

# PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering

Xiaoman Zhang<sup>\*,1,2</sup>, Chaoyi Wu<sup>\*,1,2</sup>, Ziheng Zhao<sup>1,2</sup>, Weixiong Lin<sup>1,2</sup>,  
Ya Zhang<sup>1,2</sup>, Yanfeng Wang<sup>1,2,†</sup> and Weidi Xie<sup>1,2,†</sup>

<sup>1</sup>Shanghai Jiao Tong University      <sup>2</sup>Shanghai AI Laboratory

Medical Visual Question Answering (MedVQA) presents a significant opportunity to enhance diagnostic accuracy and healthcare delivery by leveraging artificial intelligence to interpret and answer questions based on medical images. In this study, we reframe the problem of MedVQA as a generation task that naturally follows the human-machine interaction and propose a generative-based model for medical visual understanding by aligning visual information from a pre-trained vision encoder with a large language model. We establish a scalable pipeline to construct a large-scale medical visual question-answering dataset, named PMC-VQA, which contains 227k VQA pairs of 149k images that cover various modalities or diseases. We train the proposed model on PMC-VQA and then fine-tune it on multiple public benchmarks, *e.g.*, VQA-RAD, SLAKE, and Image-Clef-2019, significantly outperforming existing MedVQA models in generating relevant, accurate free-form answers. In addition, we propose a test set that has undergone manual verification, which is significantly more challenging, serving to better monitor the development of generative MedVQA methods. To facilitate comprehensive evaluation and comparison, we have maintained a leaderboard at <https://paperswithcode.com/paper/pmc-vqa-visual-instruction-tuning-for-medical>, offering a centralized resource for tracking progress and benchmarking state-of-the-art approaches. The PMC-VQA dataset emerges as a vital resource for the field of research, and the MedVInT presents a significant breakthrough in the area of MedVQA.

## 1 Introduction

Large language models (LLMs), such as GPT-4 [43], Med-PaLM [51], PMC-LLaMA [57] have recently achieved remarkable success in the field of medical natural language processing [24, 27, 42]. While recent LLMs excel in language understanding in the medical domain, they are essentially “blind” to visual modalities, such as images and videos, hindering the use of visual content as inputs. This limitation is particularly evident in the Medical Visual Question Answering (MedVQA) domain, where there is a critical need for models to interpret medical visual content to answer text-based queries accurately [33].

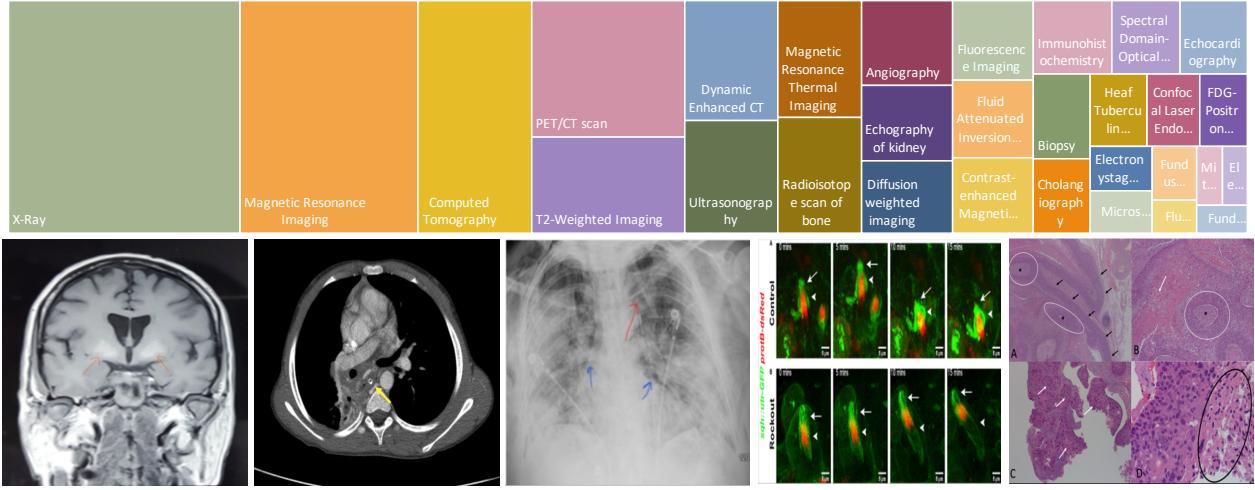
MedVQA is an important and emerging field at the intersection of artificial intelligence and healthcare, which involves developing systems that can understand and interpret medical images and provide relevant answers to questions posed about these images. By integrating AI with medical expertise, MedVQA aims to significantly impact healthcare outcomes, patient care, and medical science [60, 53]. For example, the MedVQA system can enhance diagnostic accuracy for clinicians, improve patient understanding of medical information, and advance medical education and research.

However, existing MedVQA methods [40, 34, 11, 32] typically treat the problem as a retrieval task with a limited answer base and train multi-modal vision-language models with contrastive or classification objectives. Consequently, they are only useful for limited use cases where a finite set of outcomes is provided beforehand. We propose to develop the *first* open-ended MedVQA system with a generative model as the backend, capable of handling diverse questions that arise in clinical practice, generating answers in free form without being constrained by the vocabulary. While there has been promising research in visual-language representation learning, such as Flamingo [1] and BLIP [30], these models have primarily been trained on natural language and images, with very limited application in the medical domain, due to the complex and nuanced visual concepts often found in medical scenarios.

To effectively train the generative-based models, our study reveals that existing datasets are limited in size, making them insufficient for training high-performing models. we leverage well-established medical

\* Equal contributions. Email addresses: {wtzxxwcy02, xm99sjtu, weidi}@sjtu.edu.cn

† Corresponding author.



**Figure 1 |** (a) Several examples of challenging questions and answers along with their respective images. To answer questions related to these images, the network must acquire sufficient medical knowledge, for example, for the first two images, it is essential to recognize the anatomy structure and modalities; for the third image, recognizing the X-ray image pattern of pathologies is necessary; for the final two images, apart from the basic biomedical knowledge, the model is also required to discern colors, differentiate subfigures, and perform Optical Character Recognition (OCR). (b) The top 20 figure types in PMC-VQA, cover a wide range of diagnostic procedures.

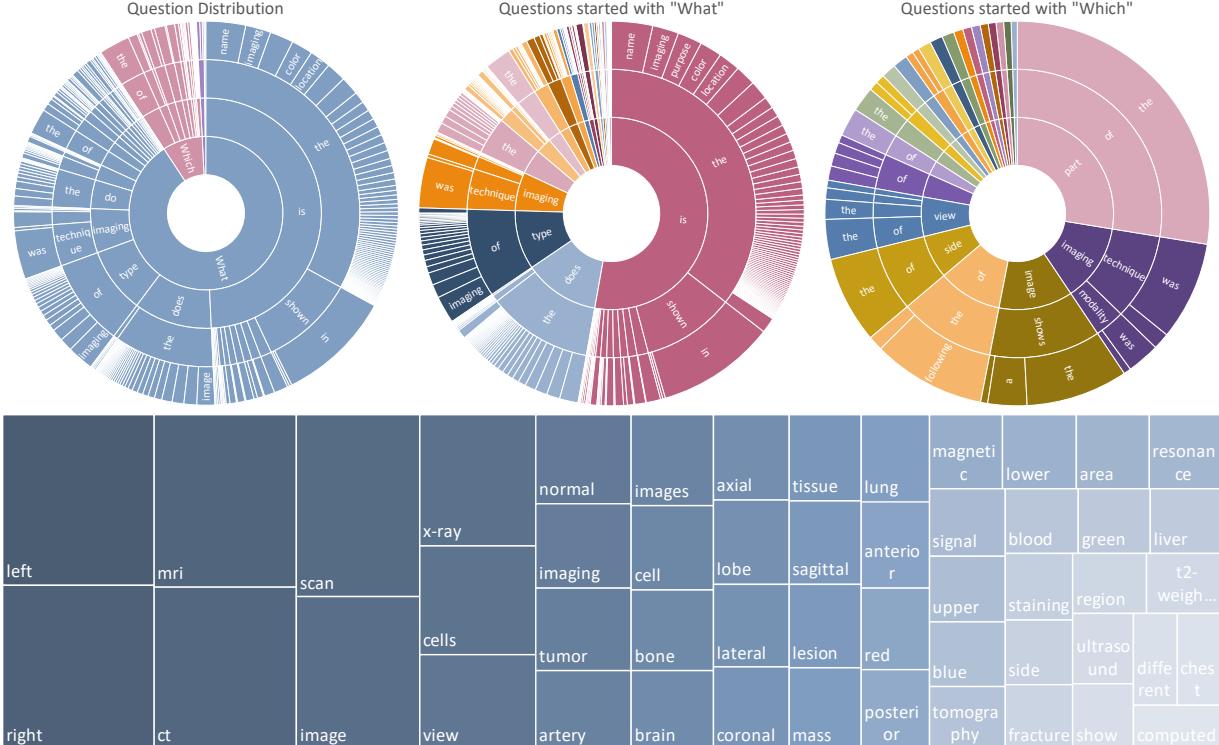
visual-language datasets [32] and initiate a scalable, automatic pipeline for constructing a new large-scale medical visual question-answering dataset. This new dataset, termed as **PMC-VQA**, contains 227k VQA pairs of 149k images, including 80% of radiological images, covering various modalities or diseases (Figure 1), surpassing existing datasets in terms of both amount and diversity.

In our experiments, we trained a generative visual-language model, termed as MedVInT, on the training set of PMC-VQA and fine-tuned it on the existing public benchmarks, *e.g.*, VQA-RAD [28], SLAKE [35], and ImageClef-VQA-2019 [6], outperforming existing models by a large margin, achieving over 80% accuracy on multi-choice selection. However, while evaluating our proposed challenging benchmark, even the state-of-the-art models struggle, showing that there is still ample room for development in this field.

