

<https://doi.org/10.1038/s43856-024-00709-2>

Development of a large-scale medical visual question-answering dataset

Check for updates

Xiaoman Zhang^{1,2,3}, Chaoyi Wu^{1,2,3}, Ziheng Zhao^{1,2}, Weixiong Lin^{1,2}, Ya Zhang^{1,2}, Yanfeng Wang^{1,2,4} ✉ & Weidi Xie^{1,2,4} ✉

Abstract

Background Medical Visual Question Answering (MedVQA) enhances diagnostic accuracy and healthcare delivery by leveraging artificial intelligence to interpret medical images. This study aims to redefine MedVQA as a generation task that mirrors human-machine interaction and to develop a model capable of integrating complex visual and textual information.

Methods We constructed a large-scale medical visual-question answering dataset, PMC-VQA, containing 227,000 VQA pairs across 149,000 images that span various modalities and diseases. We introduced a generative model that aligns visual information from a pre-trained vision encoder with a large language model. This model was initially trained on PMC-VQA and subsequently fine-tuned on multiple public benchmarks.

Results Here, we show that our model significantly outperforms existing MedVQA models in generating relevant, accurate free-form answers. We also propose a manually verified test set that presents a greater challenge and serves as a robust measure to monitor the advancement of generative MedVQA methods.

Conclusions The PMC-VQA dataset proves to be an essential resource for the research community, and our model marks a significant breakthrough in MedVQA. We maintain a leaderboard to facilitate comprehensive evaluation and comparison, providing a centralized resource for benchmarking state-of-the-art approaches.

Plain language summary

Medical images play a crucial role in healthcare, but interpreting them accurately can be challenging. This study developed an artificial intelligence system that can answer questions about medical images, similar to how a medical expert would explain findings to patients. We created a large collection of medical images paired with questions and answers to train our AI system, covering various types of medical scans and conditions. Our system can generate detailed, accurate responses to questions about medical images, performing better than existing approaches. The system and dataset we developed are freely available to researchers, which should help advance the field of medical image interpretation and ultimately improve healthcare delivery.

Large language models (LLMs), such as GPT-4¹, Med-PaLM², and PMC-LLaMA³ have recently achieved remarkable success in the field of medical natural language processing^{4–6}. 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 accurately⁷.

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^{8,9}. 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^{10–13} 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 Flamingo¹⁴ and BLIP¹⁵, 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 visual-language datasets¹³ and initiate a scalable, automatic pipeline for constructing a new large-scale medical visual question-answering dataset. This

¹Shanghai Jiao Tong University, Shanghai, China. ²Shanghai Artificial Intelligence Laboratory, Shanghai, China. ³These authors contributed equally: Xiaoman Zhang, Chaoyi Wu. ⁴These authors jointly supervised this work: Yanfeng Wang, Weidi Xie. ✉ e-mail: wangyanfeng622@sjtu.edu.cn; weidi@sjtu.edu.cn

