Med-StepBench: A Hierarchical Reasoning Framework for Evaluating Hallucinations in Medical Vision-Language Models

Minh Khoi Nguyen[1], **Dai Lam Le**[1], **Amir Reza Jafari**[2], **Tuan Dung Nguyen**[1], **Hong Son Mai**[3], **Huy Thong Mai**[3], **Quang Huy Nguyen**[4], **Thanh Trung Nguyen**[3], **Reza Farahbakhsh**[2], **Noel Crespi**[2] and **Phi Le Nguyen**[1]
[1]AI4LIFE, Hanoi University of Science and Technology, Vietnam
[2]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France
[3]108 Military Central Hospital, Vietnam
[4]Hanoi Medical University

Abstract

Large vision-language models (VLMs) demonstrate strong performance in medical image understanding, but frequently generate clinically plausible yet incorrect statements, raising significant safety concerns. Existing medical hallucination benchmarks primarily focus on 2D imaging with one-shot diagnostic questions, offering limited insight into whether predictions are grounded in correct localization and abnormality identification, allowing critical reasoning errors to remain hidden behind seemingly correct diagnoses. We introduce *Med-StepBench*, the first large-scale benchmark for step-wise hallucination detection in 3D oncological PET/CT, comprising over 12,000 images and more than 1,000,000 image–statement pairs across volumetric and multi-view 2D data, which decomposes clinical reasoning into four expert-designed diagnostic stages. Using clinician-verified annotations, we perform the first step-level evaluation of general-purpose and medical VLMs, revealing systematic failure modes obscured by aggregate accuracy metrics. Furthermore, we show that current VLMs are highly susceptible to adversarial yet clinically plausible intermediate explanations, which significantly amplify hallucinations despite contradictory visual evidence. Together, our findings highlight fundamental limitations in grounding multi-step clinical reasoning and establish *Med-StepBench* as a rigorous benchmark for developing safer and more reliable medical VLMs.

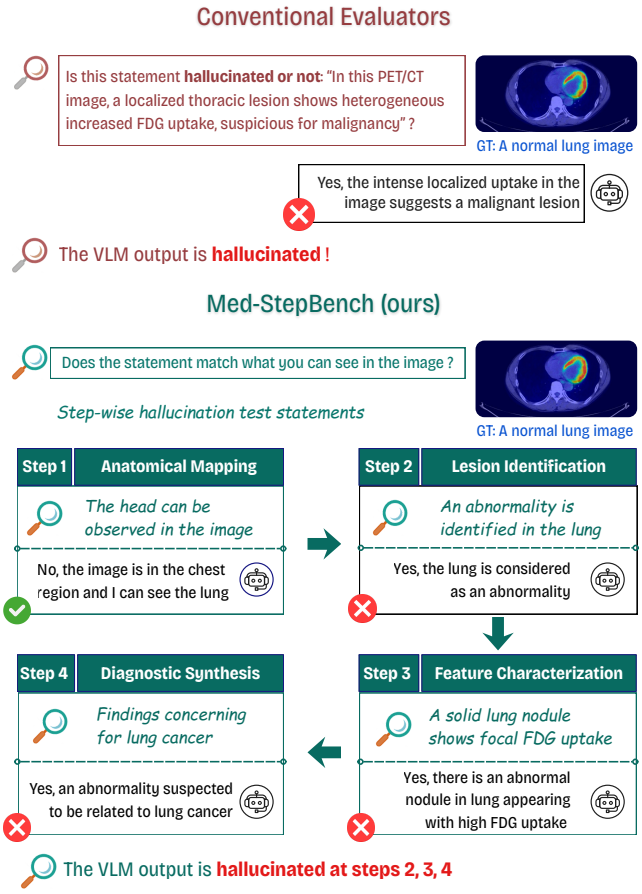

Figure 1: Qualitative examples from Med-StepBench illustrating how clinically plausible diagnoses can mask incorrect intermediate reasoning, and how step-wise evaluation reveals hallucinations.

1 Introduction

Background. Vision-Language Models (VLMs) have emerged as a transformative force in healthcare, offering the potential to revolutionize tasks such as medical visual question answering [Yan *et al.*, 2025; Wu *et al.*, 2023] and automated radiological report generation by integrating multimodal data [Hou *et al.*, 2025]. Despite the rapid scaling of VLMs, characterized by increasingly massive architectures and expansive training datasets, their internal reasoning capabilities and factual reliability remain significant concerns [Jiang *et al.*, 2025]. A primary bottleneck to their clinical integration is the phenomenon of hallucination, where models generate plausible-sounding but clinically incorrect or non-existent information [Asgari *et al.*, 2025]. In high-stakes environments like medicine, where precision is non-

negotiable, such hallucinations pose substantial risks to patient safety [Kim *et al.*, 2025]. To enhance the trustworthiness of medical VLMs, it is no longer sufficient to focus solely on model scaling. There is an urgent need for robust evaluation frameworks that can systematically diagnose and analyze these failures [Zhu *et al.*, 2025]. Beyond MedVQA-centric grounded protocols, recent benchmarks increasingly emphasize probing evaluations that expose hallucinations under distribution shifts and adversarial query designs [Gu *et al.*, 2026; Nguyen *et al.*, 2025a].

