	0.091	0.127	0.024	0.571	0.604	0.543	0.038	0.098	0.023
GPT-4-O [OpenAI *et al.*, 2023]	0.646	0.706	0.556	0.109	0.143	0.089	0.599	0.660	0.487	0.097	0.087	0.069
GPT-4-Vision [OpenAI *et al.*, 2023]	0.521	0.652	0.409	0.087	0.093	0.067	0.568	0.637	0.424	0.074	0.053	0.072
Open-source Models												
CheXagent [Chen *et al.*, 2024]	0.677	0.621	0.695	0.425	0.450	0.416	0.610	0.628	0.603	0.374	0.342	0.330
BioMedGPT [Zhang *et al.*, 2023a]	0.708	0.734	0.672	0.397	0.387	0.412	0.624	0.606	0.631	0.289	0.265	0.310
MedFlamingo [Alayrac *et al.*, 2022]	0.562	0.574	0.552	0.365	0.390	0.310	0.584	0.600	0.547	0.301	0.285	0.322
LLaVA-Med [Li *et al.*, 2024]	0.722	0.712	0.709	0.551	0.541	0.565	0.703	0.708	0.699	0.523	0.519	0.530
LISA [Lai *et al.*, 2024]	0.716	0.692	0.728	0.524	0.556	0.511	0.707	0.712	0.694	0.521	0.509	0.534
Adversarial Techniques												
PH + LLaVA-Med [Zhang *et al.*, 2024b]	0.735	0.754	0.724	0.563	0.589	0.540	0.716	0.725	0.698	0.522	0.510	0.527
VCD + LLaVA-Med [Leng *et al.*, 2024]	0.716	0.708	0.721	0.566	0.583	0.538	0.652	0.699	0.591	0.501	**0.547**	0.497
CRG + LLaVA-Med [Zhang *et al.*, 2022]	0.731	0.715	0.725	0.544	0.563	0.554	0.689	0.667	0.702	0.509	0.524	0.494
LobA w/o self-prompt	0.727	0.716	0.731	0.544	0.557	0.535	0.724	0.745	0.706	0.511	0.526	0.507
LobA (Ours)	**0.752**	**0.759**	**0.743**	**0.581**	**0.593**	**0.572**	**0.728**	**0.758**	**0.711**	**0.542**	0.533	**0.560**

Table 2: Performance comparison across different models on the proposed HEAL-MedVQA benchmark on MIMIC and VinDr datasets for Yes/No and Open-ended VQA tasks. Text in **bold** and underline highlights the best and second best results, respectively.

bone. The deep intervention on the visual backbone is more effective in mitigating the visual shortcut learning.

5 Experimental Results

In this section, we report the benchmark results of popular medical LMMs and our very own framework, as well as further analysis for our curated adversarial grounding VQA benchmark.

Method	MIMIC		VinDr	
	Anatomy	Disease	Anatomy	Disease
Gemini-1.5	0.532	0.510	0.521	0.487
GPT-4-O	0.594	0.603	0.509	0.596
CheXagent	0.621	0.650	0.689	0.650
BioMedGPT	0.687	0.610	0.623	0.612
LLaVA-Med	0.764	0.762	0.751	0.734
PH + LLaVA-Med	0.783	0.774	0.724	**0.762**
VCD + LLaVA-Med	0.744	0.788	0.740	0.750
CRG + LLaVA-Med	0.731	0.771	0.742	0.738
LobA (Ours)	**0.792**	**0.801**	**0.770**	0.748

Table 3: Textual Perturbation Test on MIMIC and VinDr datasets

	MIMIC	VinDr
Gemini-1.5	0.410	0.403
GPT-4-O	0.557	0.581
CheXagent	0.631	0.671
BioMedGPT	0.647	0.653
LLaVA-Med	0.696	0.680
PH + LLaVA-Med	0.720	0.698
VCD + LLaVA-Med	0.678	0.654
CRG + LLaVA-Med	0.682	0.693
LobA (Ours)	**0.734**	**0.701**