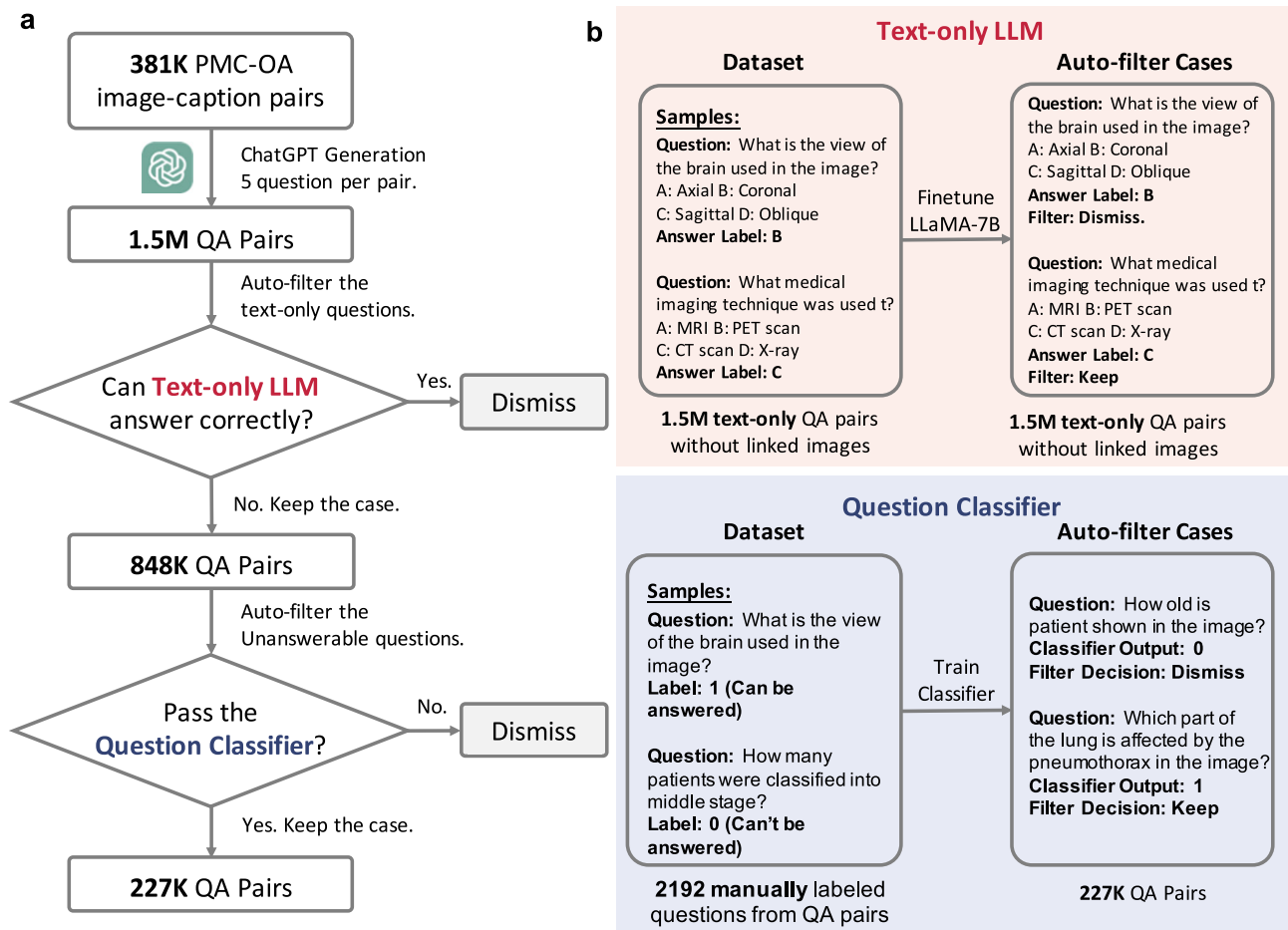


Fig. 1 | The whole flowchart demonstrating how we build up our PMC-VQA dataset. a The general progress. **b** Illustration of the two auto-filter models used in our data collection.

dataset, termed as PMC-VQA, contains 227k VQA pairs of 149k images, including 80% of radiological images, covering various modalities or diseases, surpassing existing datasets in terms of both amount and diversity. 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-RAD¹⁶, SLAKE¹⁷, and ImageClef-VQA-2019¹⁸, outperforming existing models by a large margin, achieving over 80% accuracy on multi-choice selection. Additionally, we present a more challenging benchmark for MedVQA, even the state-of-the-art models struggle, showing that there is still ample room for development in this field.

Methods

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 Fig. 1.

Source data. We start from PMC-OA¹³, which is a comprehensive biomedical dataset comprising 1.6 million image-text pairs collected from PubMedCentral (PMC)'s OpenAccess subset¹⁹, 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. Our dataset is entirely constructed from open-source resources publicly available for research and no special permissions were required to access them. These datasets were either released under licenses or conditions that comply with relevant ethical and regulatory standards or were explicitly exempted from requiring individual informed consent under their respective IRB approvals. Since our work does not involve accessing or managing personally identifiable or sensitive data beyond what is publicly available, nor does it entail direct data collection from individuals, a separate IRB review or informed consent waiver does not apply to our dataset construction process.

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)."

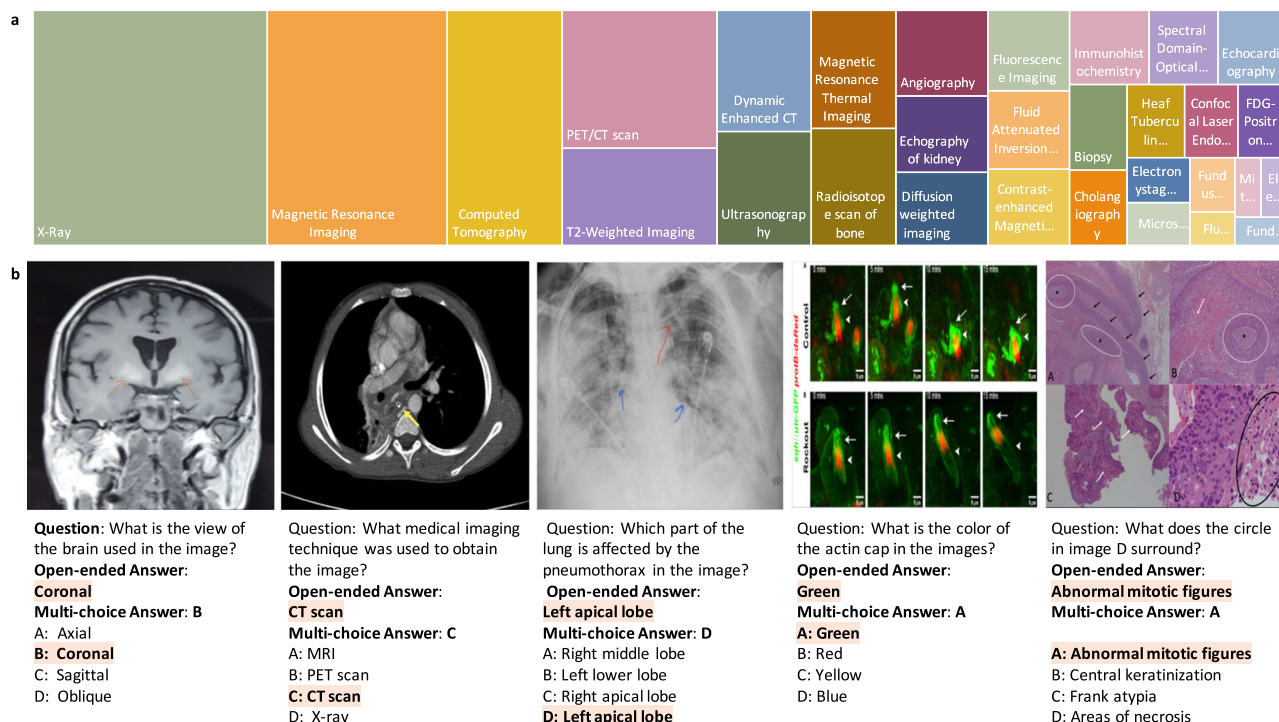


Fig. 2 | Distribution and examples of PMC-VQA. **a** The top 20 figure types in PMC-VQA, cover a wide range of diagnostic procedures. **b** Several examples of challenging questions and answers along with their respective images. To answer questions related to these images, the network must acquire sufficient 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. Images are from PubMedCentral's OpenAccess subset papers^{67–71}, which are used with permission under the PMC Open Access Subset license.