Research Gap. While recent efforts have introduced benchmarks to measure medical hallucinations [Zuo and Jiang, 2025], the current landscape suffers from two critical architectural gaps. First, existing benchmarks predominantly focus on 2D modalities, such as X-rays, leaving the evaluation of 3D volumetric imaging, such as CT and PET scans, largely neglected [Liu *et al.*, 2025]. This dimensionality gap is particularly concerning because 3D spatial reasoning is fundamentally more complex, and existing VLMs often exhibit a marked performance degradation when transitioning from 2D slices to 3D volumes. Second, the majority of current evaluation frameworks adopt a reductive "black-box" approach, relying on end-to-end comparisons between a single input and a final output [Wu *et al.*, 2024]. This methodology fails to reflect the professional reality of radiology, where clinicians follow a rigorous, hierarchical sequence: localizing anatomy, detecting abnormalities, analyzing specific features, and only then synthesizing a final conclusion. Consequently, end-to-end metrics cannot reveal the internal point of failure within a model's reasoning chain, rendering it impossible to derive deep insights into the root causes of hallucination.

Our Contribution. To bridge these gaps, we present Med-StepBench, the first evaluation framework designed to assess hallucinations in medical VLMs across both 2D and 3D modalities (specifically CT and PET imaging). Our approach moves beyond terminal outputs to evaluate the model's reasoning across multiple steps, mimicking the granular cognitive process of a radiologist. Figure 1 illustrates the differences between Med-StepBench and existing approaches. Our primary contributions are summarized as follows:

- We introduce Med-StepBench, a novel evaluation framework that shifts from "black-box" end-to-end assessment to a multi-stage diagnostic approach. By decomposing the interpretive process into four granular stages, *Anatomical Mapping, Lesion Identification, Feature Characterization, and Diagnostic Synthesis*, our framework enables systematic identification of where reasoning chains fracture and hallucinations emerge.
- We present a large-scale, expert-curated benchmark specifically designed for 2D and 3D PET/CT imaging. The dataset contains 12.5K and 1,011,847 image–statement pairs, where each statement is derived based on clinicians' meticulous annotations. Unlike existing benchmarks, Med-StepBench includes a balanced set of factual ground truths and strategically designed "hallucination distractors" tailored to each stage of the hierarchical reasoning chain, providing a rigorous testbed for model faithfulness.
- We perform an extensive empirical evaluation of current state-of-the-art medical VLMs. Our analysis provides critical insights into the limitations of current architectures, particularly regarding 3D spatial consistency and the propagation of errors from low-level visual localization to high-level clinical synthesis. These findings establish a baseline for future research into more grounded and reliable medical AI.

2 Related Work

Medical VQA datasets and benchmarks. Medical visual question answering (MedVQA) has been investigated predominantly on 2D radiology or pathology images, where the primary evaluation objective has been recognition of visible findings instead of holistic clinical verification. Classic datasets in this domain include SLAKE [Liu *et al.*, 2021], VQA-RAD [Lau *et al.*, 2018], PathVQA [He *et al.*, 2020], MIMIC-CXR-VQA [Bae *et al.*, 2023], PMC-VQA [Zhang *et al.*, 2023], and VQA-Med (ImageCLEF) [Ben Abacha *et al.*, 2021]. Beyond 2D radiology, ViMed-PET (VQA) [Nguyen *et al.*, 2025b] introduces a Vietnamese 3D PET/CT multimodal dataset and benchmarks, helping extend MedVQA research to low-resource languages. These resources have underpinned advances in medical image understanding and superficial pattern recognition rather than deeper clinical inference. Recognizing this limitation, DrVD-Bench [Zhou *et al.*, 2025] was recently proposed as the first structured benchmark for clinical visual reasoning, consisting of 7,801 image–question pairs spanning 20 task types, 17 diagnostic categories, and five imaging modalities with explicit module decomposition for visual evidence comprehension, reasoning trajectory assessment, and report generation evalua-

Dataset / Benchmark	2D/ 3D	Clinician Step Analysis	Halluc. Detection	Clinician's rationale	Images	VQA Size
Medical VQA datasets / benchmarks						
ViMed-PET (VQA)	3D	✗	✗	✗	2.8K	8.3K
SLAKE	2D	✗	✗	✗	642	14K
VQA-RAD	2D	✗	✗	✗	315	3.5K
PathVQA	2D	✗	✗	✗	5K	32.8K
MIMIC-CXR-VQA	2D	✗	✗	✗	142.8K	377.4K
PMC-VQA	2D	✗	✗	✗	149K	227K
VQA-Med 2021 (ImageCLEF)	2D	✗	✗	✗	5K	5K
DrVD-Bench	2D	✓	✗	✗	– – –	7.8K
Hallucination benchmarks (VQA evaluated)						
Med-HallMark	2D	✗	✓	✗	– – –	– – –
CARES	2D	✗	✓	✗	18K	41K
MedHEval	2D	✗	✓	✗	– – –	15.9K
HEAL-MedVQA	2D	✗	✓	✗	34K	67K
MedVH	2D	✗	✓	✗	– – –	– – –
Med-StepBench (Our)	2D, 3D	✓	✓	✓	12.5K	1,011.8K

Table 1: Comparison of medical image datasets/benchmarks for VQA and hallucination benchmarks. Med-StepBench uniquely combines 2D and 3D images with clinician-aligned stepwise analysis, step-wise hallucination detection, and hallucinated statements with clinician rationales.

tion. While DrVD-Bench reflects an explicit clinical reasoning workflow, its focus remains on alignment between images and corresponding textual reasoning rather than fully representing the complex diagnosis processes required in oncological PET/CT contexts. In contrast, Med-StepBench targets oncologic PET/CT with both 2D and 3D inputs and clinician-aligned stepwise supervision at large scale (∼1,000,000 QA), enabling evaluation beyond 2D-only pattern matching.

