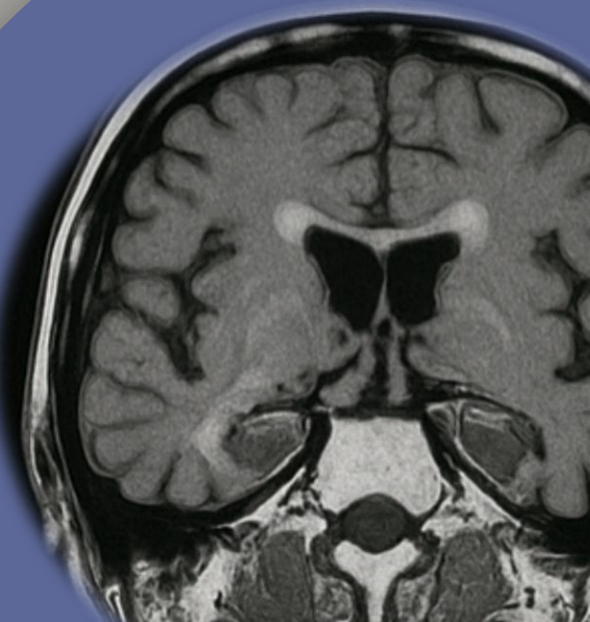
Vision-Language Models in medical image analysis

Presentation by: Sahel Sharifpoor
Supervised by: Dr. Tanhapour

June 2, 2025



Overview

- Medical Image and Analysis
- Visual-Language Models
- Multimodal Models
- VLMs in Medical Image Analysis
- Review articles

“ Medical Images ang Analysis”

What is Medical Imaging?

- **Definition:** Medical imaging refers to the process of creating images from inside the human body.
- **Primary Goal:** Its primary goal is to use these images for clinical analysis and medical intervention.
- **Application:** These images help in accurately identifying internal structures and their functions.
- **Tools and Techniques:** Medical imaging uses various tools and techniques to provide detailed visuals of tissues and organs.

What Is Medical Image Analysis?

- **Definition:** Processing and interpreting medical images to extract useful information.
- **Purpose:** Assist in diagnosis, prediction, and decision-making.
- **Importance:** Enhances accuracy and reduces human error.

Input: Image → Processing → Output :Diagnosis

A Brief History of Medical Imaging

1895

Discovery of X-rays by Wilhelm Röntgen

1970s

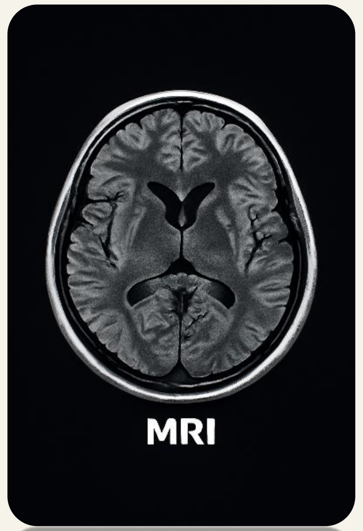
Emergence of CT Scan and MRI

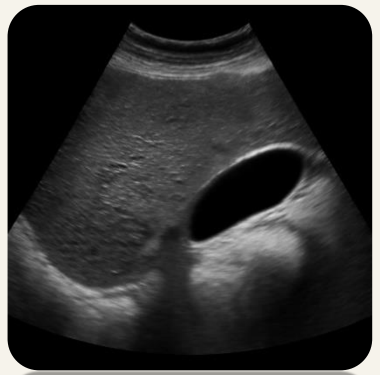
2000s Onward

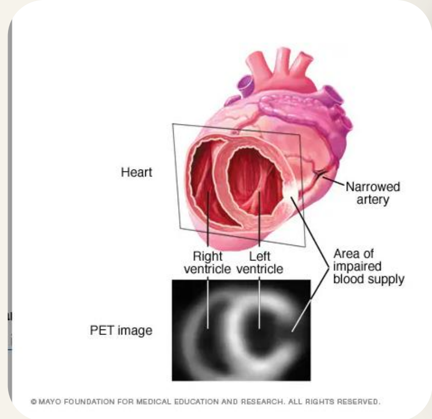
Advances in Digital Imaging and Artificial Intelligence

Types of Medical Images

- **X-ray:** For bones and fractures.
- **Computed Tomography (CT):** Detailed cross-sectional images.
- **Magnetic Resonance Imaging (MRI):** Soft tissue details.
- **Ultrasound:** Non-invasive imaging with sound waves.
- **Nuclear Imaging (PET/SPECT):** Functional analysis of organs.







Importance of Medical Image Analysis

- **Early Detection**
- **Accurate Diagnosis**
- **Treatment Planning**
- **Improved Outcomes**

Traditional Methods of Medical Image Analysis

- **Manual analysis by radiologists:**

Medical images are traditionally examined by radiologists, relying on their experience and pattern recognition skills.

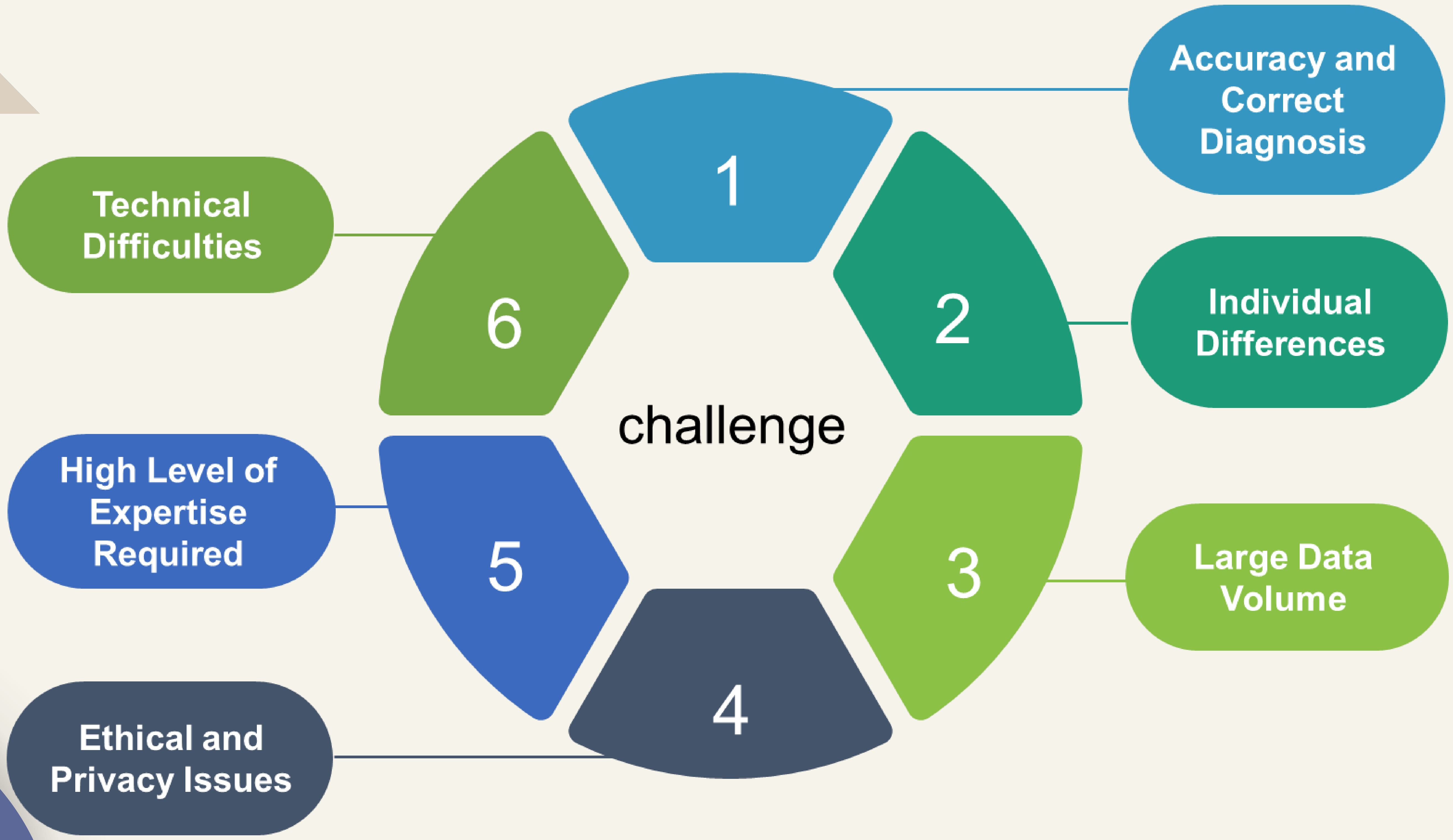
- **Use of Basic Software Tools:**

Simple tools like filtering and contrast adjustment help, but offer limited support in analysis.