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 five 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 and 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-7B²⁰ 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 with only text input following the full fine-tuning pipeline introduced in PMC-LLaMA³ 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. Examples of this dataset can be found in Fig. 2.

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 2000 samples, which were manually verified for quality, termed as PMC-VQA-test, where we mainly consider the following criteria: (i) whether questions are related to the image and can be answered via images; (ii) whether the distractor choices in the candidate list are complex enough, to avoid pure guessing from options; (iii) whether the image quality is good enough, dismissing the figures which contain too many extra elements (charts, flows or numbers). During this verification procedure, we have estimated that over 80% of cases in PMC-VQA-test can be retained.

Architecture design

We start with an introduction to the problem of generative medical visual question answering, and detail our proposed architecture for generative MedVQA (Fig. 3). 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 MedViInT-TE and MedViInT-TD.

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.

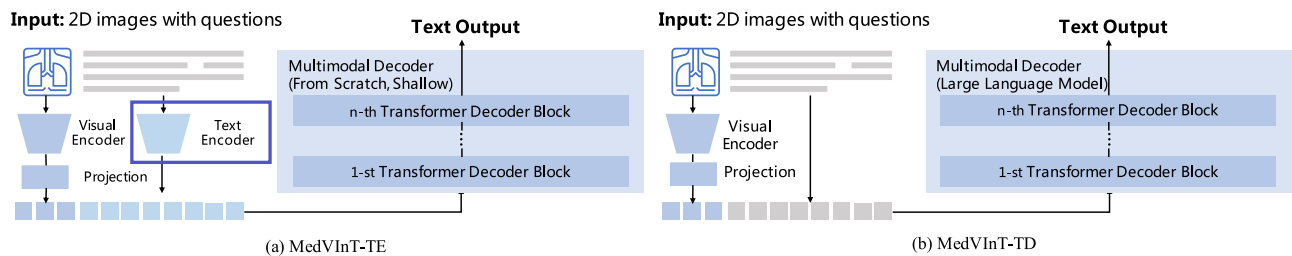


Fig. 3 | The proposed architecture, mainly consists of three components: a visual encoder to extract visual features, a text encoder to encode textual context, and a multimodal decoder to generate the answer. a MedVInT-TE, encodes textual

context (blue box) before input to the multimodal decoder. **b MedVInT-TD, concatenates text tokens with visual features as input.**

Specifically, our goal is to train a model that can output the corresponding answer for a given question, which can be expressed as

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

Here, \hat{a}_i refers to the predicted answer, $I_i \in \mathbb{R}^{H \times W \times C}$ 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 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.

MedVInT-TE

Visual encoder. Given one specific image I , we can obtain the image embedding, i.e., $v = \Phi_{\text{vis}}(I) \in \mathbb{R}^{n \times d}$, 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-CLIP¹³, 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: $q = \Phi_{\text{text}}(q) \in \mathbb{R}^{l \times d}$, where q refers to the text embedding, l represents the sequence length for the prompt, and q is the prompted question. Φ_{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 (v) and question embeddings (q), we concatenate them as the input to the multimodal decoder (Φ_{dec}). The multimodal decoder is initialized from scratch with a 4-layer transformer structure. Additionally, acknowledging that the encoder-based language models lack casual masking, we reformulate the generation task as a mask language modeling task, i.e., the question input is padded with several [MASK] tokens, and the decoder module learns to generate the prediction for the masked token.

MedVInT-TD

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

Text encoder. We design Φ_{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 LLaMA²⁰ or PMC-LLaMA³. Same with MedVInT-TE, it also encodes the question input into embedding features 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-OA¹³ dataset by captioning each image, which is similar to BLIP-2¹⁵. Then train for the MedVQA task on our PMC-VQA dataset.

Datasets and backbones

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. These datasets are freely accessible through their respective public repositories and no special permissions were required to access them. While the original publications and documentation for these datasets do not explicitly state IRB approval information, all three datasets contain de-identified medical images from public medical knowledge bases. As we are working with existing de-identified public datasets and not conducting human subjects research, our analysis did not require additional IRB review.

VQA-RAD¹⁶ is a VQA dataset specifically designed for radiology, consisting of 315 images and 3515 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.

SLAKE¹⁷ 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-2019¹⁸ is a VQA dataset constructed based on images from MedPix²¹. It comprises 4200 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.

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 scores²² and ACC scores.

We compare with strong generative models in the field of computer vision (Open-Flamingo²³ and BLIP-2¹⁵). Open-Flamingo²³ is an open-source implementation of the prior state-of-the-art generalist visual-language model, namely, Flamingo from Google DeepMind²⁴, which was trained on large-scale data from the general visual-language domain. We utilized the released checkpoint for zero-shot evaluation in our study. BLIP-2¹⁵ 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.

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.

CLIP²⁵: 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-CLIP¹³: 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 are as follows.

LLaMA²⁰: 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-LLaMA³: 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.

PubMedBERT²⁶: 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 Pile²⁷. 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.

Implementation details. Our models are all trained using the AdamW optimizer²⁸ 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 checkpointing²⁹. All models are implemented in PyTorch and trained on 8 NVIDIA A100 GPUs with 80 GB memory. The training of the models took ~128 A100 GPU h, with an average inference time of 7 s per sample.

Baseline methods. We compare our proposed model with established generative models (Open-Flamingo²³, BLIP-2¹⁵) and state-of-the-art approaches across various medical visual question-answering models (Hanlin¹⁸, MEVF-BAN¹⁰, CPRD-BAN¹¹, M3AE¹², PMC-CLIP¹³).