Table 4: Visual Perturbation Test on MIMIC and VinDr datasets

Module		Backbone	LobA w/o self-prompt	LobA (Ours)
Component	Localization	✗	✓	✓
	Self-prompt	✗	✗	✓
Perb Test	TPT	0.763	0.772	**0.797**
	VPT	0.743	0.738	**0.759**
F1 Score	MIMIC (Yes/No)	0.722	0.727	**0.752**
	MIMIC (Open-ended)	0.551	0.544	**0.581**
	VinDr (Yes/No)	0.703	0.724	**0.728**
	VinDr (Open-ended)	0.523	0.511	**0.542**

Table 5: Ablation studies of different components of LobA. The TPT and VPT scores are averages for MIMIC and VinDr datasets.

Benchmarks. We conduct a thorough and systematic evaluation of the most popular Medical Large Multimodal Models on our large-scale HEAL-MedVQA benchmark. Specifically, we utilize 7 state-of-the-art multi-modal LLMs for our evaluation protocol, 3 of which are proprietary models: GPT-4o, GPT-4 Vision [OpenAI *et al.*, 2023] and Gemini 1.5 [Team *et al.*, 2023] and 4 are open source M-LLMs of the medical domain: CheXagent [Chen *et al.*, 2024], LLaVA-Med [Li *et al.*, 2024], BioMedGPT [Zhang *et al.*, 2023a] and Med-Flamingo [Alayrac *et al.*, 2022]. We also assess the capabilities of state-of-the-art adversarial VQA methods, including VCD [Leng *et al.*, 2024], Prompt Highlighter [Zhang *et al.*, 2024b], and CRG [Wan *et al.*, 2025].

Implementation Details. Open source multi-modal LLMs are fine-tuned on our training dataset with LoRA [Hu *et al.*, 2022] before evaluation. The learning rate is tuned from the range of values $\{1e-4, 1e-5, 1e-6\}$. For LobA framework, we used LLaVA-Med [Li *et al.*, 2024] as the pre-trained MLLM and MedSAM [Kirillov *et al.*, 2023] as the segmentation model. Following [Zhang *et al.*, 2024b], we select $\alpha = 0.3$, $\beta_{ViT} = \beta_{LLM} = 2$ as the set of hyperparameters for the Attention Highlighter module. All models

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

are fine-tuned using Transformer, PyTorch and DeepSpeed frameworks on a single 80GB A100 GPU cluster. During inference, the temperature of all multi-modal LLMs is set to 0.1, and the beam size is set to 1.

Evaluation Metrics. To assess the accuracy of the multi-modal LLMs, we utilize LLaMa 3.1-8B to extract information from the M-LLM's output, and categorize them into a set of disease labels. For example, for an open-ended question, if the model returns the answer *"The heart suffers from pneumonia, pulmonary fibrosis and nodule/mass"* then the extracted labels by the LLM will be *pneumonia*, *pulmonary fibrosis*, *nodule/mass*. We report the multi-label precision, recall, and micro F1 score as the main accuracy metrics. For hallucination sensitivity analysis, we adopt our proposed evaluation metrics, Textual Perturbation Test (TPT), and Visual Perturbation Test (VPT) scores, as discussed in Sec. 3.1.

Experimental Results and Discussion. *VQA Accuracy.* As shown in Table 2, proprietary models achieve sub-optimal accuracy. Most models have accuracy less than 50% on binary Yes/No questions and less than 10% for open-ended questions. This shows the challenging nature of our HEAL-MedVQA benchmark, requiring the model to be robust against visual and textual shortcut learning. Our proposed framework consistently outperforms recent advances in M-LLMs and adversarial VQA techniques even after fine-tuning them. Notably, the proposed LoBA outperforms the state-of-the-art Prompt Highlighter [Zhang *et al.*, 2024b] by 3.19% and 3.63% on Open-ended questions for VinDr and MIMIC-CXR, respectively. Injecting grounding prior, as in our proposal, boosts the VQA accuracy.