- **Use of Traditional AI Methods:**

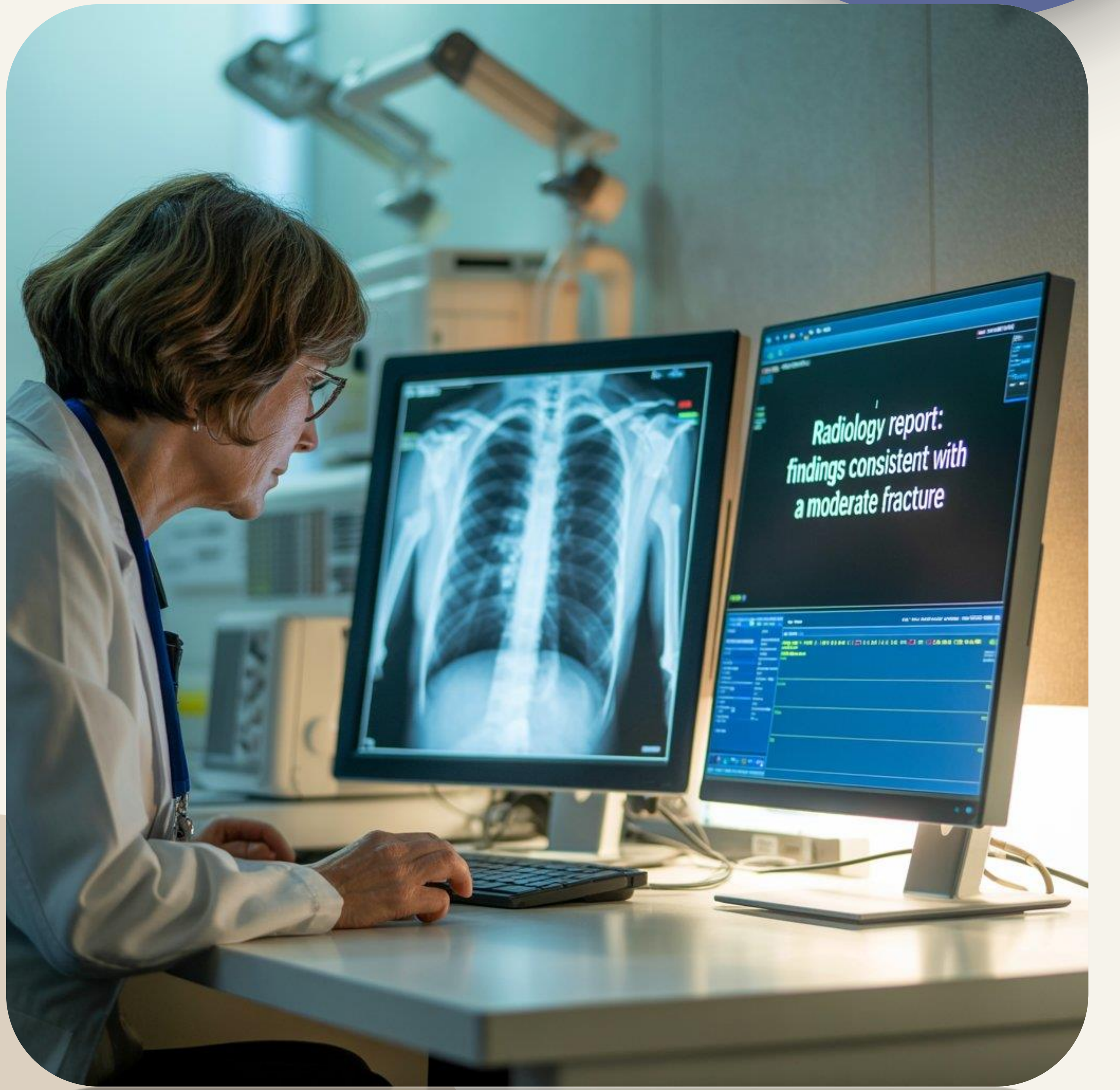
Before deep learning, algorithms like SVM, KNN, and threshold-based segmentation were used. These methods analyzed medical images in a one-dimensional way, lacking the ability to capture complex interactions. They also required manual feature extraction and had limited accuracy.

Challenges in Medical Image Analysis



Why is this matter important?

- Medical images (CT, MRI, X-ray) are central to diagnosis.
- **Challenge:** Interpretation is time-consuming, prone to human error and often one-dimensional.
- **Solution:** The Use of Multimodal Models in Medical Image Analysis.



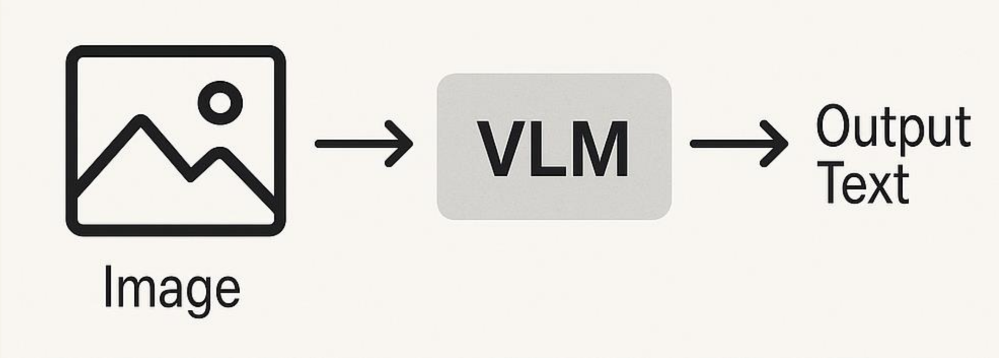
“Multimodal Models”

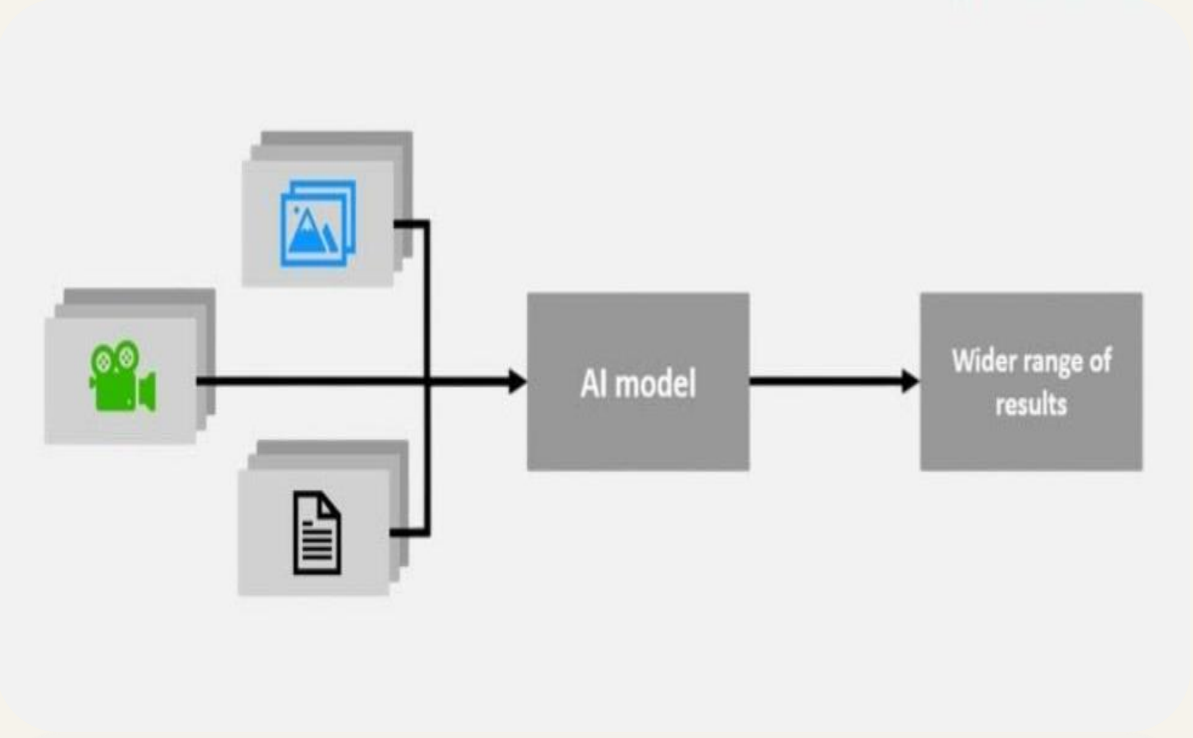
Multimodal Models

- Multimodal models refer to systems that can process and integrate data from different sources (modalities). These models combine various types of data such as images, text, audio, and video to achieve a better understanding of the data.

