

## 1 APPENDIX A ABBREVIATIONS

|    |                                                                         |
|----|-------------------------------------------------------------------------|
| 2  | AV – Average Value                                                      |
| 3  | BLEU score – Bilingual Evaluation Understudy score                      |
| 4  | CAM – Class Activation Mapping                                          |
| 5  | CLIP – Contrastive Language-Image Pre-Training                          |
| 6  | FN – False Negative                                                     |
| 7  | FP – False Positive                                                     |
| 8  | GHZ – GigaHertz                                                         |
| 9  | GPU – Graphics Processing Unit                                          |
| 10 | Grad-CAM – Gradient-weighted Class Activation Mapping                   |
| 11 | GRU – Gated Recurrent Unit                                              |
| 12 | HT – Hardsample Threshold                                               |
| 13 | ID – Improved datasets                                                  |
| 14 | IM – Images                                                             |
| 15 | L – Labels                                                              |
| 16 | LIME – Local Interpretable Model-Agnostic Explanations                  |
| 17 | LRP – Layer-wise Relevance Propagation                                  |
| 18 | LSTM – Long Short Term Memory                                           |
| 19 | MAM – Multimodal Augmented Mixup                                        |
| 20 | MII – Medical Images Improvement                                        |
| 21 | MVQA – Medical Visual Question Answering                                |
| 22 | NLP – Natural Language Processing                                       |
| 23 | Path-VQA – Pathology visual question answering                          |
| 24 | QAI – Question-Answer pairs Improvement                                 |
| 25 | QAP - Question-Answer Pairs                                             |
| 26 | RAM – Random-access memory                                              |
| 27 | ResNet – Residual Neural Network                                        |
| 28 | RGB – Red Green Blue                                                    |
| 29 | RoI – Region of Interest                                                |
| 30 | SHAP – Shapley Additive exPlanations                                    |
| 31 | SoA - State-of-Art                                                      |
| 32 | TN - True Negative                                                      |
| 33 | TP - True Positive                                                      |
| 34 | TSE – Transformer with Sentence Embedding                               |
| 35 | VGGNet – Visual Geometry Group Neural Network                           |
| 36 | VQA-MED 20XX dataset – Visual Question Answering – Medical 20XX dataset |
| 37 | VQA-RAD – Visual Question Answering in Radiology                        |
| 38 | XAI - eXplainable Artificial Intelligence                               |
| 39 | ZCA – Zero-Phase Component Analysis                                     |

40

|                                          |                                                  |
|------------------------------------------|--------------------------------------------------|
| $I_{in}$                                 | → Medical Images (Input)                         |
| $QA_{in}$                                | → Question-Answer Pairs (Input)                  |
| $E_{out}$                                | → Generated MVQA model                           |
| $p_a$                                    | → Probability of category $a$                    |
| $N$                                      | → Number of samples                              |
| $\varepsilon$                            | → Small constant                                 |
| $M$                                      | → Number of classes                              |
| $\alpha$                                 | → Shape parameter                                |
| $\gamma$                                 | → Factorial function (Gamma function)            |
| $\beta$                                  | → Target probability distribution value modifier |
| $\lambda$                                | → Random value used to create new samples        |
| $X^{(1)}, X^{(2)}, \dots, X^{(i)}$       | → Set of inputs                                  |
| $QAP^{(1)}, QAP^{(2)}, \dots, QAP^{(i)}$ | → Set of QA pairs                                |
| $Q^{(1)}, Q^{(2)}, \dots, Q^{(i)}$       | → Set of Questions                               |
| $A^{(1)}, A^{(2)}, \dots, A^{(i)}$       | → Set of Answers                                 |
| $IM^{(1)}, IM^{(2)}, \dots, IM^{(i)}$    | → Set of Images                                  |
| $L^{(1)}, L^{(2)}, \dots, L^{(l)}$       | → Set of Labels                                  |
| $C^{(1)}, C^{(2)}, \dots, C^{(i)}$       | → Set of Categories                              |
| $\wedge$                                 | → Logical AND / Conjunction                      |
| $ID_{out}$                               | → Improved Dataset                               |
| $IM_{out}$                               | → Improved Medical Images                        |
| $IQAP_{out}$                             | → Improved QA Pairs                              |
| $LQAP_{HT}$                              | → Hard Sample Threshold for QA Pairs             |
| $LQAP_{AV}$                              | → Average Value for QA Pairs                     |
| $L_{upd}$                                | → Updated Number of Labels                       |
| $Q_{in}$                                 | → Question (Input)                               |
| $A_{in}$                                 | → Answer (Input)                                 |
| $C_{in}$                                 | → Category (Input)                               |
| $\mu$                                    | → Average of total number of samples             |
| $LR_{CON}$                               | → Concatenated Latent Representation             |
| $m$                                      | → Number of $n$ -grams                           |
| $w_m$                                    | → Weight for each $n$ -gram precision            |
| $q$                                      | → Count of words in Ground Truth                 |
| $c$                                      | → Count of words in a Generated Question         |
| $PM_{PA}$                                | → Proposed Model                                 |
| $LWO_{ATV}$                              | → Layerwise Output                               |
| $C_{ATV}^{Norm}$                         | → Normalized Activation Layer                    |
| $\sigma$                                 | → Standard Deviation                             |
| $ATV_{Loss}$                             | → Activation Layer Loss                          |
| $LWL_{Loss}$                             | → Layerwise Loss                                 |
| $MG_{LL}$                                | → Model Gradient Loss                            |
| $HM^{Norm}$                              | → Normalized HeatMap                             |
| $\nabla$                                 | → Gradient                                       |

42 **3 APPENDIX C HYPERPARAMETERS AND ARCHITECTURE CONFIGURA-**  
43 **TIONS**

**Table 1.** Hyperparameters and Architecture Configurations

| Component               | Parameter           | Value / Description                             |
|-------------------------|---------------------|-------------------------------------------------|
| 7*General Settings      | Random Seed         | 42                                              |
|                         | Epochs              | 100                                             |
|                         | Batch Size          | 32                                              |
|                         | Optimizer           | Adam                                            |
|                         | Learning Rate       | 0.0001                                          |
|                         | Weight Decay        | $1 \times 10^{-4}$                              |
|                         | Dropout Rate        | 0.5                                             |
|                         | Early Stopping      | Patience = 10 (on validation loss)              |
|                         | Loss Function       | Cross-Entropy Loss                              |
| 3*Image Preprocessing   | Resize Dimensions   | $224 \times 224$                                |
|                         | Normalization       | ImageNet mean/std                               |
|                         | Augmentations       | QAI, MII                                        |
| 3*Text Preprocessing    | Tokenizer           | NLTK                                            |
|                         | Max Sequence Length | 50 tokens                                       |
|                         | Padding             | Post-padding                                    |
| 3*VGGNet Configuration  | Pretrained          | Yes (ImageNet)                                  |
|                         | Frozen Layers       | All layers including final FC layer             |
|                         | Output Feature Size | 4096                                            |
| 3*ResNet Configuration  | Pretrained          | Yes (ImageNet)                                  |
|                         | Frozen Layers       | All layers including final FC layer             |
|                         | Output Feature Size | 2048                                            |
| 4*LSTM Configuration    | Hidden Size         | 512                                             |
|                         | Number of Layers    | 1                                               |
|                         | Bidirectional       | No                                              |
|                         | Dropout             | 0.5                                             |
| 3*Fusion & Output Layer | Fusion Strategy     | Concatenation of image and text features        |
|                         | Final Output Layer  | Fully connected, output size = number of labels |
|                         | Activation          | Softmax                                         |