Hallucination Detection Benchmarks. Beyond simple QA accuracy, recent benchmarks have shifted toward evaluating the trustworthiness of medical vision-language systems by quantifying hallucination and misleading outputs. CARES [Xia *et al.*, 2024], MedHEval [Chang *et al.*, 2025], and HEAL-MedVQA [Nguyen *et al.*, 2025a] each introduce stress tests or grounded evaluation protocols that target hallucination detection within medical VQA frameworks. In particular, MedHallMark [Chen *et al.*, 2024] formulates a hierarchical categorization of hallucination types and proposes a dedicated evaluative metric (MediHall Score) that scores hallucination severity and type in a graded manner, enabling a more nuanced assessment of potential clinical impact. These benchmarks expose fundamental failure modes of current models, but they generally remain limited to 2D imagery and do not integrate multi-step clinical reasoning with hallucination evaluation under a unified protocol. Additional resources such as the Medical Visual Hallucination Test including MedVH [Gu *et al.*, 2026] offer systematic hallucination assessment across multiple task formulations including multi-choice QA and long textual response generation using chest X-ray based input to evaluate hallucination tendencies of domain-specific LVLMs. A consolidated comparison of the above MedVQA datasets/benchmarks and hallucination-focused evaluations is provided in Table 1. Compared with prior hallucination benchmarks, Med-StepBench jointly supports hallucination detection, multi-step clinician workflow analysis, and hallucinated statement with clinical rationales on PET/CT with 2D/3D views.

3 Hallucination Evaluation Framework

3.1 Overview of Med-StepBench

Med-StepBench consists of two main components: Hallucination evaluation datasets (Subsection 3.2) and a step-wise hallucination evaluation procedure (Subsection 3.4). The Med-StepBench dataset contains *1,011,847* labeled samples spanning a four-step diagnostic workflow, namely Anatomical Mapping, Lesion Identification, Feature Characterization, and Diagnostic Synthesis (Table 2). Based on this dataset, we introduce a four-step hallucination evaluation procedure that mirrors the clinical reasoning pipeline. For each step, we specify a clinically grounded hallucination definition and a step-wise verification protocol to assess whether a VLM's response remains consistent with the visual evidence.

3.2 Dataset Acquisition and Annotation

Every sample in our dataset is composed of PET/CT images and associated explanations. Specifically, the former encompasses a 2D tri-planar image set (axial, coronal, and sagittal) alongside a registered 3D PET/CT volume. The latter consists of step-by-step statements labeled as *ground truth* or *hallucination*. Furthermore, we introduce contextual augmentations by propagating prior knowledge from previous steps or retrieving external patient data via RAG. This structure allows us to analyze the influence of intermediate knowledge on hallucination detection across different reasoning stages. The dataset consists of two parallel languages: the original Vietnamese clinical text and its English counterpart. The English version was translated from Vietnamese using ChatGPT-4o under terminology-preserving constraints to ensure medical entity fidelity and standard radiology phrasing, and was subsequently validated by licensed physicians. In the following sections, we detail the methodology used to construct these two primary components.

Step	# Images	# Image–Question Pairs	Label distribution
Step 1	9,375	981,250	111 : 46
Step 2	4,917	14,751	2 : 1
Step 3	4,917	9,834	1 : 1
Step 4	3,006	6,012	1 : 1
Total	**12,500**	**1,011,847**	2.286 : 1

Table 2: Number of images and image-question pairs per step in Med-StepBench. Label distribution represents the ratio of hallucination labels vs. ground truth labels.

Image Preprocessing and Representation

We retrospectively collected whole-body oncologic PET/CT studies from a tertiary referral hospital in Vietnam as paired PET-CT DICOM series with finalized radiology reports. Under institutional ethics oversight, all scans were de-identified at the DICOM header level. We excluded studies with missing/corrupted series or incomplete whole-body coverage, retaining *12,500* matched PET-CT volume pairs with reports. For 2D standardization, CT volumes were windowed using a *lung window* for the thorax and a *mediastinal (soft-tissue) window* elsewhere, then rendered as axial slices and large-field sagittal/coronal reformats. PET and CT were fused via a clinically established overlay strategy (Fig. 2) consistent with routine FDG PET/CT reporting [Hofman and Hicks, 2016].

Statement Generation and Labeling

For each reasoning step, we first select the step-specific visual evidence to define the observable input, then generate a step-aligned factual statement grounded in that evidence.

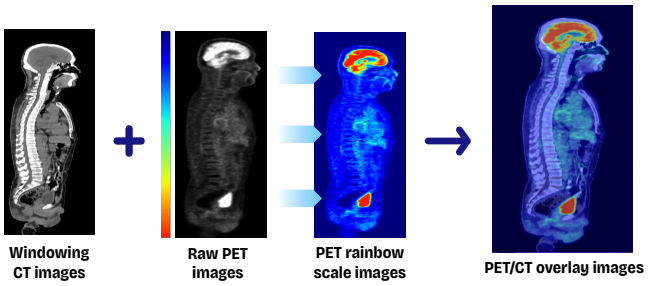

Figure 2: PET/CT overlay method used in routine oncologic FDG PET/CT reporting [Hofman and Hicks, 2016].