Types of Multimodal Models:

- Image-Text Models
- Audio-Text Models
- Video-Text Models
- Cross-modal Models



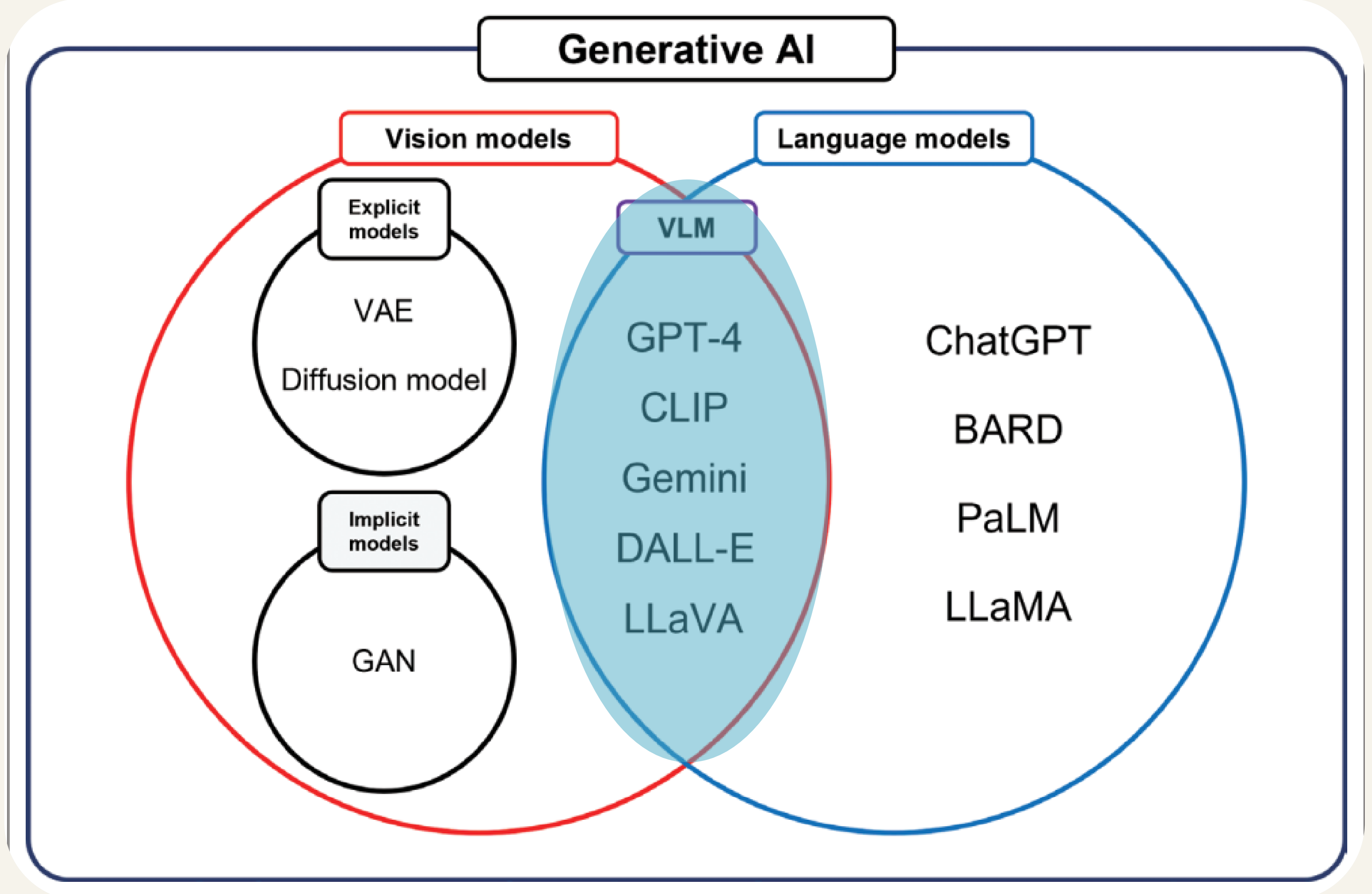


“Visual-Language Models”

What Are Vision-Language Models (VLMs)?

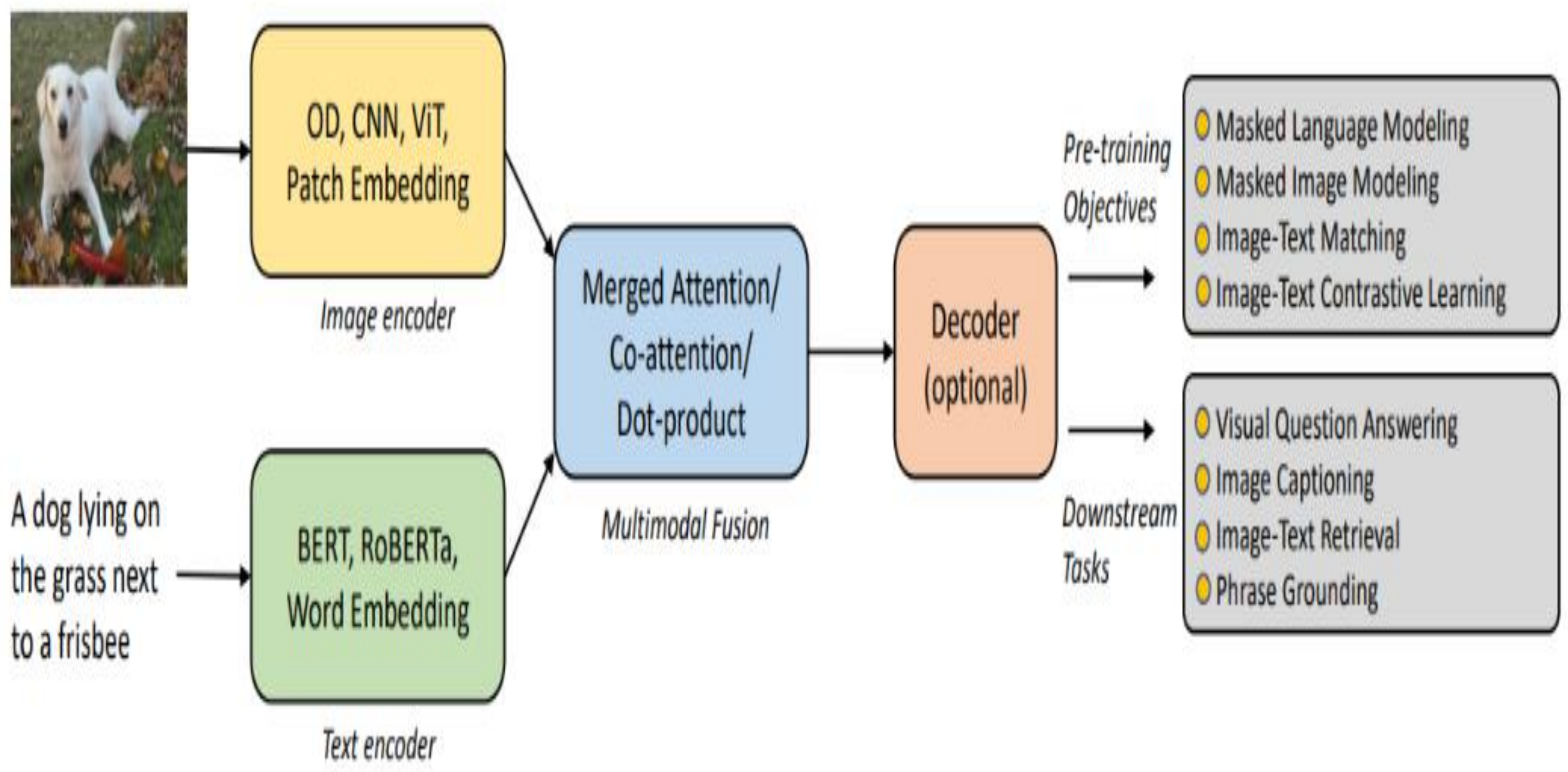
- Vision language models (VLMs) are a type of artificial intelligence (AI) model that can understand and generate text about images. They do this by combining computer vision and natural language processing models. VLMs can take image inputs and generate text outputs.

- Vision language models (VLMs) are multimodal, generative AI models capable of understanding and processing video, image, and text.



general VLM architecture

- **Architecture:** VLMs typically use a transformer-based architecture, combining components like convolutional neural networks (CNNs) or vision transformers (ViTs) for image processing and large language models (LLMs) for text processing. These are fused into a unified model, often with cross-attention mechanisms to align visual and textual features.

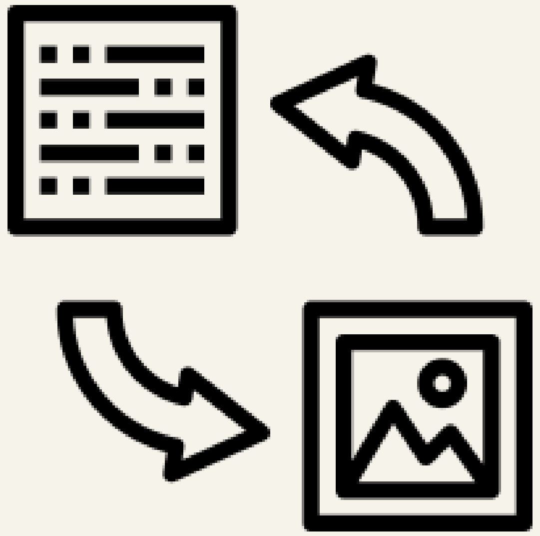


- Image Encoder
- Text Encoder
- Multimodal Fusion
- Decoder

Fusion in VLMs

- **Fusion** is the process of combining information from different modalities (e.g., images and text) to produce a unified representation.

