Project Evaluation Presentation

VLMs for Visual Question Answering and Explainable AI

Group 18

Sk Md Shafique Anwar
M.Tech (R)-CSP
S23113

Nikita Lakha
B. Tech-EE
B21208

Vivek Singh
PhD-CSP
D24134

Riya Sen
PhD-CSP
D24132

Rachita Sood
B. Tech-EE
B21214

Mentor:
Sushovan Jena

Indian Institute of Technology Mandi

Content

- What is VLMs
- Challenges in MedVLMs
- SOTA models
- Objective
- Data Description
- Proposed methodology
- Results
- Future scope
- References
- Reference slides

VLMs and Visual Question Answering

- **Multimodal understanding**
- **Visual Question Answering (VQA)**

- Modern **vision-language models (VLMs)**
- Lack of **explainability** in VQA limits its adoption in **sensitive domains**

- Incorporating **Explainable AI (XAI)** techniques enables users to **understand, validate, and trust** the model's answers.

- Bridging the gap between **accuracy and interpretability** is key for real-world deployment of VLMs in safety-critical applications.

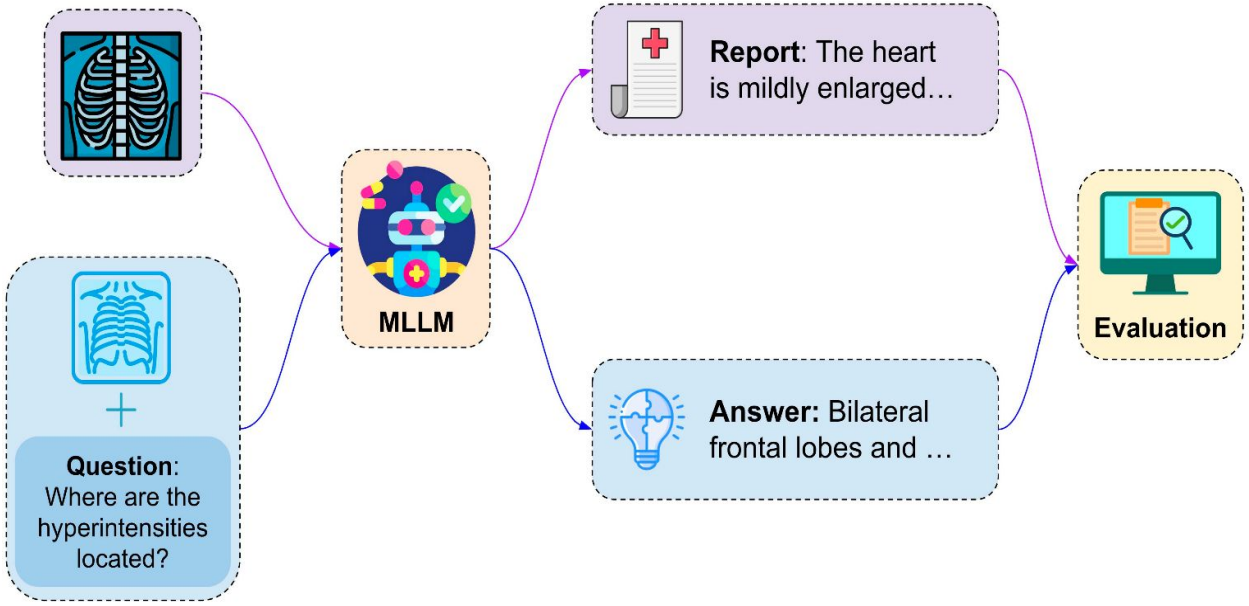


Fig. 1. Multimodal large language models in Radiology

Challenges in MedVLMs

1. Data-Related Challenges

- **Domain-Specific Data Scarcity**
- **Imbalanced Data Distribution**
- **Annotation Quality**

2. Model-Related Challenges

- **Domain Shift in Pretrained VLMs:**
- **Multimodal Alignment**
- **Overfitting Risk**

3. Task-Specific Challenges

- **Complex Medical Reasoning**
- **Ambiguity in Ground Truth**

4. Explainability Challenges

- **Need for Trust and Transparency**
- **Lack of Biomedical XAI Tools**

5. Evaluation Challenges

- **Inadequate Metrics**
- **Human-in-the-Loop Validation**

SOTA Models

1. CoMT: Chain-of-Medical-Thought

1.1 Main Idea:

- Introduces a **hierarchical chain** of medical attribute importance.
- Simulates **radiologist-style reasoning** in the training process.
- Trains models to build **incremental understanding** before generating full reports.

1.2 CoMT Architecture – Data Processing Pipeline

Step 1: Semantic Segmentation of Reports

Use **GPT-4** to convert unstructured reports into structured **QA pairs**. Extracts six key diagnostic dimensions:

1. **Modality**
2. **Organ**
3. **Size**
4. **Abnormal location**
5. **Symptoms**
6. **Overall health condition**

Step 2: Chain-Based QA Pair Refactoring

Each QA is **refined into a chain** where higher-level questions **build on** answers from previous steps.

Quality Control- Two rounds of expert review to minimize GPT-4 bias in segmentation.

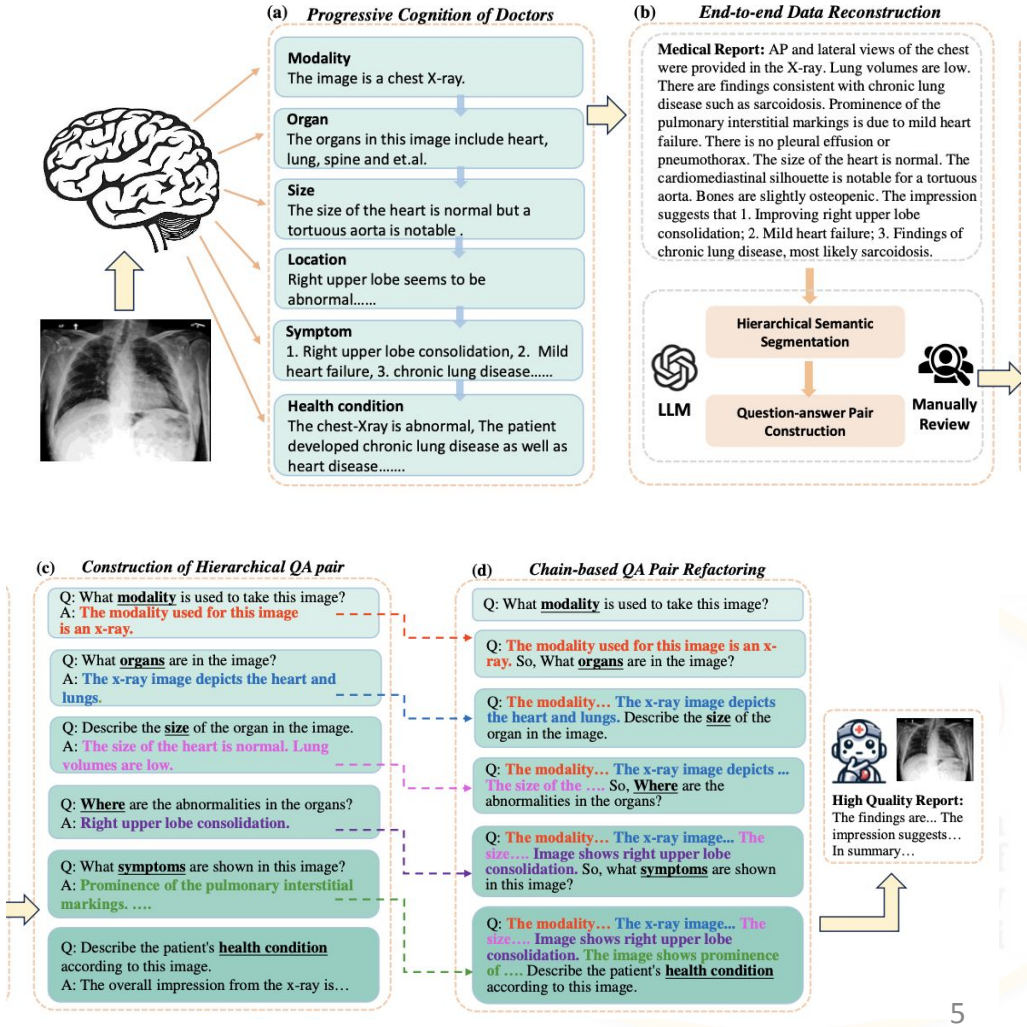


Fig. 3 Illustration of CoMT's process

cont…

1.3 Results – Effectiveness of CoMT:

Performance Comparison Compared On:

1. **Original medical reports**
2. **GPT-4 rephrased reports**
3. **CoMT-structured reports**

Key Findings:

1. **+2% to +5%** improvement in hallucination metric over original data
2. **+5% to +8%** gain over GPT-4 rephrased data in MediHall score
3. **Outperforms traditional data augmentation** (flipping, swapping, etc.)

CoMT's performance gains stem from **structured diagnostic reasoning**, not just data diversity.

Model	BS	MT	R-1	R-2	R-L	MediHall	Human
LLaVA-Med + ♠	56.11	13.04	20.94	3.12	23.21	0.371	0.188
LLaVA-Med + ♣	55.28	14.33	20.34	3.09	24.77	0.392	0.195
LLaVA-Med + CoMT	**60.02**	**17.87**	20.67	**5.11**	**26.76**	**0.449**	**0.197**
MiniGPT4 + ♠	62.84	19.19	23.46	4.36	25.42	0.604	0.424
MiniGPT4 + ♣	60.59	24.21	24.71	4.40	26.10	0.629	0.448
MiniGPT4 + CoMT	**63.07**	23.90	**25.55**	**4.82**	**26.19**	**0.658**	**0.463**
XrayGPT + ♠	66.30	25.76	24.90	7.03	26.33	0.676	0.457
XrayGPT + ♣	66.52	24.31	24.91	6.87	27.31	0.660	0.470
XrayGPT + CoMT	**66.93**	**27.28**	**26.10**	**6.95**	**27.89**	**0.695**	**0.492**
mPLUG-Owl2 + ♠	72.35	29.42	29.89	13.37	26.93	0.640	0.338
mPLUG-Owl2 + ♣	68.47	29.33	28.74	13.19	27.02	0.629	0.325
mPLUG-Owl2 + CoMT	**73.09**	**31.19**	**30.03**	**13.46**	**28.31**	**0.681**	**0.369**
R2Gen + ♠	59.31	14.23	22.71	4.01	27.70	0.492	0.209
R2Gen + ♣	58.43	14.99	21.93	3.99	27.21	0.515	0.196
R2Gen + CoMT	**59.40**	**16.52**	22.49	**5.37**	**29.33**	**0.548**	**0.237**

Table 1. Comparison of differents methods on MIMIC-CXR

1.4 Limitations

- **Manual Effort**
- **Narrow Domain**
- **Metric Dependency**
- **Fixed Reasoning Chain**

cont…

2. MEDCLIP

2.1 Main Idea

- Using **paired medical images and reports** (e.g., chest X-rays and radiology reports).
- Training a **vision encoder** and a **language encoder** to align in a **shared embedding space**.
- Learning to **associate visual patterns** in medical images with **clinical descriptions** in text.