Open-Flamingo²³: 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-2¹⁵: 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.

Hanlin¹⁸: 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 (<https://www.aicrowd.com/challenges/imageclef-2019-vqa-med/leaderboards>).

MEVF-BAN¹⁰: 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-CLIP¹³, where MEVF-BAN is finetuned on each specific dataset and evaluated on the corresponding official test set.

CPRD-BAN¹¹: 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-CLIP¹³ following the finetuning setting.

M3AE¹²: 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-CLIP¹³. The official checkpoint is finetuned on each dataset and subsequently evaluated on the official test set.

PMC-CLIP¹³: 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.

Evaluation metrics. We adopt two conventional metrics from the NLP community, BLEU-1 scores²² (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.

Table 1 | Comparison of existing medical VQA datasets with PMC-VQA, demonstrating our dataset's increase in size and diversity

Dataset	Modality	Source	Images	QA pairs
VQA-RAD ¹⁶	Radiology	MedPix® database	0.3k	3.5k
PathVQA ⁵²	Pathology	PEIR Digital Library ⁵³	5k	32.8k
SLAKE ¹⁷	Radiology	MSD ⁵⁴ , ChestX-ray8 ⁵⁵ , CHAOS ⁵⁶	0.7k	14k
VQA-Med-2021 ⁵⁷	Radiology	MedPix® database	5k	5k
PMC-VQA	Mixture* (80% Radiology)	PubMed Central®	149k	227k

Mixture refers to Radiology, Pathology, Microscopy, Signals, Generic biomedical illustrations, etc.

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` 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.

Statistics and reproducibility

All data analysis was performed with Python and is reproducible using the code linked below in the “Code availability” section. All the data used was obtained as described in the “Data availability” section. We estimate confidence intervals using non-parametric bootstrapping: for each evaluation metric, we generate 1000 bootstrap replicates by repeatedly sampling with replacement from the original dataset, where each replicate has the same size n as the original dataset. The confidence intervals are derived from the empirical distribution of the resampled estimates, specifically using the interval between the $100 \times (\alpha/2)$ and $100 \times (1 - \alpha/2)$ percentiles; we pick $\alpha = 0.05$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

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 with a comprehensive analysis of 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-RAD¹⁶, SLAKE¹⁷ and ImageClef-VQA-2019¹⁸. 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 generative MedVQA benchmark with PMC-VQA, and evaluate various pre-trained visual or language models using our framework, serving as a reference to promote future research in generative medical VQA.

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 Fig. 2, 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 Fig. 2. 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 Fig. 4. 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 Supplementary Fig. 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 answers is provided in Fig. 4. 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%).

Evaluation of 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 ~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.

Evaluation on PMC-VQA

In this section, we introduce a 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

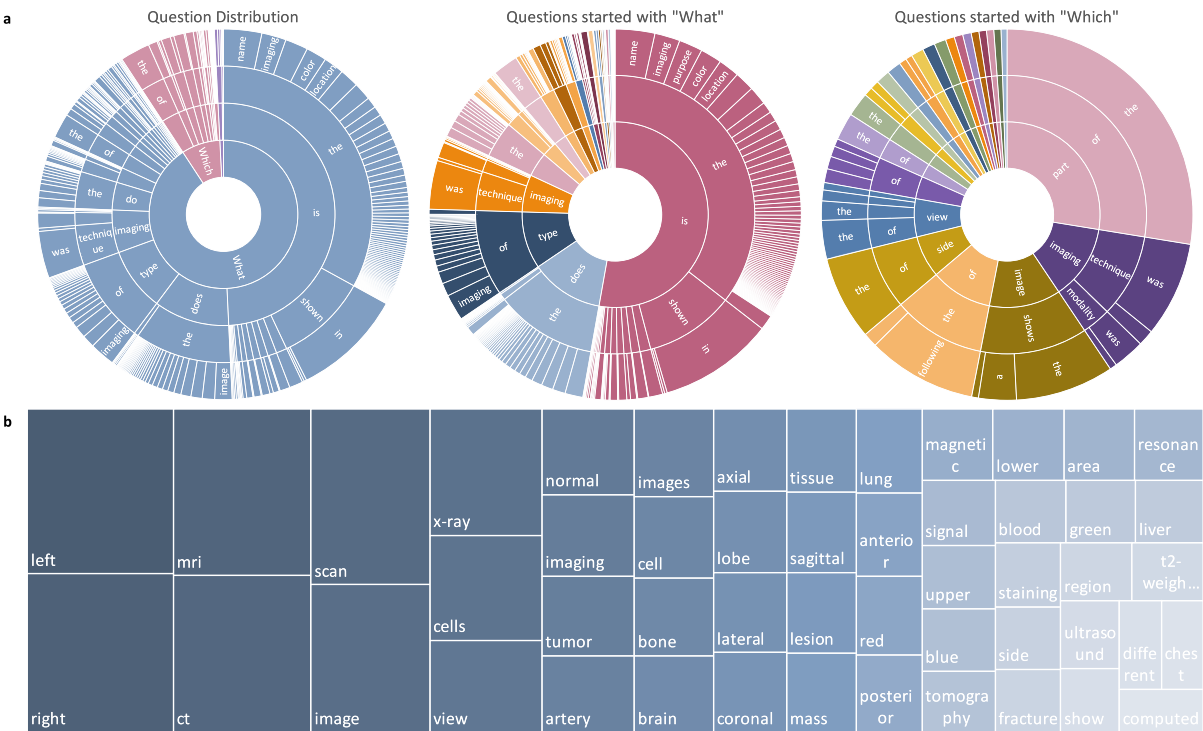


Fig. 4 | Distribution of questions and answers in PMC-VQA. **a** The question distribution of the training set by their first four words. From left to right are all questions, questions started with “What” and questions started with “Which”. The ordering of the words starts towards the center and radiates outwards. **b** Answer the distribution of the training set.

