Medical Visual Question Answering - Multimodal Fusion

brought to you by -
Anjali Mudgal, Udbhav Kush, Aditya Kumar

Vision-Language Pre-training -> Image Scene Understanding

Why Medical VQA?

- Increased Electronic Health Records -> Increased Autonomy and Accessibility

- Where’s the gap?
  1. Not accompanied by credible, reliable and accurate diagnosis.
  2. Preventing misleading or incorrect diagnosis from open-source agents.
  3. Patient DLP

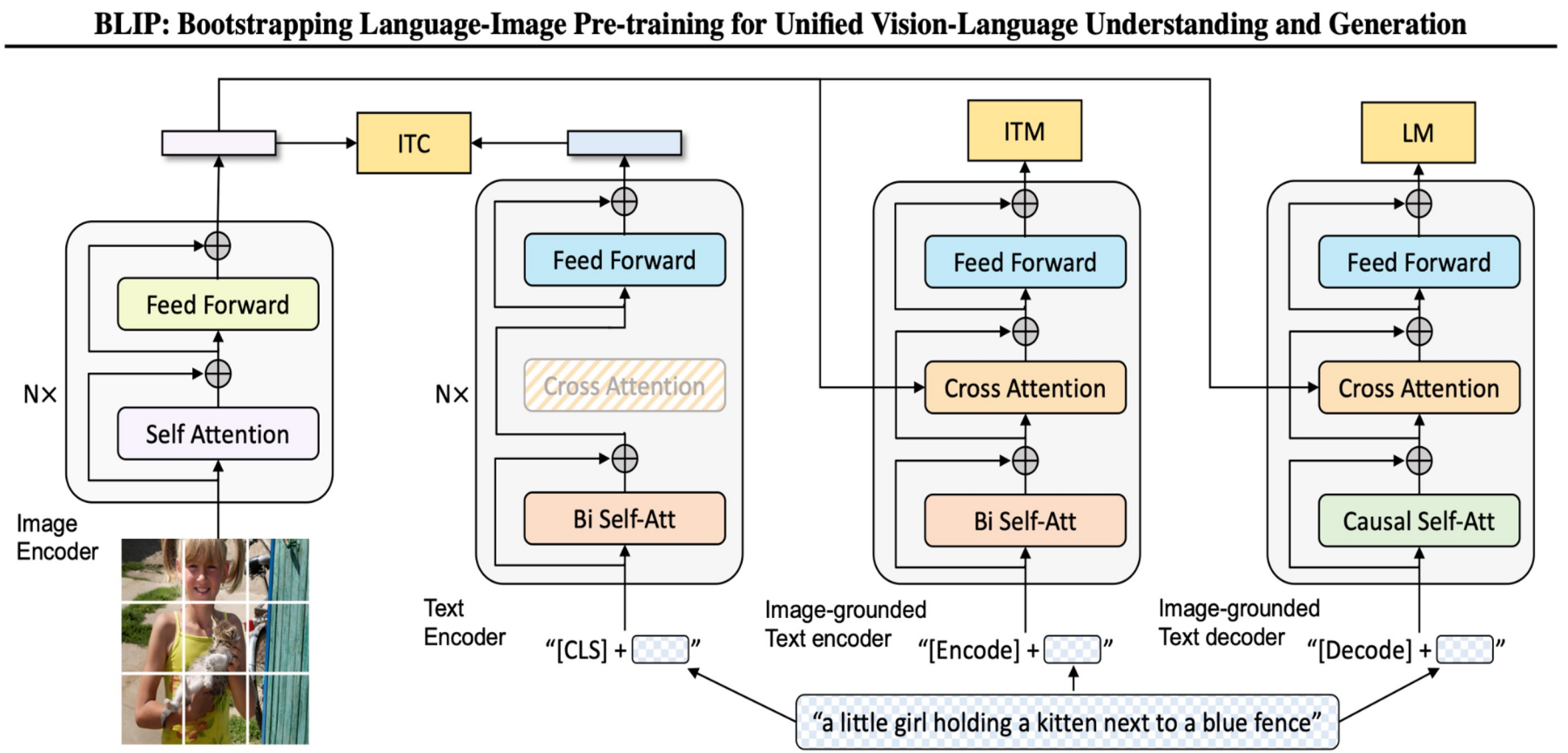
Contributions

1. Dataset limitations -> long-way from clinical applications.
   I. Combined data source from two largest annotated medical dataset and performed strategic augmentations.