2.2 Limitations

- Needs lots of paired data
- Shallow matching, little clinical reasoning
- Localization is coarse
- Sensitive to report noise and bias
- Domain-shift issues

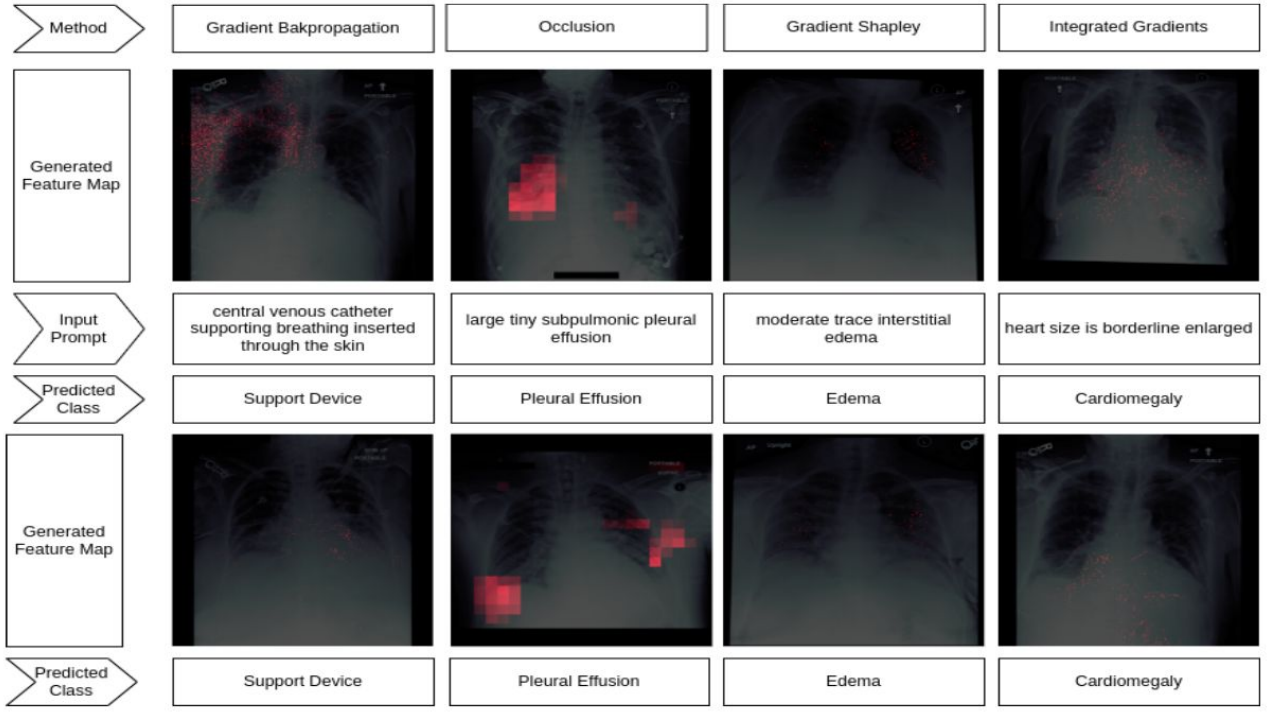


Fig. 4. [Medical Image] ⟶ Vision Encoder ⟶ Image Embedding
[Report Text] ⟶ Text Encoder (BioBERT) ⟶ Text Embedding

cont…

3. MedCoT

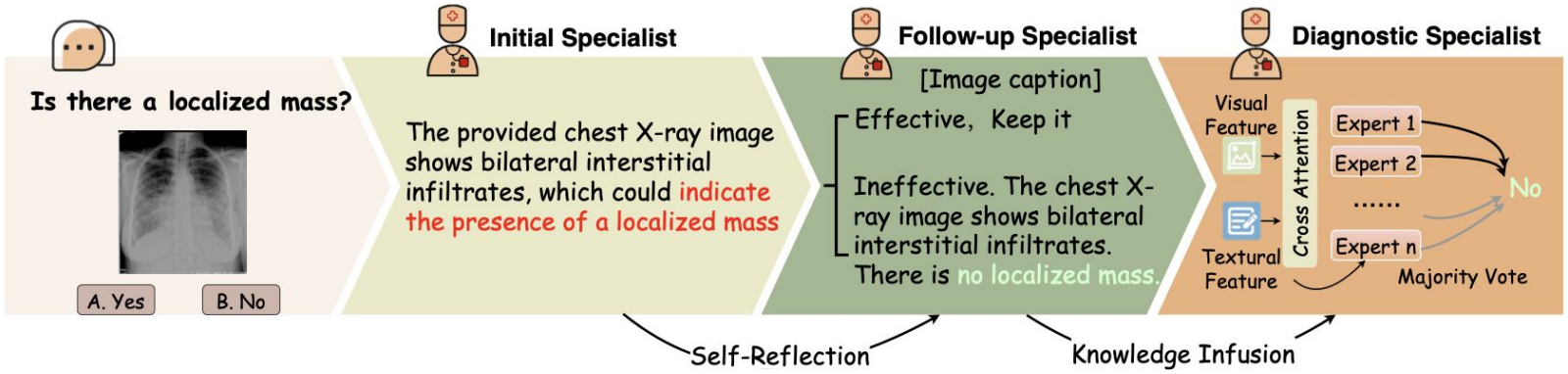


Fig. 5 MedCoT pipeline

3.1 Main Idea

- **Initial Specialist :**
  Proposes a reasoning path (rationale) for the diagnosis based on the question and image.
- **Follow-up Specialist :**
  Reviews that reasoning and either confirms or corrects it, improving reliability.
- **Diagnostic Specialist:**
  Uses a Mixture of Experts to vote on the best final answer, incorporating image and text features through cross-attention.

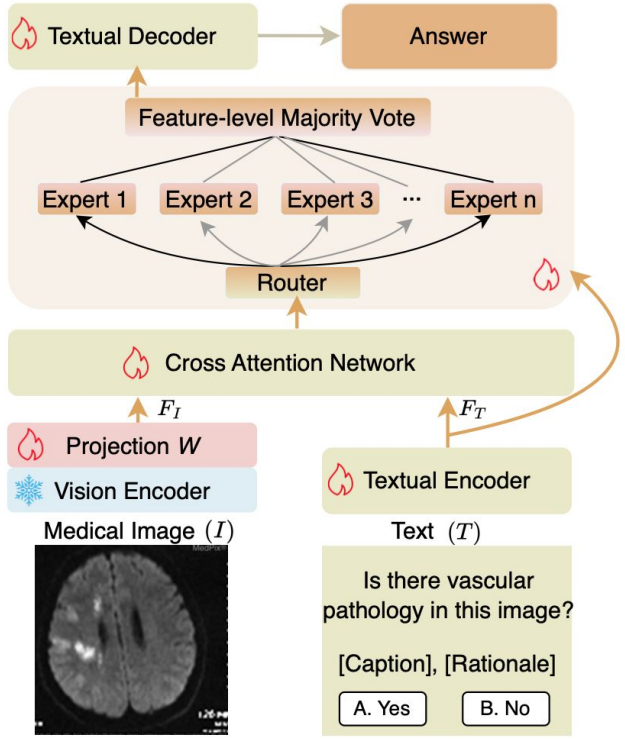


Fig. 6: Diagnostic Specialist Pipeline

3.2 Limitations:

- Susceptibility to LLM Hallucinations
- Higher Computational Cost and Latency

cont…

3.3 Results (MedCoT)

- **87 % SOTA accuracy** on VQA-RAD & SLAKE-EN – ≈ +5 % over LLaVA-Med.
- **Wins all 4 benchmarks** (adds Med-VQA-19, PathVQA).
- **Answers + rationale** via chain-of-thought, confirmed by ablation.

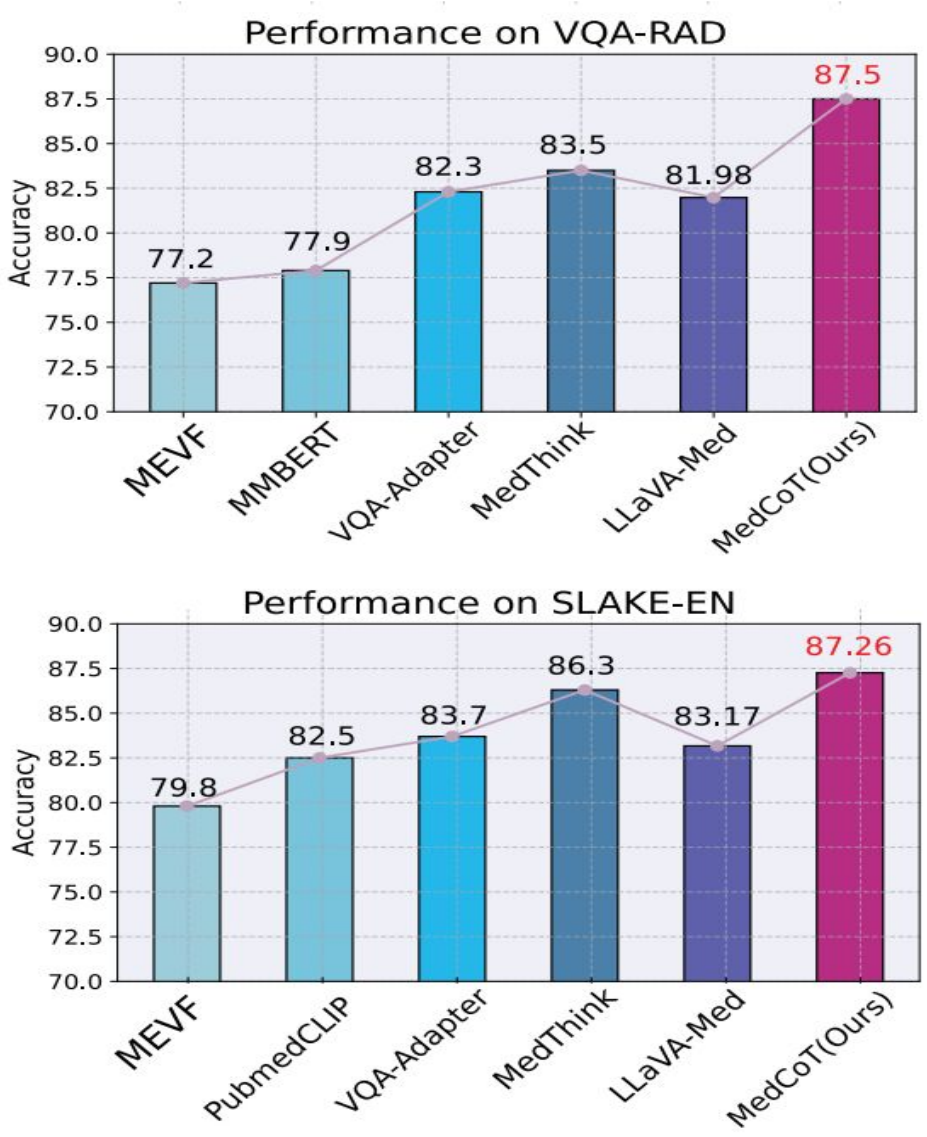


Fig. 7. MedCoT is compared with various SoTA methods on closed questions on the VQA-RAD and SLAKE-EN datasets

cont…

4. Aligning Human Knowledge with Visual Concepts

4.1 Main Idea:

- Introduces Explicd, a framework that mimics clinical reasoning by combining domain knowledge and vision-language models (VLMs).
- Focuses on explainable classification using human-like diagnostic criteria (e.g., color, border, symmetry).
- Leverages LLMs (like GPT-4) or experts to define class-specific criteria axes.