Table 2 | Comparison of ACC to SOTA approaches on VQA-RAD, SLAKE, and ImageClef-VQA-2019

Method	Pretraining data	VQA-RAD		SLAKE		VQA-2019
		Open	Close	Open	Close	Overall
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)	67.9 (60.6,74.2)
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 ^a	–	–	–	–	62.4
MEVF-BAN	VQA-RAD ^{a16}	49.2	77.2	77.8	79.8	–
CPRD-BAN	ROCO, MedlCat ^{58,59}	52.5	77.9	79.5	83.4	–
M3AE	CC12M ⁶⁰	67.2	83.5	80.3	87.8	–
PMC-CLIP	PMC-OA ¹³	67.0	84.0	81.9	88.0	–
MedViNT-TE	PMC-VQA	69.3 (55.9,79.3)	84.2 (76.8,90.4)	88.2 (84.6,92.7)	87.7 (81.3,92.8)	70.5 (62.8,78.2)
MedViNT-TD	PMC-VQA	73.7 (64.8,84.5)	86.8 (80.4,95.5)	84.5 (80.4,90.5)	86.3 (79.6,90.6)	61.0 (53.0,67.6)

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, denote 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. Numbers in parentheses represent 95% confidence intervals obtained through non-parametric bootstrapping with 1000 replicates.

^aHanlin 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.

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 possible score attainable through guessing. These observations highlight the crucial requirement of multimodal understanding in our

Table 3 | Comparison of baseline models using different pre-trained models on both open-ended (Blank) and multiple-choice (Choice) tasks

Method	Language Backbone	Vision Backbone	Choice	Blanking	
			ACC	ACC	BLEU-1
Language-only					
GPT-4-Oracle ¹	GPT-4 ¹	–	89.3 (87.7,90.8)	22.0 (19.6,24.5)	18.8 (17.6,20.2)
GPT-4 ¹	GPT-4 ¹	–	25.7 (23.5,28.1)	21.1(18.8,23.5)	3.0(2.6,3.4)
LLaMA ²⁰	LLaMA ²⁰	–	30.8 (27.4,34.8)	27.2 (23.1,31.3)	14.6 (12.7,16.6)
Zero-shot					
PMC-CLIP ¹³	PMC-CLIP ¹³	PMC-CLIP ¹³	24.7 (21.3,28.0)	-	-
BLIP-2 ¹⁵	OPT-2.7B ⁶¹	CLIP ²⁵	24.3 (20.7,27.7)	21.8 (17.2,26.4)	7.6 (5.3,9.9)
Open-Flamingo ²³	LLaMA ²⁰	CLIP ²⁵	26.4 (22.7,29.8)	26.5 (22.3,30.7)	4.1 (2.1,6.13)
LLaVA-Med ⁶²	Vicuna ⁴⁹	BioMedCLIP ⁶³	34.8 (32.2,37.8)	29.4 (26.6,32.1)	3.9(3.5,4.2)
MedICap-GPT-4	GPT-4 ¹	MedICap ³¹	27.2 (24.7,29.7)	20.9 (18.8,23.3)	4.2 (3.6,4.6)
Trained on PMC-VQA					
MedICap-PMCVQA-GPT-4	GPT-4 ¹	MedICap-PMCVQA	35.9 (33.0, 38.3)	22.4 (20.1,24.8)	3.8 (3.3,4.3)
MedVInt-TE	PubMedBERT ²⁶	Scratch	34.9 (31.7,38.5)	34.2 (31.2,37.0)	20.9 (18.9,23,2)
		CLIP ²⁵	34.3 (30.7,37.8)	34.4 (31.0,37.6)	20.8 (18.6,23.3)
		PMC-CLIP ¹³	37.6 (34.7,40.9)	<u>36.4 (32.6,39.4)</u>	23.2 (21.2,25.7)
	LLaMA-ENC ²⁰	Scratch	35.2 (31.8,38.3)	32.5 (29.6,35.9)	15.9 (12.8,16.8)
		CLIP ²⁵	36.1 (31.0,39.5)	33.4 (29.8, 36.5)	15.1 (12.8,17.5)
		PMC-CLIP ¹³	37.1 (34.0,40.1)	36.8 (33.5,40.0)	18.4 (15.6,20.5)
	PMC-LLaMA-ENC ³	Scratch	38.0 (34.9,42.2)	35.0 (31.9,38.5)	17.0 (14.5,18.9)
		CLIP ²⁵	38.5 (35.7,42.4)	34.4 (31.3,37.8)	16.5 (14.4, 18.8)
		PMC-CLIP ¹³	39.2 (36.7,41.7)	35.3 (31.4, 38.8)	18.6 (16.6,21.6)
MedVInt-TD	LLaMA ²⁰	Scratch	37.9 (34.5,41.4)	30.2 (26.9,33.8)	18.0 (16.2,20.0)
		CLIP ²⁵	39.2 (35.3,42.7)	32.2 (29.4,36.0)	20.0 (17.8,23.0)
		PMC-CLIP ¹³	<u>39.5 (35.1,42.7)</u>	33.4 (30.6,37.4)	21.3 (18.9,23.8)
	PMC-LLaMA ³	Scratch	36.9 (33.2,40.2)	29.8 (26.9,32.7)	17.4 (15.1, 19.6)
		CLIP ²⁵	36.9 (32.9,40.1)	32.6 (29.0,36.2)	20.4 (18.1,22.9)
		PMC-CLIP ¹³	40.3 (37.2,43.8)	33.6 (29.9,36.5)	21.5 (19.4,24.0)

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. The best result is bold, the second-best result is underlined. Numbers in parentheses represent 95% confidence intervals obtained through non-parametric bootstrapping with 1000 replicates.