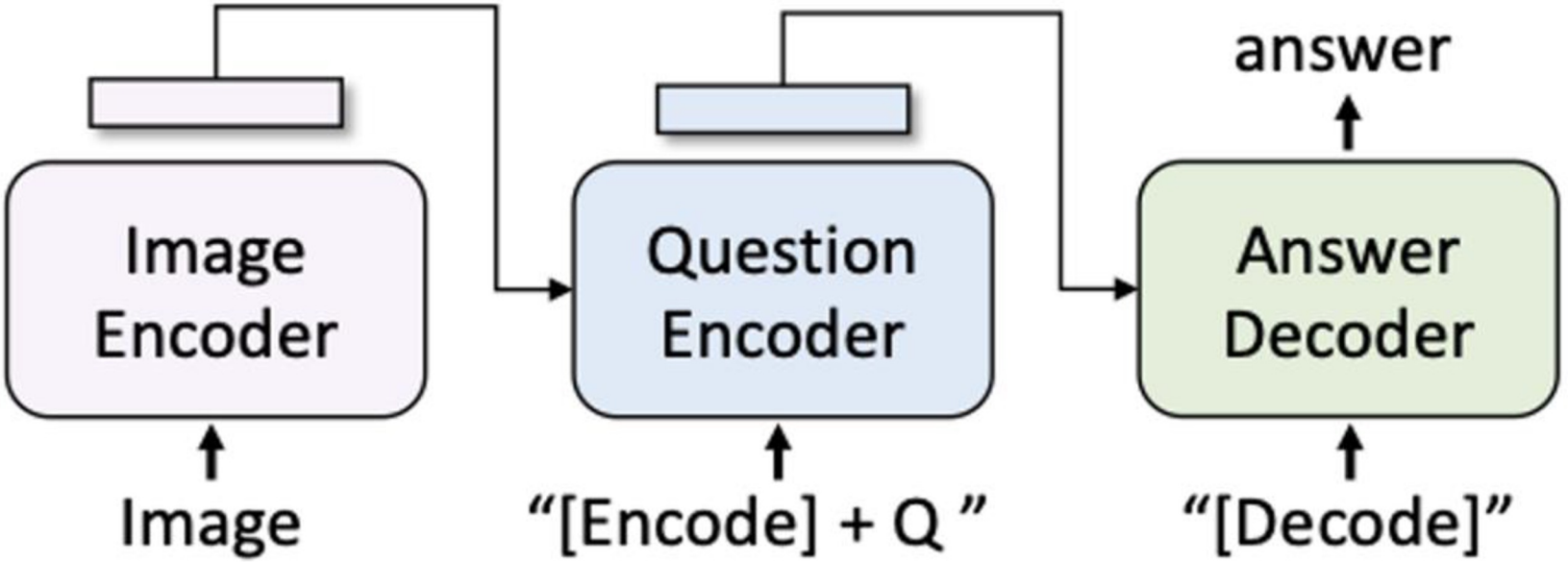
2. Medical vs general domain texts and images
   I. Hypothesized specialized pre-training / fine-tuning for improved domain-centric performance. (using BLIP) -> Faster training and improved performance
   II. Proposed unified vision-language pre-training architecture with novel fusion