Types of Fusion:



- **Early Fusion:** Information from different modalities is combined early in the process, often before feature extraction.
  - ➢ Example: Concatenating raw image pixels and text tokens together.
- **Late Fusion:** Modalities are processed separately, and their outputs are combined at a later stage, after individual processing.
  - ➢ Example: Separate image and text encoders, with outputs merged for final processing.
- **Hybrid Fusion:** A combination of early and late fusion, where some features are combined early, and others are fused later in the model.
  - ➢ Example: Early fusion of simple features, with later fusion of more complex features or predictions.

Popular Types of VLMs

CLIP (Contrastive Language-Image Pre-training)

BLIP (Bootstrapped Language Image Pretraining)

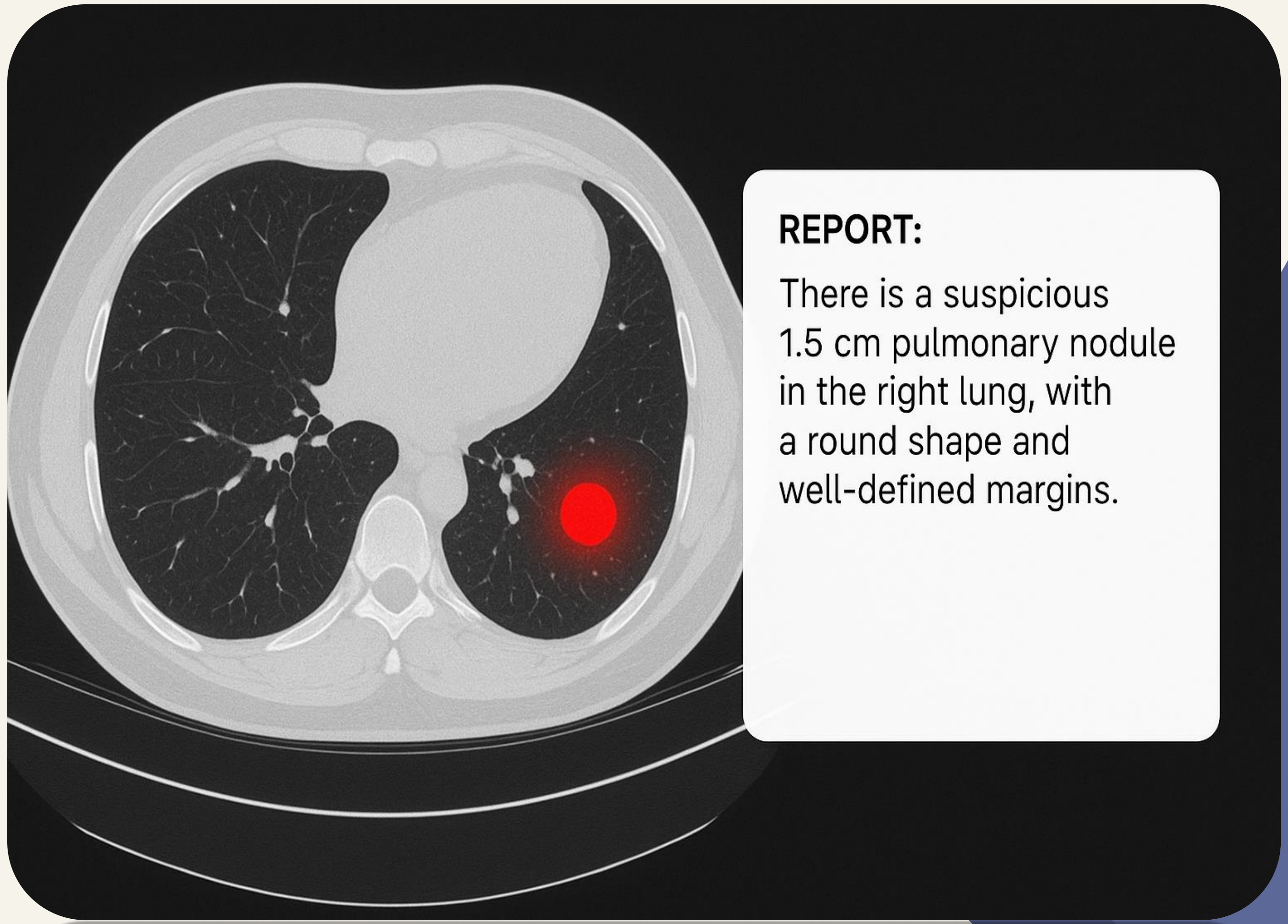
ViLBERT (Vision-and-Language BERT)

Flamingo

“Applications of VLMs in Medical Image Analysis”

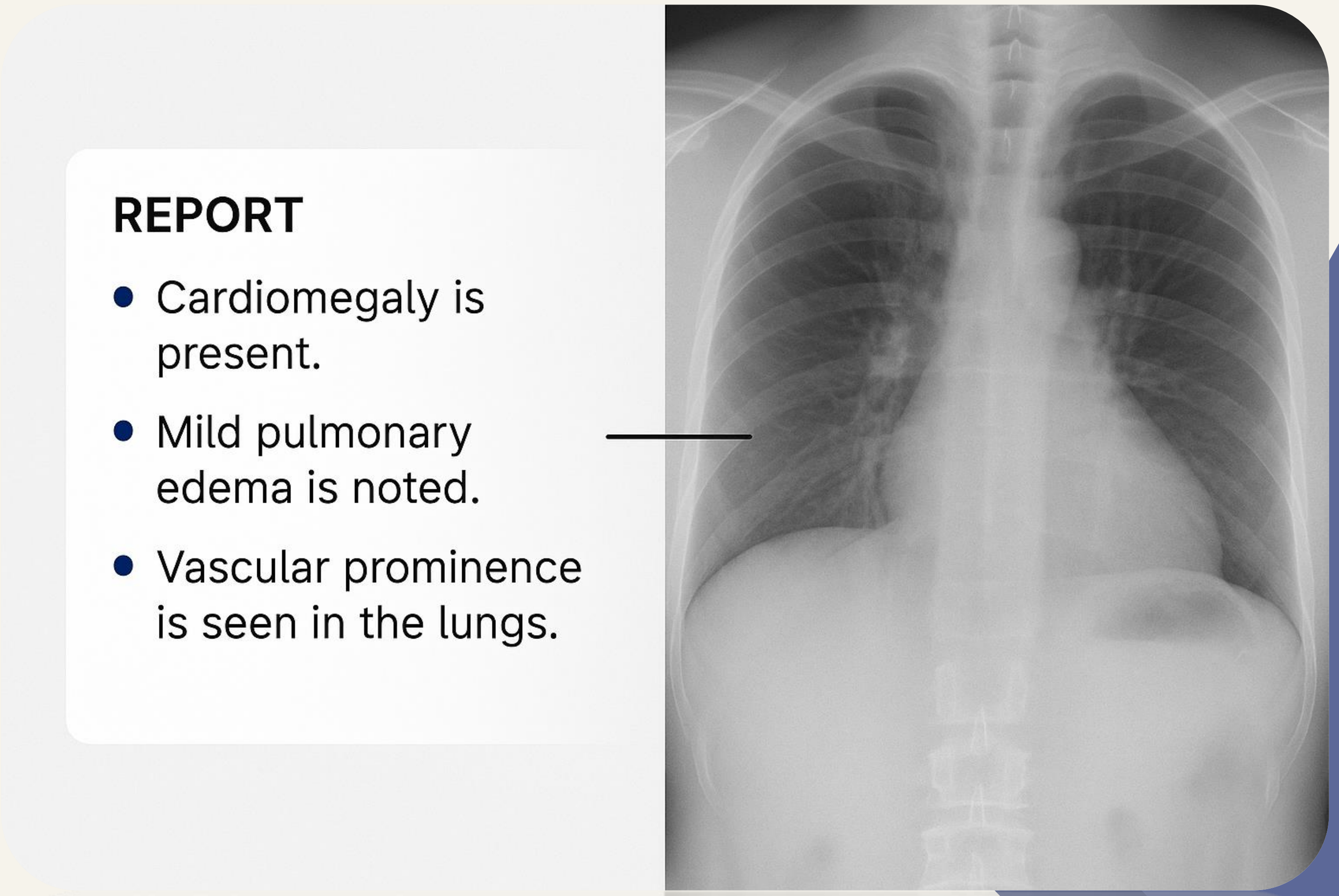
Disease Detection

- **Automated Diagnosis:** VLMs identify diseases like cancer or neurological disorders by analyzing medical images (e.g., mammograms, brain MRIs) alongside clinical text.
- **Enhanced Accuracy:** Combining imaging data with textual descriptions (e.g., patient symptoms) improves diagnostic precision.
- **Example:** Models like Med-PaLM achieve up to 20% higher accuracy in detecting lung nodules compared to traditional methods.