In summary, our contributions are as follows: **(i)** We reframe the problem of MedVQA as a generative learning task and propose MedVInT, a model obtained by aligning a pre-trained vision encoder with a large language model through visual instruction tuning; **(ii)** We introduce a scalable pipeline and construct a large-scale MedVQA dataset, PMC-VQA, which far exceeds the size and diversity of existing datasets, covering various modalities and diseases; **(iii)** We pre-train MedVInT on PMC-VQA and fine-tune it on VQA-RAD [28] and SLAKE [35], achieving state-of-the-art performance and significantly outperforming existing models; **(iv)** We propose a new test set and present a more challenging benchmark for MedVQA, to evaluate the performance of VQA methods thoroughly.

## 2 Results

The goal of our proposed model, Medical Visual Instruction Tuning (MedVInT), is to perform generative-based medical visual question answering (MedVQA). Serving for this purpose, we curate a new large-scale medical visual instruction tuning dataset, namely PMC-VQA. In this section, we start by a comprehensive analysis



**Figure 2** | The top row shows the question distribution of the training set by their first four words. From left to right are all questions, questions started with “What” and questions started with “Which”. The ordering of the words starts towards the center and radiates outwards. The bottom row show the answer distribution of the training set.

on the PMC-VQA dataset, which contains 227k VQA pairs of 149k images, covering various modalities or diseases and compare it with the existing medical VQA datasets. Then, we will evaluate our trained model on three external MedVQA benchmarks, VQA-RAD [28], SLAKE [35] and ImageClef-VQA-2019 [6]. Note that, our model has two variants, which are tailored to encoder-based and decoder-based language models, respectively, denoted as MedVInT-TE and MedVInT-TD. At last, we establish a novel generative MedVQA benchmark with PMC-VQA, and evaluate various pre-trained visual or language models using our framework, serving as a reference to promote the future research in generative medical VQA.

## 2.1 Data Analysis

This section provides an analysis of images, questions, and answers of our final proposed dataset. In detail, the dataset comprises 227k image-question pairs, some examples are presented in Figure 1, which demonstrates the wide diversity of images within our dataset. As indicated in Table 1, PMC-VQA outperforms existing MedVQA datasets in terms of data size and modality diversity. The questions in our dataset cover a range of difficulties, from simple questions such as identifying image modalities, perspectives, and organs to challenging questions that require specialized knowledge and judgment. Additionally, our dataset includes difficult questions that demand the ability to identify the specific target sub-figure from the compound figure.

Our analysis of the PMC-VQA dataset can be summarized in three aspects: (i) **Images**: We show the top 20 figure types in Figure 1. The images in the PMC-VQA are extremely diverse, ranging from Radiology to Signals. (ii) **Questions**: We clustered the questions into different types based on the words that start the question, as shown in Figure 2. The dataset covers very diverse question types, including “What is the difference...”, “What type of imaging...”, and “Which image shows...”. Most questions range from 5 to 15 words, and detailed information about the distribution of question lengths is shown in the supplementary materials A.1. (iii) **Answers**: The words in answers primarily encompass positional descriptions, image modalities, and specific anatomical regions. Detailed information about the top 50 words that appeared in the

**Table 1** | Comparison of existing medical VQA datasets with PMC-VQA, demonstrating our dataset’s significant increase in size and diversity. Mixture refers to Radiology, Pathology, Microscopy, Signals, Generic biomedical illustrations, *etc.*

Dataset	Modality	Source	Images	QA pairs
VQA-RAD [28]	Radiology	MedPix® database	0.3k	3.5k
PathVQA [22]	Pathology	PEIR Digital Library [25]	5k	32.8k
SLAKE [35]	Radiology	MSD [3], ChestX-ray8 [56], CHAOS [26]	0.7k	14k
VQA-Med-2021 [7]	Radiology	MedPix® database	5k	5k
PMC-VQA	Mixture*(80% Radiology)	PubMed Central®	149k	227k

answers is provided in Figure 2. Most answers are around 5 words, which is much shorter than the questions. The correct options were distributed as follows: A (24.07%), B (30.87%), C (29.09%), D (15.97 %).

## 2.2 Evaluation on Public Benchmarks

Table 2 presents the performance of our MedVInT model on three widely recognized MedVQA benchmarks: VQA-RAD, SLAKE, and ImageClef-VQA-2019. The results demonstrate that the MedVInT model, regardless of whether we use the “MedVInT-TE” or “MedVInT-TD” version, surpasses previous best-performing methods on the VQA-RAD and SLAKE datasets. By default, we employ PMC-CLIP as the visual backbone and PMC-LLaMA as the language backbone, as demonstrated in Table 3, models pre-trained using PubMed Central data generally yield superior performance.

It is important to note that both the VQA-RAD and SLAKE datasets include questions that are categorized as either open-ended or close-ended. Close-ended questions restrict answers to a predefined set of options, whereas open-ended questions allow for free-from text responses. Specifically, for open-ended questions, the accuracy rates were enhanced from 67.2% to 73.7% on VQA-RAD and from 81.9% to 88.2% on SLAKE. For close-ended questions, the MedVInT model improved the accuracy from 84.0% to 86.8%. On the ImageCLEF benchmark, the “MedVInT-TE” version of our model achieved a significant improvement with an accuracy rate of 70.5%, significantly higher than the previous state-of-the-art (SOTA) accuracy of 62.4%.

Beyond comparing baselines with their default settings, we also consider an architecture-specific comparison where all models are directly trained from scratch on the downstream tasks. To distinguish from the default setting, our models here are denoted as “MedVInT-TE-S” and “MedVInT-TD-S”. As shown by the results, our proposed two variants can both surpass the former “M3AE” and “PMC-CLIP” architectures in most cases.

Additionally, when comparing the performance of the MedVInT model with and without pre-training on the PMC-VQA-train dataset, using the same architectural framework, it becomes evident that pre-training plays a crucial role in enhancing model performance. Specifically, the “MedVInT-TE” version, when pre-trained, showed a remarkable increase of approximately 16% in accuracy for open-ended questions on VQA-RAD and a 4% increase on SLAKE, compared to the “MedVInT-TE-S” version, which denotes training the model from scratch. Similar enhancements were observed with the “MedVInT-TD” version.

## 2.3 Evaluation on PMC-VQA

In this section, we introduce a new MedVQA benchmark, termed as PMC-VQA-test. We evaluate different models for both open-ended (Blanking) and multiple-choice (Choice) tasks. The results are summarized in Table 3. GPT-4-Oracle refers to the use of GPT-4 to answer questions based on the original captions of figures in academic papers. This approach represents the upper bound of model performance, as it leverages the most accurate and comprehensive information available about each figure. As shown in the tables, when only using language, the model is unable to provide accurate answers and give nearly random outcomes, with an accuracy of only 27.2% in Blanking and 30.8% in Choice for LLaMA and enhancing the language model from LLaMA to latest GPT-4 still cannot improve the results, *i.e.*, 21.1% in Blanking and 25.7% in Choice for GPT-4. The lower score in Blanking is due to the language model’s tendency to output longer sentences that cannot be correctly matched to a specific choice, which affects the calculation of model’s accuracy. It is worth noting that around 30% of the questions have “B” answers, making the 30.8% score nearly equivalent to the highest

**Table 2** | Comparison of ACC to SOTA approaches on VQA-RAD, SLAKE, and ImageClef-VQA-2019. We use the blank model for evaluation which provides output as free text answers rather than multiple-choice options. Pre-training data indicates whether the model is pre-trained on the medical multi-modal dataset before training on the target dataset. “MedVInT-TE-S” and “MedVInT-TD-S” respectively denotes we train the same architecture as “MedVInT-TE” or “MedVInT-TD” from scratch without pre-training on PMC-VQA. The best result is bold, the second-best result is underlined.

Method	Pretraining Data	VQA-RAD		SLAKE		VQA-2019 Overall
		Open	Close	Open	Close	
M3AE	/	66.5	79.0	79.2	83.4	-
PMC-CLIP	/	52.0	75.4	72.7	80.0	-
MedVInT-TE-S	/	53.6 (41.3,64.8)	76.5 (69.1,84.9)	84.0 (80.4,88.4)	85.1 (79.3,90.1)	<u>67.9 (60.6,74.2)</u>
MedVInT-TD-S	/	55.3 (45.4,69.4)	80.5 (74.3,89.4)	79.7 (74.6,85.3)	85.1 (78.2,89.3)	58.4 (50.6,66.2)
Hanlin	Unknown*	-	-	-	-	62.4
MEVF-BAN	VQA-RAD [28]	49.2	77.2	77.8	79.8	-
CPRD-BAN	ROCO, MedICaT [46, 52]	52.5	77.9	79.5	83.4	-
M3AE	CC12M [9]	67.2	83.5	80.3	<u>87.8</u>	-
PMC-CLIP	PMC-OA [32]	67.0	84.0	81.9	<b>88.0</b>	-
MedVInT-TE	PMC-VQA	69.3 (55.9,79.3)	84.2 (76.8,90.4)	<b>88.2 (84.6,92.7)</b>	87.7 (81.3,92.8)	<b>70.5 (62.8,78.2)</b>
MedVInT-TD	PMC-VQA	<b>73.7 (64.8,84.5)</b>	<b>86.8 (80.4,95.5)</b>	84.5 (80.4,90.5)	86.3 (79.6,90.6)	61.0 (53.0,67.6)

\*“Hanlin” is a solution in VQA-2019 challenge instead of a detailed scientific paper and, thus, no more details are provided. The numbers are directly copied from challenge papers. “MEVF-BAN” views the images in the train set of VQA-RAD as a pretraining dataset, performs image-wise self-supervised learning on it, and finetunes the model with VQA cases on each dataset. . We utilize the results of MEVF-BAN on various VQA benchmarks as reported by PMC-CLIP.

possible score attainable through guessing. These observations highlight the crucial requirement of multimodal understanding in our dataset and emphasize the strong relationship between images and the questions posed. In contrast to the training split, PMC-VQA-test has undergone thorough manual checking (Check Sec. 4.1 for more details), ensuring the credibility of the evaluation. We also report the experimental results on the original randomly split test set PMC-VQA-test-initial, which is larger but lacks further manual checking, in the supplementary materials A.2.