Textual Perturbation Test. Table 3 reports the sensitivity of each model to textual language bias. The higher the percentage is, the more robust a model is against textual shortcut bias. Our proposed framework achieves the highest sensitivity measure overall, with the ratio of changed answers drastically increasing by up to 3.89% compared to CRG [Wan *et al.*, 2025]. Enforcing visual reasoning via highlighted attention to the queried regions, as in our proposal, alleviates the shortcut learning. Lastly, fine-tuned open-source models are generally more robust than closed-source models like GPT-4-O, showing that fine-tuning LLMs on our adversarial data remedies hallucination.

Visual Perturbation Test. Table 4 reports the visual sensitivity score, with the score defined as the number of changed answers when the original image is overlaid with a healthy localized anatomy. Most multi-modal LLMs in our benchmark shift their answer around 40-70% of the time. Our method is the most sensitive to the change in the localized area, increasing the number of visually dependent answers by 1.94% compared to Prompt Highlighter [Zhang *et al.*, 2024b].

Ablation studies. Table 5 reports the effects of each component in our framework: the LLaVA-Med backbone, Grounded Text Generation (GTG) module, and the full framework with self-prompting. Applying GTG improves the model's performance in some categories, showing the effectiveness of localization-aware training. With the self-prompting framework, the accuracy and textual/visual robustness improve greatly, most evidently increasing 4.45% in TPT score, and 2.73% on MIMIC-CXR dataset Yes/No F1 Score.

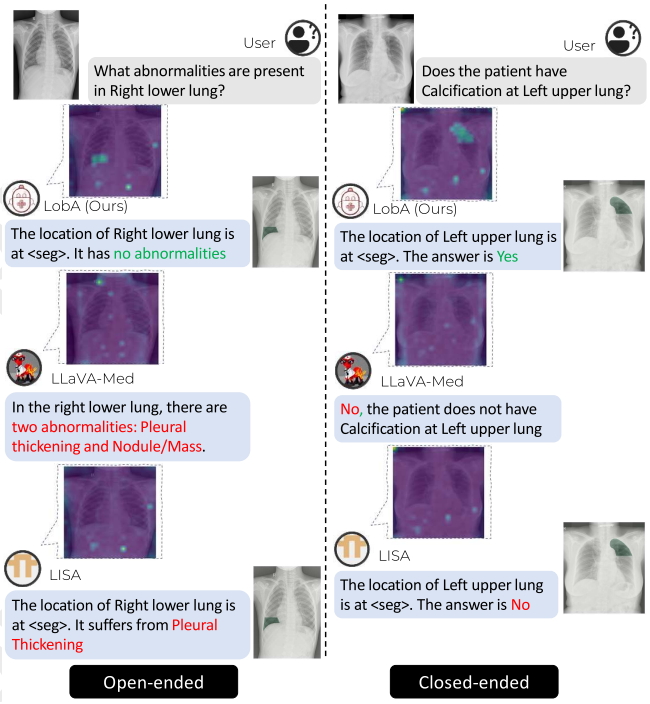

Figure 6: Qualitative case study of both question types

Qualitative analysis. Fig. 6 showcases our *LobA* and some other LMM's response on some qualitative examples, as well as the visual attention maps on the input images. We can see that *LobA* manages to query the attention weights much more intuitively compared to LLaVA-Med or LISA. While others' attentions are scattered into many irrelevant areas, in the first example LobA's main attention weights are allocated around the right lower lung area, and the second example is near the left upper lung. Thanks to the grounding generation module, followed by LobA's self-prompting module to automatically re-weight the attention to the area of interest, the model is able to have better visual reasoning capabilities, leading to more accurate answers compared with other methods.

6 Conclusion

This paper introduces Heal-MedVQA, a new large-scale Medical Visual Question Answering Benchmark with over 67,000 question-answer pairs, which queries diseases at local anatomies, evaluates LMM's capabilities to localize at grounded visual evidence when answering. In addition, we present the Localize-before-Answer (LobA) framework, which trains the LMMs to segment the region of interest and re-adjust their attention for more emphasis on the segmented pathological areas, leading to more reliable answers. Our experimental results showed that our framework LobA outperformed state-of-the-art medical LMMs on Heal-MedVQA, proving its robustness and localization capabilities.