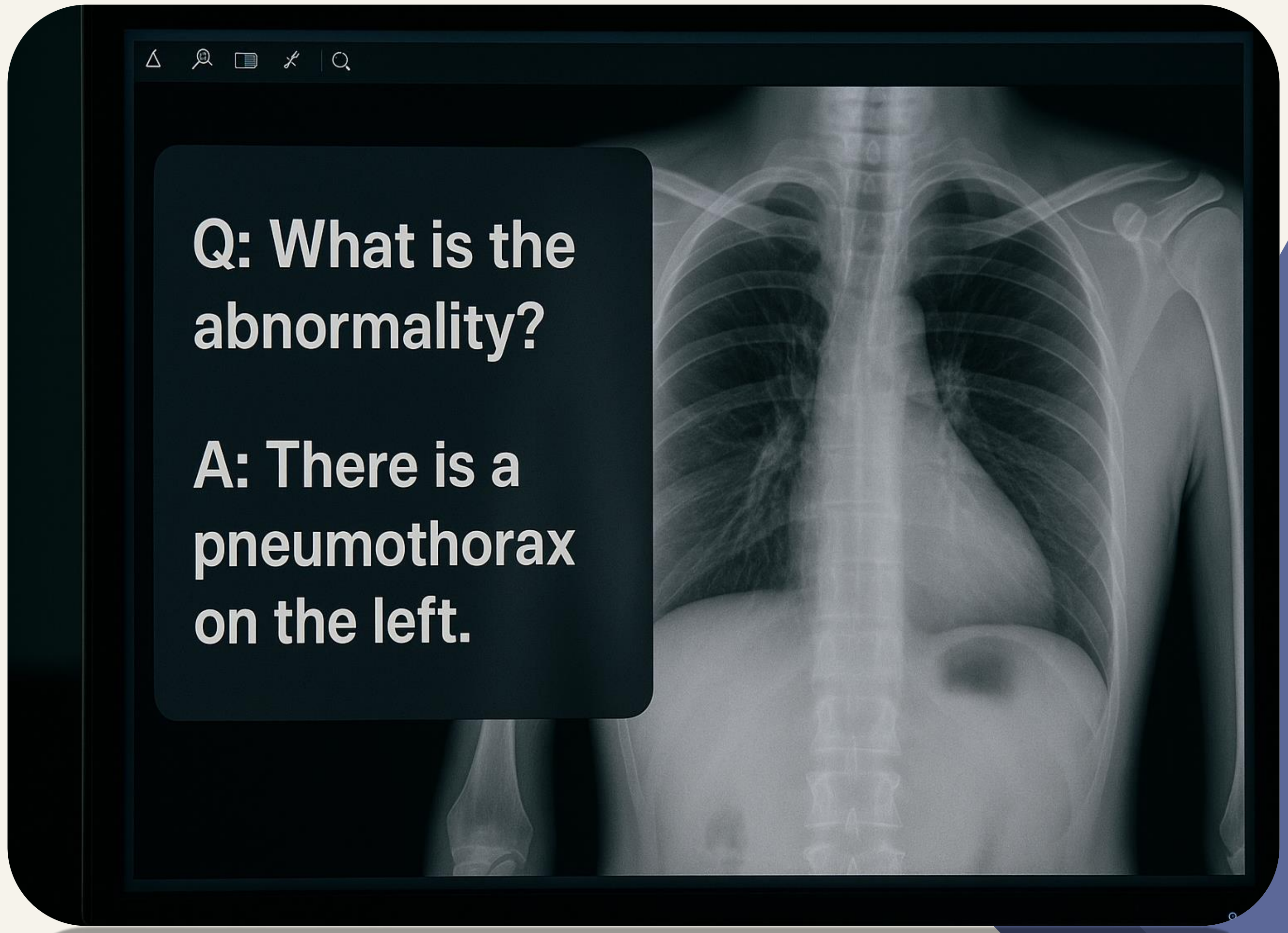
Automated Report Generation

- **Report Writing**: VLMs generate detailed radiology reports from medical images, reducing workload for radiologists.
- **Contextual Understanding**: Models interpret images in the context of patient history or clinical notes, producing coherent reports.
- **Case Study**: CheXzero generates chest X-ray reports with 90% alignment to expert radiologist reports.



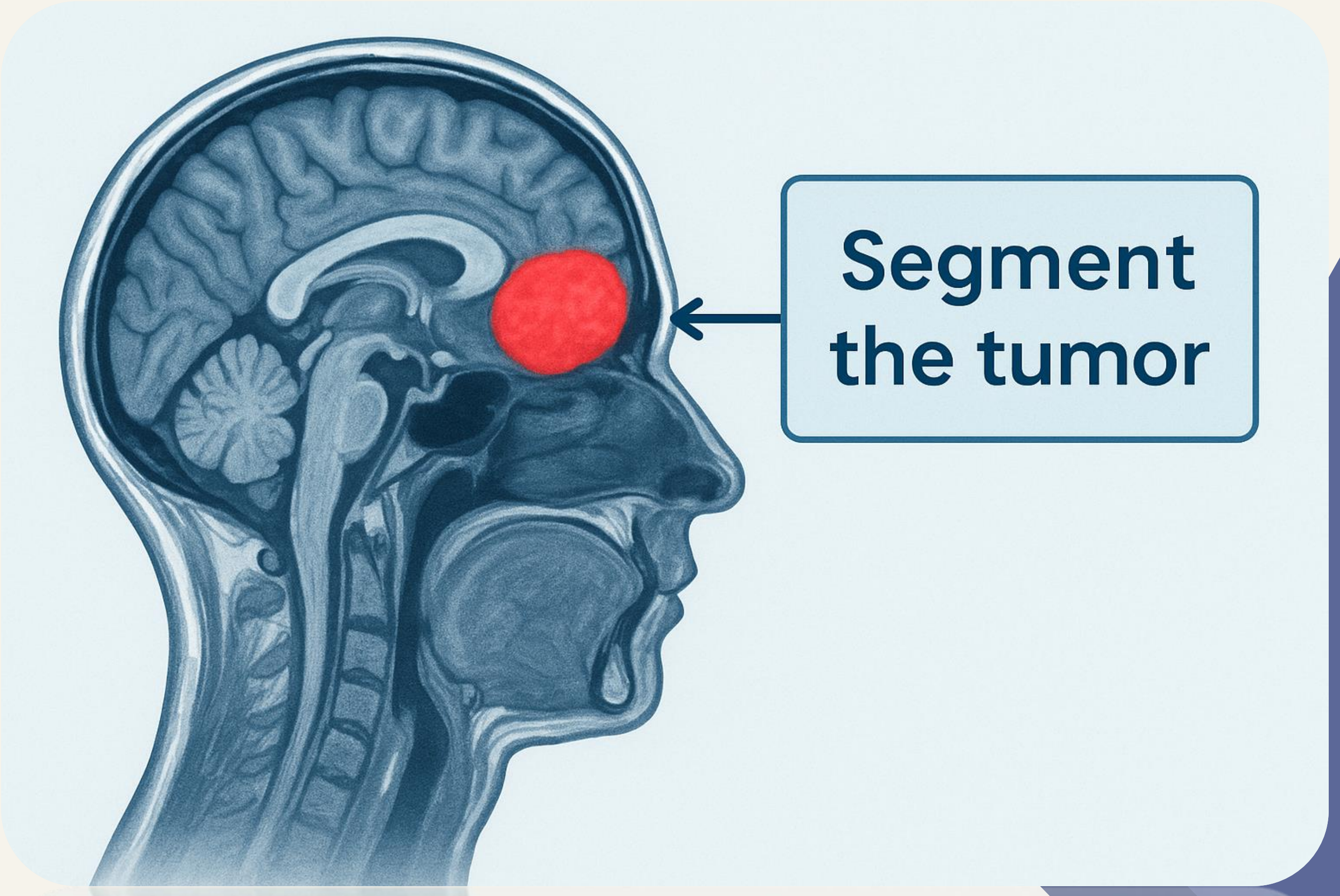
Visual Question Answering (VQA) in Medicine

- **Interactive Diagnostics:** VLMs answer specific questions about medical images (e.g., “Is there a fracture in this X-ray?”).
- **Training and Education:** Supports medical students by providing real-time feedback on image-based queries.
- **Application:** VLM-based tools assist in telemedicine, enabling remote consultations with precise image analysis.



Medical Image Segmentation

- **Precise Segmentation:** VLMs perform accurate segmentation of medical images (e.g., identifying tumors or organs in MRI/CT scans) using natural language prompts to guide the process.
- **Context-Aware Analysis:** Language inputs (e.g., "Segment the tumor in this MRI") enable precise delineation of regions of interest.
- **Example:** Models like LLaVA-Med achieve high Dice scores (e.g., 0.85) for tumor segmentation, outperforming traditional methods.



Limitations of VLMs

Requirement for large training datasets

Computational complexity

Generalization challenges with new data

“Review articles”

JMIR Publications
Advancing Digital Health & Open Science

Articles

Career Center Login Register

JMIR Formative Research | Journal Information | Browse Journal | Submit Article

Published on 08.02.2024 in Vol 8 (2024)

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/32690, first published October 17, 2023.

Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study

Jia Ji[1]; Yongshuai Hou[2]; Xinyu Chen[3]; Youcheng Pan[2]; Yang Xiang[2]

Citation

Please cite as:

Ji J, Hou Y, Chen X, Pan Y, Xiang Y
Vision-Language Model for Generating Textual Descriptions From Clinical Images: Model Development and Validation Study
JMIR Form Res 2024;8:e32690
doi: 10.2196/32690
PMID: 38329788
PMCID: 10884898

Copy Citation to Clipboard

Export Metadata

goal

- Develop ClinicalBLIP model that produces precise radiology reports from clinical images using prior medical knowledge.

Methods

- **Model**: Built on InstructBLIP, it includes a visual encoder, a query transformer (Q-Former) to connect images and text, and a large language model (LLM) for report generation.
- **Fine-Tuning**:

  .Stage 1: Enhances image understanding without extra data.
  .Stage 2: Combines image and text data with medical knowledge (e.g., medical tags)

- **Datasets**: Tested on IU X-RAY (3337 reports) and MIMIC-CXR (152,173 training reports).
- **Metrics**: Evaluated with BLEU, METEOR, and ROUGE-L for report quality.

Results

- **IU X-RAY**: Achieved METEOR (0.570) and ROUGE-L (0.534), outperforming other models. BLEU-A (0.296) was also strong.
- **MIMIC-CXR**: Scored METEOR (0.365) and ROUGE-L (0.313), better than competitors, but BLEU-A (0.162) was slightly lower.
- **Case Studies**: Reports were mostly accurate but had minor differences in details or wording.

Conclusion

- ClinicalBLIP generates reliable radiology reports quickly, showing great potential to assist doctors. It needs further refinement for precise terminology and details.

SPRINGER NATURE Link

Log in

Find a journal Publish with us Track your research Search Cart

Home > Artificial Intelligence Review > Article

Pathologyvlm: a large vision-language model for pathology image understanding

Open access | Published: 28 March 2025

Volume 58, article number 186, (2025) Cite this article

Artificial Intelligence Review

Aims and scope →

Submit manuscript →

Download PDF

You have full access to this open access article

Dawei Dai, Yuanhui Zhang, Qianlan Yang, Long Xu, Xiaojing Shen, Shuyin Xia & Guoyin Wang

Use our pre-submission checklist →

Avoid common mistakes on your manuscript.

goal

- Create PathologyVLM model to understand pathology images and answer questions using public data.

Methods

Data:

- Training Data: The researchers created a large dataset consisting of pathology images and their corresponding textual descriptions, which were used to train the model. This dataset includes 1.4 million image-description pairs.

- **Model**:
- PLIP: A pathology-specific encoder, replacing CLIP, for better image understanding.
- Connector: Prevents image detail loss by avoiding scaling.
- Language Model: LLaMA3, fine-tuned for pathology.

Training: Three stages: PLIP pretraining, data alignment, and VQA fine-tuning.

Results

Supervised VQA:

- PathVQA: 92.51% accuracy (closed-set), 67.43% recall (overall), outperforming LLaVA (63.20%) and LLaVA-Med (91.09%).
- PMC-VQA: 39.89% accuracy, better than Quilt-LLaVA (33.34%).

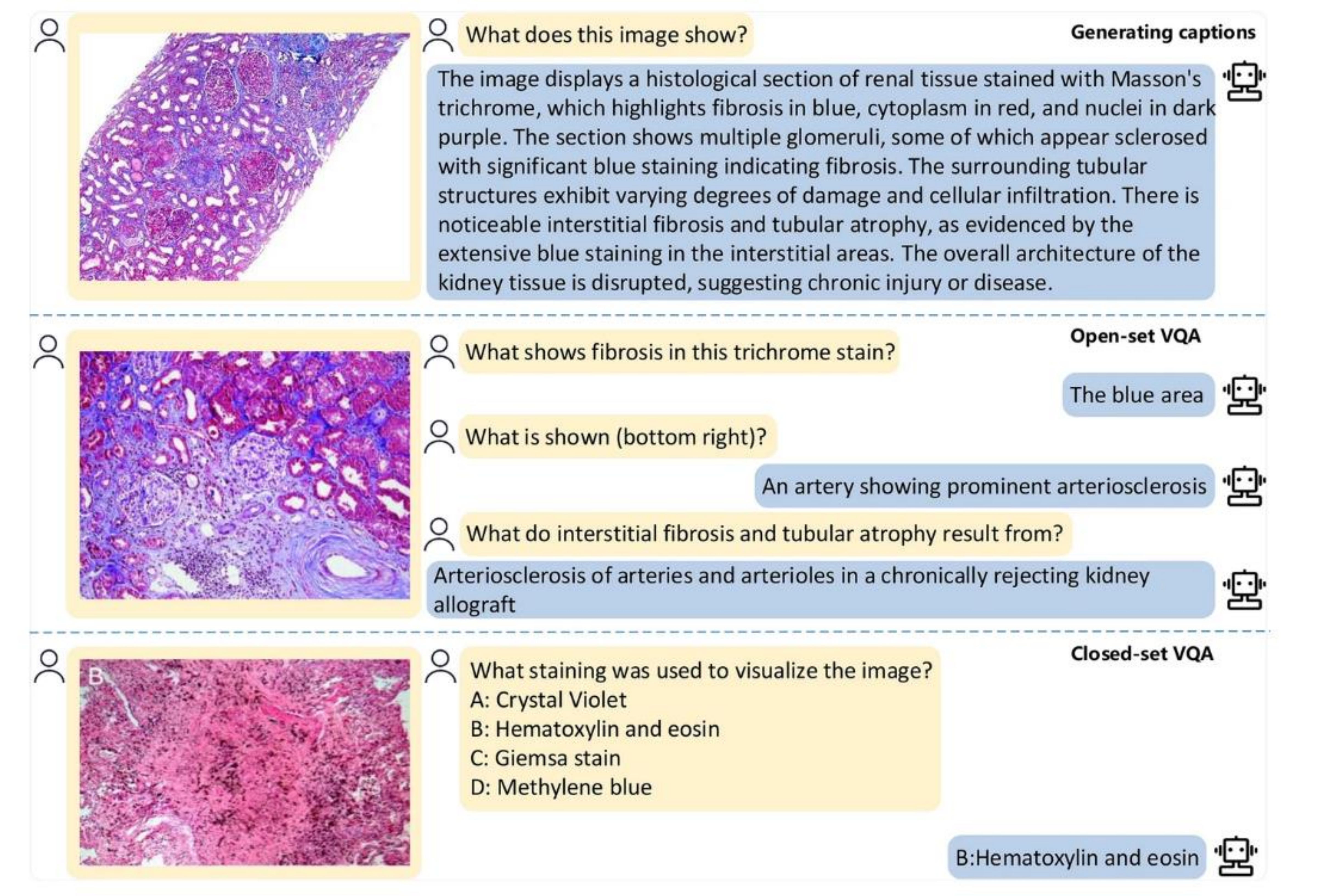
Zero-Shot: Excelled on BACH, OSCC, and ColonPath compared to other models.

Manual Evaluation: Beat doctors in supervised VQA (92.51% vs. 71.15%) but lagged in zero-shot (62.98% vs. 83.30%) due to image quality.

Conclusion

PathologyVLM is strong at understanding pathology images and answering questions. Its open-source data and code support research, but image quality and data balance need improvement.

Pathology VLM



nature medicine

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Sign up for alerts | RSS feed

nature > nature medicine > articles > article

Article | Open access | Published: 07 November 2024

Collaboration between clinicians and vision–language models in radiology report generation

Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, Karan Singhal, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Sara Mahdavi, Zahra Ahmed, Yossi Matias, Joelle Barral, S. M. Ali Eslami, Danielle Belgrave, Yun Liu, Sreenivasa Raju Kalidindi, Shravya Shetty, ... Ira Ktena + Show authors

Nature Medicine **31**, 599–608 (2025) | Cite this article

Download PDF

Sections | Figures | References

goal

- The study aimed to develop and evaluate a cutting-edge AI system named Flamingo-CXR for automatic radiology report generation from chest X-rays (CXR). The primary objective was to assess the clinical quality of AI-generated reports, both as standalone outputs and in collaboration with radiologists, and to compare them to human-written reports.

Methods

- **Model**: Fine-tuning the Flamingo vision-language model on two large deidentified datasets — MIMIC-CXR (from the U.S.) and IND1 (from India).

- **Evaluation Framework**:

1. Pairwise Preference Test: Certified radiologists compared AI-generated reports with human-written ones and indicated preference or equivalence.
2. Error Correction Task: Radiologists reviewed and edited reports (either AI or human-written) to identify clinically significant errors.

- **Metrics**: Performance was measured using clinical metrics like CheXpert F1 and RadGraph F1, along with natural language generation metrics (BLEU, ROUGE, CIDEr).