We also present the zero-shot evaluation results of the general VQA models like PMC-CLIP, BLIP-2, and Open-Flamingo which show relatively lower performance on the choice task. For instance, in the choice task, the model Open-Flamingo only achieved a 26.4% accuracy rate, significantly lower performance than our model at 40.3%. We also evaluate the medical-specific generative-based VQA model, *e.g.*, LLaVA-Med. Though it is better than the general models, it still lags behind our proposed MedVInT. It’s worth noting that LLaVA-Med is a work after our first announcement. This contrasts with the trained models on PMC-VQA, where we see notable improvements. Specifically, the MedVInT-TE and MedVInT-TD models, when paired with the PMC-CLIP vision backbone, demonstrate superior performance. For the open-ended task, the PMC-CLIP vision backbone again proves beneficial, with the MedVInT-TE model reaching the highest accuracy (36.4%) and BLEU-1 score (23.2%) when combined with the PubMedBERT language backbone. Moreover, the comparison between models trained from scratch and those utilizing CLIP or PMC-CLIP as vision backbones across different configurations of language backbones (PubMedBERT, LLaMA-ENC, and PMC-LLaMA-ENC) reveals a consistent trend: pre-trained models, especially those pre-trained with domain-specific data (PMC-CLIP), tend to outperform their counterparts trained from scratch. This emphasizes the importance of pre-training in achieving higher accuracies and better natural language generation metrics in MedVQA tasks. We then prompted a Large Language Model (LLM) to answer questions based on these generated captions. We also compared our approach with two-stage Visual Question Answering (VQA) models, which employ image captioning followed by a large language model for question answering. We experimented with a two-stage VQA method similar to Chatcad [55]. We first used MedICap [41], a state-of-the-art medical image captioning model, to interpret the given images into captions. The results showed poor performance on the test set. We then trained MedICap on the original image-caption pairs from the PMC-VQA training set to mitigate the domain gap. As shown, MedICap-PMCVQA-GPT-4 still shows inferior performance, which highlights key challenges in the two-stage approach: Captioning models need to anticipate potential questions in their descriptions. There’s often a mismatch between caption content and question focus. For example, a caption might state “This is an MRI image of a brain.” while the question asks “Is there a mass in the image?”. To provide a more comprehensive understanding of the dataset, we offer additional examples illustrated in

**Table 3** | Comparison of baseline models using different pre-trained models on both open-ended (Blank) and multiple-choice (Choice) tasks. We reported the results of the PMC-VQA-test. “Scratch” means to train the vision model from scratch with the same architecture as PMC-CLIP.

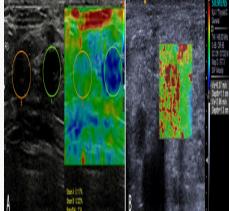
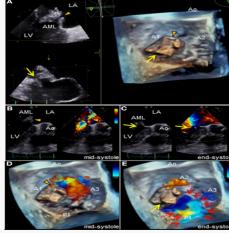
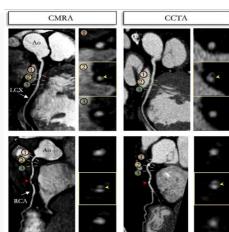
Method	Language Backbone	Vision Backbone	Choice	Blanking	
			ACC	ACC	BLEU-1
<b>Language-only</b>					
GPT-4-Oracle [43]	GPT-4 [43]	–	89.3 (87.7,90.8)	22.0 (19.6,24.5)	18.8 (17.6,20.2)
GPT-4 [43]	GPT-4 [43]	–	25.7 (23.5,28.1)	21.1 (18.8,23.5)	3.0 (2.6,3.4)
LLaMA [54]	LLaMA [54]	–	30.8 (27.4,34.8)	27.2 (23.1,31.3)	14.6 (12.7,16.6)
<b>Zero-shot</b>					
PMC-CLIP [32]	PMC-CLIP [32]	PMC-CLIP [32]	24.7 (21.3,28.0)	–	–
BLIP-2 [30]	OPT-2.7B [63]	CLIP [47]	24.3 (20.7,27.7)	21.8 (17.2,26.4)	7.6 (5.3,9.9)
Open-Flamingo [4]	LLaMA [54]	CLIP [47]	26.4 (22.7,29.8)	26.5 (22.3,30.7)	4.1 (2.1,6.13)
LLaVA-Med [29]	Vicuna [14]	BioMedCLIP [64]	34.8 (32.2,37.8)	29.4 (26.6,32.1)	3.9 (3.5,4.2)
MedICap-GPT-4	GPT-4 [43]	MedICap [41]	27.2 (24.7,29.7)	20.9 (18.8,23.3)	4.2 (3.6,4.6)
<b>Trained on PMC-VQA</b>					
MedICap-PMCVQA-GPT-4	GPT-4 [43]	MedICap-PMCVQA	35.9 (33.0, 38.3)	22.4 (20.1,24.8)	3.8 (3.3,4.3)
		Scratch	34.9 (31.7,38.5)	34.2 (31.2,37.0)	20.9 (18.9,23.2)
	PubMedBERT [20]	CLIP [47]	34.3 (30.7,37.8)	34.4 (31.0,37.6)	20.8 (18.6,23.3)
		PMC-CLIP [32]	37.6 (34.7,40.9)	36.4 (32.6,39.4)	<b>23.2 (21.2,25.7)</b>
MedVInT-TE		Scratch	35.2 (31.8,38.3)	32.5 (29.6,35.9)	15.9 (12.8,16.8)
	LLaMA-ENC [54]	CLIP [47]	36.1 (31.0,39.5)	33.4 (29.8, 36.5)	15.1 (12.8,17.5)
		PMC-CLIP [32]	37.1 (34.0,40.1)	<b>36.8 (33.5,40.0)</b>	18.4 (15.6,20.5)
		Scratch	38.0 (34.9,42.2)	35.0 (31.9,38.5)	17.0 (14.5,18.9)
	PMC-LLaMA-ENC [57]	CLIP [47]	38.5 (35.7,42.4)	34.4 (31.3,37.8)	16.5 (14.4,18.8)
		PMC-CLIP [32]	39.2 (36.7,41.7)	35.3 (31.4, 38.8)	18.6 (16.6,21.6)
		Scratch	37.9 (34.5,41.4)	30.2 (26.9,33.8)	18.0 (16.2,20.0)
MedVInT-TD	LLaMA [54]	CLIP [47]	39.2 (35.3,42.7)	32.2 (29.4,36.0)	20.0 (17.8,23.0)
		PMC-CLIP [32]	39.5 (35.1,42.7)	33.4 (30.6,37.4)	21.3 (18.9,23.8)
		Scratch	36.9 (33.2,40.2)	29.8 (26.9,32.7)	17.4 (15.1,19.6)
	PMC-LLaMA [57]	CLIP [47]	36.9 (32.9,40.1)	32.6 (29.0,36.2)	20.4 (18.1,22.9)
		PMC-CLIP [32]	<b>40.3 (37.2,43.8)</b>	33.6 (29.9,36.5)	21.5 (19.4,24.0)

Figure 3. This figure showcases random instances of the original image and corresponding captions, along with multiple-choice questions generated from them.

## 2.4 Evaluation of Visual Backbone Performance

We conducted additional experiments on standard medical image classification tasks to demonstrate the visual backbone’s performance and its improvement through the VQA pre-training. We evaluated our model on the MedMNIST dataset [61], which provides a diverse set of medical imaging modalities and classification tasks.

As shown in Table 4, our MedVInT models demonstrate competitive performance across all three tasks. Notably, MedVInT-TE achieves the best performance on DermaMNIST and the second-best performance on PneumoniaMNIST and BreastMNIST, only slightly behind PMC-CLIP. The results are impressive considering that MedVInT was pre-trained on only 177K images, compared to PMC-CLIP’s 1.6M image-caption pairs. Our results demonstrate the effectiveness of VQA-based pre-training compared to CLIP-style training. While both approaches aim to align visual and textual information, VQA requires a deeper understanding of the image content to answer specific questions. This difference in training objectives appears to lead to more robust visual representations, as evidenced by our model’s competitive performance despite being trained on significantly fewer images. These results demonstrate that our MedVQA task not only “standardizes” data into QA pairs but also substantially improves the visual backbone’s performance on various medical image classification tasks.

Image Caption	Image	Generated QA Pair	Model Prediction
Magnetic resonance imaging coronal view of the brain showing T1 weighted image revealing hyperintensity in bilateral basal ganglia due to mineral deposition. Both the arrows point out to the hypertense foci at the basal ganglia bilaterally on a T1-weighted MRI image that suggests mineral deposition.		<b>Question:</b> What is the name of the medical imaging technique used in this case? A: X-ray B: Magnetic resonance imaging C: Computed tomography D: Ultrasound	<b>MedVInT-TE Prediction :</b> Magnetic resonance imaging  <b>MedVInT-TD Prediction:</b> MRI
Malignant lymph nodes (LN) (carcinoma infiltration). The strain elastography reveals typically harder (blue) area in the LN than the surrounding tissues (green); strain ratio = 2.74 (A). The shear wave-based virtual touch tissue imaging quantification reveals a harder (red) area in the LN, and the maximum shear wave velocity (6.37 m/s) is much higher than that of surrounding tissues (2.96 m/s) (B).		<b>Question:</b> What color represents the harder area in the strain elastography image? A: Blue B: Red C: Green D: Yellow  The answer is: A: Blue	<b>MedVInT-TE Prediction:</b> Blue  <b>MedVInT-TD Prediction:</b> Blue
Pre-operative (1) and most recent post-operative (2) standing lateral pelvic radiographs.		<b>Question:</b> What type of radiographs are shown in the image? A: AP radiographs B: Lateral pelvic radiographs C: Oblique radiographs D: PA radiographs  The answer is: B: Lateral pelvic radiographs	<b>MedVInT-TE Prediction:</b> <b>Postalpalic radiograph</b>  <b>MedVInT-TD Prediction:</b> Lateral radiographs
(A) Three-dimensional TEE of the mitral valve. Note the two distinct ruptured perforations through the MVA (arrowhead and arrow, respectively). (B-E) Two or three-dimensional color Doppler TEE reveals that severe MR with two different jets communicate with the LA through the MVA: a superior jet (arrowhead) and a posterior jet (arrow), respectively. Note the MR with posterior jet heading toward the LA via the PML surface (dotted arrows). MR, mitral regurgitation.		<b>Question:</b> How many jets of mitral regurgitation are seen in images B-E? A: One jet B: Two jets C: Three jets D: Four jets  The answer is: B: Two jets	<b>MedVInT-TE Prediction:</b> 2  <b>MedVInT-TD Prediction:</b> Two
Reformatted non-contrast whole-heart sub-millimeter isotropic CMRA (left) and CCTA (right) images along the LCX (top) and RCA (bottom) are shown for a 54 year-old male patient. The CMRA dataset was acquired in 9 min with 100% scan efficiency (heart rate of 57 bpm). The CCTA images demonstrate mild (25–49%) disease with a calcified plaque within the proximal RCA and severe disease (70–90%) with a partially calcified plaque in the mid-segment of RCA (red arrows), and minimal (0–24%) disease with calcified plaque in the mid-segment of the LCX.		<b>Question:</b> Which arteries are shown in the top and bottom images of the CCTA, respectively? A: LAD and RCA B: RCA and LAD C: LCX and LAD D: RCA and LCX  The answer is: A: LAD and RCA	<b>MedVInT-TE Prediction:</b> <b>Left and artery</b>  <b>MedVInT-TD Prediction:</b> Left anterior descending artery and right circumflex artery

**Figure 3** | Examples of image captions, images, the generated question-answer pairs, and model prediction. The wrong predictions are highlighted in red.

