Hybrid Reinforced Medical Report Generation With M-Linear Attention and Repetition Penalty

Zhenghua Xu, Wenting Xu, Ruizhi Wang, Junyang Chen, *Member, IEEE*, Chang Qi, and Thomas Lukasiewicz

***Abstract*— To reduce doctors' workload, deep-learning-based automatic medical report generation has recently attracted more and more research efforts, where deep convolutional neural networks (CNNs) are employed to encode the input images, and recurrent neural networks (RNNs) are used to decode the visual features into medical reports automatically. However, these state-of-the-art methods mainly suffer from three shortcomings: 1) incomprehensive optimization; 2) low-order and unidimensional attention; and 3) repeated generation. In this article, we propose a hybrid reinforced medical report generation method with m-linear attention and repetition penalty mechanism (HReMRG-MR) to overcome these problems. Specifically, a hybrid reward with different weights is employed to remedy the limitations of single-metric-based rewards, and a local optimal weight search algorithm is proposed to significantly reduce the complexity of searching the weights of the rewards from exponential to linear. Furthermore, we use m-linear attention modules to learn multidimensional high-order feature interactions and to achieve multimodal reasoning, while a new repetition penalty is proposed to apply penalties to repeated terms adaptively during the model's training process. Extensive experimental studies on two public benchmark datasets show that HReMRG-MR greatly outperforms the state-of-the-art baselines in terms of all metrics. The effectiveness and necessity of all components in HReMRG-MR are also proved by ablation studies. Additional experiments are further conducted and the results demonstrate that our proposed local optimal weight search algorithm can significantly reduce the search time while maintaining superior medical report generation performances.**

***Index Terms*— Attention mechanism, hybrid reward, medical report generation, reinforcement learning (RL).**

Nomenclature

f	Combined regional visual features.
f_{global}	Global averaged visual features.
$\hat{f}$	M-linear attended features.
$\tilde{f}$	Combination of different order $\hat{f}$ and f_{global}.
$g^{(n+1)}$	Final attended visual feature.
M_i^k	Joint bilinear representation of query-key.
M_i^v	Joint bilinear representation of query-value.
$\bar{M}$	Channelwise descriptor.
a^s	Spatial attention weights.
A^s	Spatial attention distribution.
a^c	Channelwise attention weights.
A^c	Channelwise attention distribution.

I. Introduction

IN RECENT years, medical imaging has become the most commonly used medical examination method in disease diagnosis. It produces reports that are paragraph-based documents issued by radiologists after examinations. These reports describe the important medical findings observed on the corresponding medical images and emphasize the abnormalities, along with the sizes and locations of detected lesions. However, due to the increasing number of patients and the shortage of experienced radiologists, a radiologist may have to conduct dozens or sometimes even hundreds of medical imaging examinations and then write the same number of reports every day, which makes the radiologists overloaded and may lead to increasing misdiagnosis. Therefore, there is a compelling demand to find promising methods to generate medical reports automatically.

Existing deep-learning-based medical report generation methods mainly adopt the encoder–decoder architecture [1], [2], [3] where deep convolutional neural networks (CNNs) encode the input medical images, and then recurrent neural networks (RNNs), e.g., long short-term memory (LSTM), as decoder generate medical reports automatically. However, such encoder–decoder models inevitably suffer from the problem of sentence coherence: Cross-entropy is widely used in these methods for optimization, however, it only focuses on word-level errors but ignores the interword connections; since the generated reports consist of long sentences, the coherence of their resulting sentences is usually not satisfactory. Some recent researches [4] have proved that reinforcement learning (RL) can overcome this problem and improve the performances of sequence-to-sequence models (such as LSTM), so RL is used in many recent works to enhance the presentation of the generated medical reports [5], [6].