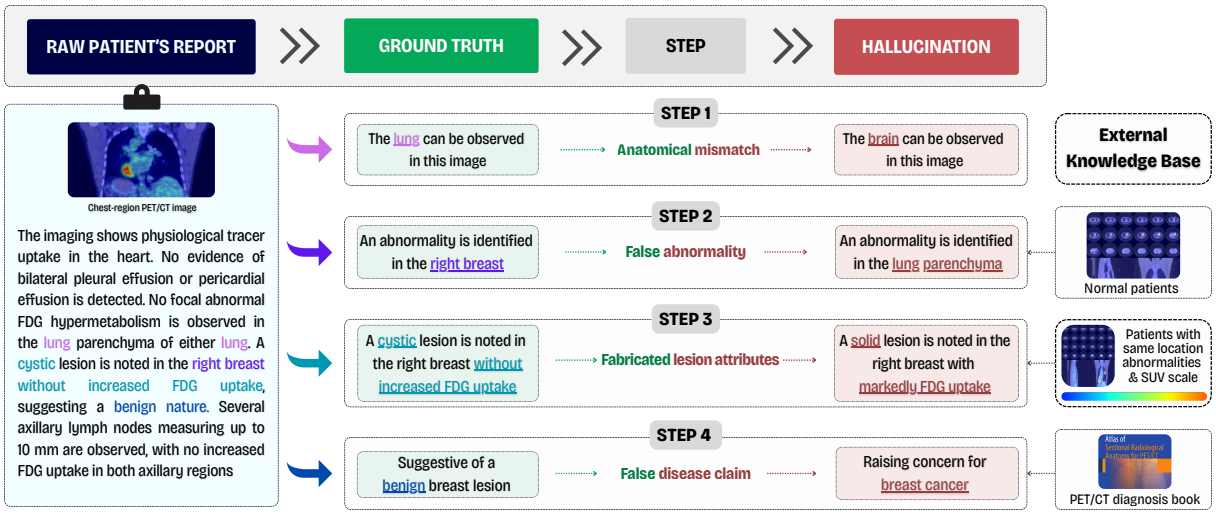

Figure 3: *Step-wise hallucination evaluation framework*. Decomposing each PET/CT case into four sequential reasoning stages and extracting clinician-verified step-level ground-truth statements. For each step, we generate paired hallucinated statements by introducing anatomically plausible but incorrect edits (mismatch, false abnormality, fabricated attributes, false disease claim), enabling fine-grained evaluation of grounding across the full reasoning pipeline.

The resulting statement is subsequently assigned as either *ground truth* (unchanged) or *hallucination* (corrupted by a step-specific operator) to yield a plausible but incorrect variant consistent with the step's hallucination definition. For later steps, each sample can be further augmented with contextual information: *prior knowledge* is derived from the previous step's ground-truth statement to ensure step-appropriate alignment before being propagated to the next step, whereas *external knowledge* is obtained from images and reports of referenced patient cases or via retrieval-augmented generation (RAG) over curated medical documents. Depending on the evaluation setting, an optional rationale (LLM-generated or clinician-written) may also be attached. The procedure for generating hallucinated statements comprises four steps, as illustrated in Fig. 3 and detailed in the following.

(1) Anatomical mismatch: We create region-incompatible statements by pairing images from one anatomical field-of-view with organ mentions from a different region, and ask whether that structure appears in the provided image or not.

(2) False abnormality / Missing abnormality: Using the report to identify truly abnormal regions for each patient, we generate statements that either claim that a different pathology is observed in the same or nearby region, or deny any abnormality despite documented disease.

(3) Fabricated lesion attributes: We extract lesion-level attributes described in the report (e.g. SUV value, size, uptake descriptors,..) and replace them with incompatible values chosen to induce a large discrepancy.

(4) False disease claim: Starting from the report-confirmed diagnosis or disease entities, we synthesize statements that assert entirely different diseases that are not supported by the imaging evidence for that case.

3.3 Step-wise Hallucination Definition

Leveraging the aforementioned evaluation dataset, we present a comprehensive procedure to assess hallucination susceptibility in VLMs. Our approach mirrors the multi-stage diagnostic reasoning of a radiologist, consisting of four steps:

Step 1: Anatomical Mapping Hallucination The first step in our procedure assesses whether VLMs can correctly ground anatomical region recognition in the visual evidence. This step is motivated by the observation that many hallucinations in VLMs arise not from a lack of domain knowledge, but from failures to correctly align textual outputs with the spatial and anatomical constraints of the image [Sun *et al.*, 2025]. Hallucination in this step is defined as follows:

Spatial Mislocalization: The model correctly identifies a real anatomical structure but assigns it to an incorrect spatial region. For instance, asserting the presence of abdominal organs (e.g., kidneys) in images belonging to the head–neck region reflects a failure to map anatomical entities to their correct body locations.

Visual Confusion: The model visually confuses morphologically similar patterns (e.g., soft tissue contours, high FDG uptake) and maps them to an incorrect anatomical class.

Step 2: Lesion Identification Hallucination We systematically evaluate the hallucinatory tendencies of medical VLMs regarding pathological lesion detection and the verification of their absence. This step targets a more advanced level of visual–semantic grounding than Step 1, as lesion identification requires not only correct anatomical localization but also discrimination between pathological findings and normal or physiological imaging patterns. Errors at this stage are particularly critical in medical settings, as they may lead to false positives or false negatives in downstream clinical reasoning.

Lesion Fabrication: The model asserts the presence of a lesion in an image where no annotated lesion exists. This

reflects over-generation of pathological findings driven by lesion-rich training data or an implicit abnormality bias.
Phantom Lesion Hallucination: Physiological uptake or imaging artifacts are hallucinated as pathological findings. The model incorrectly interprets physiological tracer uptake, normal anatomical variations, or imaging artifacts as pathological lesions.
Lesion Mislocalization: The model correctly recognizes that a pathological lesion exists. However, it assigns the lesion to an incorrect anatomical structure or organ.