3. Classification heads for Med-VQA
   I. Generation-based solution

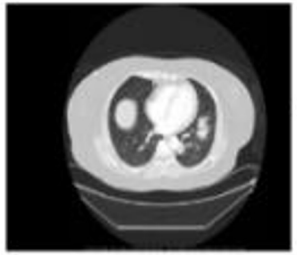
BLIP, BLIP, BLIP



VQA Using BLIP



Medical VQA Dataset



Q) what kind of image is this?
Ans) cta - ct angiography

a) Modality Category

Q) what plane is this ultrasound in?
Ans) longitudinal

b) Plane Category

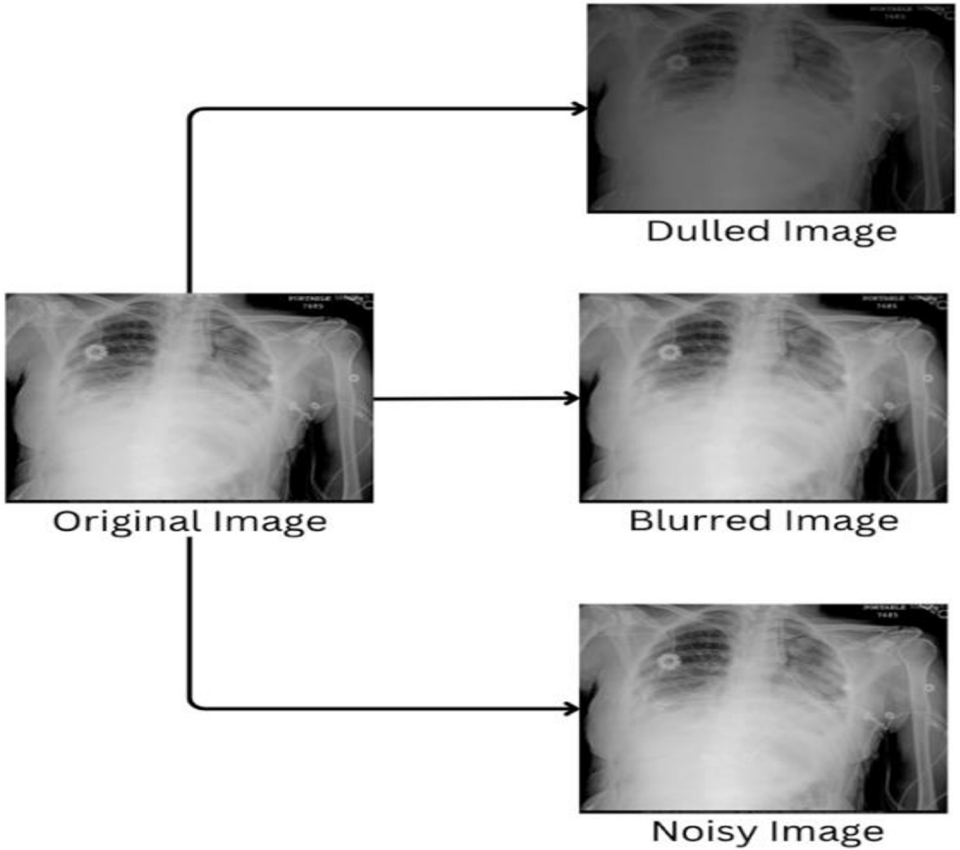
Q) what organ system is visualized?
Ans) heart and great vessels