Manuscript received 26 January 2022; revised 28 March 2023 and 30 August 2023; accepted 10 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62276089, Grant 61906063, and Grant 62102265; in part by the Natural Science Foundation of Hebei Province, China, under Grant F2021202064; in part by the "100 Talents Plan" of Hebei Province under Grant E2019050017; in part by the Key Research and Development Project of Hainan Province, China, under Grant ZDYF2022SHFZ015; in part by the Open Research Fund from the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant GML-KF-22-29; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2022A1515011474; and in part by the AXA Research Fund. *(Corresponding authors: Zhenghua Xu; Ruizhi Wang.)*

Zhenghua Xu, Wenting Xu, and Ruizhi Wang are with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment and the Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin 300130, China (e-mail: zhenghua.xu@hebut.edu.cn; 202131403021@stu.hebut.edu.cn).

Junyang Chen is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.

Chang Qi is with the Department of Information Data, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510275, China.

Thomas Lukasiewicz is with the Institute of Logic and Computation, Vienna University of Technology, 1040 Vienna, Austria, and also with the Department of Computer Science, University of Oxford, OX1 3QD Oxford, U.K.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3343391.

Digital Object Identifier 10.1109/TNNLS.2023.3343391

Despite achieving some improvements, the existing RL methods still suffer from the following three problems.

1) *Incomprehensive Optimization Goals:* Medical report generation is a cross-modal long text generation task, so multiple metrics are needed to evaluate the quality of the generated report comprehensively. So when adopting evaluation metrics as the rewards of RL to optimize the model, simply using one or two metrics (as the existing works [5], [6] that use CIDEr only) is not enough to optimize the model comprehensively; and studies in the existing RL-based dialogue generation work [7] also show that adding more semantic-related rewards will enhance the model's textual generation capability.
2) *Low-Order and Unidimensional Attention:* The organs and tissues in medical images usually have similar characteristics and complex shape changes; however, the medical images in the report generation dataset are usually unlabeled (i.e., without annotations for the potential diseases or lesion areas), making the model difficult to identify suspicious areas and also difficult to match the textual description in the report with the corresponding region in the image. Therefore, the medical report generation models need to be able to not only extract the key regions in the images but also associate them with the corresponding semantic features of the textual descriptions. Although the existing works usually use attention mechanisms to meet these requirements, they only adopt the single low-level visual and/or semantic attention, which, however, cannot exploit channelwise information and high-order feature interactions of medical images and texts to obtain a fine-grained visual and semantic information; therefore, the existing attention mechanisms fail to catch high-order and multimodal information [2], [3], and thus cannot be adapted to the complex multimodal task of medical report generation.
3) *Repeated Generation:* Different from the image captioning [8] or video captioning [9] tasks which usually aim to generate only a short sentence for a given natural image or video frame, the medical report generation task has to generate a paragraph with tens or even hundreds of words, making it a more difficult generation task and more feasible to generate repeated phrases, which thus weakens the coherence and readability of the generated medical reports. Also, as the amount of normal findings is usually much higher than that of abnormal findings in clinical practices, the learned model will tend to generate repeated normal descriptions, which thus reduces the generation accuracy of abnormal findings. Although Melas-Kyriazi et al. [10] also identifies this problem and introduces a penalty weight on the generated words, it is simply an artificially preset hyperparameter that requires carefully manual design and cannot automatically adapt to the data changes, making it hard to assure the correctness and robustness of the penalty in clinical practices.

In this article, to overcome the above problems, we propose a **H**ybrid **Re**inforced **M**edical **R**eport **G**eneration method with **M**-linear attention and **R**epetition penalty mechanisms (abbreviated as HReMRG-MR). Compared to the existing medical report generation works, the proposed HReMRG-MR consists of three improvements.

First, there exist two technical difficulties in solving the problem of incomprehensive optimization goals: 1) how to introduce a hybrid reward that can use all the existing evaluation metrics to adaptively optimize the RL processes according to the different importance of these metrics and 2) how to search the importance weights of these evaluation metrics in a reasonable time. To deal with these technical difficulties, we first propose a hybrid reinforced medical report generation method, HReMRG, where a weighted hybrid reward is used. Specifically, the hybrid reward consists of all seven metrics that are widely used in the existing works to evaluate the quality of the generated reports, making it capable of achieving comprehensive optimization goals. Also, since the evaluation metrics have different importance for the final generated report, an importance weight is assigned to each metric, resulting in a weighted hybrid reward. However, the complexity of finding the optimal weights using grid search is exponential and unacceptable for practical usage, so we further propose a new local optimal weight search algorithm based on a greedy algorithm to approximate the optimal weights in linear complexity. Consequently, compared to the existing RL-based medical report generation works [5], [6], the proposed HReMRG can efficiently optimize the RL process more comprehensively, resulting in much higher quality medical reports in all evaluation aspects.