### 3 Discussion

In this study, we target the challenge of MedVQA, where even the strongest VQA models trained on natural images yield results that closely resemble random guesses. To overcome this, we propose MedVInT, a generative model tailored to advance this crucial medical task. MedVInT is trained by aligning visual data from a pre-trained vision encoder with language models. Additionally, we present a scalable pipeline for constructing PMC-VQA, a comprehensive VQA dataset in the medical domain comprising 227k pairs across 149k images, spanning diverse modalities and diseases. Our proposed model delivers state-of-the-art performance on existing

**Table 4** | Classification results on three representative subsets of MedMNIST: PneumoniaMNIST (chest X-ray), BreastMNIST (ultrasound), and DermaMNIST (dermatoscopy). The best results are in **bold**, and the second-best are in underlined.

Methods	PneumoniaMNIST		BreastMNIST		DermaMNIST	
	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
ResNet50 [21]	96.20	88.40	86.60	84.20	91.20	73.10
DWT-CV [13]	95.69	<u>88.67</u>	89.77	85.68	91.67	74.75
SADAE [19]	98.30	91.80	91.50	87.80	92.70	75.90
PMC-CLIP	<b>99.02</b>	<b>95.35</b>	<b>94.56</b>	<b>91.35</b>	<u>93.41</u>	79.80
MedVInT-TE	<u>98.49</u>	<u>94.87</u>	<u>93.44</u>	<u>90.38</u>	<b>93.71</b>	<b>80.00</b>
MedVInT-TD	97.39	94.71	90.04	87.82	93.43	78.30

datasets, providing a new and reliable benchmark for evaluating different methods in this field.

**Significance of Medical VQA for Medical Imaging Ecosystem.** The development of advanced MedVQA systems has far-reaching implications for various stakeholders in the medical imaging ecosystem [5, 60, 15]. For radiologists and referring physicians, MedVQA can serve as a powerful decision-support tool, potentially enhancing diagnostic precision and streamlining image interpretation processes [16]. This could lead to more efficient clinical workflows and allow healthcare professionals to dedicate more time to direct patient care. For patients, MedVQA systems can significantly improve the communication of complex medical information. By translating intricate radiology reports into more comprehensible language, these systems can enhance patient understanding and engagement in their healthcare journey. This aligns with the growing emphasis on patient-centered care and shared decision-making in modern healthcare practices [45]. From a research and education perspective, MedVQA systems like MedVInT, trained on comprehensive datasets such as PMC-VQA, can serve as valuable tools for medical students and researchers [49]. They can provide interactive learning experiences, assist in the design of research plans, and offer insights into complex medical imaging concepts, thereby contributing to the advancement of medical knowledge and skills.

**PMC-VQA Act as a Valuable Resource for Medical VQA Domain.** Previous MedVQA datasets are usually limited in size and diversity, as demonstrated in Table 1. In contrast, PMC-VQA represents a pivotal advancement, offering an extensive resource that addresses the diverse and complex needs of the medical VQA domain. PMC-VQA facilitates the development of models capable of understanding and interpreting medical imagery with unprecedented accuracy and detail. Moreover, comparing results using the same architecture, with and without PMC-VQA (Table 3), it is clear that pre-training with PMC-VQA significantly outperforms. These results highlight the critical role that our PMC-VQA plays in addressing the major challenges that hinder the development of a generative MedVQA system. The pre-training enables models to gain a deep understanding of medical visuals and their associated questions, significantly enhancing their predictive capabilities.

**General Visual-language Models Struggle on MedVQA.** We evaluated the zero-shot performance of existing SOTA multimodal models, BLIP-2 and open-source version of Flamingo [30, 4]. As shown, even the best-performing models in natural images struggle to answer our questions, demonstrating the challenging nature of our dataset and its strong biomedical relevance. These results highlight the critical role that our PMC-VQA-train plays in addressing the major challenges that hinder the development of a generative MedVQA system.

**MedVInT Achieves State-of-the-art Performance of Generative MedVQA.** As demonstrated in the results, both MedVInT-TE and MedVInT-TD perform well on the MedVQA tasks. We compared it against various baselines that use different generative model backbones. Our results show that replacing the general visual backbone with a specialized medical one leads to improved performance, highlighting the importance of visual understanding in our test set. Additionally, we observed that replacing the language backbone with a domain-specific model also leads to some improvements, although not as significant as those achieved in the visual domain. In addition, the gap between the two training styles mainly exists in open-ended questions, with “MedVInT-TD” performing better on VQA-RAD and “MedVInT-TE” being more effective on SLAKE. This difference can be attributed to the fact that the VQA-RAD answers are typically longer than those in SLAKE, making the “MedVInT-TD” model more suitable. Conversely, SLAKE questions often require short responses, making the “MedVInT-TE” model more appropriate for such retrieve-like tasks.

**PMC-VQA-test Presents a Significantly More Challenging Benchmark.** Notably, the previous SOTA medical multimodal model, PMC-CLIP [32], struggles on our dataset. Not only does it fail to solve the blanking task, but it also significantly underperforms on multi-choice questions, with accuracy close to random. These findings underline the difficulty of our proposed benchmark and its capacity to provide a more rigorous evaluation of VQA models. However, while evaluating our proposed challenging benchmark, even the state-of-the-art models struggle, showing that there is still ample room for development in this field.

**Impacts of Our Work.** Since released to the public, we are delighted to observe the rapid adoption and extensive utilization of the PMC-VQA dataset, across a diverse range of research endeavors since its release. The dataset has served as a foundational resource for the development of numerous generative models, demonstrating its significant impact on the field. Notable examples include MathVista [38], RadFM [58], Qilin-Med-VL [36], SILKIE [31], CheXagent [12], UniDCP [62], and Quilt-LLaVA [50]. In addition, the methodology employed in constructing the dataset and the innovative prompt strategies we introduced have also inspired a series of works [59] and [10]. Furthermore, many studies have compared with our proposed MedVInT, recognizing it as the pioneering medical generative foundation model, such as Med-flamingo [39], OmniMedVQA [23]. This widespread adoption not only validates the robustness and utility of our dataset but also highlights its role in the scientific community.

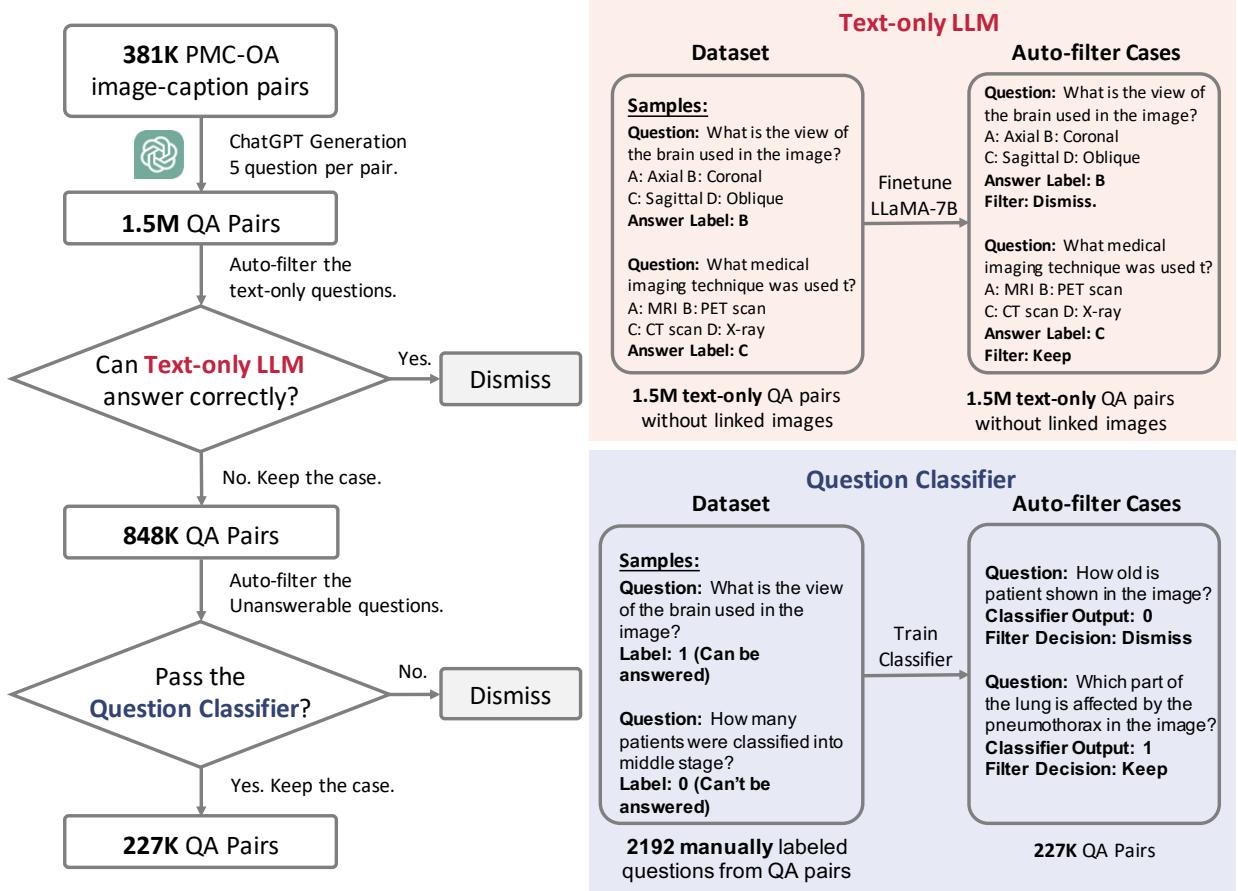
**Limitations.** The proposed PMC-VQA, while comprehensive, is subject to several limitations. First, similar to all existing datasets, there might be potential distribution biases in the images included in PMC-VQA compared to clinical practice. Specifically, our data is curated from academic papers, where there may be selective use of images to illustrate typical cases or slices, along with additional annotations such as arrows to aid understanding, resulting in our data being simpler compared to clinical scenarios. Nevertheless, for training purposes, the data from PMC-VQA remains crucial to help models better understand real clinical imaging data, as shown by the performance on public benchmarks in Table 2. On the other hand, for testing, *i.e.*, the benchmark we propose as shown in Table 3, even in such relatively simple scenarios, current methods still face significant challenges. Hence, for the ongoing advancement of MedVQA, conducting assessments in such an experimental playground to steer the emergence of more potent methodologies for the future still holds significance. On evaluation metrics, measuring the results from generative models poses a general challenge in the entire AI community [14], and this holds true for our testing as well. Although both the ACC score and Bleu score are used in our benchmark for assessing open-ended blanking results, these two metrics fail to capture the fluency of the generated sentence since they measure string similarity irrespective of word order. The encoder-based model thus significantly underperforms the decoder-based model in this regard. To address this issue, we plan to explore more accurate and effective evaluation metrics in our benchmark in future work. Lastly, as a starting point for generative-based MedVQA methods, our models may still suffer from hallucinations in non-sensical or adversarial cases with huge domain gaps (more case studies in our supplementary). Thus this paper is more as a **proof-of-concept** for building generative-based medical VQA models and needs more future efforts for real clinical applications.