Step 3: Feature Characterization Hallucination This step moves beyond lesion existence and localization (Step 2) to examine whether models can faithfully ground fine-grained lesion attributes in measurable imaging signals. Feature characterization is particularly prone to hallucination, as models may generate plausible yet unverified numerical values or transfer attributes across parts of an organ based on learned priors rather than image-derived measurements.
Numeric Hallucination: The model generates quantitative values (e.g., lesion diameter, SUV_{max}, radiomic features) that do not correspond to any measured or inferable data from the image. This includes reporting specific measurements when such values cannot be reliably extracted from the provided visual input.
Attribute Omission: The relevant feature (increased FDG uptake) is genuinely present in the image. However, the model fails to recognize or acknowledge this feature and instead asserts its absence.
Spatial Hallucination: This type of hallucination reflects a failure in fine-grained spatial grounding, rather than errors in organ recognition or lesion existence. It occurs when a model correctly identifies an anatomical structure or lesion within the correct organ, but misrepresents its relative spatial position, such as confusing *left vs. right* or *superior vs. inferior* within that organ.

Step 4: Diagnostic Synthesis Hallucination The final step represents the most complex reasoning stage in the proposed framework. Unlike earlier steps, which focus on perception, localization, or feature attribution, diagnostic synthesis requires multi-hop reasoning and adherence to clinical plausibility constraints. Hallucinations at this stage often emerge when models overextend beyond observable findings, inferring disease entities or clinical stages that are not justified by the image.
Reasoning Hallucination: The model correctly perceives key visual elements but produces conclusions or interpretations that are not supported by the observed evidence.
Cognitive Bias Hallucination: The model disproportionately predicts certain diseases or conditions regardless of whether they are supported by the visual evidence.

3.4 Hallucination Evaluation Procedure

Building on the dataset design and the step-specific hallucination definitions, we evaluate VLMs' hallucination by measuring their ability to recognize and reject hallucinations during clinical image interpretation. Each medical image is paired with a statement that is either factual or hallucinated. Models are required to judge whether the statement is consistent with the visual evidence. A prediction is considered correct if the model's judgment matches the ground-truth consistency label, and incorrect otherwise. Evaluation is conducted independently at four sequential clinical reasoning steps. Each step is associated with its own hallucination definition, reflecting increasing clinical abstraction and reasoning complexity. Optionally, during testing we provide additional knowledge context, including either the model's prior knowledge or external retrieved knowledge, to examine how knowledge injection affects visual-textual consistency. Performance is reported using Precision, Recall, and F1-score, capturing hallucination propensity, detection sensitivity, and their overall balance.

4 Benchmarking Hallucinations in Contemporary Vision-Language Models

To investigate the problem of hallucination in sequential clinical reasoning, we design a set of controlled experiments on *Med-StepBench*. The goal is to use our evaluation framework to systematically evaluate how VLMs detect hallucinations across different stages of clinical reasoning, and to analyze the factors that influence visual–textual consistency.

4.1 Hallucination Evaluation Setup

Benchmarked Models. We evaluate a diverse model suite spanning 2D general-purpose VLMs, 2D medical VLMs, and 3D medical VLMs. The benchmarked 2D general VLMs include frontier natively multimodal systems (GPT-4o [OpenAI, 2024] & Gemini 2.5 [Comanici *et al.*, 2025]) that represent strong state-of-the-art baselines for multimodal reasoning. We additionally include competitive open-source VLMs (LLaVA-7B [Liu *et al.*, 2023], Qwen3-VL [Bai *et al.*, 2025]) to assess robustness across model families with strong visual recognition and grounding capabilities. For domain-specialized evaluation, we consider medical-adapted VLMs (LLaVA-Med [Li *et al.*, 2023] & Med-Flamingo [Moor *et al.*, 2023]) that are trained or fine-tuned to better handle biomedical imagery and medical dialog. Finally, we benchmark 3D medical VLMs (MedM-VL [Shi *et al.*, 2025], M3D-LaMed [Bai *et al.*, 2024]) to study hallucination behaviors when reasoning over volumetric representations. Overall, this model suite enables a comprehensive comparison between general vs. medical adaptation and 2D vs. 3D perception under our step-wise hallucination evaluation protocol.
Research Questions. Our experiments are designed to answer the following main research questions:
RQ1: Step-wise Hallucination in 2D and 3D VLM Clinical Reasoning. How does hallucination detection performance differ between 2D and 3D VLMs across sequential clinical reasoning steps?
RQ2: Vulnerability to Plausible Rationales. To what extent are VLMs misled by clinically plausible but incorrect rationales during consistency judgment?
RQ3: Prior Knowledge vs. External Knowledge. Does injecting intermediate or external knowledge into prompts improve hallucination detection over no-knowledge settings?
RQ4: Effect of PET/CT Fusion. How does PET/CT fusion contribute to hallucination mitigation at the region level?

