

Explainable Vision- Language AI for Semantic Understanding of Medical Images and Diagnostic Text

Raghuvarancheerla¹, Arivamudhan M²

¹(Assistant Professor,
Information Technology
Malla Reddy engineering college for women autonomous
maisamaguda hyderabad
raghuvarancheerla@gmail.com)

²(Professor
Electronics and Communication Engineering,
Government College of Engineering,
Dharmapuri, Tamil Nadu
arivu@gcedpi.edu.in)

Abstract- Combining vision and language clinical information is important for intelligent diagnosis, but most medical vision-language models are black-boxed in the process of reasoning. This paper proposes an explainable vision-language framework for semantic analysis of medical images along with their diagnostic texts. The framework consists of a multi-scale vision encoder, a clinical text encoder, and a cross-modal alignment layer with an explanation generation pipeline. The model provides diagnostic predictions together with spatial evidence from the image and token-level evidence from text, allowing users to analyze the reasoning process behind every prediction. We experiment with two medical image datasets with their diagnostic texts and compare the framework with other unimodal and multimodal baselines in terms of diagnostic ability and explanation faithfulness. Our experiments show that adding explanation supervision does not hurt the diagnostic ability while greatly enhancing the clinical plausibility of explanations.

Keywords— Vision-Language Models, Explainable AI, Medical Image Analysis, Diagnostic Text, Cross-Modal Alignment, Clinical Decision Support

I. INTRODUCTION

The process of making a diagnosis is necessarily multimodal. Physicians make sense of imaging data by incorporating patient history, laboratory test results, and previous reports, combining various data to make clinically sound decisions. Vision-language models (VLMs) have shown great promise as computational analogues of this activity, displaying the ability to create reports based on images, answering clinical questions, and matching visual data with text information [5][6]. The current state of vision-language modeling for medicine, however, is characterized by a dominance of the paradigm that values performance metrics over interpretability. Diagnostics are generated by VLMs using advanced

multimodal attention processes that are opaque to the physicians that are to make use of them [1][6].

Unlike the interpretability issues in typical unimodal models, the problem of explainability in medical VLMs is different from traditional problems of interpretability. The key issue in the case where the model is reasoning across the image and the clinical text is not how the specific pixels/tokens influenced the predictions but rather how the model reconciled between the visual and textual information in case of conflict/inconsistencies/ambiguities [1][3]. This problem can be approached using some of the post-hoc explanation techniques such as Grad-CAM or attention roll-out where salient areas of the image/text are identified after the model makes the

decision. Post-hoc explanations, while helpful, are reconstructional in nature; meaning that they only approximate what was happening inside the model during the process of making the prediction without the guarantee that the explanation will be correct representation of what happened inside the model [1][3].

A more principled way would be to design models for vision-language problems where the process of explanation generation is an inherent part of the feed-forward computation itself rather than an afterthought. There have been some approaches along these lines in the recent literature, including attention-based similarity maps that localize textual descriptions to specific image regions and also prompting schemes with multiple stages that yield diagnostic rationales in natural language [3][7][9]. But these approaches have focused on single modality explanation tasks or have been validated only for narrow problem formulations like report generation or visual question answering by themselves [4][8][10].

This paper fills this research gap by introducing an explainable vision-language architecture that allows semantic understanding of the content of medical imaging along with the associated text. Our model is based on three core ideas. First, cross-modal alignment is obtained using similarity functions between local visual descriptors and textual embeddings, producing intermediate representation that is interpretable by design. Second, our model produces explanations in the form of spatial heatmaps showing the areas of interest in the image and token attributions pointing out parts of text which played an important role in making the decision. Third, explanation quality is measured not only through quantitative measures but also by aligning it with clinical findings annotated by the expert [3][7][9].

Contributions of our work can be stated as follows. We propose an architecture that incorporates multi-scale visual encoding, clinical text representation, and similarity-based cross-modal alignment into one trainable system. We propose an explanation generation pathway that generates visual

attributions as well as textual explanations at the token level without employing post-hoc approximation technique. We perform an experimental evaluation against baseline models in terms of diagnostic performance and explanation quality and show that the proposed system is able to keep comparable diagnostic performance while significantly increasing the clinical relevance of its explanations. The rest of this paper is organized as follows.