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, 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 Supplementary Table 1.

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, a 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 is 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. In addition, the comparison of baseline models using different projection modules (MLP or Transformer) on both open-ended and multiple-choice tasks is shown in Supplementary Table 2.

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 Chatcad³⁰. We first used MedICap³¹, 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

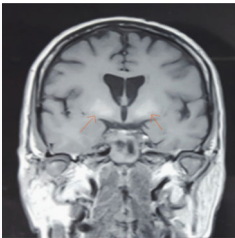
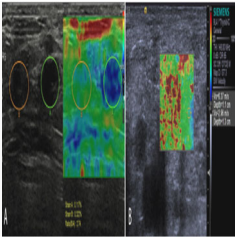
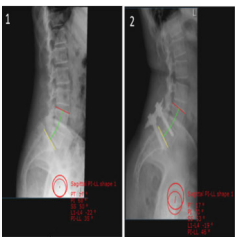
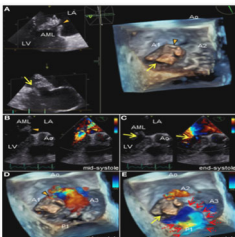
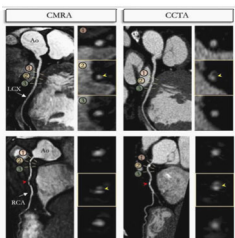
Image Caption	Image	Generated QA Pair	Model Prediction
Magnetic resonance imaging coronal view of the brain showing T1 weighted image revealing hyperintensity in bilateral basal ganglia due to mineral deposition. Both the arrows point out to the hypertense foci at the basal ganglia bilaterally on a T1-weighted MRI image that suggests mineral deposition.		Question: 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 The answer is: B: Magnetic resonance imaging	MedVInT-TE Prediction : Magnetic resonance imaging MedVInT-TD Prediction: MRI
Malignant lymph nodes (LNs) (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).		Question: What color represents the harder area in the strain elastography image? A: Blue B: Red C: Green D: Yellow The answer is: A: Blue	MedVInT-TE Prediction: Blue MedVInT-TD Prediction: Blue
Pre-operative (1) and most recent post-operative (2) standing lateral pelvic radiographs.		Question: 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	MedVInT-TE Prediction: Postapical radi radiograph MedVInT-TD Prediction: 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.		Question: How many jets of mitral regurgitation are seen in images B-E? A: A:One jet B: B:Two jets C: C:Three jets D: D:Four jets The answer is: B:Two jets	MedVInT-TE Prediction: 2 MedVInT-TD Prediction: 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.		Question: 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	MedVInT-TE Prediction: Left and and artery MedVInT-TD Prediction: Left anterior descending artery and right circumflex artery

Fig. 5 | Examples of image captions, images, the generated question–answer pairs, and model prediction. The wrong predictions are highlighted in red. Images are from PubMedCentral's OpenAccess subset papers^{67,72–75}, which are used with permission under the PMC Open Access Subset license.

Fig. 5. This figure showcases random instances of the original image and corresponding captions, along with multiple-choice questions generated from them. Additionally, we employed GPT-4 to categorize the modality of each caption into four categories: radiology, pathology, both, and others. The detailed results for each category are provided in Table 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 dataset³², which provides a diverse set of medical imaging modalities and classification tasks.

As shown in Table 5, 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

Table 4 | Performance of MedVInT-TE-Transformer using PMC-LLaMA-Enc and PMC-CLIP as backbones on both open-ended (Blank) and multiple-choice (Choice) tasks

Modality	Count	Choice	Blanking	
		ACC	ACC	BLEU-1
Radiology	1434	40.10	36.47	21.93
Pathology	245	30.61	31.02	8.26
Others	196	32.14	33.67	9.86
Both	125	32.00	32.00	13.99
All	2000	37.65	35.25	18.58

Table 5 | Classification results on three representative subsets of MedMNIST: PneumoniaMNIST (chest X-ray), BreastMNIST (ultrasound), and DermaMNIST (dermatoscopy)

Methods	PneumoniaMNIST		BreastMNIST		DermaMNIST	
	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
ResNet50 ⁶⁴	96.20	88.40	86.60	84.20	91.20	73.10
DWT-CV ⁶⁵	95.69	88.67	89.77	85.68	91.67	74.75
SADAE ⁶⁶	98.30	91.80	91.50	87.80	92.70	75.90
PMC-CLIP	99.02	95.35	94.56	91.35	<u>93.41</u>	<u>79.80</u>
MedVInT-TE	<u>98.49</u>	<u>94.87</u>	<u>93.44</u>	<u>90.38</u>	93.71	80.00
MedVInT-TD	97.39	94.71	90.04	87.82	93.43	78.30

The best results are in bold, and the second-best are underlined.

standardizes data into QA pairs but also substantially improves the visual backbone's performance on various medical image classification tasks.

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

The development of advanced MedVQA systems has far-reaching implications for various stakeholders in the medical imaging ecosystem^{8,33,34}. For radiologists and referring physicians, MedVQA can serve as a powerful decision-support tool, potentially enhancing diagnostic precision and streamlining image interpretation processes³⁵. 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 practices³⁶. 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 researchers³⁷. 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.

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.

We evaluated the zero-shot performance of existing SOTA multimodal models, BLIP-2 and open-source version of Flamingo^{15,23}. 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.

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.

Notably, the previous SOTA medical multimodal model, PMC-CLIP¹³, struggles with 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.