Acknowledgements

We would like to thank Jiangyu Zhou for assisting with the annotation of chest X-ray images for this project.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

References

[Abacha *et al.*, 2019] Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF*, 2019.

[Abacha *et al.*, 2020] Asma Ben Abacha, Vivek Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *CLEF*, 2020.

[Abacha *et al.*, 2021] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF*, 2021.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[Chen and others, 2024] Zhe Chen et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[Chen *et al.*, 2023] Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *arXiv preprint arXiv:2305.10799*, 2023.

[Chen *et al.*, 2024] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[Favero *et al.*, 2024] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *CVPR*, 2024.

[Hasan *et al.*, 2018] Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Henning Müller, and Matthew P. Lungren. Overview of imageclef 2018 medical domain visual question answering task. In *CLEF*, 2018.

[He *et al.*, 2024] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv preprint arXiv:2401.03105*, 2024.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[Hu *et al.*, 2024] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *CVPR*, pages 22170–22183, 2024.

[Huang *et al.*, 2024] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

[Huy *et al.*, 2025] Ta Duc Huy, Duy Anh Huynh, Yutong Xie, Yuankai Qi, Qi Chen, Phi Le Nguyen, Sen Kim Tran, Son Lam Phung, Anton van den Hengel, Zhibin Liao, et al. Seeing the trees for the forest: Rethinking weakly-supervised medical visual grounding. *arXiv preprint arXiv:2505.15123*, 2025.

[Jain *et al.*, 2024] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *CVPR*, pages 27992–28002, 2024.

[Jocher *et al.*, 2020] Glenn Jocher, Alex Stoken, Jirka Borovec, et al. ultralytics/yolov5: v3.1 - bug fixes and performance improvements. 2020.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *CVPR*, 2023.

[Lai *et al.*, 2024] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

[Lau *et al.*, 2018] Jason Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5, 2018.

[Leng *et al.*, 2024] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[Li *et al.*, 2023b] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*. Association for Computational Linguistics, 2023.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Li *et al.*, 2024] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2024.

[Liu *et al.*, 2021] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ISBI*. IEEE, 2021.

[Liu *et al.*, 2023] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024.

[Moor *et al.*, 2023] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*. PMLR, 2023.

[Nguyen *et al.*, 2020] Ha Q. Nguyen, Khanh Lam, Linh T. Le, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2020.

[OpenAI *et al.*, 2023] OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Seibold and others, 2022] Constantin Seibold et al. Detailed annotations of chest x-rays via ct projection for report understanding. In *BMVC*, 2022.

[Stan *et al.*, 2024] Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, et al. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*, 2024.

[Tascon-Morales *et al.*, 2023] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Localized questions in medical visual question answering. In *MICCAI*, pages 361–370. Springer Nature Switzerland, 2023.

[Tascon-Morales *et al.*, 2024] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Targeted visual prompting for medical visual question answering. *arXiv preprint arXiv:2408.03043*, 2024.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wan *et al.*, 2025] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer, 2025.

[Wu *et al.*, 2023a] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical knowledge enhanced language-image pre-training. In *ICCV*, 2023.

[Wu *et al.*, 2023b] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

[Wu *et al.*, 2024] Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. In *ICLR*, 2024.

[Xia *et al.*, 2024] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In *NeurIPS*, 2024.

[Yan and Pei, 2022] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI*, 2022.

[Yan *et al.*, 2024] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*, 2024.

[Zhang *et al.*, 2022] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[Zhang *et al.*, 2023a] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pages arXiv–2305, 2023.

[Zhang *et al.*, 2023b] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, et al. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[Zhang *et al.*, 2024a] Yue Zhang, Wanshu Fan, Peixi Peng, Xin Yang, Dongsheng Zhou, and Xiaopeng Wei. Dual modality prompt learning for visual question-grounded answering in robotic surgery. *Visual Computing for Industry, Biomedicine, and Art*, 7(1):9, 2024.

[Zhang *et al.*, 2024b] Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. In *CVPR*, 2024.