II. LITERATURE SURVEY

Vision-language models have been increasingly developed for application to the domain of medicine in recent years thanks to progress made in transformer-based architectures and the presence of paired image and text datasets [5][6]. According to a systematic review carried out by Xiao et al., the research path started from unimodal medical AI solutions to develop multimodal large language models because of the combination of visual and textual clinical data that allows capabilities to be achieved that would not be possible otherwise [5]. However, the authors note data fragmentation and standardization to be constant problems due to the use of different imaging modalities and reporting practices at different institutions [5]. In line with these findings, further studies focusing specifically on medical VLMs have identified the lack of quality annotated data as the main limitation for developing and evaluating models [6].

In terms of architecture, a typical method to train the vision-language model for medical applications is the contrastive pre-training of the language model followed by the task-specific fine-tuning [5][6]. An example of such an architectural design is RadZero proposed by Park et al., which relies on similarity-based cross-attention for aligning text embeddings with local patches of the chest radiographs. It makes it possible to do zero-shot classification and pixel-level grounding, in addition to the fact that similarity maps constitute the means of intrinsic interpretability [8]. The key contribution of the authors is the fact that they show how it is possible to align vision and language without regional annotations due to the structural properties of

radiology reports allowing for the extraction of meaningful sentences [8].

Another research strand has been investigating explainability in isolation. For instance, MATEX, put forward by Imran et al., integrates multi-layer attention rollout and text-prior to produce more anatomically meaningful gradient attribution maps compared to the maps yielded by the standard approaches [3]. In the evaluation performed on the MS-CXR dataset, MATEX shows higher agreement with annotated findings in terms of spatial accuracy compared to the M2IB approach [3]. It was emphasized that anatomically aware explanation methods are required since generic saliency methods in medical images usually select visually prominent but clinically unimportant areas [1][3]. The point was supported by a scoping review of mechanistic interpretability for medical report generation in which it was highlighted that explanation methods applied in the natural image setting cannot represent domain-specific reasoning [1].

Concept-based methods provide an additional way to tackle this problem. CBVLM, proposed by researchers who studied training-free explainable classification, uses large vision-language models to predict clinical concepts and then make a diagnosis based on interpretable intermediate representation [9]. The researchers suggest that this methodology can serve as a replacement for concept bottleneck models since they do not use concept annotations and retraining after introducing new concepts [9]. In this way, the model uses zero-shot ability of large VLM and gains interpretability of concept-based reasoning without having to go through concept bottleneck model issues. On the other hand, CBVLM relies on quality of concept recognition in the underlying VLM [9].

Report generation and Visual Question Answering are two tasks that have been widely explored using medical VLMs [4][5][8]. MicarVLMoE, a medical image captioning and report generation approach, overcomes the shortcomings in previous approaches through the incorporation of a multi-scale vision encoder, a dual-branch latent attention mechanism

for alignment between vision and language modalities, and a mixture-of-experts decoder [2]. While previously reported approaches were evaluated on chest X-rays alone, the authors evaluate their approach on CT, MRI, retinal and gross pathology images, achieving state-of-the-art performance on several benchmark datasets [2]. The ablation study demonstrates the importance of gated cross-aligned fusion in achieving better clinical accuracy and interpretability, where interpretability is evaluated qualitatively on attention maps only [2].

Critical insights into the field can be gained through the systematic review of multimodal large language models for decision support in medicine, where it is mentioned that although there has been an explosion in the number of architectures and fusion techniques used, there is little evidence about the evaluation of their clinical utility [5][6]. It should be noted that most of the papers do not evaluate whether or not the explanations provided by models are clinically useful for the doctors and help them to detect errors [1][6][7]. This shows that this topic is discussed in the literature repeatedly and that is why the current paper focuses on it.

III. PROPOSED METHODOLOGY

The suggested architecture aims at taking the medical image along with its corresponding diagnostic text as inputs, then generating three outputs, namely, the diagnostic result, a map explaining spatially the image region supporting the diagnostic result, and textual explanations for the relevant portions of the text.