Since released to the public, we have been 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 MathVista³⁸, RadFM³⁹, Qilin-Med-VL⁴⁰, SILKIE⁴¹, CheXagent⁴², UniDCP⁴³, and Quilt-LLaVA⁴⁴. In addition, the methodology employed in constructing the dataset and the innovative prompt strategies we introduced have also inspired a series of works⁴⁵ and⁴⁶. Furthermore, many studies have compared with our proposed MedVInT, recognizing it as the pioneering medical generative foundation model, such as Med-flamingo⁴⁷, OmniMedVQA⁴⁸. This widespread adoption not only validates the robustness and utility of our dataset but also highlights its role in the scientific community.

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. This selection bias could lead to models that perform well on academic examples but fail to generalize to clinical practice. Researchers and developers using this dataset should be aware of these limitations and consider them when

designing or training models, to prevent the propagation of errors and to ensure that medical AI advancements are built on a foundation that truly reflects the diverse and complex nature of clinical environments. 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 community⁴⁹, 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, as shown in Supplementary Fig. 2. This paper is more like a proof-of-concept for building generative-based medical VQA models and needs more future efforts for real clinical applications.

Data availability

The proposed dataset is freely available at <https://huggingface.co/datasets/xmcmic/PMC-VQA>⁵⁰. SLAKE is freely available at <https://huggingface.co/datasets/BoKelvin/SLAKE>. VQA-RAD is freely available at <https://paperswithcode.com/dataset/vqa-rad>. ImageClef-VQA-2019 is freely available at <https://zenodo.org/records/10499039>. MedMNIST is freely available at <https://medmnist.com/>. The source data for Tables 2–5, Supplementary Fig. 1 and Supplementary Tables 1, 2 is in Supplementary Data 1.

Code availability

All codes associated with this study are available at <https://github.com/xiaoman-zhang/PMC-VQA>⁵¹.

Received: 7 March 2024; Accepted: 12 December 2024;

Published online: 21 December 2024

References

- OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
- Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
- Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Pmc-llama: towards building open-source language models for medicine. *JAMIA* **31**, 1833–1843 (2024).
- Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
- Kung, T. H. et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
- Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023).
- Lin, Z. et al. Medical visual question answering: a survey. *Artif. Intell. Med.* **143**, C (2023).
- Yang, J., Li, H. B. & Wei, D. The impact of chatgpt and llms on medical imaging stakeholders: perspectives and use cases. *Meta-Radiology* **1**, 100007 (2023).
- Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
- Nguyen, B. D. et al. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention* (ed. Shen, D.) 522–530 (Springer, 2019).
- Liu, B., Zhan, L.-M. & Wu, X.-M. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention* (ed. de Bruijne, M.) 210–220 (Springer, 2021).
- Chen, Z. et al. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention* (ed. Wang, L.) 679–689 (Springer, 2022).
- Lin, W. et al. Pmc-clip: contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention* (ed. Greenspan, H.) 525–536 (Springer, 2023).
- Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
- Li, J., Li, D., Savarese, S. & Hoi, S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML* **2022**, 814, 19730–19742 (2023).
- Lau, J. J., Gayen, S., Ben Abacha, A. & Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **5**, 1–10 (2018).
- Liu, B. et al. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (ed. Ludivine, F.) 1650–1654 (IEEE, 2021).
- Ben Abacha, A., Hasan, S. A., Datla, V. V., Demner-Fushman, D. & Müller, H. Vqa-med: overview of the medical visual question answering task at imageclef 2019. In *Proc. Conference and Labs of the Evaluation Forum (CLEF) 2019 Working Notes*, 9–12 September 2019 (2019).
- Roberts, R. J. Pubmed central: the genbank of the published literature. *Proc. Natl Acad. Sci. USA* **98**, 381–382 (2001).
- Touvron, H. et al. Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
- Bethesda, M. Medpix™ receives patent (2006).
- Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (2002).
- Awadalla, A. et al. *Openflamingo* https://github.com/mlfoundations/open_flamingo (2023).
- Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022).
- Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
- Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* **3**, 1–23 (2021).
- Gao, L. et al. The pile: an 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020).
- Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *ICLR* (2019).
- Feng, J. & Huang, D. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11433–11442 (2021).
- Wang, S., Zhao, Z., Ouyang, X., Wang, Q. & Shen, D. Interactive computer-aided diagnosis on medical image using large language models. *Commun Eng* **3**, 133 (2024).