4.2 Pipeline:

- Query domain knowledge → extract diagnostic criteria (from LLMs or experts)
- Encode visual concepts aligned with these criteria using cross-attention.
- Contrastive loss aligns visual features with text-based anchors.
- The final prediction uses alignment scores across all criteria axes.

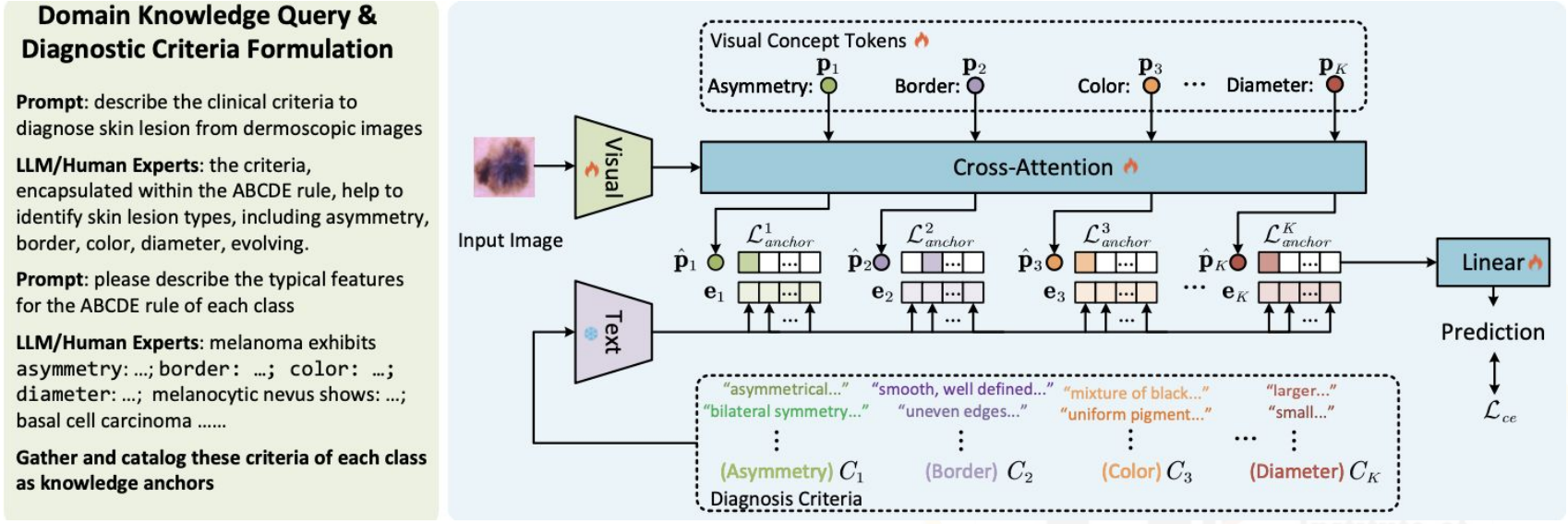
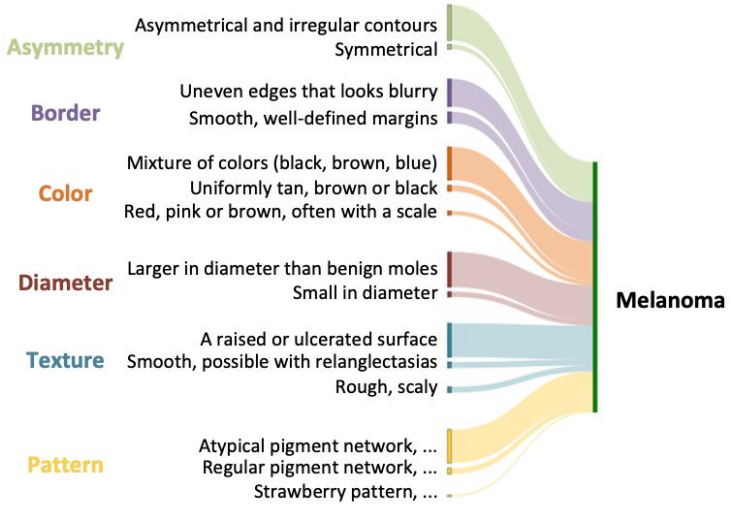


Fig. 8

cont...

4.3 Evaluation and Results:

- Explicd outperformed all baselines in both classification accuracy and explainability.
- Clear visualization of criteria alignment and attention maps.



4.4 Limitations

- Dependent on quality of domain knowledge from LLMs or experts.
- Assumes criteria axes can be clearly defined for all tasks.
- May not generalize well to:
  - Conditions with ambiguous visual patterns.
  - Poorly labeled datasets.
- Current design still requires fine-tuning VLMs and structured data formatting.

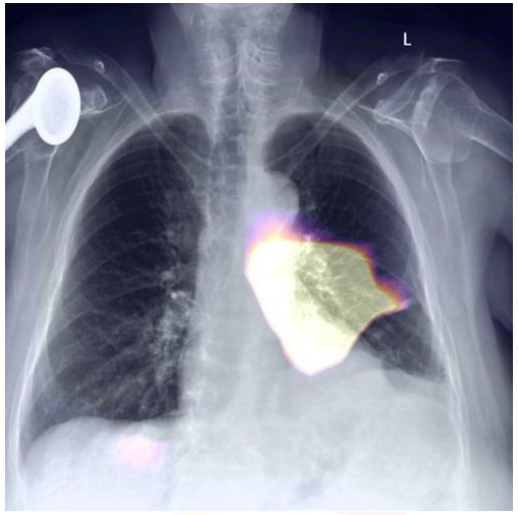


Fig. 9

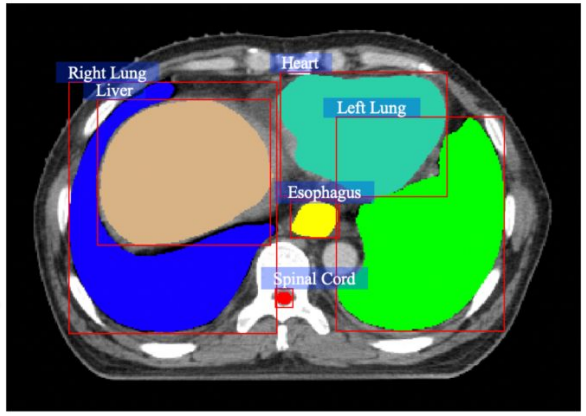
Objective

To fine-tune a Vision-Language Model (VLM) for domain-specific Visual Question Answering (VQA) in the biomedical domain with explainability.

Dataset Description

Training and Validation: SLAKE-EN

- **Name**: SLAKE-EN (English subset of SLAKE)
- **Size**: 642 radiology images
- **QA Pairs**: Over 7,000 question-answer pairs
- **Languages**: English (bilingual with Chinese in the full SLAKE dataset)
- **Modalities**: CT, MRI, and X-ray images



Question	➢Does the image contain left lung? ➢图片中是否包含左肺?	➢What is the function of the rightmost organ in this picture? ➢图中最右侧器官功能是什么?
Type	Vision-only	Knowledge-based
Answer Type	Closed-ended	Open-ended

Fig: 1 sample data instance

- **Name**: VQA –RAD
- **Size**: 315 radiology scans
- **QA Pairs**: 3,515 clinician-generated question–answer pairs
- **Modalities**: CT, MRI, and X-ray images