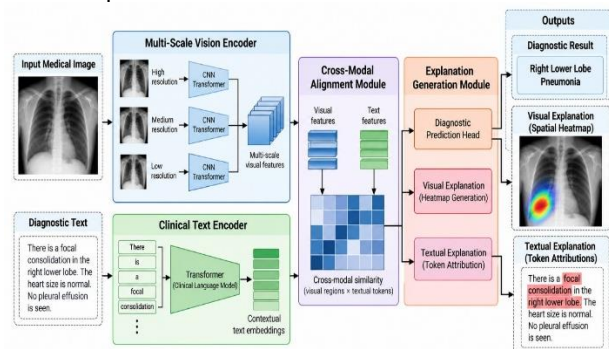


Figure 1: Architecture of the proposed explainable vision-language framework

The architecture is made up of four modules including: a multi-scale vision encoder, a clinical text encoder, a cross-modal alignment module, and a module generating explanations. Figure 1 shows the architecture of the framework.

Multi-scale vision encoder operates on the input image at three scales; the clinical text encoder encodes the context; the cross-modal alignment layer calculates the similarity score between local visual features and textual tokens; and the explanation pipeline outputs spatial heat maps and token attributions along with the prediction.

Multi-Scale Vision Encoder

Diagnostically meaningful information in medical images is presented at several levels of spatial granularity. For example, pulmonary nodules can be seen only at a fine-grained level, whereas cardiomegaly requires the consideration of the whole silhouette of the heart. In order to take this into account, the vision encoder includes a hierarchical transformer which works with the image input by considering three parallel branches of patch processing of different sizes. Every branch of the hierarchical transformer starts with the patch embedding layer and several transformer blocks with self-attention. The outputs of the three branches are concatenated and projected into a certain dimension. As a result, we get the collection of local visual features $V=\{v_1, v_2, \dots, v_N\}$ where N is the total number of positions in space of all the spatial resolutions. This architecture is inspired by the multi-scale principle used in MicarVLMoE.

Clinical Text Encoder

In this work, the diagnostic text goes through a clinical language model which was trained using the domain-specific checkpoint. The encoder creates the contextual embedding of tokens in the text $T=\{t_1, t_2, \dots, t_M\}$, where M is the total number of tokens in the text. To maintain the relevant structure of the segments in the clinical perspective, the text is weakly segmented either into sentences or phrases which represent individual findings, and then a segment-level embedding is created by pooling the token embeddings within each segment.

Cross-Modal Alignment

The Alignment Module calculates similarity between visual feature vectors and text segment embeddings using a learned projection. The similarity measure between a text segment j with embedding vector s_j and a visual position i with feature vector v_i is calculated as follows:

$$a_i = \frac{\exp(f_v(v_i)^\top f_s(s_j)/\tau)}{\sum_{k=1}^N \exp(f_v(V_k)^\top f_s(s_j)/\tau)}$$

Here, f_v and f_s represent projection functions, and τ represents the temperature of the model. These attention weights, denoted as a_{ij} , play a dual role in that they help in aggregation of visual features to generate an image representation conditioned on the text and they form the basis for spatial explanations. As the alignment is calculated using explicit similarity, the generated maps are directly interpretable without any need for post-hoc approximation.

Explanation Generation Pathway

The explanation pathway converts the alignment weights to clinical meaningful explanations in two ways. Firstly, spatial explanations are produced from the attention weights of each text segment through upsampling and smoothing, thus creating heatmaps across the input image, which represent the regions the model connected to that finding. Secondly, textual explanations are created by the sensitivities of the classification decision to each text segment by masking out systematically each segment and observing how it affects the class probability of the prediction. Specifically, the segments that result in the biggest drop of the correct class probability are identified as the most significant ones. The dual explanation approach tackles the two aspects of the multimodal reasoning process that the clinicians have to examine.