Second, the technical difficulty in solving the low-order and unidimensional attention problem is how to further propose a high-order multidimensional attention mechanism to extract the features of the key areas of the image and align them with the descriptions to improve the model's ability to learn the features of both images and texts. To deal with this technical difficulty, in this work, we propose a novel m-linear attention module, which contains a set of m-linear attention blocks based on bilinear pooling and stacking architectures. Specifically, comparing to the cross-modality attention [11] and multihead attention [12] that are proposed in the existing image captioning tasks to better align the visual and semantic features, and the attention layers [13] and adaptive distilling attention [14] in the existing medical report generation works that are used to jointly attend information from images and texts, the proposed m-linear attention module has the following advantages: 1) it takes into account not only the spatial but also the channelwise interactions of the features to achieve multidimensional attention processing; 2) to capture the high-order feature interactions, we stack four m-linear attention blocks in each attention module, where the output features of the previous block are used as the inputs of the following block, making the resulting m-linear attention module capable of learning high-order feature interactions (since each m-linear attention block is second-ordered, the resulting module can actually learn eighth-order feature interactions); and 3) besides integrating the m-linear modules into the encoder to learn intramodal features, they are also incorporated into the decoder to learn intermodal dependencies between text and images (i.e., associating the visual features with the corresponding semantic features) to achieve multimodal high-order multidimensional attention. Consequently, the proposed m-learning attention mechanism is capable of overcoming the problem of low-order and unidimensional attention and helping the model generate more accurate medical reports.

Third, the technical difficulty in solving the repeated generation problem is how to propose a new word generation penalty that can automatically adapt to the data changes to ensure its correctness and robustness in practical usage. Consequently, in this work, we propose a new repetition penalty that adaptively suppresses the probability of generating the words with different strengths, i.e., the log-probability of the output word is iteratively updated by subtracting a value proportional to the number of times the word has been generated. Although some existing medical report generation works [15], [16] have utilized topic representation or relational memory network to improve coherence, they are not specifically designed to resolve the repeated generation problem so can not overcome this problem satisfactory. Furthermore, compared to using preset penalty weight as in [10], the proposed adaptive repetition penalty can dynamically adjust the weight according to the change in the number of word repetitions, so it can provide a penalty more accurately to generate more readable and coherent reports.

Overall, the contributions of this article are as follows.

1) We identify three problems of the existing RL-based medical report generation methods and propose an HReMRG-MR to overcome these problems and generate more accurate and readable medical reports.
2) The improvements in the proposed HReMRG-MR are threefold: a) to overcome the weakness of incomprehensive optimization goals, we first propose a hybrid reward with different weights to help measure the quality of the generated report more comprehensively, and a local optimal weight search algorithm is proposed to greatly reduce the rewards' weight searching complexity from exponential to linear; b) then, m-linear attention modules are proposed to help the model learn multidimensional high-order feature interactions and also achieve multimodal reasoning, which thus remedies the low-order and unidimensional attention problem; and 3) finally, an adaptive repetition penalty is proposed to apply penalties to repeated terms adaptively during the model's training process, which thus enhances the coherence and readability of generated reports and is robust in practical usage.
3) Extensive experiments have been conducted on two publicly available medical image report benchmark datasets. The experimental results show that our proposed HReMRG-MR model greatly outperforms the state-of-the-art baselines in terms of all metrics. We have also conducted ablation studies to prove that both the m-linear attention and the repetition penalty mechanism are effective and essential for the model to achieve superior performances. Additional experiments are further conducted to demonstrate that our proposed local optimal weight search algorithm can significantly reduce the search time while maintaining superior medical report generation performances. Furthermore, we also compare the performances of m-linear attention with those of the state-of-the-art attention solutions to show its superiority.