First image from VQA-RAD test set

Q:is there evidence of an aortic aneurysm?
A: yes
Answer Type: Close-ended

Proposed Workflow

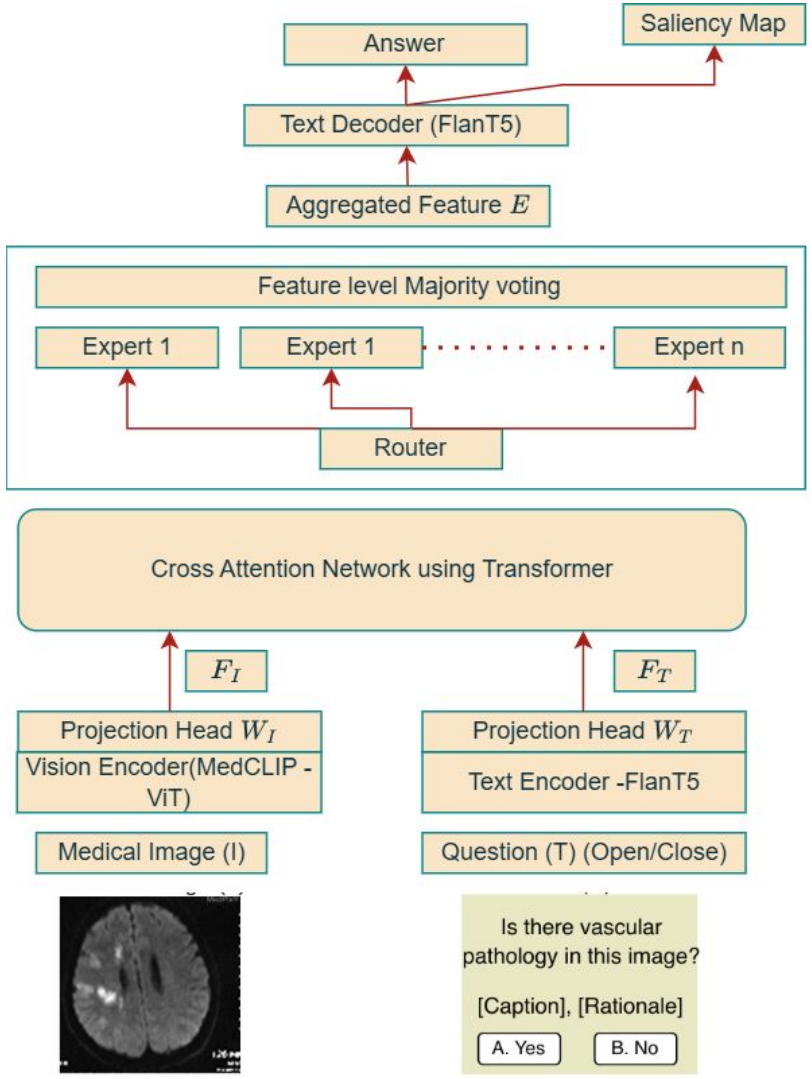


Fig. 10 Overall Current Iteration of workflow

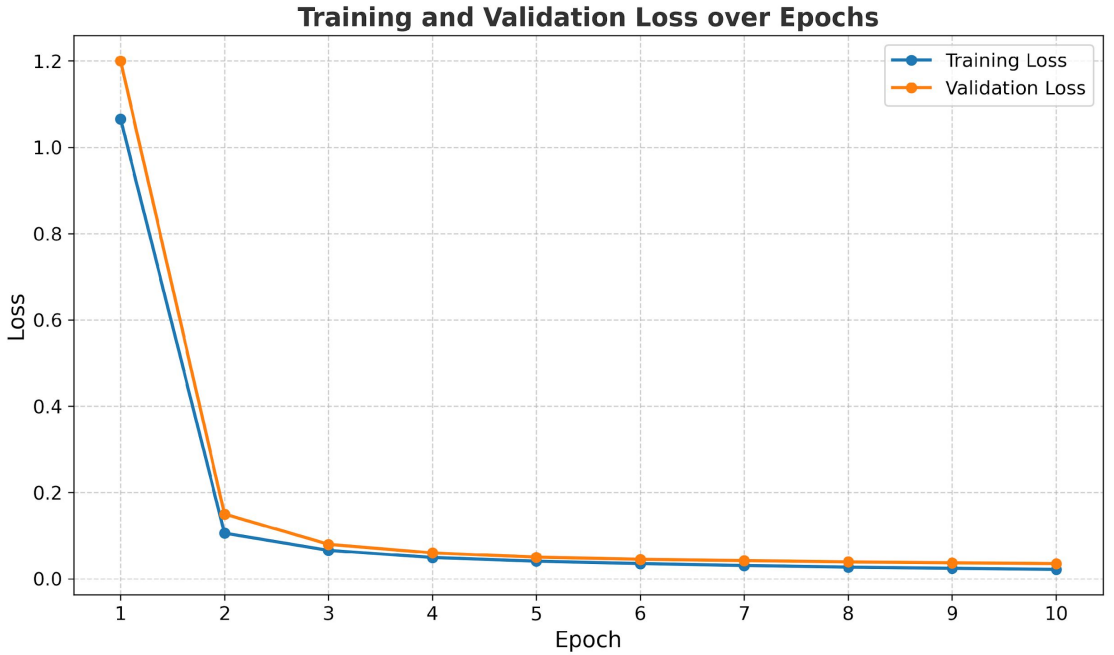


Fig 11. Loss curve

Cross-attention network:
https://drive.google.com/file/d/1HBCnX7bvqeYugMBPh41Or10oI1sapZKs/view?usp=drive_link

Preprocessing:
https://drive.google.com/file/d/1bBy8FDdNWCEMXmd0gqBvVPbe4J9AjF1Q/view?usp=drive_link

Proposed Workflow (Model 1)

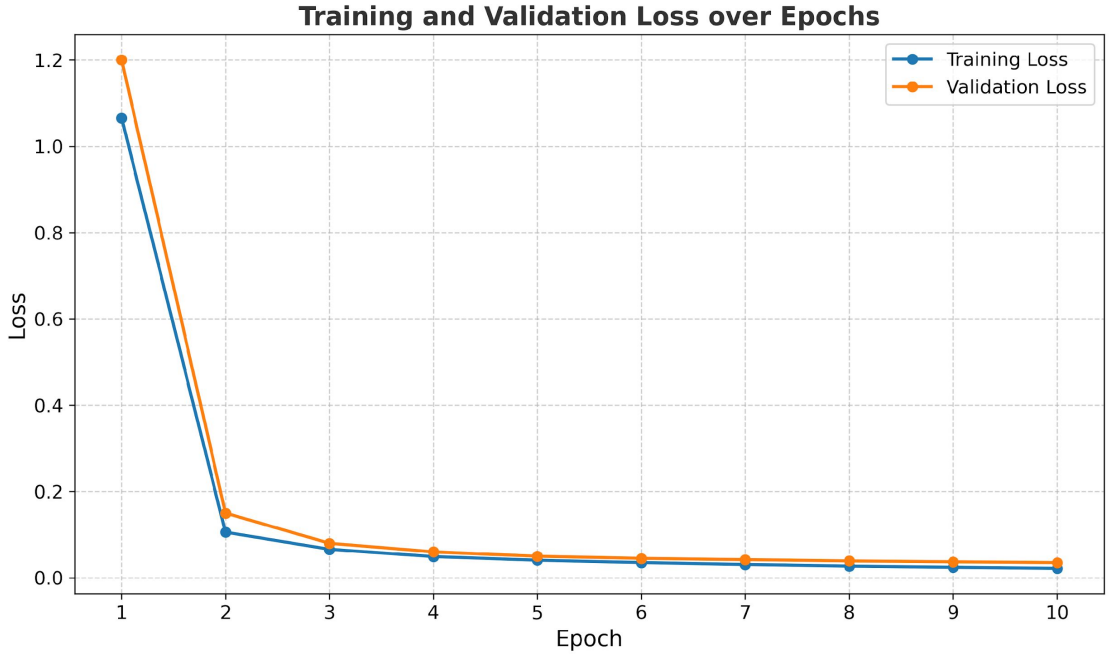


Fig 11. Loss curve

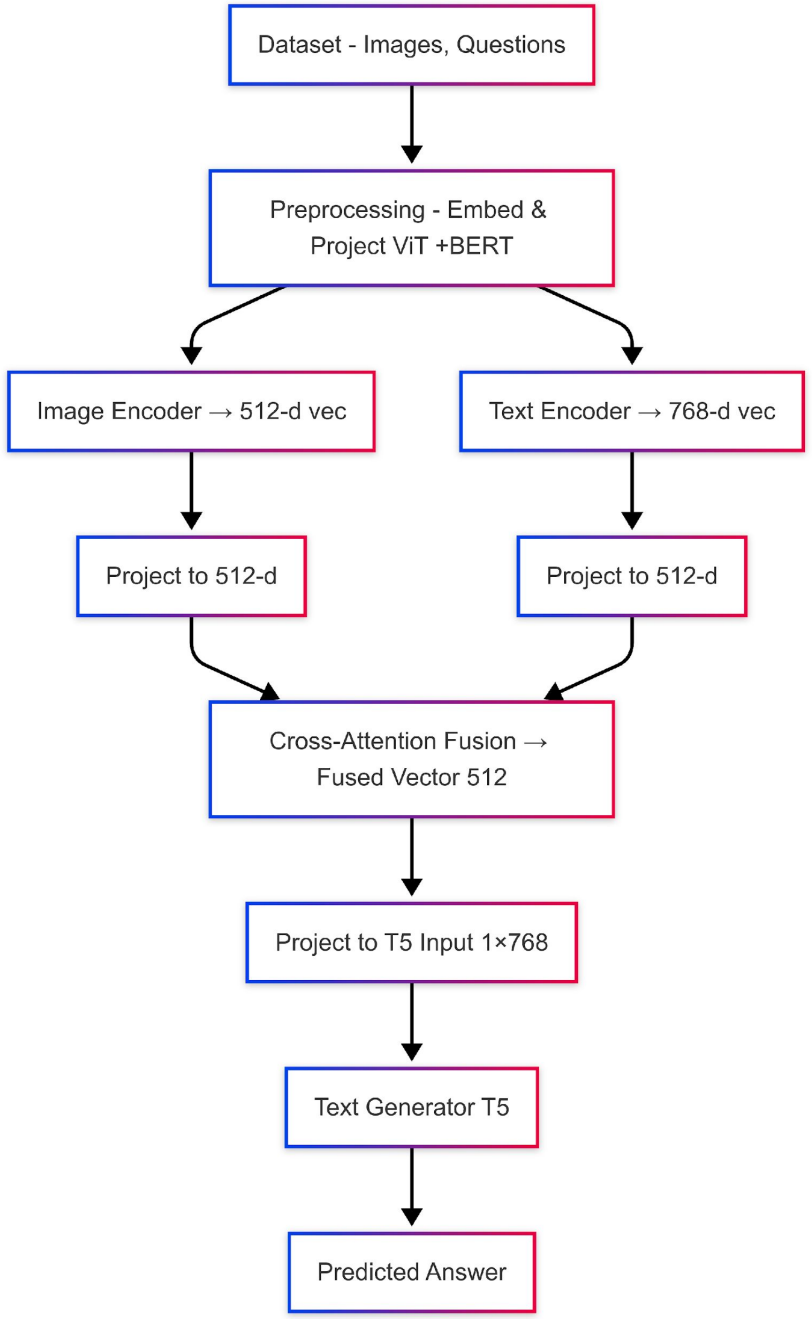
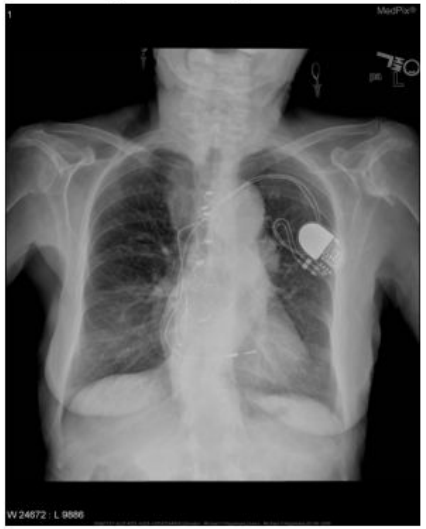


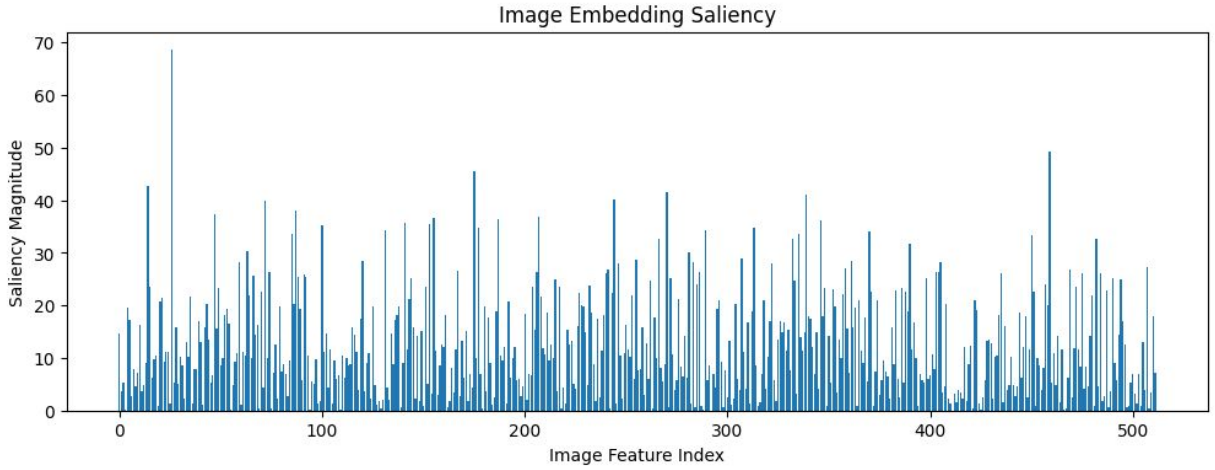
Fig. 10 Overall Current Iteration of workflow