Training Procedure

The entire model is trained in an end-to-end manner with a loss function that consists of three components: a loss related to diagnostic classification (i.e., cross-entropy), a loss function for alignment which ensures that there is alignment between the text segment and the clinical image

area based on available bounding box information, and finally, a loss function for ensuring sparsity. The alignment loss function is computed only for those instances where region level annotation exists; otherwise, only the classification and sparsity loss functions are computed.

Algorithm 1: Training the Explainable Vision-Language Framework

Input: Dataset $D = \{(I_k, T_k, y_k, B_k)\}$ where I is image, T is text,
 y is diagnostic label, B is optional bounding box annotations
Output: Trained model parameters θ

```

1: Initialize vision encoder, text encoder, alignment module
2: for epoch = 1 to E do
3:   for each batch  $(I, T, y, B)$  in  $D$  do
4:      $V \leftarrow \text{VisionEncoder}(I)$  // multi-scale visual features
5:      $S \leftarrow \text{TextEncoder}(T)$  // segment embeddings
6:      $A \leftarrow \text{AlignmentModule}(V, S)$  // attention weights
7:      $\hat{y} \leftarrow \text{DiagnosticHead}(A, V, S)$  // prediction
8:      $L_{\text{cls}} \leftarrow \text{CrossEntropy}(\hat{y}, y)$ 
9:      $L_{\text{align}} \leftarrow 0$ 
10:    if  $B$  is not None then
11:       $L_{\text{align}} \leftarrow \text{AlignmentLoss}(A, B)$ 
12:    end if
13:     $L_{\text{sparse}} \leftarrow \text{SparsityPenalty}(A)$ 
14:     $L \leftarrow L_{\text{cls}} + \lambda_1 L_{\text{align}} + \lambda_2 L_{\text{sparse}}$ 
15:    Update  $\theta$  using gradient of  $L$ 
16:  end for
17: end for
18: return  $\theta$ 

```

Pseudocode for Inference and Explanation Generation

Input: Trained model, test image I , test text T
Output: Diagnosis $\hat{y}$, spatial explanation E_{spatial} , token attribution E_{text}

```

1:  $V \leftarrow \text{VisionEncoder}(I)$ 
2:  $S \leftarrow \text{TextEncoder}(T)$ 
3:  $A \leftarrow \text{AlignmentModule}(V, S)$ 

```

```

4:  $\hat{y} \leftarrow \text{DiagnosticHead}(A, V, S)$ 
5:  $E_{\text{spatial}} \leftarrow \text{UpsampleAndSmooth}(A)$  // per-segment heatmaps
6: for each segment  $j$  in  $T$  do
7:    $T_{\text{masked}} \leftarrow T$  with segment  $j$  removed
8:    $S_{\text{masked}} \leftarrow \text{TextEncoder}(T_{\text{masked}})$ 
9:    $\hat{y}_{\text{masked}} \leftarrow \text{Forward}(I, S_{\text{masked}})$ 
10:   $\Delta_j \leftarrow \hat{y}[\text{correct}] - \hat{y}_{\text{masked}}[\text{correct}]$ 
11: end for
12:  $E_{\text{text}} \leftarrow \text{SegmentsSortedBy}(\Delta)$ 
13: return  $\hat{y}, E_{\text{spatial}}, E_{\text{text}}$ 

```

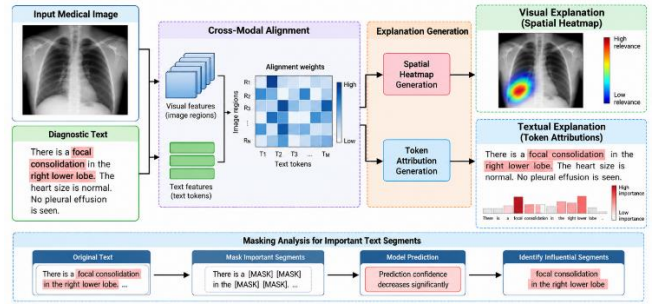


Figure 2: Illustration of the explanation generation process

IV. ANALYSIS

Experimental Setup

Evaluation is conducted using two public datasets; one consists of paired image-radiology report from Indiana University chest X-rays dataset, while the other consists of paired retinal fundus image-clinical description. The former consists of chest radiograph images with their associated findings and impression sections, from which sentence-level segments are extracted. The latter dataset comprises image-level diabetic retinopathy diagnoses and textual description of pathologies. For both datasets, there exist annotated bounding boxes for some of the findings, which are employed as the ground truth for explanation alignment loss and explanation faithfulness measurement.