c) Organ Category

Q) what abnormality is seen in the image?
Ans) ankylosing spondylitis

d) Abnormality Category

Augmentations



Experimentations

1. BLIP Pre-trained – Benchmarking

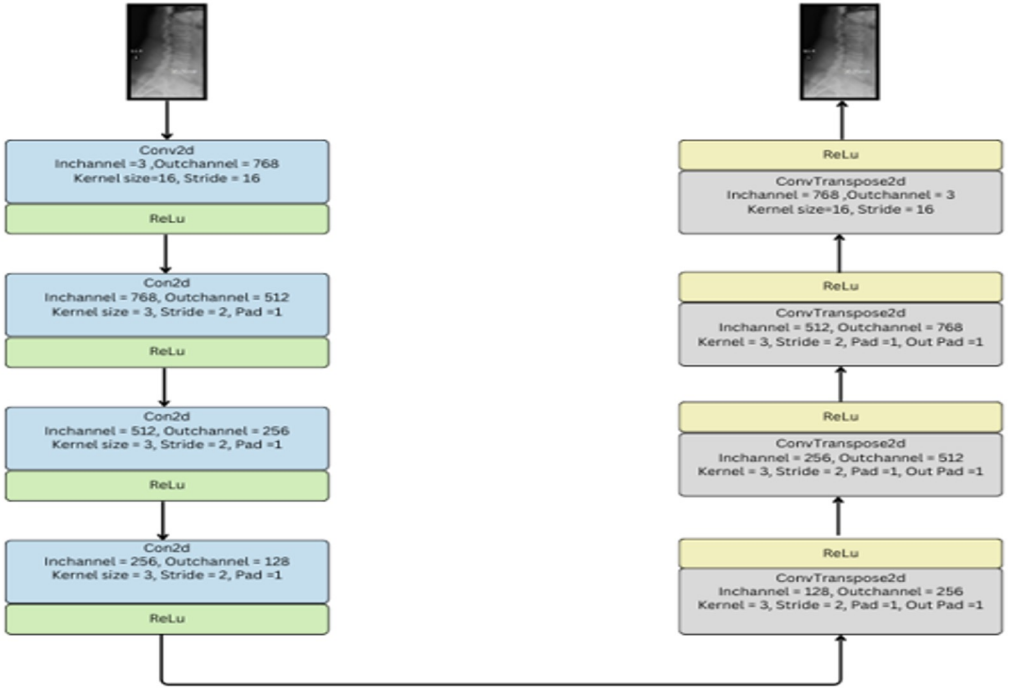
2. BLIP end-to-end Fine-tuning

   I. Time consuming + Dataset limitations

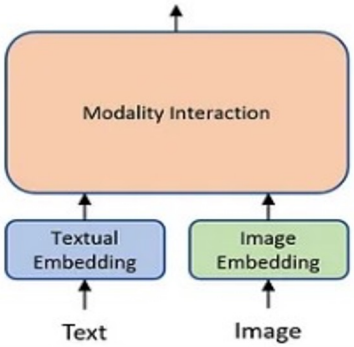
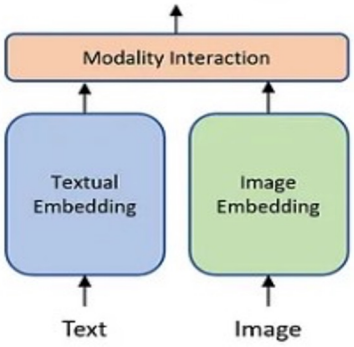
   II. Why not try specialized fine-tuning?

BLIP Vision Encoder – Incorporating Medical Domain Knowledge

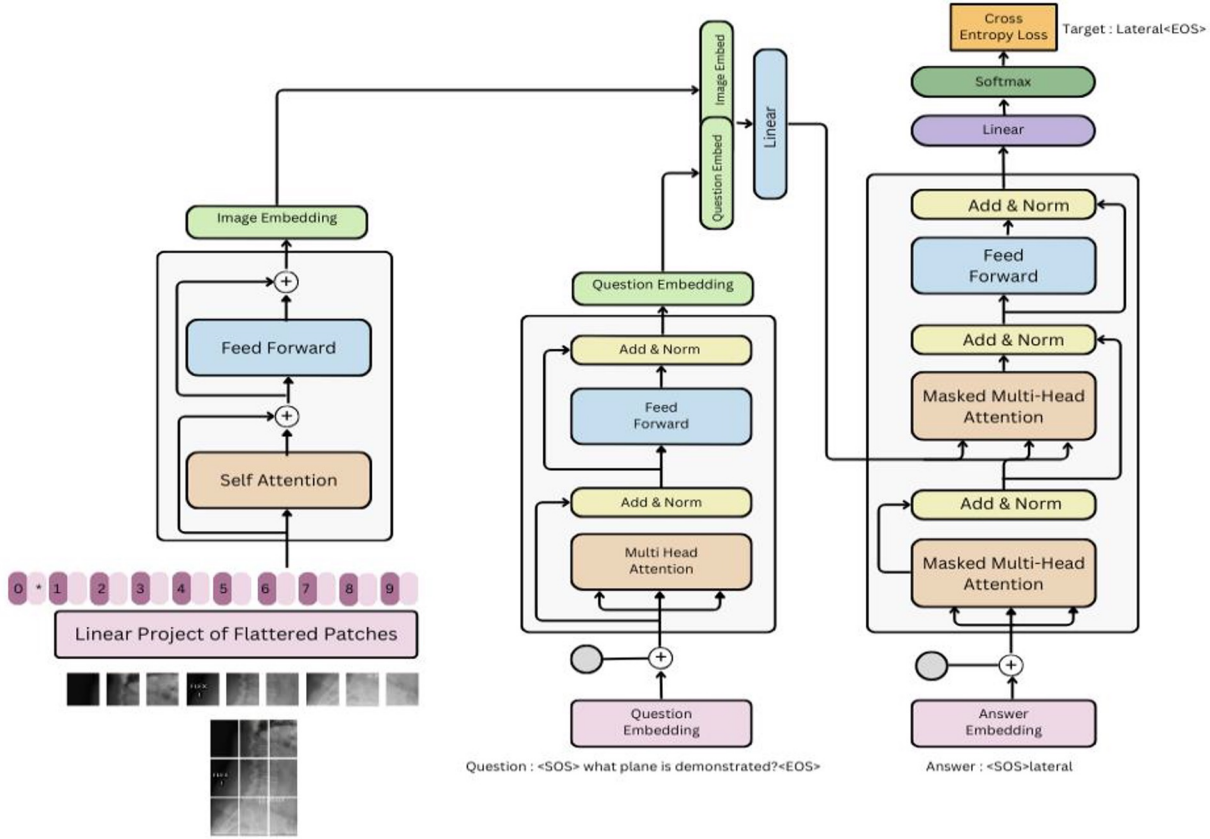
3. Dissecting Vision Encoder -
Convolution Patch Embedding layer was selected