## 4 Method

### 4.1 The PMC-VQA Dataset

Our study has identified the lack of large-scale, multi-modal MedVQA datasets as a significant obstacle to the development of effective generative MedVQA models. In this section, we provide a detailed description of our dataset collection process, starting with the source data and continuing with the question-answer generation and data filtering procedures. Finally, we analyze the collected data from various perspectives to gain insights into its properties and potential applications. The main data collection flow can be found in Figure 4.

**Source Data.** We start from PMC-OA [32], which is a comprehensive biomedical dataset comprising 1.6 million image-text pairs collected from PubMedCentral (PMC)'s OpenAccess subset [48], covering 2.4 million papers. The pipeline of creating PMC-OA consists of three major stages: i) medical figure-caption collection; (ii) subfigure separation; (iii) subcaption separation & alignment. To maintain the diversity and complexity of PMC-VQA, we have used a version of **381K image-caption pairs** obtained from the first stage of the medical figure collection process without subfigure auto-separation.



**Figure 4** | The whole flowchart demonstrating how we build up our PMC-VQA dataset. In left, we show the general progress and in right we show how we build up the two auto-filter models used in our data collection.

**Question-Answer Generation.** To automatically generate high-quality question-answer pairs, we input the image captions of PMC-OA, and prompt ChatGPT to generate 5 question-answer pairs for each caption. We use the following prompt to generate 5 question-answer pairs for each caption.

Ask 5 questions about the content and generate four options for each question. The questions should be answerable with the information provided in the caption, and the four options should include one correct and three incorrect options, with the position of the correct option randomized. The output should use the following template: i:'the question index' question:'the generate question' choice: 'A:option content B:option content C:option content D:option content' answer: The correct option(A\B\C\D).

This approach allows us to generate a large volume of diverse and high-quality questions that cover a wide range of medical topics. Considering some captions are too short to ask 5 questions, ChatGPT will repeat generated question-answer pairs or refuse to generate new pairs halfway and we dismissed the dummy cases. After generating the question-answer pairs using ChatGPT, we applied a rigorous filtering process to ensure that the pairs met our formatting requirements. As a result, we obtained 1,497,808 question-answer pairs, and since the original captions are linked with images, the pairs can naturally find corresponding images, resulting in an average of 3.93 pairs per image.

**Automatic & Manual Data Filtering.** As the questions are sourced from image captions, some of them can be answered correctly using biomedical knowledge alone, *i.e.*, without the need for a specific image, for example, question: “which type of MRI sequence shows high signal in the marrow edema?”. To address this issue, we trained a question-answer model using LLaMA-7B [54] with **text data only** and eliminated all questions that could be potentially answered by the language model. Specifically, we first split the dataset into

two parts, then we train a LLaMA-7B model only text input following the full fine-tuning pipeline introduced in PMC-LLaMA [57] in each part and do inference on the other part. To avoid that sometimes language model may make the correct choice by randomly guessing, for each case, we will shuffle the choice list and do inference five times. The questions the language model can make the right choice three times out of five will be dismissed. This filtering process resulted in 848,433 question-answer pairs that are unanswerable by the language-only model.

Furthermore, some questions in our data rely on additional information in the caption that cannot be answered with only the corresponding image, such as "How many patients were classified into the middle stage?" To identify these questions, we manually annotated 2192 question-answer pairs with binary labels, using '1' for answerable based on images and '0' otherwise. Then we train and evaluate a question classification model on these labeled data, specifically 1752 pairs for training and 440 for testing, and the model can achieve an accuracy of 81.77% on this binary classification task. We then used this model for data cleaning, resulting in a total of 226,946 question-answer pairs corresponding to 149,075 images, termed as **PMC-VQA** dataset.

From this cleaned dataset, we randomly selected 50,000 image-question pairs to create an initial test set, **PMC-VQA-test-initial**. The same image is guaranteed to not appear in both the training and testing sets. Additionally, we manually checked some test samples again, resulting in a small clean test set of 2,000 samples, which were **manually** verified for quality, termed as **PMC-VQA-test**, where we mainly consider the following criteria:

- whether questions are related to the image and can be answered via images;
- whether the distractor choices in the candidate list are complex enough, to avoid pure guessing from options;
- whether the image quality is good enough, dismissing the "paper images" which contain too many extra elements (charts, flows or numbers).

During this verification procedure, we have estimated that over 80% cases in **PMC-VQA-test** can be retained.

## 4.2 Architecture Design

We start with an introduction to the problem of generative medical visual question answering in Sec. 4.2.1, and detail our proposed architecture for generative MedVQA (Figure 5). We mainly focus on leveraging the pre-trained uni-model model to build up a multi-modal generative VQA architecture. Specifically, we offer two model variants, that are tailored to encoder-based and decoder-based language models, respectively, denoted as MedVInT-TE (Sec. 4.2.2) and MedVInT-TD (Sec. 4.2.3).

### 4.2.1 Problem Formulation

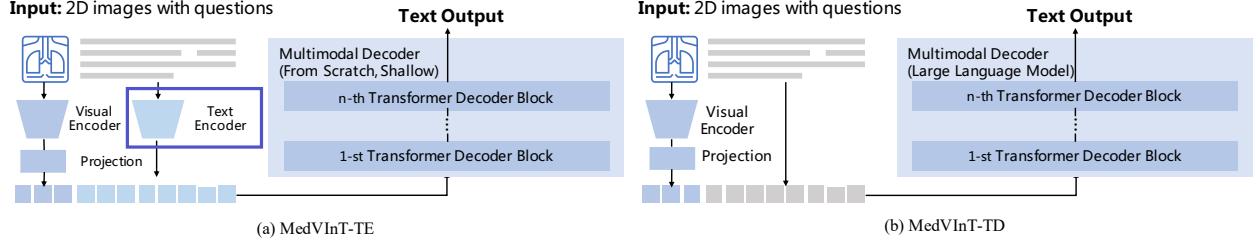
MedVQA is a task of answering natural language questions about medical visual content, typically images or videos obtained from medical devices like X-ray, CT, MRI, or microscopy, *etc*. Specifically, our goal is to train a model that can output the corresponding answer for a given question, which can be expressed as:

$$\hat{a}_i = \Phi_{\text{MedVQA}}(\mathcal{I}_i, q_i; \Theta) = \Phi_{\text{dec}}(\Phi_{\text{vis}}(\mathcal{I}_i; \theta_{\text{vis}}), \Phi_{\text{text}}(q_i; \theta_{\text{text}}); \theta_{\text{dec}}) \quad (1)$$

Here,  $\hat{a}_i$  refers to the predicted answer,  $\mathcal{I}_i \in \mathbb{R}^{H \times W \times C}$  refers to the visual image,  $H, W, C$  are height, width, channel respectively. The posed question and corresponding ground-truth answer in the form of natural language are denoted as  $q_i$  and  $a_i$ , respectively.  $\Theta = \{\theta_{\text{vis}}, \theta_{\text{text}}, \theta_{\text{dec}}\}$  denote the trainable parameters.

Existing approaches have primarily treated medical VQA as a classification problem, with the goal of selecting the correct answer from a candidate set, *i.e.*,  $a_i \in \Omega = \{a_1, a_2, \dots, a_N\}$ , where  $N$  represents the total number of answers within the dataset. Consequently, this approach limits the system's utility to predefined outcomes, hampering its free-form user-machine interaction potential.