Model	Know.	Step 1				Step 2				Step 3				Step 4			
		EN		VI		EN		VI		EN		VI		EN		VI	
		R	F1	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1	R	F1
2D General VLM																	
GPT 4o	w/o	**0.98**	**0.98**	**1.00**	**0.98**	0.76	0.73	0.80	0.73	0.85	0.69	0.96	0.70	0.65	0.71	0.59	0.71
	w	–	–	–	–	**0.83**↑	**0.74**↑	**0.84**↑	**0.74**↑	0.85	0.72↑	0.96	0.70	0.78↑	0.72↑	0.83↑	0.72↑
Gemini 2.5 flash	w/o	0.96	0.98	0.97	0.98	0.82	0.72	0.90	0.71	0.84	0.67	0.74	0.67	0.58	0.61	0.50	0.62
	w	–	–	–	–	0.80↓	0.72	0.76↓	0.68↓	0.78↓	0.67	0.74	0.69↑	0.77↑	0.65↑	0.79↑	0.68↑
Gemini 2.5 pro	w/o	1.00	0.98	1.00	0.97	0.72	0.72	0.72	0.72	0.70	0.72	0.68	0.70	0.72	0.73	0.71	0.69
	w	–	–	–	–	0.70↓	0.73↑	0.71↓	0.69↓	**0.76**↑	**0.73**↑	**0.79**↑	**0.69**↓	**0.77**↑	**0.76**↑	**0.65**↓	**0.75**↑
Qwen3-VL-8B	w/o	0.76	0.86	0.68	0.81	0.80	0.69	0.80	0.70	0.76	0.59	0.65	0.57	0.69	0.65	0.54	0.63
	w	–	–	–	–	0.82↑	0.73↑	0.92↑	0.73↑	0.74↓	0.64↑	0.76↑	0.64↑	0.83↑	0.65	0.81↑	0.68↑
Qwen3-VL-4B	w/o	0.96	0.92	0.92	0.89	0.74	0.65	0.73	0.64	0.70	0.62	0.86	0.63	0.65	0.63	0.65	0.62
	w	–	–	–	–	0.84↑	0.69↑	0.84↑	0.69↑	0.82↑	0.67↑	0.78↓	0.67↑	0.88↑	0.67↑	0.79↑	0.72↑
LLaVA-7B	w/o	0.53	0.61	–	–	0.58	0.50	–	–	0.63	0.52	–	–	0.71	0.57	–	–
	w	–	–	–	–	0.85↑	0.68↑	–	–	0.55↓	0.56↑	–	–	0.76↑	0.62↑	–	–
Medical VLM																	
LLaVA-Med	w/o	0.50	0.62	–	–	0.42	0.54	–	–	0.64	0.56	–	–	0.32	0.42	–	–
	w	–	–	–	–	0.36↓	0.46↓	–	–	0.40↓	0.43↓	–	–	**0.61**↑	**0.65**↑	–	–
MedM-VL 2D	w/o	0.80	0.64	0.56	0.50	**0.75**	**0.71**	**0.70**	**0.69**	0.38	0.40	0.54	0.53	0.74	0.61	0.79	0.61
Med-Flamingo	w/o	**0.82**	**0.77**	–	–	0.69	0.66	–	–	**0.82**	**0.63**	–	–	0.67	0.59	–	–
	w	–	–	–	–	0.74↑	0.69↑	–	–	0.82	0.59↓	–	–	0.71↑	0.62↑	–	–
3D Medical VLM																	
MedM-VL 3D	w/o	**0.76**	**0.66**	**0.90**	**0.63**	**0.70**	**0.67**	**0.96**	**0.65**	**0.42**	**0.51**	**0.50**	**0.54**	**0.56**	**0.59**	**0.26**	**0.39**
M3D	w/o	0.54	0.60	–	–	0.52	0.55	–	–	0.32	0.38	–	–	0.54	0.57	–	–

Table 3: Hallucination results over Med-StepBench (Recall and F1 score) with/without knowledge in both English & Vietnamese; arrows indicate change from w/o to w. The performance generally decreases from Step 1 to Step 4, reflecting the increasing difficulty of later-stage clinical reasoning. ↑ and ↓ denote performance increase and decrease, respectively, when comparing with-knowledge to without-knowledge settings. **Bold** and underline indicate the best and second best performing models per each group of models for each step.

Implementation Details. Experiments are conducted in English and Vietnamese using identical images and ground-truth labels, with only the language of the statements and prompts changed, enabling controlled analysis of language effects on hallucination detection. All models are evaluated using a temperature in the range of 0.2–0.3 and a top-p value of 0.9.

4.2 Result and Analysis

Step-wise Hallucination Analysis

RQ1. Table 3 reports hallucination detection results for popular VLMs, across 2D/3D architectures, as well as general/medical models. Results show a consistent performance drop from Step 1 to later stages, indicating increasing hallucination effects as tasks progress from simple anatomical recognition to more fine-grained analysis and clinical reasoning. We observe that Step 3, which involves describing basic attributes of identified lesions, constitutes a pronounced bottleneck for volumetric and PET/CT-focused models. For example, MedM-VL 2D drops to an F1 score of 0.40 (in EN) at Step 3, and then recovers to 0.61 at Step 4, corresponding to final diagnostic reasoning. Similarly, MedM-VL 3D and M3D achieve their lowest F1 scores at Step 3 (0.51 and 0.38, respectively). In contrast, several 2D models exhibit their weakest performance at Step 4, the diagnostic hallucination recognition stage, including Gemini 2.5 Flash (0.61), Qwen3-VL-4B (0.63), LLaVA-Med (0.42), and Med-Flamingo (0.59). Overall, Step 3 exposes failures in fine-grained attribute and quantitative grounding, while Step 4 stresses long-horizon clinical synthesis. Both stages represent significant bottlenecks in medical hallucination, substantially reducing VLMs' ability to identify incorrect diagnoses. We observe notable differences between English and Vietnamese settings. Although overall trends are similar, performance on Vietnamese prompts shows less stability across steps and models, particularly at Step 3. In several cases, English prompts yield more stable and slightly higher F1 scores across reasoning stages, while Vietnamese performance fluctuates more. This suggests that hallucination detection is more sensitive to Vietnamese prompt formulation, especially for fine-grained attribute descriptions.