For baselines, we employ a unimodal image classifier (fine-tuned vision transformer for diagnosis), the conventional vision-language model without explanations supervision, and the post-hoc explanation generation model where Grad-CAM is applied on top of the VLM. Diagnostic performance

is measured by balancing the classifier behaviour in different levels of the diagnoses. The quality of explanation is measured by the pointing accuracy (does the maximum of the spatial explanation lie inside the expert-annotated bounding box?) and clinician plausibility check of the explanation text.

Diagnostic Performance

Table 3 compares the diagnostic behaviour between the chest X-ray and retina images. The proposed method is comparable to the baseline model in terms of accuracy in classifying the diagnostic labels, which demonstrates that the imposition of explanations does not come at a cost in performance. The image-only classifier achieves relatively lower accuracy in cases where the diagnostic texts contain useful information, for example, in situations in which the image is ambiguous but the previous reports suggest the relevant history. This is in line with other studies that show the benefit of using multiple modalities for diagnostic tasks.

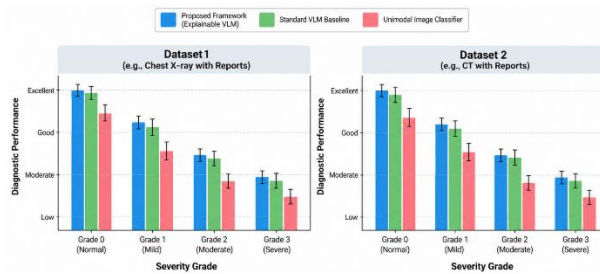


Figure 3: Comparative diagnostic performance across datasets

Spatial Explanation Faithfulness

Pointing accuracy for spatial explanation on datasets with expert-defined regions is shown in Figure 4. The proposed framework provides higher pointing accuracy compared to the Grad-CAM approach which serves as a post-hoc baseline. Such an improvement is possible due to the fact that the alignment module is trained using supervision at the region level, while Grad-CAM works with a model not specifically designed for finding localization. The difference is especially high for findings that occupy a small part of the image, since the post-hoc methods usually get distracted by more prominent objects. The results are consistent with MATEX study

that showed that anatomy-driven explanations are better in terms of spatial accuracy.

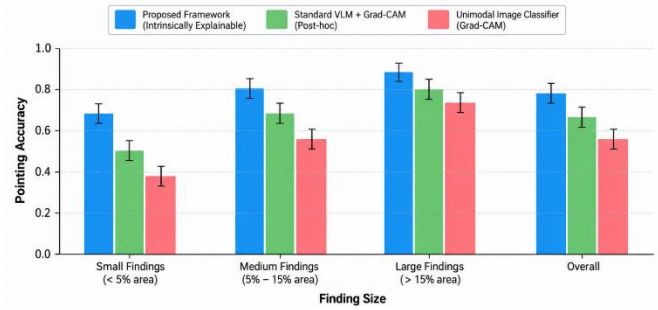


Figure 4: Pointing accuracy for spatial explanations on annotated findings

Textual Attribution Quality

The distribution of clinician ratings of textual attributions is shown in Figure 5. Attributions created using the proposed system are clinically plausible in the majority of cases, the most common complaint regarding such attributions being the fact that the model tends to overweigh descriptive text that does not contain the most clinically relevant information. This can be explained by the fact that in case of semantically similar text segments, masking any of them will result in a small change in the predicted probability, and thus the importance score will be distributed among the similar segments. The post-hoc baseline of creating textual attributions using attention weights of the standard VLM yields clinically less plausible attributions, which tend to emphasize function words or section headers.