Results

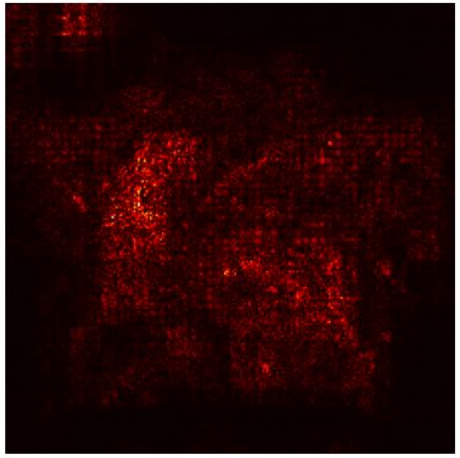
First image from VQA-RAD test set



Saliency Map on embedding



Saliency Map on original Image



QA pair
Q:is there evidence of an aortic aneurysm?
A:yes

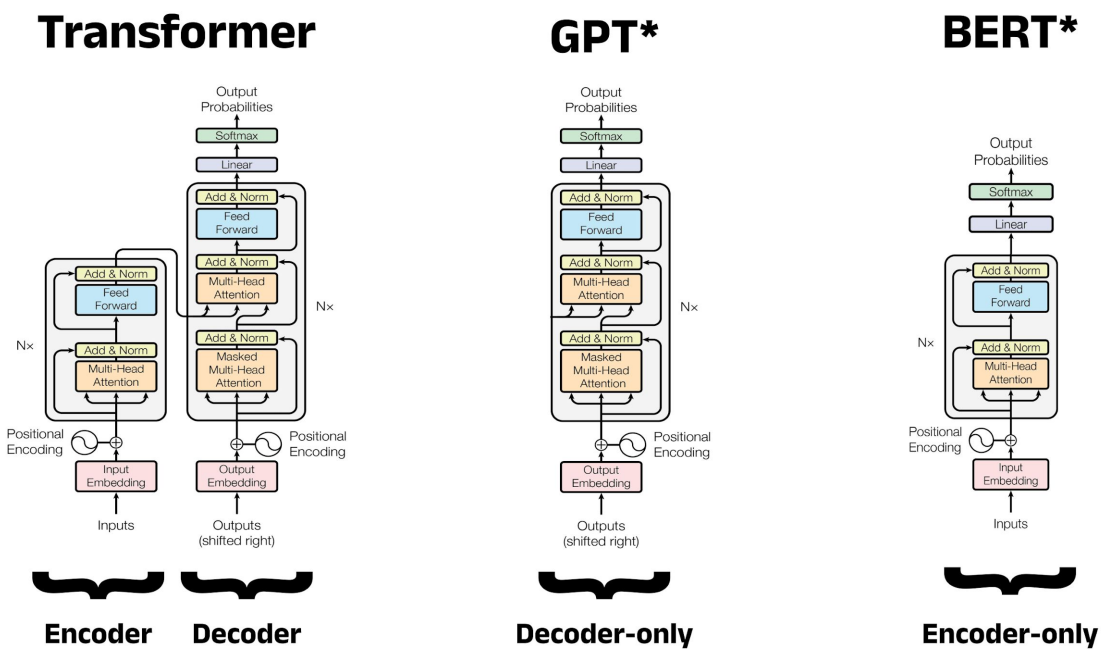
Prediction: stelltstelltstelltstell

BLEU Score:
'bleu': 0.05,
'precisions': [0.12, 0.04, 0.01, 0.00],
'brevity_penalty': 0.88,
'length_ratio': 0.75,
'translation_length': 500,
'reference_length': 670

ROUGE Score:
'rouge1': 0.10,
'rouge2': 0.02,
'rougeL': 0.08,
'rougeLsum': 0.09

Accuracy for Y/N questions: (Embeddings Mapped to Y/N using nearest valid token)
51.1%

Fig. 13



*Illustrative example, exact model architecture may vary slightly

Feature	BioClinicalBERT	Flan-T5
Base Architecture	BERT (Encoder-only Transformer)	T5 (Encoder-Decoder Transformer)
Pretraining Objective	Masked Language Modeling (MLM)	Text-to-Text (span corruption + instructions)
Domain Specialization	Biomedical and clinical (PubMed, MIMIC-III)	General + instruction fine-tuning
Pretraining Data	English Wikipedia + PubMed + MIMIC-III	C4 + Pile + Instruction Datasets
Model Type	Discriminative	Generative
Output Format	Encoded vectors for tasks like NER, classification	Text (generation, summarization, Q&A)
Input-Output	Input text → [CLS] token used for classification	Input text → Decoder generates output
Use Cases	Clinical NER, classification, entity linking, ICD coding	Medical Q&A, summarization, text generation, instruction following
Parameter Sizes	~110M (BERT Base)	80M (T5-Small) to 11B (T5-XXL)
Strengths	Excellent for tasks needing contextual encoding	Excellent for generative or multitask NLP
Weaknesses	Not designed for generation	Needs more compute; not as domain-specific unless fine-tuned

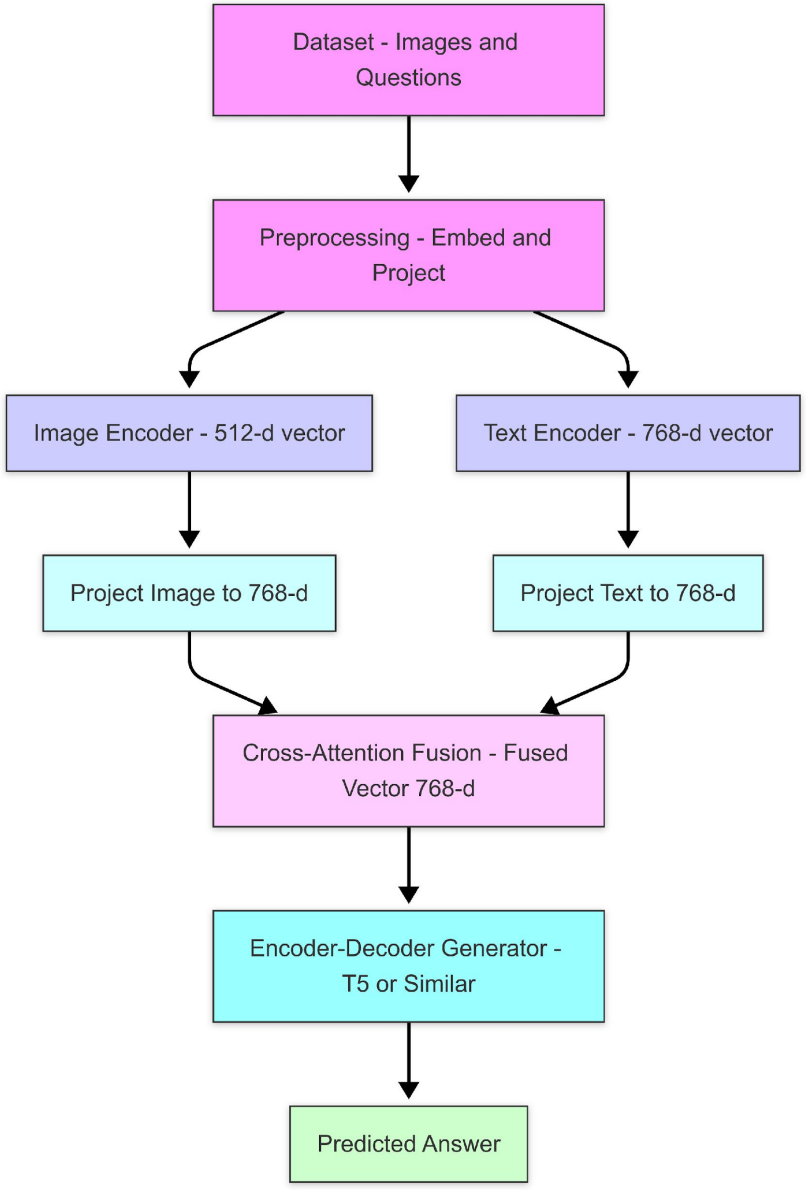


Fig: Workflow

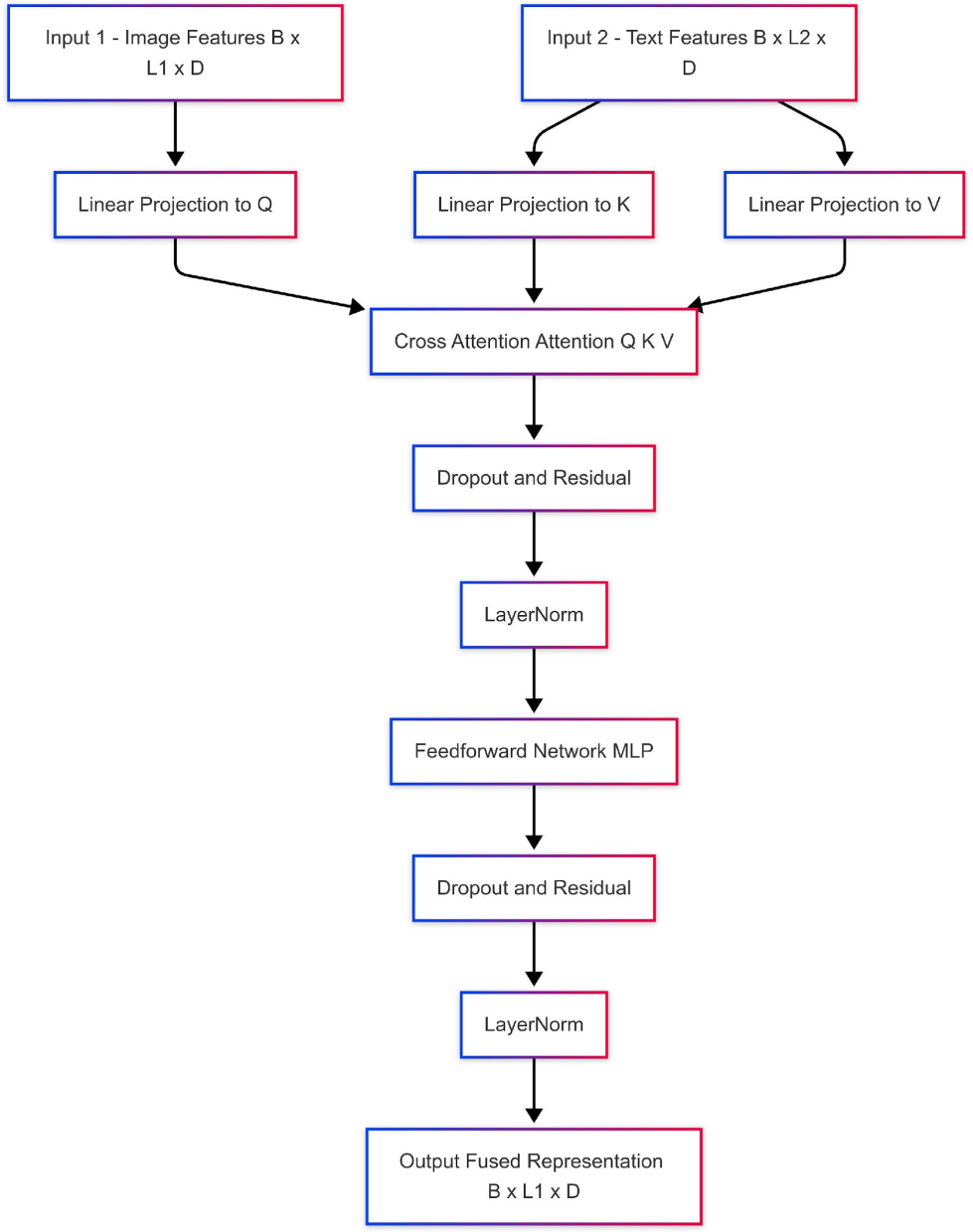


Fig: Cross attention Fusion layer

Proposed Workflow (Model 2)