Vulnerability to Plausible Rationales

RQ2. When a textual rationale is introduced, hallucination detection is no longer driven purely by visual evidence; instead, the additional language steers the consistency judgment, weakening image grounding across models (Table 4). This mainly reduces recall, indicating that models increasingly accept incorrect statements as image-consistent when a rationale is present. Clinician-written rationales are gener-

Setting	Med-Flamingo			Gemini 2.5 Pro			Gemini 2.5 Flash			MedM-VL 3D			MedM-VL 2D		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
No rationale	0.49	0.84	0.62	0.50	0.88	0.64	0.50	0.96	0.66	0.48	0.90	0.63	0.38	0.60	0.46
LLM rationale	**0.20**	**0.24**	**0.22**	0.51	0.96	0.67	0.53	0.98	0.69	0.47	0.86	0.60	**0.36**	**0.57**	**0.45**
Clinician's rationale	0.43	0.76	0.55	**0.48**	**0.70**	**0.57**	**0.47**	**0.74**	**0.57**	**0.42**	**0.72**	**0.53**	0.37	0.58	**0.45**

Table 4: Comparison of VLMs' performance across different input rationales. Models become more susceptible to hallucinations when conditioned on clinician-written rationales, likely because these explanations closely match medical styles and priors learned during training.

ally more harmful than LLM-generated ones. Their clinically standard, fluent, and report-like style makes them more persuasive to the model's language prior, leading to missed hallucinations. This is evident for Gemini 2.5 Pro/Flash, where F1 decreases from 0.67/0.69 to 0.57/0.57 under clinician rationales, or an F1 drop from 0.60 to 0.53 for MedM-VL 3D. In contrast, LLM-generated rationales do not induce a consistent degradation. Gemini 2.5 Pro/Flash maintain high recall even under LLM rationales (0.96/0.98), achieving F1 scores up to 0.69. This robustness likely stems from LLM rationales being less tightly constrained by real clinical practice, often appearing generic or insufficiently persuasive to override visual evidence. Med-Flamingo is an exception, showing strong sensitivity even to LLM rationales, however the model still suffered a major decrease in hallucination detection capabilities. These results reveal a persuasion-over-evidence failure mode, where clinically plausible narratives override visual verification, exposing a weakness in visual grounding.

Effects of Prior and External Knowledge

RQ3. Table 5 shows that knowledge injection for both types only yields marginal F1 changes for most models in most steps, and does not meaningfully alleviate hallucination. Step 2 generally yields small F1 gains or remains unchanged, indicating that contextual constraints align well with lesion presence verification, though occasional degradations (e.g., LLaVA-Med) suggest limited robustness. In contrast, Step 3 shows little improvement and frequent drops, highlighting a persistent bottleneck where lightweight prompts fail to support fine-grained attribute grounding and models revert to textual priors. Step 4 typically improves or stays flat, but gains are largely driven by a precision–recall trade-off, with higher sensitivity accompanied by increased false positives. Overall, intermediate knowledge primarily redistributes errors rather than achieving consistent hallucination reduction, benefiting lesion identification more than feature characterization.

Clinically Aligned PET/CT Fusion

RQ4. As shown in Table 6, clinically fused PET/CT inputs consistently outperform simple PET+CT concatenation across models. For Gemini 2.5 Flash, PET/CT fusion yields a large gain at Step 1 (VI), improving F1 from 0.79 to 0.98, and also improves fine-grained reasoning at Step 3 (EN) from 0.62 to 0.67. Similarly, Qwen3-VL-4B benefits substantially from PET/CT fusion, with Step 3 (EN) F1 increasing from 0.20 to 0.63, more than a threefold improvement. These results indicate that clinically grounded PET/CT fusion significantly enhances cross-modal alignment and fine-grained reasoning compared to naïve modality concatenation.

Model	Knowledge		Step 2			Step 3			Step 4		
	Prior	Ext	P	R	F1	P	R	F1	P	R	F1
Qwen3 VL-4B			0.58	0.74	0.65	0.56	0.70	0.62	0.62	0.65	0.63
	✓		0.53	**0.94**	0.68	0.58	0.72	0.64	**0.63**	0.83	**0.72**
		✓	**0.60**	0.88	**0.71**	0.54	0.78	0.64	0.55	0.50	0.52
	✓	✓	0.59	0.84	0.69	**0.59**	0.78	**0.67**	0.55	**0.88**	0.67
Med Flamingo			0.63	0.69	0.66	**0.51**	0.82	0.63	0.53	0.67	0.59
	✓		0.55	0.94	0.70	0.50	**0.88**	**0.64**	0.54	**0.79**	**0.64**
		✓	0.55	**0.96**	**0.70**	0.47	0.80	0.59	0.52	0.77	0.62
	✓	✓	**0.65**	0.74	0.69	0.47	0.82	0.59	**0.55**	0.71	0.62
Gemini 2.5 Flash			0.65	0.82	**0.72**	0.56	**0.84**	**0.67**	**0.64**	0.58	0.61
	✓		0.52	0.65	0.58	0.50	0.58	0.54	0.62	0.63	0.63
		✓	0.63	**0.85**	**0.72**	0.57	0.82	**0.67**	0.60	0.72	**0.65**
	✓	✓	**0.66**	0.80	**0.72**	**0.58**	0.78	**0.67**	0.57	**0.77**	**0.65**

Table 5: Different knowledge input across steps. Prior/Ext indicate whether prior knowledge or external knowledge is provided (both checked means +Both; none checked means no knowledge).