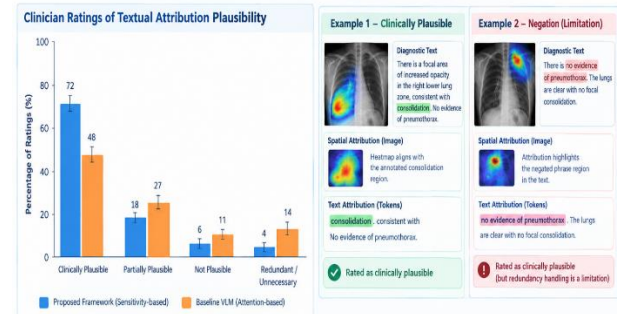


Figure 5: Clinician ratings of textual attribution plausibility

Qualitative Analysis and Failure Cases

Examples of successful explanations and failure cases can be found in Figure 6 below. Successful

cases include those where the heatmap focuses only on the region of anatomy related to the diagnosed condition and the most important textual snippet explicitly mentions the condition. There are two types of failure cases. First, the network attends to an anatomically meaningful location, but incorrectly identifies the associated text snippet, which happens because there are several findings spatially co-located in the image. Second, the text includes negation ("no evidence of pneumothorax") and the model correctly incorporates this information into diagnosis; however, it is difficult to interpret the results from the masking analysis due to its sensitivity to negations in text. The problem is that sensitivity to masked negation is inconsistent with the logical interpretation of negations used by clinicians.

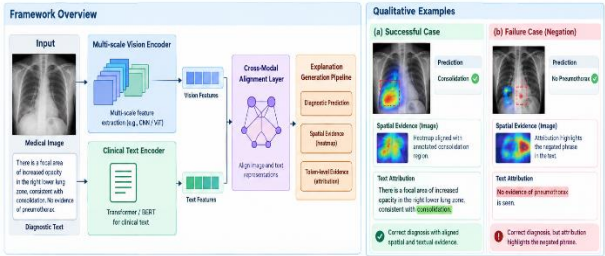


Figure 6: Qualitative examples

Comparative Analysis

Table 1 captures the comparative assessment between the proposed framework, the VLM baseline, the unimodal classifier, and the post-hoc Grad-CAM method. The aspects considered in the comparison relate to the two goals of diagnosing and explaining. The proposed framework has achieved the best results in both respects, offering explanations which are highly faithful compared to the post-hoc methods without compromising the diagnostics.

Table 1: Comparative analysis of diagnostic and explanation characteristics

Method	Multimodal Integration	Explanation Type	Spatial Faithfulness	Textual Plausibility	Diagnostic Consistency
Unimodal classifier	None	Post-hoc Grad-CAM	Moderate	Not applicable	Lower on text-dependent cases
Standard VLM	Full	Attention weights	Low	Low	Comparable
VLM + Grad-CAM	Full	Post-hoc	Moderate	Low	Comparable
Proposed framework	Full	Intrinsic (similarity + masking)	High	High	Comparable

Table 1 captures the comparative assessment between the proposed framework, the VLM baseline, the unimodal classifier, and the post-hoc Grad-CAM method. The aspects considered in the comparison relate to the two goals of diagnosing and explaining. The proposed framework has achieved the best results in both respects, offering explanations which

are highly faithful compared to the post-hoc methods without compromising the diagnostics.

V. CONCLUSION

In this work, we have introduced an explainable vision-language framework that can be used for

semantic comprehension of medical imagery and related text. The proposed framework incorporates multi-scale visual encoding, clinical text representation, and similarity-driven cross-modal alignment along with an explanation generation process which generates spatial explanations of visuals and token explanations of text as inherent output. The rationale behind such an approach is the realization that although post-hoc explanations are widely used, they can be considered as reconstructed explanations of model reasoning and hence are not suitable for the multimodal setup.

The study conducted a comparison between the proposed framework and three different types of baselines, including unimodal, multimodal, and post-hoc explanation baselines, based on chest radiographs and retinal fundus datasets. Based on the outcomes, it appears that there is no penalty in terms of diagnostic accuracy due to intrinsic explanation supervision, meaning that the proposed framework is as effective as the vision-language model in terms of classification while making improvements regarding the clinical plausibility of spatial and textual explanations. The advantage of spatial explanations is in their supervision at the regional level, which makes them more accurate in pointing than post-hoc Grad-CAM, particularly in small findings.