In this paper, we take an alternative approach, with the goal of generating an open-ended answer in natural language. Specifically, we train the system by maximizing the probability of generating the ground-truth answer given the input image and question. The loss function used to train the model is typically the negative



**Figure 5** | The proposed architecture, mainly consists of three components: a visual encoder to extract visual features, a text encoder to encode textual context, and a multimodal decoder to generate the answer. (a) MedVInT-TE, encodes textual context (blue box) before input to the multimodal decoder; (b) MedVInT-TD, concatenates text tokens with visual features as input.

log-likelihood of correctly inferring the next token in the sequence, summed over all token steps, expressed as:

$$\mathcal{L}(\Theta) = - \sum_{t=1}^T \log p(a^t | \mathcal{I}, q^{1:T}, a^{1:t-1}; \Theta) \quad (2)$$

where  $T$  is the length of the ground-truth answer, and  $p(a^t | \mathcal{I}, q^{1:T}, a^{1:t-1}; \Theta)$  is the probability of generating the  $t$ -th token in the answer sequence given the input image  $\mathcal{I}$ , the question sequence  $q^{1:T}$ , and the previous tokens in the answer sequence  $a^{1:t-1}$ . This formulation allows the model to generate diverse and informative answers, which can be useful in a wider range of scenarios than traditional classification-based methods.

#### 4.2.2 MedVInT-TE.

**Visual Encoder.** Given one specific image  $\mathcal{I}$ , we can obtain the image embedding, *i.e.*,  $\mathbf{v} = \Phi_{\text{vis}}(\mathcal{I}) \in \mathbb{R}^{n \times d}$ , where  $d$  denotes the embedding dimension,  $n$  denotes the patch number. The vision encoder is based on a pre-trained ResNet-50 adopted from PMC-CLIP [32], with a trainable projection module. We propose two distinct variants for this projection module. The first variant, MLP-based, employs a two-layer Multilayer Perceptron (MLP), while the second variant, transformer-based, employs a 12-layer transformer decoder supplemented with several learnable vectors as query input.

**Text Encoder.** Given one question on the image, we append a fixed prompt with the question to guide the language model with desirable output, *i.e.*, “Question: {question}, the answer is: ”, and encode it with the language encoder:  $\mathbf{q} = \Phi_{\text{text}}(q) \in \mathbb{R}^{l \times d}$ , where  $\mathbf{q}$  refers to the text embedding,  $l$  represents the sequence length for the prompt, and  $q$  is the prompted question.  $\Phi_{\text{text}}$  is initialized with the pre-trained language model. Note that our model can also be applied to multiple-choice tasks, by providing options and training it to output the right choice as “A/B/C/D”. The prompt is then modified as “Question:  $q$ , the options are:  $a_1, a_2, a_3, a_4$ , the answer is: ”, where  $a_i$  refers to the  $i$ -th option.

**Multimodal Decoder.** With encoded visual embeddings ( $\mathbf{v}$ ) and question embeddings ( $\mathbf{q}$ ), we concatenate them as the input to the multimodal decoder ( $\Phi_{\text{dec}}$ ). The multimodal decoder is initialized from scratch with a 4-layer transformer structure. Additionally, acknowledging that the encoder-based language models lack causal masking, we reformulate the generation task as a mask language modeling task, *i.e.*, the question input is padded with several ‘[MASK]’ token and the decoder module learns to generate the prediction for the masked token.

#### 4.2.3 MedVInT-TD.

**Visual Encoder.** The visual encoder is the same as in MedVInT-TE.

**Text Encoder.** We design  $\Phi_{\text{text}}$  as a simple tokenization embedding layer, similar to the primary GPT-like LLMs, and the tokenization layer can be initialized with the corresponding layer of any chosen pre-trained LLM, like LLaMA [54] or PMC-LLaMA [57]. Same with MedVInT-TE, it also encodes the question input into embedding features  $\mathbf{q}$  and can perform multi-choice or blank through different prompts.

**Multimodal Decoder.** For the Transformer decoder-based language model, with its output format already being free-form text, we directly use its architecture as the multimodal decoder initialized with the pre-trained weights. Specifically, we concatenate the image and text features as the input. However, directly using the text decoder as a multimodal decoder, may lead to significant mismatching between the image encoding space and the decoder input space. Therefore, to further fill the gap between the image embedding space, here, we pre-train the whole network with the PMC-OA [32] dataset by captioning each image, which is similar to BLIP-2 [30]. Then train for the MedVQA task on our PMC-VQA dataset.

### 4.3 Datasets and Backbones

#### 4.3.1 Existing MedVQA Datasets

In the paper, we evaluate our final model MedVInT on three main public benchmarks, namely VQA-RAD, SLAKE, and ImageClef-VQA-2019.

**VQA-RAD** [28] is a VQA dataset specifically designed for radiology, consisting of 315 images and 3,515 questions with 517 possible answers. The questions in VQA-RAD are categorized as either close-ended or open-ended, depending on whether the answer choices are limited or not. We follow the official dataset split for our evaluation.

**SLAKE** [35] is an English-Chinese bilingual VQA dataset composed of 642 images and 14k questions. The questions are categorized as close-ended if answer choices are limited, otherwise open-ended. There are 224 possible answers in total. We only use the “English” part, and follow the official split.

**ImageClef-VQA-2019** [6] is a VQA dataset constructed based on images from MedPix [8]. It comprises 4,200 radiological images accompanied by 15,292 question-answer pairs. These questions are categorized into four types: modality, plane, organ system, and abnormality. We follow the official dataset split for our evaluation.

#### 4.3.2 Proposed PMC-VQA Dataset

The dataset can be used for both multiple-choice and open-ended tasks.

**Multi-choice Answering.** Four candidate answers are provided for each question as the prompt. The model is then trained to **select the correct option** among them. The accuracy (ACC) score can be used to evaluate the performance of the model on this task.

**Open-ended Answering.** The total possible answers for PMC-VQA are over 100K, which challenges the traditional retrieval-based approach for the answer set of such a level. Therefore, we provide another training style, called “blank”, where the network is not provided with options in input and is required to **directly generate answers**. For evaluation, we adopt two metrics, Bleu scores [44] and ACC scores.

We compare with strong generative models in the field of computer vision (Open-Flamingo [4] and BLIP-2 [30]). Open-Flamingo [4] is an open-source implementation of the prior state-of-the-art generalist visual-language model, namely, Flamingo from Google DeepMind [2], which was trained on large-scale data from general visual-language domain. We utilized the released checkpoint for zero-shot evaluation in our study. BLIP-2 [30] is a pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. We utilized their off-shelf checkpoint for zero-shot evaluation.

#### 4.3.3 Pre-trained Backbones

In this section, we introduce the pre-trained models used in our experiments. We separate them into language and vision backbones. Notably, while all the following models can be used in our architecture, by default, we use the “PMC-LLaMA” (or “PMC-LLaMA-ENC”) and “PMC-CLIP” as backbones since they are known to be more suitable for medical data according to previous works. The vision models are as follows.

**CLIP** [47]: This model is trained from scratch on a dataset of 400 million image-text pairs collected from the internet with contrastive loss. We use its “ViT-base-patch32” version as our visual encoder with 12 transformer layers, pre-trained on natural images.

**PMC-CLIP** [32]: This model is a medical-specific visual model based on CLIP architecture, which was

trained on a dataset of 1.6 million biomedical image-text pairs collected from PubMed open-access papers using cross-modality contrastive loss. Compared to the pre-trained visual model on natural images, PMC-CLIP is specifically designed to handle medical images and text.

Our experimental approach encompasses a range of language models, enabling us to explore the pivotal role of medical knowledge and the significance of its integration into this complex task. Specifically, the language models as follows.

**LLaMA [54]:** This is a state-of-the-art large-scale language model, pre-trained on trillions of tokens and widely used in the research community. We adopt the 7B version, which consists of 32 transformer layers, as our language backbone.

**PMC-LLaMA [57]:** This is an open-source language model that is acquired by fine-tuning LLaMA-7B on a total of 4.8 million biomedical academic papers with auto-regressive loss. Compared to LLaMA, PMC-LLaMA demonstrates stronger fitting capabilities and better performance on medical tasks.

**PubMedBERT [20]:** This is an encoder-based BERT-like model that is trained from scratch using abstracts from PubMed and full-text articles from PubMedCentral in the corpus “The Pile” [18]. It has 12 transformer layers and 100 million parameters. Such domain-specific models proved to yield excellent text embedding capability before the era of large language models.

**LLaMA-ENC and PMC-LLaMA-ENC.:** While LLaMA and PMC-LLaMA are known for their performance in text generation tasks, we also experiment with them as encoder models by passing a full attention mask and sampling the embedding from the last token. This allows for a direct comparison to be made with the aforementioned BERT-like models, which are also encoder-based.

#### 4.3.4 Implementation Details

Our models are all trained using the AdamW optimizer [37] with a learning rate of 2e-5. The max context length is set as 512, and the batch size is 128. To improve the training speed of our models, we adopt the Deepspeed acceleration strategy, together with Automatic Mixed Precision (AMP) and gradient checkpointing [17]. All models are implemented in PyTorch and trained on 8 NVIDIA A100 GPUs with 80 GB memory.

#### 4.3.5 Baseline Methods

We compare our proposed model with established generative models (Open-Flamingo [4], BLIP-2[30]) and state-of-the-art approaches across various medical visual question answering models (Hanlin [6], MEVF-BAN [40], CPRD-BAN [34], M3AE [11], PMC-CLIP [32]).

**Open-Flamingo [4]:** This is an open-source version of Google DeepMind’s cutting-edge visual language model, Flamingo. Trained on a vast corpus of general visual-language data, Open-Flamingo represents a benchmark in the field. We utilized the released checkpoint for zero-shot evaluation in our study.

**BLIP-2 [30]:** This is a robust visual-language generative model developed by Salesforce, surpassing Flamingo in reported capabilities. For our study, we utilized the released checkpoint for zero-shot evaluation.

**Hanlin [6]:** This approach denotes the best overall result of the 17 participating teams in the VQA-Med 2019 task. Considering the VQA-Med 2019 dataset shares an official test split, we directly borrow the results reported in the public leaderboards\*.

**MEVF-BAN [40]:** This approach introduces a framework that combines an unsupervised denoising auto-encoder with supervised Meta-Learning to quickly adapt to the VQA problem in scenarios with limited labeled data. We utilize the results of MEVF-BAN on various VQA benchmarks as reported by PMC-CLIP [32], where MEVF-BAN is finetuned on each specific dataset and evaluated on the corresponding official test set.