Results

- Flamingo-CXR achieved:
  - ➢ CheXpert F1: 0.519
  - ➢ RadGraph F1: 0.205 (Both metrics showed ~33% improvement over previous state-of-the-art).
- In 77.7% of IND1 cases and 56.1% of MIMIC-CXR cases, the AI reports were rated as equivalent or superior to human reports by at least half of the radiologists.
- In 94% of normal IND1 cases, AI reports were considered as good as or better than human reports.
- Both AI and human-written reports contained errors, but many were non-overlapping, suggesting complementary strengths.
- In a collaborative setting, clinician-AI reports outperformed standalone AI reports:
  - ➢ 71.2% (IND1) and 53.6% (MIMIC-CXR) of such reports were preferred or deemed equivalent.

Conclusion

Flamingo-CXR demonstrates strong potential in generating radiology reports that are clinically useful, particularly in normal cases. The model's performance improves further when used in collaboration with clinicians, underscoring its value as an assistive tool rather than a replacement. This study highlights the importance of human-AI synergy and sets a benchmark for future AI applications in clinical radiology.

d Examples

Image	Clinician's report	AI report	No. of votes for	Preference reasons
	FINDINGS: Basilar opacity seen on the lateral view best corresponds to a retrocardiac opacity suspicious for developing left lower lobe pneumonia or aspiration event in the setting of altered mental status. Chronic peribronchiolar opacities seen bilaterally are similar in distribution and slightly more apparent due to lower lung volumes and AP technique. There is no pleural effusion or pneumothorax. The heart size is normal with normal cardiomediastinal silhouette. IMPRESSION: Left lower lobe opacity likely reflects pneumonia or aspiration.	FINDINGS: Lung volumes are low, accentuating the cardiac silhouette and bronchovascular structures. With this limitation in mind, there are no definite new areas of consolidation to suggest the presence of pneumonia. If clinical suspicion persists, repeat radiograph with improved inspiratory level may be helpful for more complete assessment of the lung bases.	4/4	*For AI report:* - No definite retrocardiac opacity identified, so I chose report B. - More accurate. - Findings are consistent with the radiograph. - I feel that labeling a left lower lobe pneumonia/aspiration is overenthusiastic in the limited setting and if there is a clinical suspicion then a repeat radiograph is warranted.
	FINDINGS: Portable AP chest radiograph demonstrates worsening bilateral pleural effusions and associated atelectasis, greater on the right. There is also worsening pulmonary vascular congestion. There is no pneumothorax. Right internal jugular catheter probably terminates in the right atrium. IMPRESSION: Worsening pulmonary edema and bilateral pleural effusions. Results were relayed to __, PA-C by phone at approximately 6:00 p.m. on __.	FINDINGS: The patient is status post median sternotomy and CABG. A right internal jugular central venous catheter tip terminates at the junction of the SVC and right atrium No pneumothorax is identified. Moderate to large layering bilateral pleural effusions are again demonstrated, slightly larger on the right and slightly improved on the left. Bibasilar airspace opacities likely reflect atelectasis. IMPRESSION: Right internal jugular central venous catheter tip at the junction of the SVC and right atrium. No pneumothorax. Moderate to large layering bilateral pleural effusions and bibasilar atelectasis.	2/4	*For AI report:* - [Radiologist's] Report does not mention post-sternotomy status. - Findings are consistent with radiograph. *For clinician's report:* - Right IJ line likely terminates in the right atrium so I chose report X. Pulmonary findings in reports are similar. - More concise. Both fail to mention small lung volumes and chin flexed significantly limiting exam.
	FINDINGS: Low lung volumes are present. This accentuates the size of the cardiac silhouette which is likely mildly enlarged. Mediastinal and hilar contours are likely within normal limits. A right brachiocephalic venous stent is re-demonstrated. There is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. No pleural effusion or pneumothorax is identified. IMPRESSION: Low lung volumes with mild pulmonary vascular congestion.	FINDINGS: The heart is mildly enlarged. The mediastinal and hilar contours are unremarkable. There is no pleural effusion or pneumothorax. The lungs appear clear within the limitations of technique. IMPRESSION: No evidence of acute disease.	0/4	*For clinician's report:* - Report is more detailed and accurate. - Probable mild pulmonary vascular congestion as suggested in report / Also report mentions right brachiocephalic venous stent. - I agree with the report findings of mild vascular congestion. - Describes the positive findings in detail.

nature medicine

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Sign up for alerts | RSS feed

nature > nature medicine > articles > article

Article | Open access | Published: 30 April 2024

Vision–language foundation model for echocardiogram interpretation

Matthew Christensen, Milos Vukadinovic, Neal Yuan & David Ouyang

Nature Medicine **30**, 1481–1488 (2024) | Cite this article

Abstract

The development of robust artificial intelligence models for echocardiography has been

Download PDF

Associated content

Adapting vision–language AI models to cardiology tasks

Rima Arnaout

Nature Medicine | **News & Views** | 01 May 2024

Sections | Figures | References

Abstract

goal

- To develop EchoCLIP, a vision-language foundation model for echocardiogram interpretation that can understand and match cardiac ultrasound images with expert cardiologist reports without task-specific training.

Methods

- **Dataset**: 1,032,975 echocardiogram videos from 224,685 studies across 99,870 patients, collected over a decade at Cedars-Sinai Medical Center.
- **Architecture**:
  - ➢ Image encoder: ConvNeXt.
  - ➢ Text encoder: Transformer-based decoder (CLIP-style).
- Training Strategy:Trained via self-supervised learning using paired image-text data.
- Evaluation: Performance tested internally and on external datasets (EchoNet-Dynamic), with classification (AUC) and regression (MAE) metrics, image-text retrieval (MCMRR), and interpretability analyses (PromptCAM and UMAP).

Results

- High accuracy in identifying cardiac devices (AUC up to 0.97).
- Left ventricular ejection fraction (LVEF) predicted with 7.1% MAE on external data.
- Strong performance in image-text retrieval, patient identification, and tracking clinical changes.
- Interpretability through PromptCAM showing relevant image-text associations.

Conclusion

- EchoCLIP effectively performs multiple echocardiographic tasks without dedicated training, offering a scalable tool for automated cardiac imaging interpretation. It could enhance clinical decision-making, especially in underserved settings. Future work will include video-based encoding and multi-view analysis.

MDPI Journals Topics Information Author Services Initiatives About Sign In / Sign Up Submit

Search for Articles: Title / Keyword Author / Affiliation / Email Bioengineering All Article Types Search Advanced

Journals / Bioengineering / Volume 10 / Issue 3 / 10.3390/bioengineering10030380

bioengineering

Submit to this Journal

Review for this Journal

Propose a Special Issue

Article Menu

Academic Editor

Alan Wang

Subscribe SciFeed

Order Article Reprints

Share Help Cite Discuss in SciProfiles

Open Access Article

Vision–Language Model for Visual Question Answering in Medical Imagery

by Yakoub Bazi [1,*], Mohamad Mahmoud Al Rahhal [2], Laila Bashmal [1] and Mansour Zuair [1]

1 Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

2 Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia

* Author to whom correspondence should be addressed.

Bioengineering **2023**, *10*(3), 380; https://doi.org/10.3390/bioengineering10030380

Submission received: 22 February 2023 / Revised: 15 March 2023 / Accepted: 17 March 2023 / Published: 20 March 2023

goal

- The aim of this study is to develop a transformer-based model for Visual Question Answering (VQA) in the medical imaging domain. The model is designed to generate accurate, informative answers to questions posed about medical images, ultimately supporting and enhancing clinical diagnosis.