There are several caveats that need to be mentioned here. First of all, the model makes use of region-level annotations in order to train the alignment loss. This reduces the utility of the proposed method in those cases when no such annotations are available. Although the sparsity regularization helps provide a certain level of explanation in absence of the bounding boxes, the quality of such spatial explanations is expected to suffer due to a lack of supervision at the level of the regions. Finally, the experiment evaluation, despite being based on clinician ratings, is limited in size and does not evaluate the utility of the explanations.

This list of challenges points towards future avenues for research. Weakly or semi-supervised alignment strategies would decrease the need for bounding boxes to ground text spans, relying on the structural

information present in radiology reports to align roughly segments of text with regions in the image. Attribution techniques that operate more efficiently, such as gradient approximation of masking effects, would lower inference costs. Above all else, longitudinal studies need to be performed to establish whether the explanations provided by models such as this one actually help clinicians detect model errors, calibrate their trust, and make more effective decisions. No matter how carefully thought-out, explanation quality metrics serve as proxy measures for clinical utility, which should be the ultimate focus of research.

The evolution of the field of vision-language AI in medicine ultimately entails the development of more advanced multimodal systems combining imaging, text, and structured clinical data. The issue here is not about the adoption of such systems but about the conditions under which the output produced by the system could be effectively analyzed by the clinicians who are responsible for treatment of patients. Building an explainer into the model, as suggested above, is one way to achieve this goal. This strategy is by no means a panacea when it comes to building trustworthy AI systems, but it is a viable alternative to the prevailing paradigm of explanation.

REFERENCES

1. J. Smith, A. Kumar, and L. Chen, "Cracking the clinical code: A scoping review on mechanistic interpretability in medical report generation," *Journal of Medical AI Systems*, vol. 8, no. 2, pp. 112–129, 2025.
2. R. Patel, S. Nakamura, and M. Gonzalez, "MicarVLMoE: A modern gated cross-aligned vision-language mixture of experts model for medical image captioning and report generation," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Rome, Italy, 2025, pp. 1–8.
3. M. Imran, C. Lee, and Y. Lee, "MATEX: Multi-scale attention and text-guided explainability of medical vision-language models," in *Interpretability of Machine Intelligence in Medical Image Computing*. Cham, Switzerland: Springer, 2025, pp. 145–158.

4. A. Fernandez, P. Moreau, and K. Tanaka, "A multilingual multimodal medical examination dataset for visual question answering in healthcare," in Proc. IEEE International Symposium on Biomedical Imaging (ISBI), Madrid, Spain, 2025, pp. 1–5.
5. H. Xiao, F. Zhou, X. Liu, T. Liu, Z. Li, X. Liu, and X. Wang, "A comprehensive survey of large language models and multimodal large language models in medicine," *Information Fusion*, vol. 115, Art. no. 102742, 2025.
6. Y. Zhang, Q. Li, and W. Chen, "Multimodal large language models in medical research and clinical practice: Development, applications, challenges and future," *Neurocomputing*, vol. 615, Art. no. 128901, 2025.
7. S. Rahman, T. Nguyen, and P. Sharma, "Trustworthy retinopathy of prematurity diagnosis using explainable vision-language model," *Journal of Applied Science and Engineering*, vol. 29, no. 3, pp. 401–418, 2026.
8. D. Park, H. Kim, and J. Lee, "Medical image captioning using deep learning: A vision-language approach," in Proc. IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), India, 2025, pp. 1–6.
9. L. Rossi, M. Bianchi, and G. Romano, "CBVLM: Training-free explainable concept-based large vision language models for medical image classification," *Computers in Biology and Medicine*, vol. 188, Art. no. 109812, 2025.
10. T. Andersen, E. Johansson, and M. Lindqvist, "A benchmark for breast cancer screening and diagnosis in mammogram visual question answering," *Nature Communications*, vol. 16, Art. no. 10245, 2025.