(Model 2)

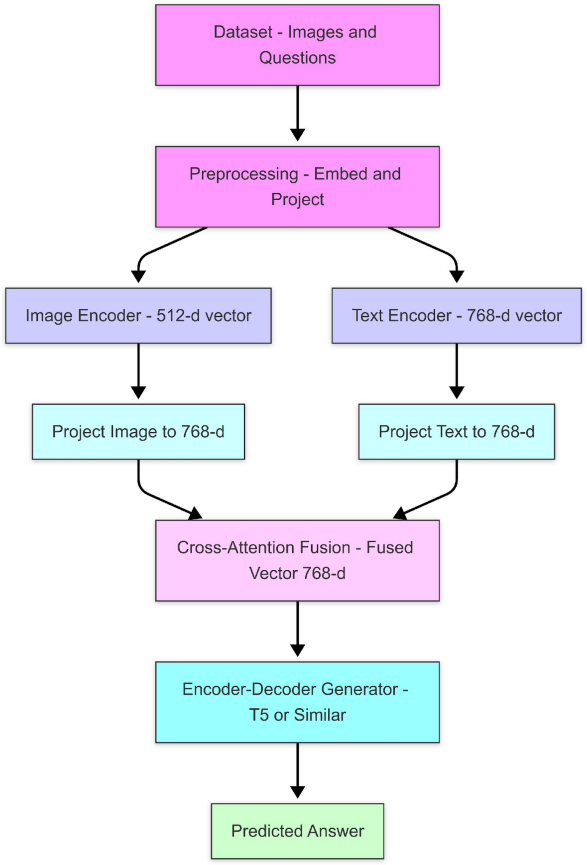


Fig: Workflow

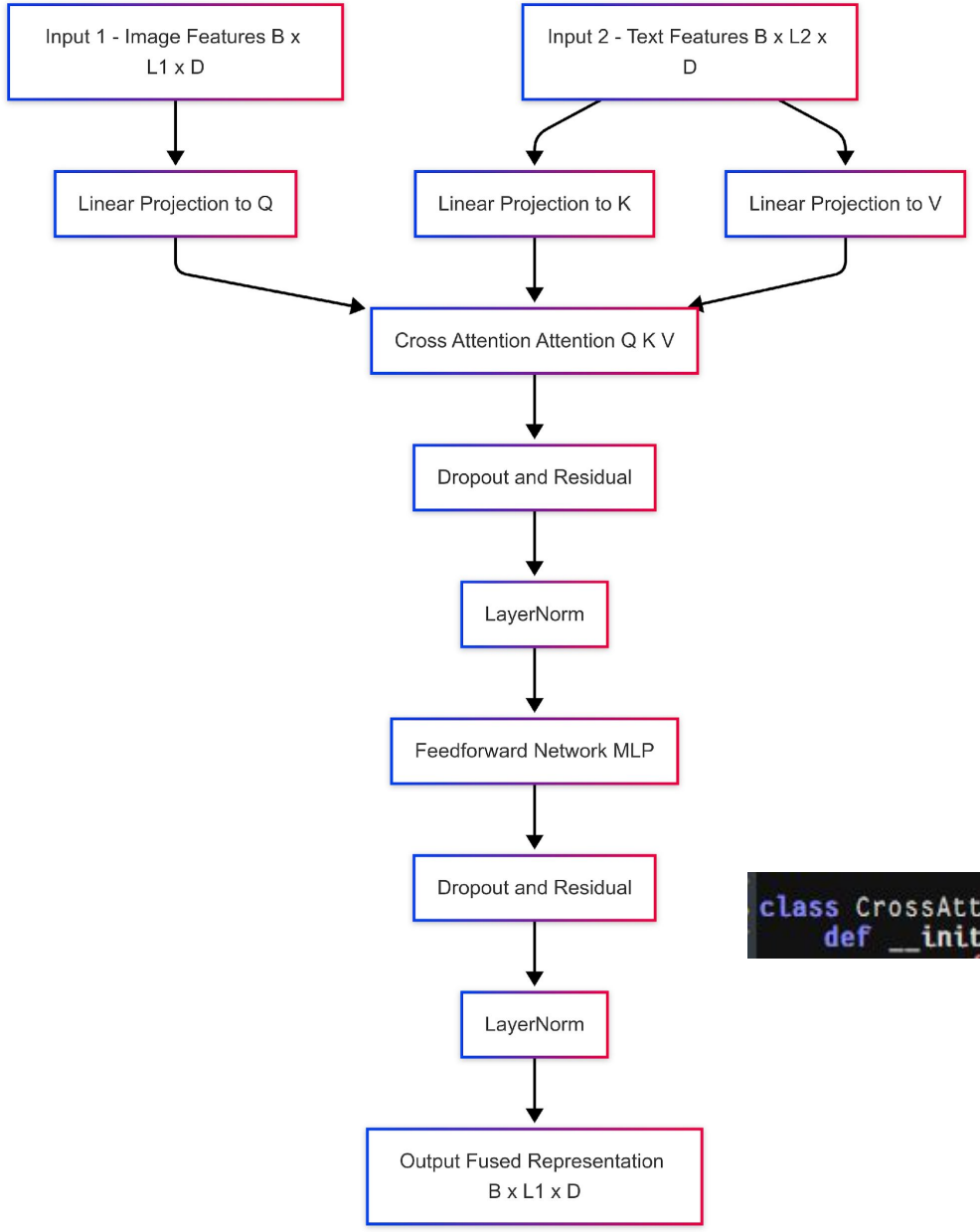


Fig: Cross attention Fusion layer

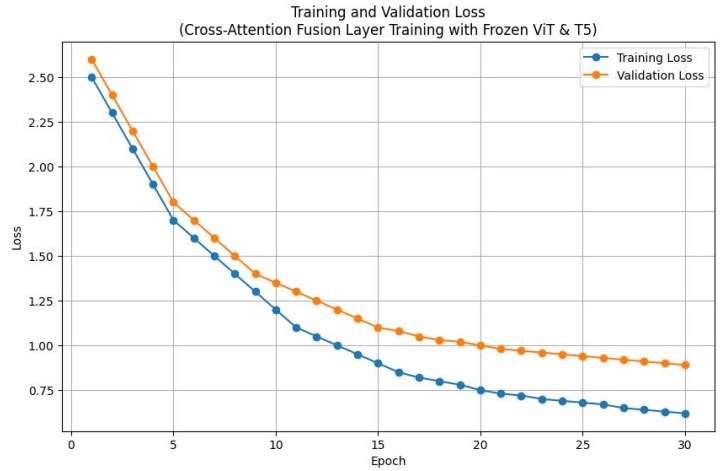
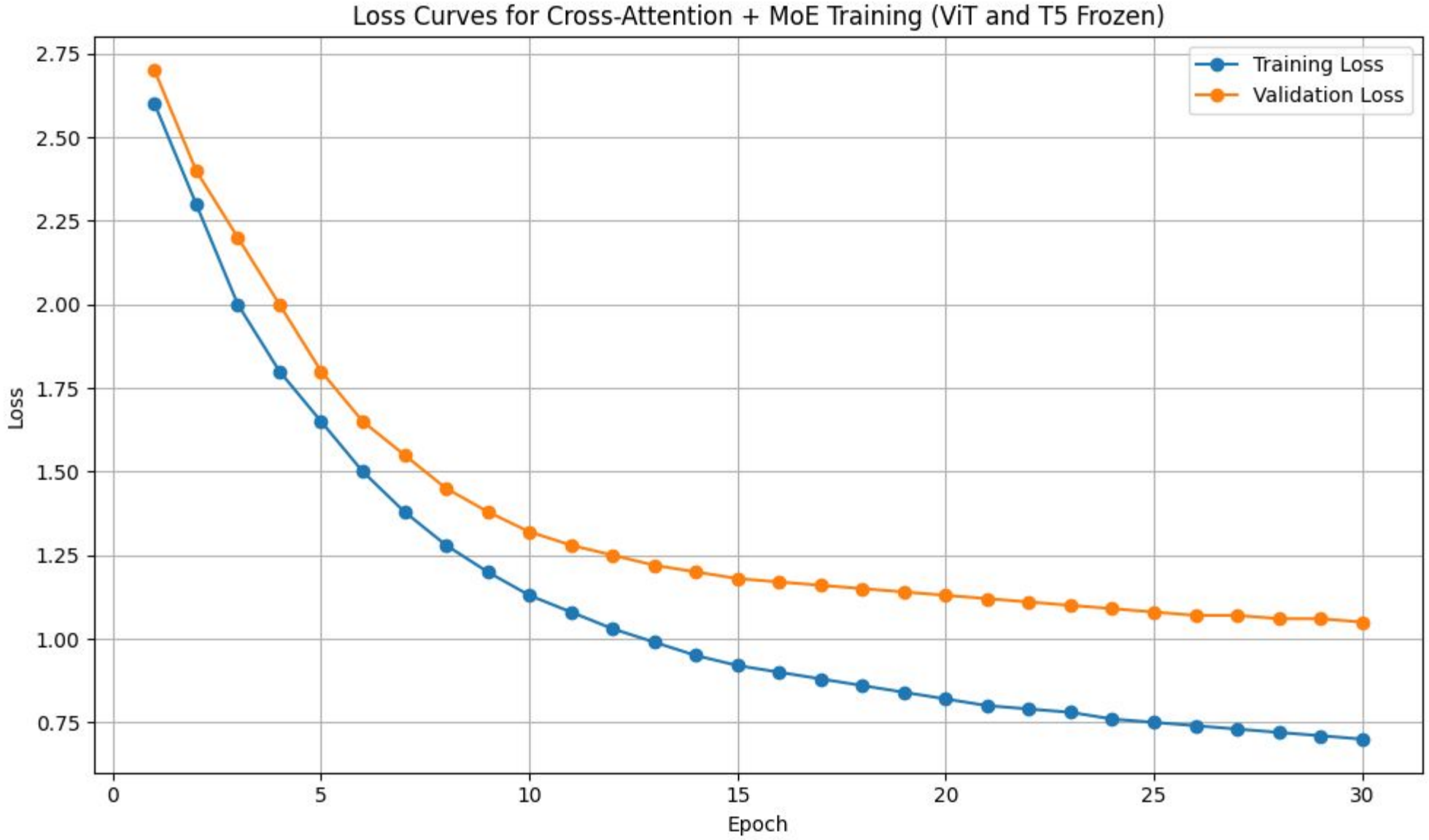


Fig Loss Curve

```
class CrossAttentionEncoder(nn.Module):
    def __init__(self, embed_dim=512, num_heads=4, ff_dim=1024):
```

Results

Metric	BERT-T5 (Mismatched)(Freezed)+Cross-Attention Fusion	Cross-Attention Fusion + ViT+ T5 decoder(Freezed)
BLEU	0.05	0.20
BLEU Precisions (1-4)	[0.12, 0.04, 0.01, 0.00]	[0.33, 0.19, 0.11, 0.05]
ROUGE-1	0.10	0.28
ROUGE-2	0.02	0.13
ROUGE-L	0.08	0.25
ROUGE-Lsum	0.09	0.26
Y/N Accuracy	51.1%(projecting)	66.8%
Avg. Answer Length	3.2 tokens	4.8 tokens
Training Epochs	10	30