Methods

- **The proposed architecture is an encoder-decoder transformer model:**

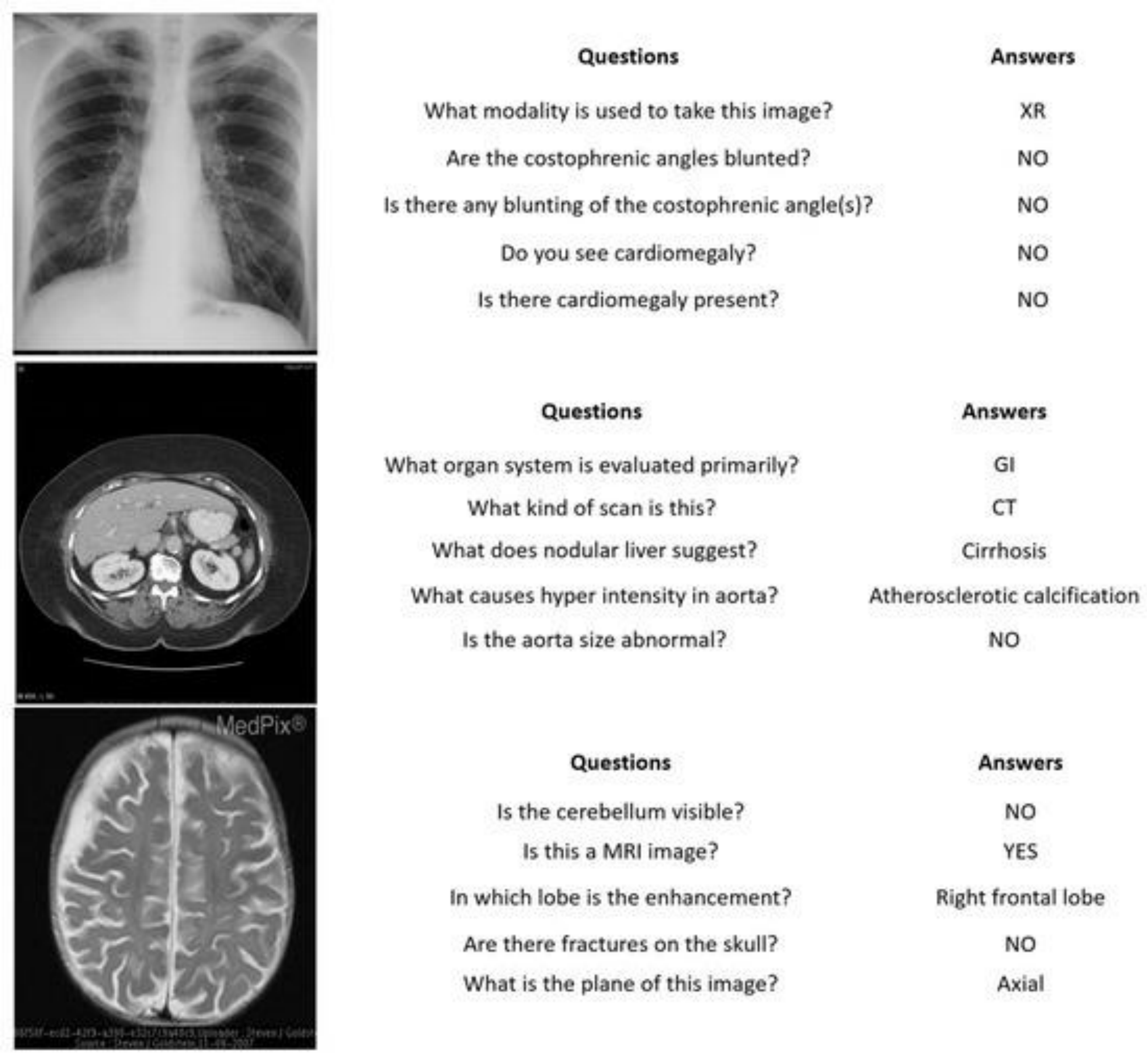
➢ Visual features are extracted using the Vision Transformer (ViT).
➢ Textual features from the question are encoded using a BERT-like language transformer.
➢ The visual and textual embeddings are concatenated and passed into a multi-modal decoder that generates the answer autoregressively.
➢ The model is trained and evaluated on two datasets: VQA-RAD and PathVQA.

Results

- VQA-RAD: Closed-ended accuracy: 84.99%, Open-ended accuracy: 72.97%, BLEU-1: 71.03%
- **PathVQA**: Closed-ended accuracy: 83.86%, Open-ended accuracy: 62.37%, BLEU-1: 61.78%
- The model outperforms existing state-of-the-art methods, especially in open-ended question answering, which is more challenging.

Conclusion

The proposed transformer-based VQA model demonstrates strong performance on two benchmark medical VQA datasets. It effectively integrates vision and language inputs to generate accurate answers. Future work includes improving performance through data augmentation and exploring more advanced multi-modal transformer architectures.

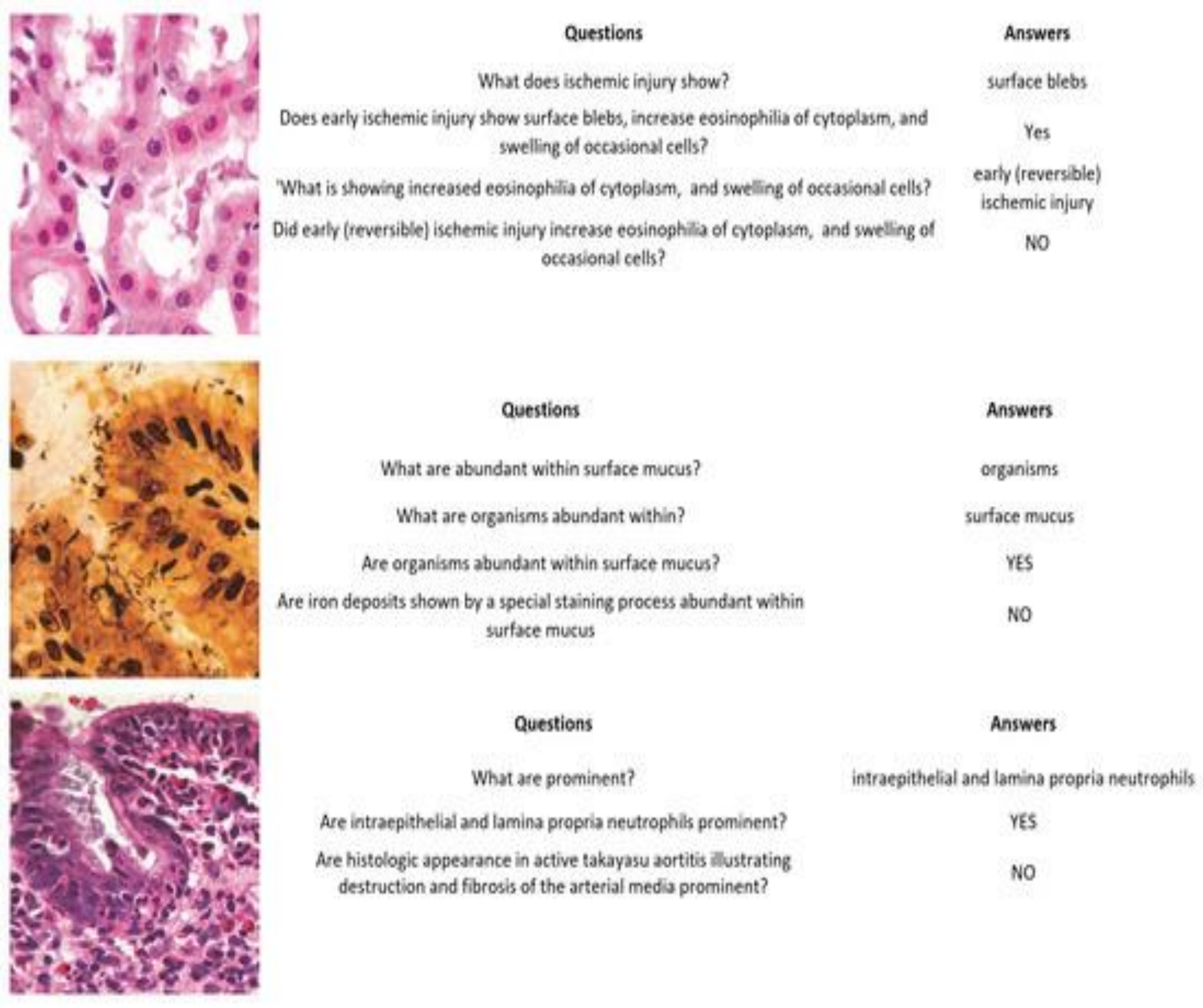


Questions	Answers
What modality is used to take this image?	XR
Are the costophrenic angles blunted?	NO
Is there any blunting of the costophrenic angle(s)?	NO
Do you see cardiomegaly?	NO
Is there cardiomegaly present?	NO

Questions	Answers
What organ system is evaluated primarily?	GI
What kind of scan is this?	CT
What does nodular liver suggest?	Cirrhosis
What causes hyper intensity in aorta?	Atherosclerotic calcification
Is the aorta size abnormal?	NO

Questions	Answers
Is the cerebellum visible?	NO
Is this a MRI image?	YES
In which lobe is the enhancement?	Right frontal lobe
Are there fractures on the skull?	NO
What is the plane of this image?	Axial

VQA-RAD images, questions, and answers.



Questions	Answers
What does ischemic injury show?	surface blebs
Does early ischemic injury show surface blebs, increase eosinophilia of cytoplasm, and swelling of occasional cells?	Yes
'What is showing increased eosinophilia of cytoplasm, and swelling of occasional cells?	early (reversible) ischemic injury
Did early (reversible) ischemic injury increase eosinophilia of cytoplasm, and swelling of occasional cells?	NO

Questions	Answers
What are abundant within surface mucus?	organisms
What are organisms abundant within?	surface mucus
Are organisms abundant within surface mucus?	YES
Are iron deposits shown by a special staining process abundant within surface mucus	NO

Questions	Answers
What are prominent?	intraepithelial and lamina propria neutrophils
Are intraepithelial and lamina propria neutrophils prominent?	YES
Are histologic appearance in active takayasu aortitis illustrating destruction and fibrosis of the arterial media prominent?	NO

PathVQA images, questions, and answers.

THANK YOU

For your attention