II. Related Works

With rapid advances in RL, recent automatic medical report generation works have used RL techniques to boost their performances [5], [6], [17], [18]. Specifically, by identifying that descriptions of normal organs in medical reports are highly similar, Li et al. [17] designs some templates for common descriptions and proposes a hybrid retrieval-based model using RL to determine the ways of generating sentences (via template retrieval or LSTMs). Then, to optimize the nondifferentiable and sequence-based test metrics directly, Xiong et al. [5] and Liu et al. [6] adopt RL to directly optimize the score of CIDEr. Besides of optimizing the natural language metrics, Liu et al. [6] additionally focused on improving the clinical accuracy of medical reports. A self-critical sequence training (SCST) method is adopted by most of these methods [5], [6], [17], [18] for RL; SCST is an advanced reinforced algorithm that utilizes the output of its own testing inference algorithm to normalize the rewards and tackle the high variance problem of other reinforced algorithms during training [19]. Although SCST has achieved great success in the area of image captioning, its generation capability is still unsatisfactory for the medical report generation, where SCST tends to generate repeated phrases in such a paragraph-based long texts' generation process, i.e., the repeated generation problem. To alleviate this problem, Melas-Kyriazi et al. [10] introduces a penalty weight on the generated words to reduce the occurrence of recurring terms; however, this penalty weight is simply an artificially preset hyperparameter that requires carefully manual design and cannot automatically adapt to the data changes, making it hard to assure the correctness and robustness of the penalty in clinical practices. Compared to the existing RL-based methods, our proposed HReMRG-MR has three advantages: 1) it proposes a novel weighted hybrid reward to utilize all the existing evaluation metrics to adaptively optimize the RL processes according to the different importance of these metrics, where an efficient weight search algorithm is also developed to ensure the weight searching is tractable; 2) an adaptive repetition penalty is proposed to apply penalties to repeated terms adaptively during the model's training process, which thus enhances the coherence and readability of the generated reports and is robust in practical usage; and 3) m-linear attention blocks are also employed in HReMRG-MR to enhance the model's capability in learning the multidimensional high-order feature interactions and also achieve multimodal reasoning.

Motivated by the recent success of using attention mechanisms in artificial medical image analysis [20], recent medical report generation works have started to explore more interactions between images and sentences via attention mechanisms [1], [2], [3], [16], [21], [22]. The classic image captioning model [22] adopts conventional attention to calculate the contribution of regional features to the generated text. To apply attention to both visual and semantic features simultaneously, Jing et al. [2] proposes a coattention mechanism that builds attention distributions separately for visual and semantic domains, but it ignores to consider the multimodal interactions, resulting in bad contextual coherence, i.e., the report is feasible for using repeated words. To guarantee the coherence among sentences, Xue et al. [3] develops a sentence-level attention mechanism to explore multimodal interactions, which compute the attention distribution over visual regions according to sentence-level semantic features; by generating the report sentence by sentence, and generating the long paragraphs in a circular manner, the semantic coherence is well improved in [3]. However, all these attention methods solely explore first-order feature interactions, ignoring higher-order features that are important for guiding medical image understanding and report generation. Therefore, Pan et al. [23] proposes x-linear attention to overcome this problem, where a bilinear attention block is developed to learn the second-order features interactions. Different from the above attention methods, our proposed m-linear attention has the following advantages: 1) it can achieve multidimensional attention by