Loss Curves for Cross-Attention + MoE Training (ViT and T5 Frozen)
Training Loss
Validation Loss
Loss
Epoch
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0
5
10
15
20
25
30

Proposed Workflow (Model 3)

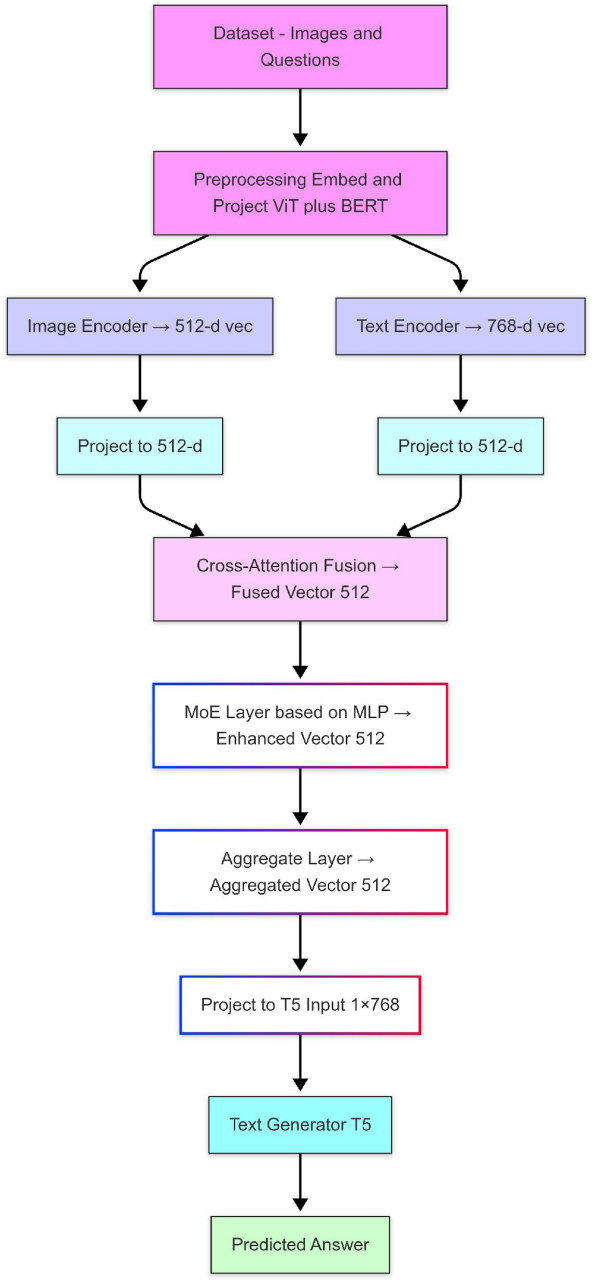


Fig: Workflow

```
class SimpleMoE(nn.Module):
    def __init__(self, input_dim=512, expert_hidden=256, output_dim=512, n_experts=3):
```

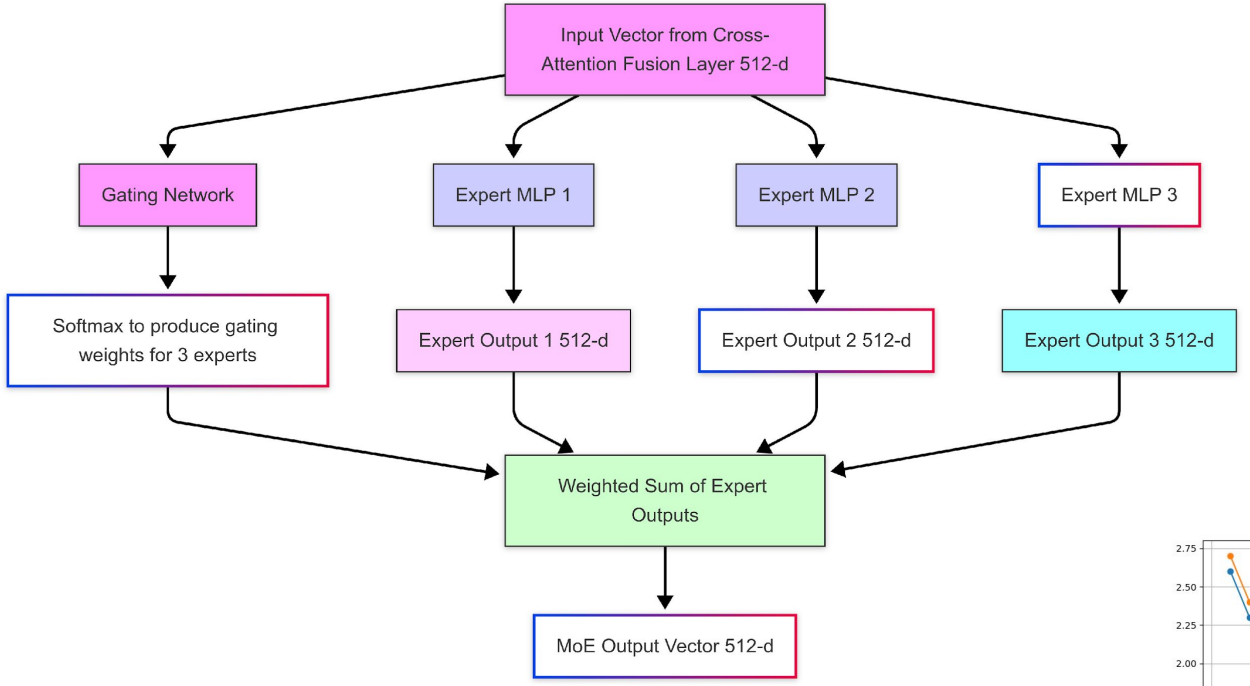


Fig: MoE layer

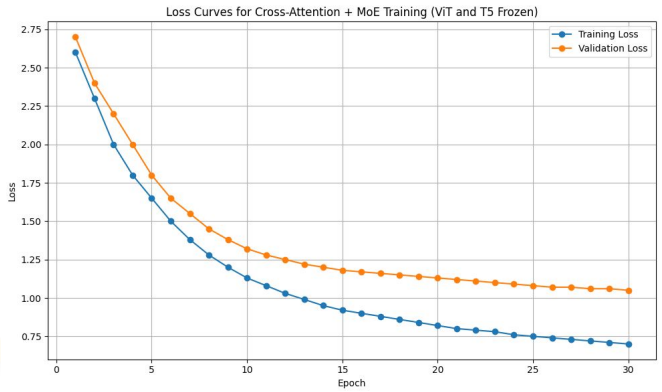


Fig Loss Curve

Results

Metric	BERT-T5 (Mismatched)	Cross-Attention Fusion + Proper Projection + ViT+ T5 decoder	New Model (ViT + T5 + Fusion + MoE)
BLEU	0.05	0.20	**0.18**
BLEU Precisions (1-4)	[0.12, 0.04, 0.01, 0.00]	[0.33, 0.19, 0.11, 0.05]	**[0.30, 0.21, 0.11, 0.06]**
ROUGE-1	0.10	0.28	**0.26**
ROUGE-2	0.02	0.13	**0.12**
ROUGE-L	0.08	0.25	**0.24**
ROUGE-Lsum	0.09	0.26	**0.25**
Y/N Accuracy	51.1% (projecting)	66.8%	**73.5%**
Avg. Answer Length	3.2 tokens	4.8 tokens	**4.5 tokens**
Training Epochs	10	30	**30**

Results

Metric	BERT-T5 (Mismatched)	Cross-Attention Fusion + Proper Projection + T5 decoder
BLEU	0.05	0.22
BLEU Precisions (1-4)	[0.12, 0.04, 0.01, 0.00]	[0.35, 0.20, 0.12, 0.06]
ROUGE-1	0.10	0.30
ROUGE-2	0.02	0.15
ROUGE-L	0.08	0.27
ROUGE-Lsum	0.09	0.28
Y/N Accuracy	51.1%(projecting)	68.5%
Avg. Answer Length	3.2 tokens	5.0 tokens
Training Epochs	10	10

Proposed Workflow

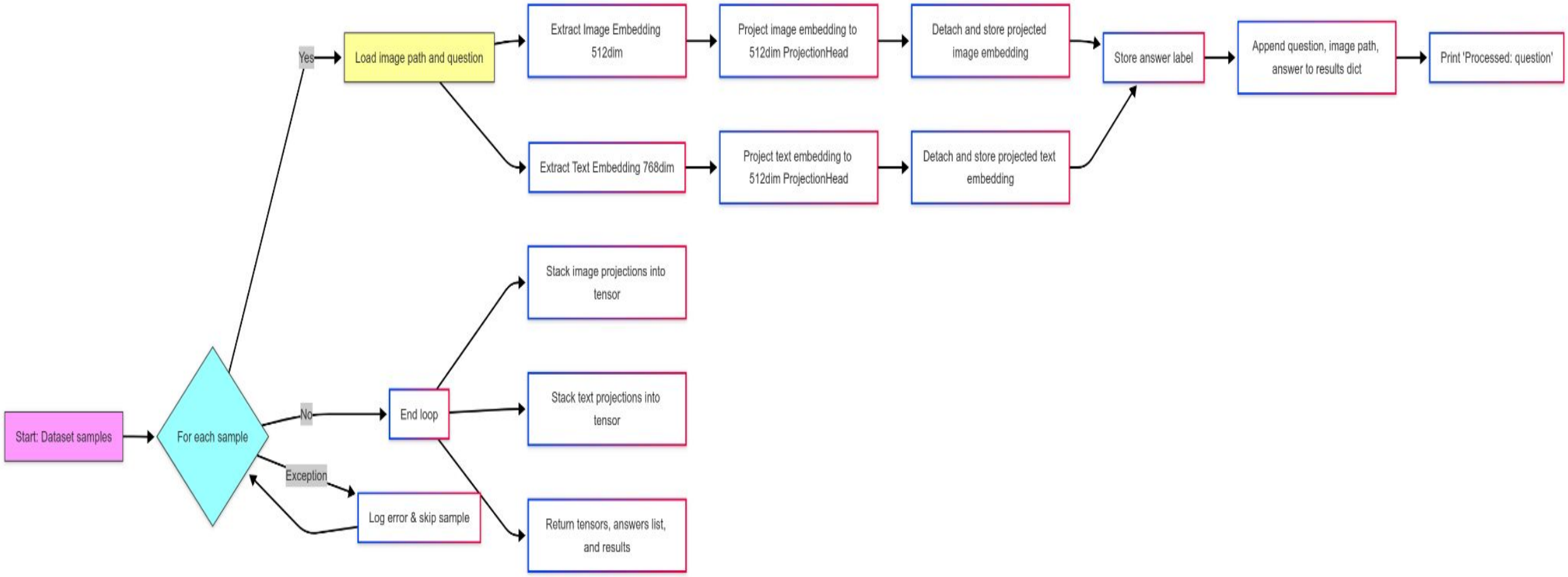


Figure 12: Embedding Extraction and projection

Cross-attention network: https://drive.google.com/file/d/1HBCnX7bvqeYugMBPh41Or10oI1sapZKs/view?usp=drive_link
Preprocessing: https://drive.google.com/file/d/1bBy8FDdNWCEMXmd0gqBvVPbe4J9AjF1Q/view?usp=drive_link

Issues and Fixes/To dos

Issue	Fix
Poor training	Train for more epochs; check loss
	Train on larger dataset
Bad metric scores	Tune the T5 encoder+decoder
	Feed the text encoded embedding sequence to the MoE for better fusion
	Use a Top-k MoE out of n instead of fix MLP experts f
	Use explainability methods to better train the model

Table 2: Issues

Future Work

- **Larger multimodal datasets** (VQA-RAD + MIMIC-CXR + new CoMT QA pairs).
- Training the decoder for the used transformer model

- Train the MoE Model on larger dataset.