Early Fusion vs Late Fusion

Model Type	Architecture	Single Modality Features	Modality Interaction	Training
Early Fusion		Very simple Example: GLoVe, CNNs	Complex Examples: Cross-Modal Transformers	Slow & difficult to train
Late Fusion		Heavy Example: Pretrained Transformers	Very simple Example: Concatenation + Single hidden layer	Simple to train

4. Proposed Architecture



Metrics

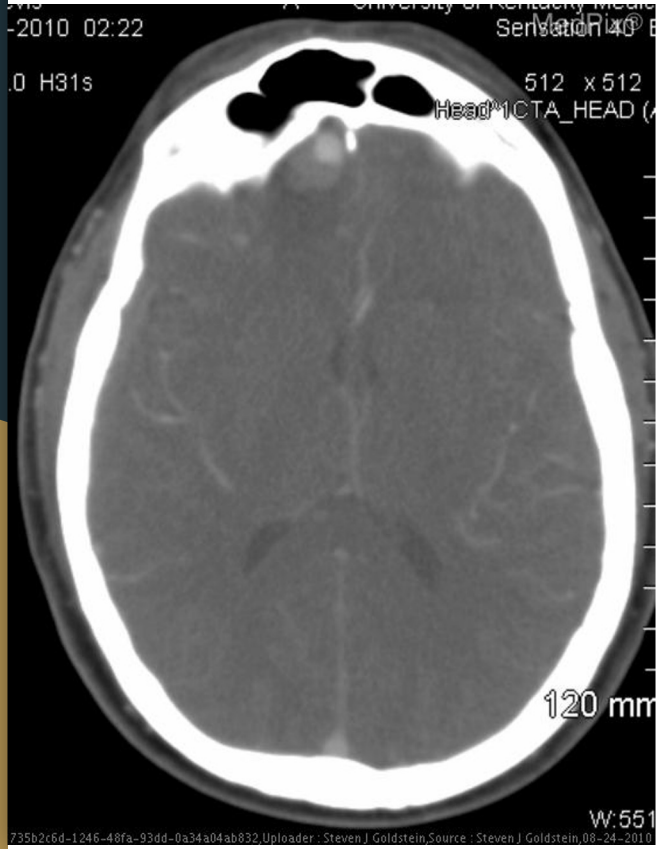
1. BLEU Score

2. ROUGE Score

Results - Quantitative

S no.	Model Version	BLEU score	ROUGE score
1.	Base BLIP	0.12	0.15
2.	Fine-tuned BLIP - end-to-end	0.37	0.40
3.	Convolution Patch Embedding Fine-tuned + Base BLIP	0.13	0.17
4.	Convolution Patch Embedding Fine-tuned + Fine-tuned BLIP model	0.38	0.42
5.	Proposed Architecture	0.41	0.44