31. Nicolson, A., Dowling, J. & Koopman, B. A concise model for medical image captioning. In *CLEF (Working Notes)* 1611–1619 (2023).
32. Yang, J., Shi, R. & Ni, B. Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (ed. Ludivine, F.) 191–195 (IEEE, 2021).
33. Bajwa, J., Munir, U., Nori, A. & Williams, B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc. J.* **8**, e188–e194 (2021).
34. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
35. Demirhan, H. & Zadrozny, W. Survey of multimodal medical question answering. *BioMedInformatics* **4**, 50–74 (2023).
36. Park, J., Oh, K., Han, K. & Lee, Y. H. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Sci. Rep.* **14**, 13218 (2024).
37. Safranek, C. W., Sidamon-Eristoff, A. E., Gilson, A. & Chartash, D. The role of large language models in medical education: applications and implications. *JMIR Med. Educ.* **14**, e50945 (2023).
38. Lu, P. et al. Mathvista: evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR* (2024).
39. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023).
40. Liu, J. et al. Qilin-med-vl: towards Chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956 (2023).
41. Li, L. et al. Silk: preference distillation for large visual language models. arXiv preprint arXiv:2312.10665 (2023).
42. Chen, Z. et al. Chexagent: towards a foundation model for chest x-ray interpretation. In *AAAI Spring Symposium Series* (2024).
43. Zhan, C., Zhang, Y., Lin, Y., Wang, G. & Wang, H. Unidcp: unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts. *IEEE Transactions on Multimedia* **26**, 9736–9748 (2023).
44. Seyfioglu, M. S., Ikezogwo, W. O., Ghezloo, F., Krishna, R. & Shapiro, L. Quilt-llava: visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *CVPR* (2024).
45. Wu, J., Kim, Y. & Wu, H. Hallucination benchmark in medical visual question answering. In *ICLR Workshop* (2024).
46. Chen, X. et al. Chatffa: interactive visual question answering on fundus fluorescein angiography image using chatgpt. Available at SSRN 4578568.
47. Moor, M. et al. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)* 353–367 (PMLR, 2023).
48. Hu, Y. et al. Omnimedvqa: a new large-scale comprehensive evaluation benchmark for medical llm. arXiv preprint arXiv:2402.09181 (2024).
49. Chiang, W.-L. et al. Vicuna: An Open-source Chatbot Impressing gpt-4 with 90%* chatgpt Quality. <https://lmsys.org/blog/2023-03-30-vicuna/> (2023).
50. Zhang, X. *Pmc-vqa Dataset* <https://doi.org/10.5281/zenodo.14286358> (2024).
51. Zhang, X. & Wu, C. *xiaoman-zhang/pmc-vqa: Release pmc-vqa* <https://doi.org/10.5281/zenodo.14286350> (2024).
52. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. Towards visual question answering on pathology images. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, August 1–6, 2021 708–718 (Association for Computational Linguistics, 2020).
53. Jones, K. N., Woode, D. E., Panizzi, K. & Anderson, P. G. Peir digital library: online resources and authoring system. In *Proc. AMIA Symposium* Vol. 1075 (American Medical Informatics Association, 2001).
54. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
55. Wang, X. et al. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 2097–2106 (2017).
56. Kavur, A. E. et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Med. Image Anal.* **69**, 101950 (2021).
57. Ben Abacha, A., Sarrouiti, M., Demner-Fushman, D., Hasan, S. A. & Müller, H. Overview of the vqa-med task at imageclef 2021: visual question answering and generation in the medical domain. In *Proc. CLEF 2021 Conference and Labs of the Evaluation Forum-working Notes*, 21–24 September 2021 (2021).
58. Pelka, O., Koitka, S., Rückert, J., Nensa, F. & Friedrich, C. M. 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* (ed. Stoyanov, D.), 180–189 (Springer, 2018).
59. Subramanian, S. et al. Medicat: A dataset of medical images, captions, and textual references. In *Findings of EMNLP* (2020).
60. Changpinyo, S., Sharma, P., Ding, N. & Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3558–3568 (2021).
61. Zhang, S. et al. Opt: open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022).
62. Li, C. et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36**, (2024).
63. Zhang, S. et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023).
64. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
65. Cheng, J. et al. Dwt-cv: dense weight transfer-based cross validation strategy for model selection in biomedical data analysis. *Future Gener. Comput. Syst.* **135**, 20–29 (2022).
66. Ge, X., Qu, Y., Shang, C., Yang, L. & Shen, Q. A self-adaptive discriminative autoencoder for medical applications. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 8875–8886 (2022).
67. John, R., Datta, A. & Owallath, S. A case of euthyroid steroid-responsive encephalopathy with subacute dementia. *Cureus* **13**, e17689 (2021).
68. Chaya, S., De Decker, R., Zampoli, M. & Vanker, A. An unusual cause of haemoptysis in childhood: a case report and literature review. *Afr. J. Thorac. Crit. Care Med.* **24**, 104–106 (2018).
69. Suresh, K., Figart, M. W., Mehmood, T., Butt, A. & Sherwal, A. Covid-19-associated spontaneous pneumomediastinum and pneumopericardium: review of case series. *Cureus* **13**, e19546 (2021).
70. Kapoor, T., Dubey, P. & Ray, K. Time-lapse imaging of drosophila testis for monitoring actin dynamics and sperm release. *STAR Protoc.* **3**, 101020 (2022).
71. Joshi, T. P., Marchand, S. & Tschen, J. Malignant proliferating trichilemmal tumor: a subtle presentation in an African American woman and review of immunohistochemical markers for this rare condition. *Cureus* **13**, e17289 (2021).
72. Wang, B. et al. Ultrasound elastography for the evaluation of lymph nodes. *Front. Oncol.* **11**, 714660 (2021).
73. Bakouny, Z. et al. Normative spino-pelvic sagittal alignment of Lebanese asymptomatic adults: comparisons with different ethnicities. *Orthop. Traumatol.: Surg. Res.* **104**, 557–564 (2018).
74. Yamamoto, H. et al. Miniature erupting volcano-shaped mitral valve aneurysm secondary to *Streptococcus agalactiae* st1656 endocarditis: a case report. *Front. Cardiovasc. Med.* **8**, 728792 (2021).

75. Hajhosseiny, R. et al. Coronary magnetic resonance angiography in chronic coronary syndromes. *Front. Cardiovasc. Med.* **8**, 682924 (2021).

Author contributions

X.Z., C.W., and W.X. designed the study, X.Z. and C.W. conducted the implementation and experiments, analyzed the results and wrote the manuscript. Z.Z. and W.L. collected the original dataset and provided results of baseline models. W.X. revised the manuscript. Y.Z., Y.W., and W.X. provided supervision. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s43856-024-00709-2>.

Correspondence and requests for materials should be addressed to Yanfeng Wang or Weidi Xie.

Peer review information *Communications Medicine* thanks Jiancheng Yang and Vishwa Parekh for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024