**CPRD-BAN [34]:** This approach proposes a two-stage pre-training framework that focuses on learning transferable features from radiology images and distilling a compact visual feature extractor tailored for Med-VQA tasks. Similarly to MEVF-BAN, we adopt the results of CPRD-BAN reported in PMC-CLIP [32] following the finetuning setting.

---

\*<https://www.aicrowd.com/challenges/imageclef-2019-vqa-med/leaderboards>

**M3AE** [11]: This approach is a self-supervised learning approach using multimodal masked autoencoders to learn cross-modal knowledge by reconstructing missing information from partially masked images and texts. Similarly, we adopt the results of M3AE on various MedVQA datasets as reported in PMC-CLIP [32]. The official checkpoint is finetuned on each dataset and subsequently evaluated on the official test set.

**PMC-CLIP** [32]: For the VQA task under zero-shot settings, we directly employed it to match image embeddings with the most similar text embeddings obtained from question-and-answer choices and then calculated the accuracy.

#### 4.3.6 Evaluation Metrics

We adopt two conventional metrics from the NLP community, BLEU-1 scores [44] (BiLingual Evaluation Understudy) and ACC scores (Accuracy).

**BLEU-1.** BLEU-1 scores focus on the precision of unigrams, or single words, by comparing the model prediction to reference texts, yielding a score between 0 and 1.

**ACC.** ACC scores refer to the percentage of correctly answered questions out of the total number of questions. For the generative model, we calculate ACC scores by matching the model’s output with the options using `difflib.SequenceMatcher`<sup>†</sup> and choosing the most similar one, which is more difficult than the evaluation for retrieval-based methods due to the unlimited output space. Note that, `difflib.SequenceMatcher` is a class in the `difflib` module of the Python Standard Library. It is based on the Ratcliff-Obershelp algorithm, to compare sequences of elements, such as strings, lists, or any other iterable objects, and find the similarities and differences between them.

## 5 Conclusion

In conclusion, this paper addresses the challenge of Medical Visual Question Answering (MedVQA). Specifically, we reframe the problem of MedVQA as a generation task that naturally mirror the human-machine interactions. We introduce a generative model for medical visual understanding by aligning visual information from a pre-trained vision encoder with a large language model. To facilitate the model training, we present a scalable pipeline for constructing PMC-VQA, a comprehensive MedVQA dataset comprising 227k VQA pairs across 149k images, spanning diverse modalities and diseases. Our proposed model delivers state-of-the-art performance on existing MedVQA datasets, providing a new and reliable benchmark for evaluating different methods in this field.

## 6 Code Availability

Our model checkpoint can be found in <https://huggingface.co/xmcmic/MedVInT-TE> and <https://huggingface.co/xmcmic/MedVInT-TD>, and our codes can be found in <https://github.com/xiaoman-zhang/PMC-VQA>.

## 7 Data Availability

The proposed dataset PMC-VQA can be found in <https://huggingface.co/datasets/xmcmic/PMC-VQA>. The papers used for developing PMC-VQA are from the “Commercial Use Allowed” split of PMC Open Access Subset<sup>‡</sup>. We provide the detailed PubMed Central ID for each paper and corresponding licenses on [huggingface](https://huggingface.co)<sup>§</sup>, which are all under CC0 or CC BY licenses. Our final dataset PMC-VQA is under CC BY-SA licenses so that it can be widely used to support the development of medical generative-based VQA models. The other used public dataset can be found as follows. SLAKE is available at <https://www.med-vqa.com/slake/>. VQA-RAD is available at <https://osf.io/89kps/>. Image-Clef-2019 is available at <https://www.imageclef.org/2019>.

<sup>†</sup><https://docs.python.org/3/library/difflib.html>

<sup>‡</sup><https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>

<sup>§</sup>[https://huggingface.co/datasets/xmcmic/PMC-VQA/blob/main/oa\\_comm\\_use\\_file\\_list.csv](https://huggingface.co/datasets/xmcmic/PMC-VQA/blob/main/oa_comm_use_file_list.csv)

## References

- [1] 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. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.
- [4] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, et al. Openflamingo, 2023.
- [5] Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2):e188–e194, 2021.
- [6] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.
- [7] Asma Ben Abacha, Mourad Sarouti, 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 *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.
- [8] Md BETHESDA. Medpix™ receives patent, 2006.
- [9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.
- [10] Xiaolan Chen, Pusheng Xu, Yao Li, Weiyi Zhang, Fan Song, Ying-Feng Zheng, Danli Shi, and Mingguang He. Chatffa: Interactive visual question answering on fundus fluorescein angiography image using chatgpt. Available at *SSRN 4578568*.
- [11] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention*, pages 679–689. Springer, 2022.
- [12] 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.
- [13] Jianhong Cheng, Hulin Kuang, Qichang Zhao, Yahui Wang, Lei Xu, Jin Liu, and Jianxin Wang. Dwt-cv: Dense weight transfer-based cross validation strategy for model selection in biomedical data analysis. *Future Generation Computer Systems*, 135:20–29, 2022.
- [14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023.
- [15] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.
- [16] Hilmi Demirhan and Wlodek Zadrozny. Survey of multimodal medical question answering. *BioMedInformatics*, 4(1):50–74, 2023.

- [17] Jianwei Feng and Dong Huang. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11442, 2021.
- [18] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [19] Xiaolong Ge, Yanpeng Qu, Changjing Shang, Longzhi Yang, and Qiang Shen. A self-adaptive discriminative autoencoder for medical applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8875–8886, 2022.
- [20] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [22] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. pages 708–718, 2020.
- [23] 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 lvm. *arXiv preprint arXiv:2402.09181*, 2024.
- [24] Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- [25] Kristopher N Jones, Dwain E Woode, Kristina Panizzi, and Peter G Anderson. Peir digital library: Online resources and authoring system. In *Proceedings of the AMIA Symposium*, page 1075. American Medical Informatics Association, 2001.
- [26] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.
- [27] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
- [28] 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.
- [29] 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. *Advances in Neural Information Processing Systems*, 36, 2024.
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [31] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.
- [32] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. 2023.
- [33] Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*, 2022.
- [34] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention*, pages 210–220. Springer, 2021.

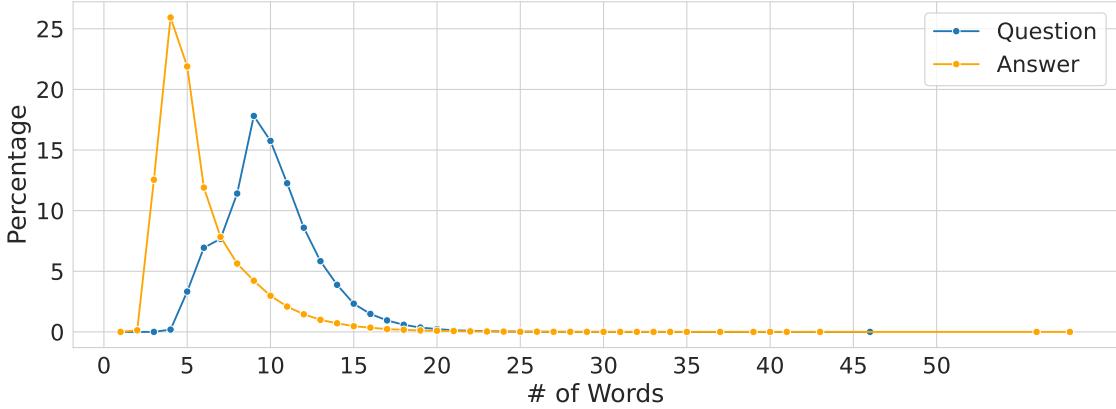
- [35] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
- [36] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023.
- [37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [38] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- [39] 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 (ML4H)*, pages 353–367. PMLR, 2023.
- [40] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention*, pages 522–530. Springer, 2019.
- [41] Aaron Nicolson, Jason Dowling, and Bevan Koopman. A concise model for medical image captioning. In *CLEF (Working Notes)*, pages 1611–1619, 2023.
- [42] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- [43] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [45] Jiwoo Park, Kangrok Oh, Kyunghwa Han, and Young Han Lee. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Scientific Reports*, 14(1):13218, 2024.
- [46] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *MICCAI Workshop on Large-scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS) 2018*, pages 180–189. Springer, 2018.
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [48] Richard J Roberts. Pubmed central: The genbank of the published literature. volume 98, pages 381–382. National Acad Sciences, 2001.
- [49] Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. The role of large language models in medical education: applications and implications, 2023.
- [50] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *arXiv preprint arXiv:2312.04746*, 2023.
- [51] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.
- [52] Sanjay Subramanian et al. Medicat: A dataset of medical images, captions, and textual references. In *Findings of EMNLP*, 2020.
- [53] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [55] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023.
- [56] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammad Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [57] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine. *arXiv preprint arXiv:2304.14454*, 2023.
- [58] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.
- [59] Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*, 2024.
- [60] Jiancheng Yang, Hongwei Bran Li, and Donglai Wei. The impact of chatgpt and llms on medical imaging stakeholders: perspectives and use cases. *Meta-Radiology*, page 100007, 2023.
- [61] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195. IEEE, 2021.
- [62] Chenlu Zhan, Yufei Zhang, Yu Lin, Gaoang Wang, and Hongwei Wang. Unidcp: Unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts. *arXiv preprint arXiv:2312.11171*, 2023.
- [63] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [64] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

## A Supplemental Materials

### A.1 Data Analysis

Fig. 6 shows the percentage of questions and answers with different word lengths. Most questions range from 5 to 15 words, and most answers are around 5 words.



Supplementary Fig. 6 | Percentage of questions and answers with different word lengths.

### A.2 Evaluation on Original Split Test Set

In this section, we report the experimental results on the original randomly split test set in Supplementary Table 5. This test set is more extensive but did not undergo additional manual verification. As indicated, the performance experienced a slight decline when compared with the PMC-VQA-test, yet the reduction was minimal. For instance, the accuracy (ACC) for the choice task decreased from 40.3 to 39.2. This slight variation underscores the inherent high quality and robustness of our dataset.