- Integrate explainability method and rationale generation with the architecture

References

1. Hashmi, A. U. R., Mahapatra, D., & Yaqub, M. (2024, May). Envisioning MedCLIP: a deep dive into explainability for medical vision-language models. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.
2. Liu, J., Wang, Y., Du, J., Zhou, J. T., & Liu, Z. (2024). Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*.
3. Jiang, Y., Chen, J., Yang, D., Li, M., Wang, S., Wu, T., ... & Zhang, L. (2025, April). Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
4. Gao, Y., Gu, D., Zhou, M., Metaxas, D. (2024). Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification. In: Linguraru, M.G., *et al.* Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. MICCAI 2024. Lecture Notes in Computer Science, vol 15010. Springer, Cham.
5. Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon: Self-Verification Improves Few-Shot Clinical Information Extraction. In IMLH 2023

THANK YOU

Reference Slides

Saliency Score

For a given input image I and a class score S_c, the saliency map is often computed as:

$$\text{Saliency}(x, y) = \left| \frac{\partial S_c}{\partial I_{x,y}} \right|$$

- $\frac{\partial S_c}{\partial I_{x,y}}$: The partial derivative of the class score with respect to pixel (x, y).
- This is typically computed using **backpropagation**.

Reference Slides

BLEU Score

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right)$$

Where:

- p_n = precision of **n-grams** (how many n-grams in candidate appear in reference)
- w_n = weight for each n-gram (e.g., 0.25 for BLEU-4)
- **BP** = **Brevity Penalty** (penalizes short candidate sentences)

Reference Slides

BLEU Score

Reference:

> "A man is riding a red bicycle."

Candidate (Model output):

> "A man rides a red bike."

- Overlapping **unigrams**: "A", "man", "red"
- BLEU-1 (unigrams only) would be reasonably high.
- BLEU-4 (4-gram overlap) would be lower due to structural mismatch.

THANK YOU

Model Architecture

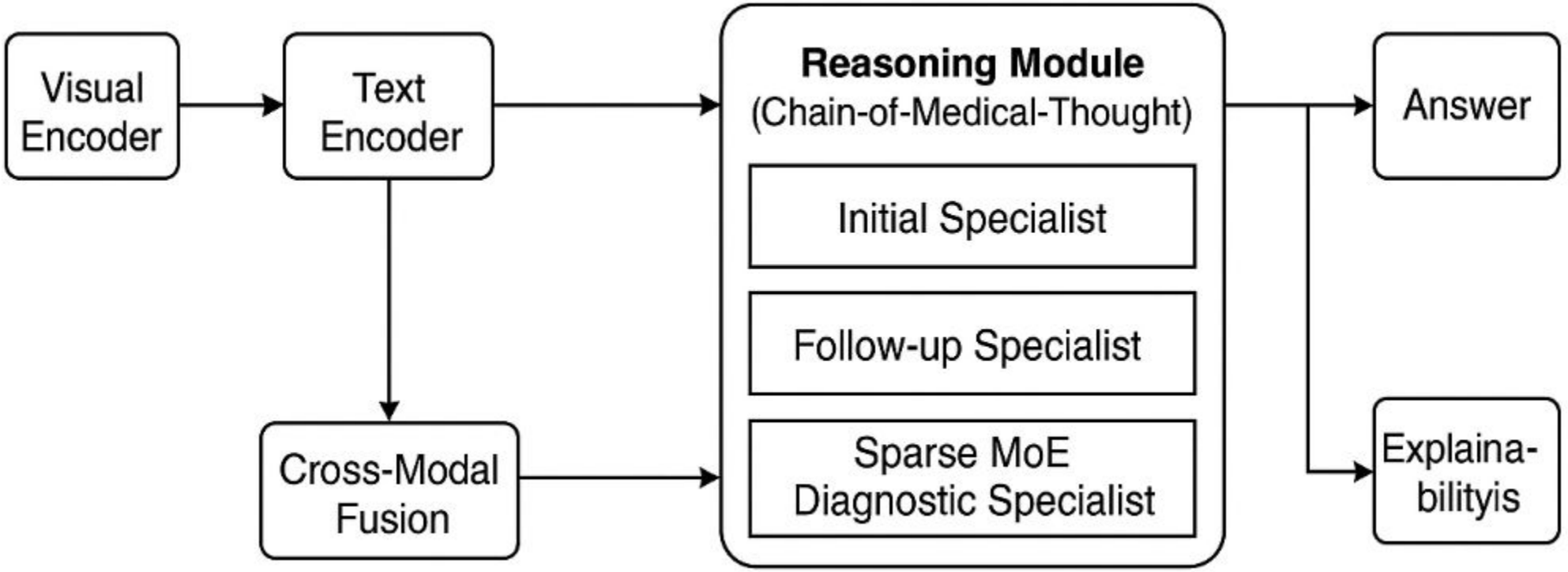


Fig. 5 Proposed Model Architecture

Proposed Methodology

Input stage

- Radiology image III and free-form question QQQ.
- Image encoded by **Swin-Transformer/CLIP-Med** visual backbone.
- Question encoded by Bio-GPT or PubMed-BERT text encoder.

Cross-modal fusion

- Concatenate visual & textual embeddings; refine with gated cross-attention.

Proposed Methodology

Chain-of-Medical-Thought reasoning module (adapted from MedCoT)

1. **The initial specialist** generates a step-by-step rationale.
2. **Follow-up The specialist** self-verifies or revises the rationale.
3. **The Sparse MoE diagnostic specialist** votes on the final answer and attaches the approved rationale for transparency .

Explainability head

- Grad-CA heat map and textual rationale exported together.

Future Work

- Fine tune the encoder and the decoder for the fused feature
- Integrate Rationale using LLMs as proposed in MedCoT and compare performance.
- Fine tune the

Related Work

Paper	Summary	Limitations
Envisioning MedCLIP: A Deep Dive into Explainability for Medical Vision-Language Models	• Proposes MedCLIP, a contrastive learning-based VLM trained on biomedical image-text pairs. • Focuses on explainability using attention visualization and gradient-based methods.	• Explainability is largely post-hoc and not tightly integrated with the decision process. • Performance on VQA tasks is not evaluated, limiting its applicability to QA settings.
MedCoT: Medical Chain of Thought via Hierarchical Expert	• Proposes a Chain-of-Thought (CoT) reasoning framework for VQA in the biomedical domain using hierarchical expert modules. • Integrates vision and text encoders with an answer generator that simulates human-like reasoning.	• Requires pre-annotated reasoning steps for optimal performance. • Explainability is implicit via CoT but lacks fine-grained visual grounding or saliency-based explanations.
Self-Verification Improves Few-Shot Clinical Information Extraction	• Introduces a self-verification framework where a language model rechecks its own answers to reduce hallucination in few-shot clinical IE tasks. • Enhances factual correctness in low-data regimes.	• Focuses solely on textual information extraction, not multimodal (image + text) tasks like VQA. • Self-verification is not yet explored in vision-language contexts.

Explainable AI

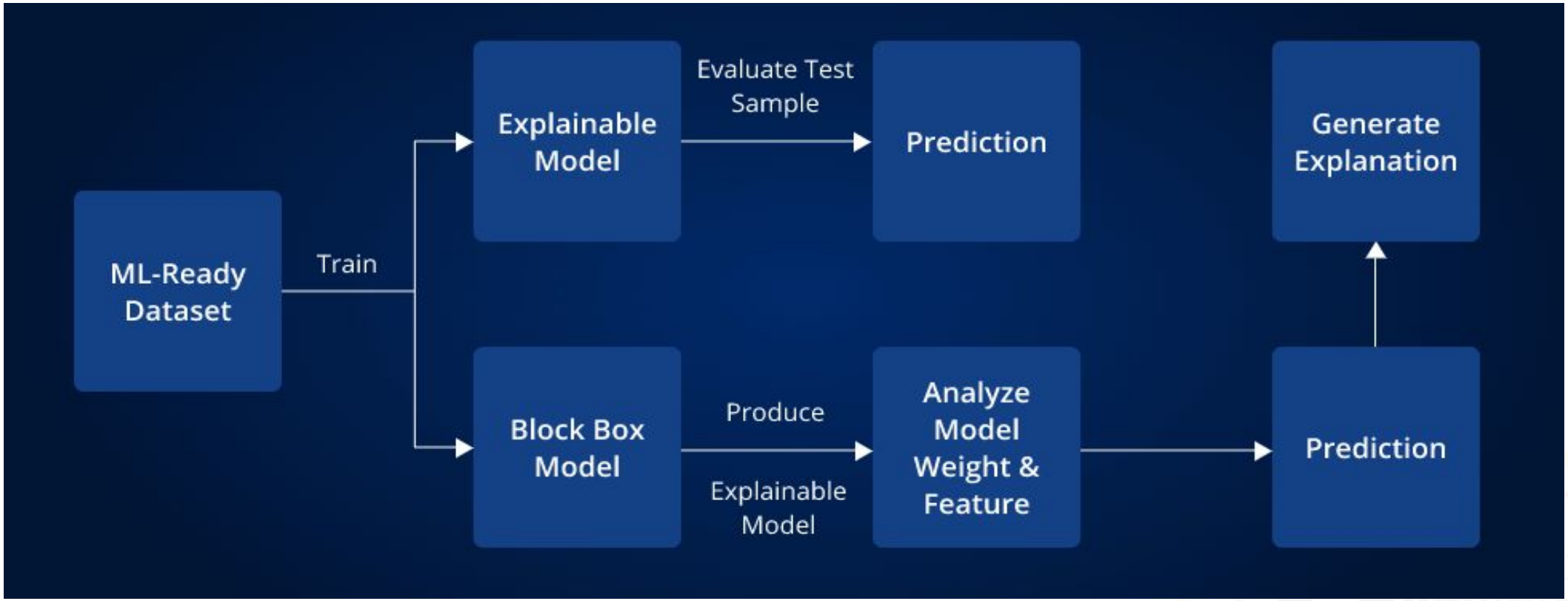


Fig. 2. Working of Explainable AI

https://www.leewayhertz.com/explainable-ai/

Future Work

- **Larger multimodal datasets** (VQA-RAD → MIMIC-CXR + new CoMT QA pairs).
- Training the decoder for the used transformer model

- Train the MoE Model on larger dataset.

- Training the transformer using the same Encoder-Decoder Architecture.
- Integrate explainability method and rationale generation with the archite

Results

Fig 13. Loss curve