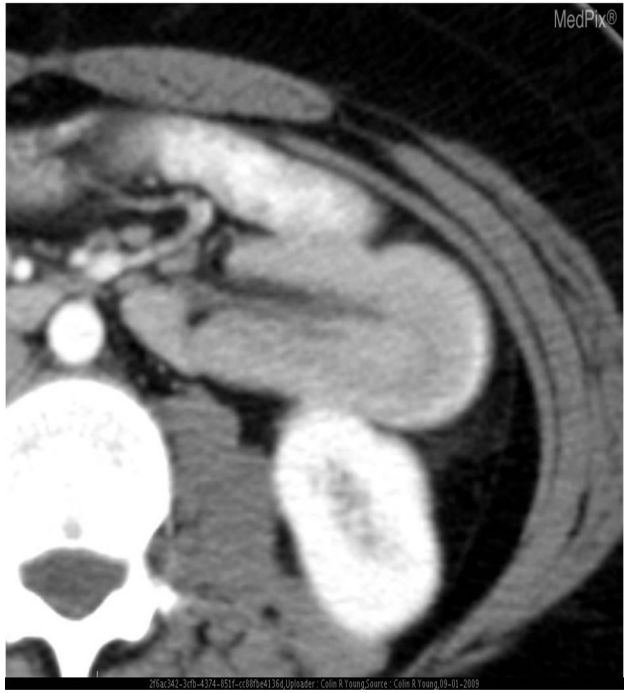
Results - Qualitative



Question - what modality is shown?
Ground Truth - cta - ct angiography
Answers -

1. Pre-trained BLIP: no
2. Fine-tuned BLIP: ct noncontrast
3. Convolution Fine-tuned BLIP: cta - ct angiography
4. Proposed Architecture: cta - ct angiography

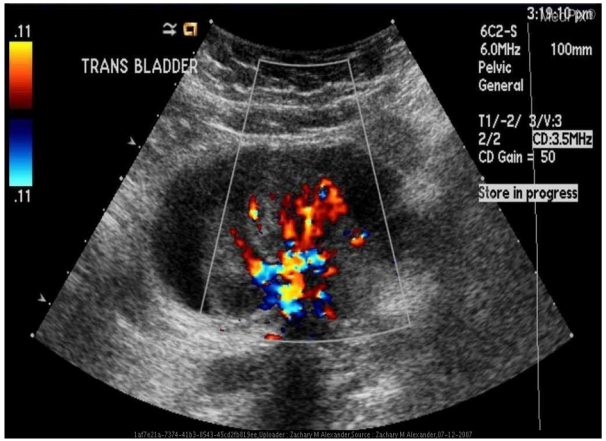
Results - Qualitative



Question: what type of contrast did this patient have?
Ground Truth: iv
Answers:

1. Pre-trained BLIP: no
2. Fine-tuned BLIP: iv
3. Convolution Fine-tuned BLIP: iv
4. Proposed Architecture: iv

Results - Qualitative



Question: what imaging method was used?
Ground Truth: us-d - doppler ultrasound
Answers:

1. Pre-trained BLIP: yes
2. Fine-tuned BLIP: us - ultrasound
3. Convolution Fine-tuned BLIP: us - ultrasound
4. Proposed Architecture: us - ultrasound

Results - Qualitative

Question: is this a noncontrast mri?
Ground Truth: yes
Answers:

1. Pre-trained BLIP: yes
2. Fine-tuned BLIP: no
3. Convolution Fine-tuned BLIP: yes
4. Proposed Architecture: yes

Conclusion

1. Explored medical VQA on radiology scans using enhanced BLIP architecture, showing promising improvements

2. Key limitations: lack of large medical datasets, need for data augmentation/custom tokenizers and further medical expert validation

3. VQA technology has potential to revolutionize radiology analysis but requires continued dataset growth, architecture innovations, clinical evaluations