### A.3 Ablation Study

In this section, we add the comparison of baseline models using different projection modules (MLP or Transformer) on both open-ended and multiple-choice tasks. MLP-based projection module, employs a two-layer Multilayer Perceptron (MLP), while the second variant, transformer-based projection modules, employs a 12-layer transformer decoder supplemented with several learnable vectors as query input. As shown in Table 6, different projection modules demonstrate comparable performance across various evaluation tasks. Both architectures can effectively reconcile the diversity in the embedding dimensions arising from different pre-trained visual models, making our architecture adaptable to various visual foundation model designs, regardless of whether they are based on ViT or ResNet.

### A.4 Fail Case Study

In this section, we explore the hallucinations exhibited by the proposed MedVInT models. As a starting point for generative-based MedVQA methods, for now, our models still suffer from hallucinations in nonsensical or adversarial cases with huge domain gaps. As illustrated in Fig 7, for out-of-scope tasks such as report generation, the model may not produce radiology reports in a structured format. However, it sometimes provides reasonable answers. For nonsensical questions, such as inquiring about lung nodules in an abdomen CT image, the model cannot refuse to answer nor highlight the mistake in the question.

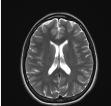
**Supplementary Table 5** | Comparison of baseline models using different pre-trained models on both open-ended and multiple-choice tasks. We reported the results on PMC-VQA-test-initial. ‘Scratch’ means to train the vision model from scratch with the same architecture as ‘PMC-CLIP’.

Method	Language Backbone	Vision Backbone	Choice	Blanking	
			ACC	BLEU-1	ACC
<b>Zero-shot</b>					
PMC-CLIP [32]	PMC-CLIP [32]	PMC-CLIP [32]	24.0 (23.4,24.6)	-	-
BLIP-2 [30]	OPT-2.7B [63]	CLIP [47]	24.6 (23.9,25.2)	22.5 (21.9,23.2)	5.2 (4.8,5.7)
Open-Flamingo [4]	LLaMA [54]	CLIP [47]	25.0 (24.5, 25.6)	26.1 (25.6,26.7)	4.1 (3.7, 4.6)
LLaVA-Med [29]	Vicuna [14]	BioMedCLIP [64]	32.6 (32.1,33.2)	28.3 (27.8,28.8)	3.7(3.7,3.8)
<b>Trained on PMC-VQA</b>					
LLaMA [54]	LLaMA [54]	-	30.6 (30.0,31.2)	26.1 (25.7,26.8)	14.2 (13.9,14.6)
		Scratch	34.4 (33.7,35.1)	33.7 (33.0, 34.5)	20.4 (19.9,20.9)
	PubMedBERT [20]	CLIP [47]	34.5 (33.8,35.1)	33.7 (32.9,34.3)	20.4 (20.0,20.9)
		PMC-CLIP [32]	37.1 (36.4,37.9)	35.2 (34.6,36.0)	<b>22.0 21.6,22.4</b>
MedVInT-TE	LLaMA-ENC [54]	Scratch	35.2 (34.5,35.9)	32.5 (31.7,33.1)	15.3 (14.8,15.7)
		CLIP [47]	35.3 (34.7,35.9)	32.3 (31.5,33.0)	15.6 (14.8,15.7)
		PMC-CLIP [32]	36.9 (36.2,37.6)	<b>35.4 (34.8,36.1)</b>	18.2 (17.7,18.6)
MedVInT-TD	PMC-LLaMA-ENC [57]	Scratch	37.0(36.3,37.6)	32.6 (32.0,33.3)	16.2 (15.7,16.6)
		CLIP [47]	37.1(36.4,37.9)	33.0 (32.1,33.7)	16.6 (16.2,17.0)
		PMC-CLIP [32]	38.2 (37.5,38.9)	34.8 (34.0,35.3)	18.1 (17.7,18.6)
MedVInT-TD	LLaMA [54]	Scratch	36.2 (35.7,36.9)	29.1 (28.1,29.7)	17.4 (17.2,17.9)
		CLIP [47]	38.2 (37.5,38.9)	31.3 (30.6,32.0)	19.5 (19.1,20.0)
		PMC-CLIP [32]	37.3 (36.8,38.0)	31.9 (31.2,32.6)	20.0 (19.6,20.5)
MedVInT-TD	PMC-LLaMA [57]	Scratch	36.8 (36.2,37.6)	28.6 (27.8,29.1)	16.8 (16.4,17.1)
		CLIP [47]	36.8 (36.2,37.5)	31.4 (30.8,32.1)	19.5 (19.1,20.0)
		PMC-CLIP [32]	<b>39.4 (38.7,40.0)</b>	32.7 (31.1,33.2)	20.3 (19.9,20.7)

**Supplementary Table 6** | Ablation study of baseline models using different projection modules and pre-trained models on open-ended and multiple-choice tasks. We reported the results of the original test set of the PMC-VQA/PMC-VQA test. “Scratch” means to train the vision model from scratch with the same architecture as “PMC-CLIP”.

Method	Language Backbone	Vision Backbone	Blanking		Choice ACC
			ACC	Bleu-1	
MedVInT-TE-MLP	PubMedBERT [20]	Scratch	33.7 / 34.2	20.4 / 20.9	34.4 / 34.9
		CLIP [47]	32.3 / 34.4	15.6 / 20.8	34.5 / 34.3
		PMC-CLIP [32]	35.2 / 36.4	<b>22.0 / 23.2</b>	37.1 / 37.6
	LLaMA-ENC [54]	Scratch	32.5 / 32.5	15.3 / 15.9	35.2 / 35.2
		CLIP [47]	32.3 / 33.4	15.6 / 15.1	35.3 / 36.1
		PMC-CLIP [32]	35.4 / <b>36.8</b>	18.2 / 18.4	36.9 / 37.1
	PMC-LLaMA-ENC [57]	Scratch	32.6 / 35.0	16.2 / 17.0	37.0 / 38.0
		CLIP [47]	33.0 / 34.4	16.6 / 16.5	37.1 / 38.5
		PMC-CLIP [32]	34.8 / 35.3	18.1 / 18.6	38.2 / 39.2
MedVInT-TE-Transformer	PubMedBERT [20]	Scratch	34.1 / 36.2	<b>21.0 / 21.9</b>	39.8 / 40.6
		CLIP [47]	33.9 / 34.6	20.6 / 21.8	<b>39.9 / 40.9</b>
		PMC-CLIP [32]	33.7 / 35.4	20.3 / 21.2	<b>40.2 / 40.9</b>
	LLaMA-ENC [54]	Scratch	32.0 / 33.5	15.1 / 15.3	38.4 / 39.7
		CLIP [47]	32.3 / 34.3	15.5 / 15.7	38.4 / 38.7
		PMC-CLIP [32]	<b>35.9 / 37.1</b>	19.0 / 19.3	38.9 / 39.4
	PMC-LLaMA-ENC [57]	Scratch	33.2 / 34.7	16.6 / 16.5	38.1 / 39.8
		CLIP [47]	33.6 / 35.1	16.7 / 17.2	38.7 / 38.9
		PMC-CLIP [32]	<b>35.5 / 36.0</b>	18.4 / 18.6	38.2 / 37.7
MedVInT-TD-MLP	LLaMA [54]	Scratch	28.1 / 30.6	16.5 / 16.9	35.8 / 37.4
		CLIP [47]	30.2 / 32.7	18.6 / 18.5	35.8 / 37.1
		PMC-CLIP [32]	31.3 / 32.6	19.5 / 19.8	38.4 / <b>41.0</b>
	PMC-LLaMA [57]	Scratch	28.3 / 30.6	16.4 / 17.3	35.8 / 37.0
		CLIP [47]	31.4 / 31.8	19.2 / 19.5	36.2 / 37.9
		PMC-CLIP [32]	32.1 / 31.7	19.7 / 20.2	38.4 / <b>42.3</b>
MedVInT-TD-Transformer	LLaMA [54]	Scratch	29.1 / 30.2	17.4 / 18.0	36.2 / 37.9
		CLIP [47]	31.3 / 32.2	19.5 / 20.0	38.2 / 39.2
		PMC-CLIP [32]	31.9 / 33.4	20.0 / 21.3	37.3 / 39.5
	PMC-LLaMA [57]	Scratch	28.6 / 29.8	16.8 / 17.4	36.8 / 36.9
		CLIP [47]	31.4 / 32.6	19.5 / 20.4	36.8 / 36.9
		PMC-CLIP [32]	32.7 / 33.6	20.3 / 21.5	39.4 / 40.3

## Out of Scope Examples



**Question:** Please write a radiology report that explains this brain MRI image.

**MedVInT:** The brain MRI image shows a normal brain.

**Case analysis:** Radiology report generation is not a task trained in PMC-VQA, but the given image is a normal brain.



**Question:** Please write a radiology report that explain the explains this abdomen CT image.

**MedVInT:** The abdomen CT image shows a large mass in the left upper quadrant of the abdomen. 1.1 cm 2.1 cm.

**Case analysis:** Radiology report generation is not a task trained in PMC-VQA, but the given answer is reasonable.



**Question:** Please give the caption of this ultrasound image.

**MedVInT:** Ultrasound image of the right kidney.

**Case analysis:** Radiology report generation is not a task trained in PMC-VQA, the given image is thyroid ultrasound.

## Non-sensical Examples



**Question:** Is there any abnormalities in the given brain CT image?

**MedVInT:** No.

**Case analysis:** The image presented is actually a brain MRI. While the model did not identify the error in the question, the response remains accurate.



**Question:** Does this patient have lung nodules?

**MedVInT:** No.

**Case analysis:** The image presented is actually an abdomen CT. The model did not identify the error in the question.



**Question:** Where does the arrow point to in the image?

**MedVInT:** The kidney.

**Case analysis:** There exist no arrow in the image.



**Question:** What is the modality of this image?

**MedVInT:** CT scan.

**Case analysis:** The input image is not a medical image.

Supplementary Fig. 7 | Examples of “out of scope” and “non-sensical” questions.