Model	Input	Step 1		Step 2		Step 3		Step 4	
		VI	EN	VI	EN	VI	EN	VI	EN
Med Flamingo	PET + CT	–	0.64	–	0.64	–	**0.66**	–	**0.66**
	PET/CT	–	**0.77**↑	–	**0.66**↑	–	0.63↓	–	0.59↓
Gemini 2.5 Flash	PET + CT	0.79	0.81	0.68	**0.72**	0.62	0.62	0.51	0.52
	PET/CT	**0.98**↑	**0.98**↑	**0.72**↑	0.71↓	**0.67**↑	**0.67**↑	**0.61**↑	**0.62**↑
Qwen3 VL-4B	PET + CT	0.60	0.79	0.31	0.34	0.29	0.20	**0.67**	**0.64**
	PET/CT	**0.92**↑	**0.89**↑	**0.68**↑	**0.65**↑	**0.66**↑	**0.62**↑	0.63↓	0.62↓

Table 6: Comparison of F1-score between PET+CT (separate) and fused PET/CT across steps in both English and Vietnamese.

5 Conclusion

This paper introduced *Med-StepBench*, a large-scale step-wise hallucination detection benchmark for 2D and 3D oncological PET/CT, comprising over 1 million image–statement pairs across four clinically grounded diagnostic stages. Med-StepBench enabled detailed examination of whether model predictions were supported by correct anatomical localization, lesion identification, feature characterization, and diagnostic reasoning. Through extensive experiments on general-purpose and medical VLMs, we showed that hallucinations emerged and propagated differently across reasoning stages, and that current models were particularly susceptible to clinically plausible but incorrect intermediate explanations. Our results highlighted critical limitations in visual grounding and multi-step clinical reasoning, positioning *Med-StepBench* as a practical benchmark for advancing safer and more reliable medical VLMs.

Ethical Statement

This study was conducted in accordance with established medical research ethics standards and institutional regulations. All imaging data were retrospectively collected and fully de-identified under approval of the Institutional Review Board (IRB), with a formally granted waiver of informed consent due to minimal risk. Personal protected information (PPI) was removed following international privacy standards (e.g., HIPAA Safe Harbor). The dataset will be publicly released for non-commercial research purposes under specified usage terms upon paper acceptance.

Acknowledgements

This research is partly supported by Toray Industries (H.K.) Vietnam Co., Ltd. This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2024-TĐ-002.

Contribution Statement

Minh Khoi Nguyen and Dai Lam Le are co-first authors. Noel Crespi and Phi Le Nguyen are corresponding authors.

References

[Asgari *et al.*, 2025] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274, 2025.

[Bae *et al.*, 2023] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880, 2023.

[Bai *et al.*, 2024] Fan Bai, Yuxin Du, Tiejun Huang, Max Qinghu Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024.

[Bai *et al.*, 2025] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025.

[Ben 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 Conference and Labs of the Evaluation Forum - Working Notes*, 2021.

[Chang *et al.*, 2025] Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. Medheval: Benchmarking hallucinations and mitigation strategies in medical large vision–language models, 2025.

[Chen *et al.*, 2024] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models, 2024. Med-HallMark benchmark.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[Gu *et al.*, 2026] Zishan Gu, Jiayuan Chen, Fenglin Liu, Changchang Yin, and Ping Zhang. Medvh: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems*, 8(1):2500255, 2026.

[He *et al.*, 2020] Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint*, 2020.

[Hofman and Hicks, 2016] Michael S. Hofman and Rodney J. Hicks. How we read oncologic FDG PET/CT. *Cancer Imaging*, 16(1):35, 2016.

[Hou *et al.*, 2025] Xiaodi Hou, Xiaobo Li, Mingyu Lu, Simiao Wang, and Yijia Zhang. Rrg-mamba: Efficient radiology report generation with state space model. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7410–7418. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

[Jiang *et al.*, 2025] Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 403–411. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

[Kim *et al.*, 2025] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025.

[Lau *et al.*, 2018] Jason J 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(1):1–10, 2018.

[Li *et al.*, 2023] 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, 2023.

[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 *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Liu *et al.*, 2025] Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16448–16460, 2025.

[Moor *et al.*, 2023] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Medflamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[Nguyen *et al.*, 2025a] Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7670–7678. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

[Nguyen *et al.*, 2025b] Huu Tien Nguyen, Dac Thai Nguyen, Duc Nguyen The Minh, Trung Thanh Nguyen, Thao Nguyen Truong, Hieu Pham, Johan Barthelemy, Tran Minh Quan, Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Mai Hong Son, Chau Quynh Anh, Thanh Trung Nguyen, and Phi Le Nguyen. Toward a vision-language foundation model for medical data: Multimodal dataset and benchmarks for vietnamese PET/CT report generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[OpenAI, 2024] OpenAI. Gpt-4o system card, 2024.

[Shi *et al.*, 2025] Yiming Shi, Shaoshuai Yang, Xun Zhu, Haoyu Wang, Xiangling Fu, Miao Li, and Ji Wu. Medm-vl: What makes a good medical lvlm?, 2025.

[Sun *et al.*, 2025] Xiaoxi Sun, Jianxin Liang, Yueqian Wang, Huishuai Zhang, and Dongyan Zhao. Understanding visual detail hallucinations of large vision-language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1900–1908, 2025.

[Wu *et al.*, 2023] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

[Wu *et al.*, 2024] Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. In *The Second Tiny Papers Track at ICLR 2024*, 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. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

[Yan *et al.*, 2025] 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19188–19205, 2025.

[Zhang *et al.*, 2023] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint*, 2023.

[Zhou *et al.*, 2025] Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. DrVD-bench: Do vision-language models reason like human doctors in medical image diagnosis? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[Zhu *et al.*, 2025] Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Zuo and Jiang, 2025] Kaiwen Zuo and Yirui Jiang. Medhallbench: A new benchmark for assessing hallucination in medical large language models. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pages 205–213. PMLR, 2025.