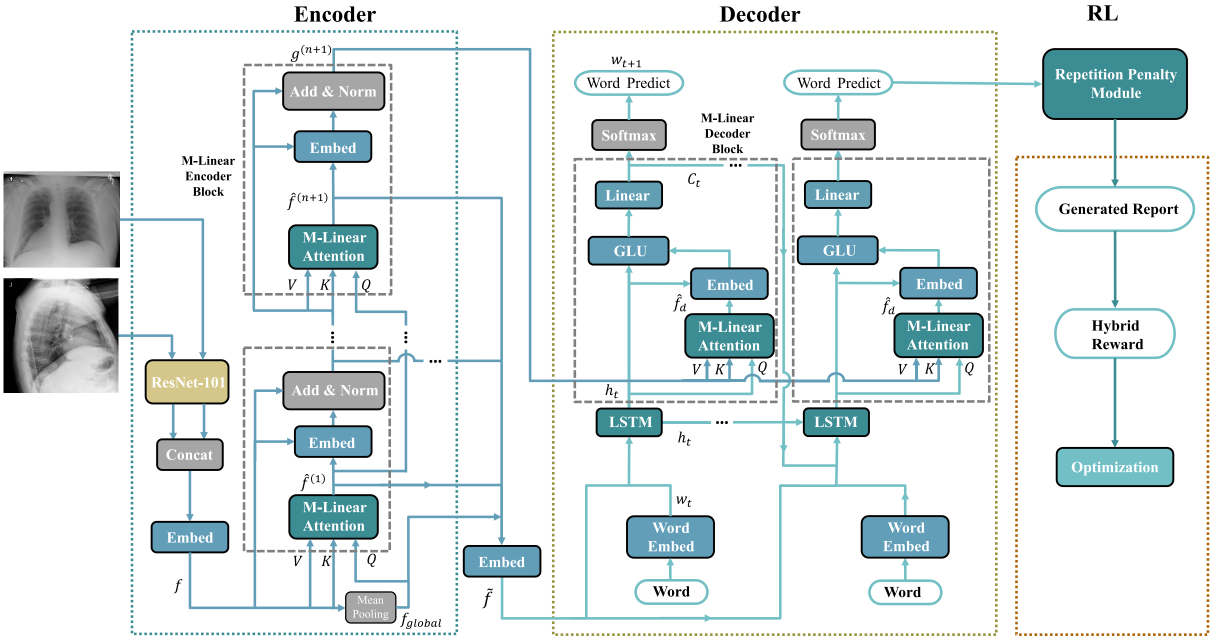


Fig. 1. Architecture of our proposed HReMRG-MR, where Embed denotes a nonlinear projection operation, and GLU denotes gated linear units. Different modules have been marked with dashed boxes of different colors, the m-linear block is marked with a gray dashed box, the encoder is marked with a blue dashed box, the decoder is marked with a green dashed box, and the hybrid RL optimizer is marked with an orange dashed box.

capturing both spatial and channelwise feature interactions; 2) it stacks multiple second-order attention blocks to achieve higher-order attention operations and learn higher order feature interactions; and 3) it can learn both intramodal and intermodal dependencies. Consequently, the proposed m-linear attention mechanism is capable of overcoming the low-order and unidimensional attention problem in the existing attention methods and achieving better report generation performances. Additional experiments are conducted to prove the superiority of the proposed m-linear attention by comparing it with the state-of-the-art attention methods [19], [23], [24].

III. Methodology

We propose an HReMRG-MR. Intuitively, we believe that the use of a hybrid weighted reward will fill the gap of a single CIDEr-based reward, thus generating more readable reports and making the evaluated performance more balanced. Besides, we believe that the use of high-order feature interactions will strengthen the model's capacity in single- and multimodal reasoning, which will enhance the model's performance in terms of the accuracy of the generated reports. Moreover, applying penalties on the repetition terms will produce much more diverse sentences and help increase the coherence and readability of the generated diagnosis reports.

Specifically, as shown in Fig. 1, images of both frontal and lateral (LL) views are encoded through a pretrained ResNet-101 [25] network and concatenated for visual feature extraction. After that, to localize prominent abnormalities of chest X-rays (CXRs) and attach the right descriptions to them, we embed them and put them into blocks similar to a Transformer encoder named m-linear encoder block recursively. Taking the global visual features extracted from medical images f_{global} as input query $\mathbf{Q}$ and the regional features f as input value $\mathbf{V}$ and key $\mathbf{K}$. A stack of m-linear encoder blocks is used to calculate the outer product between two feature vectors and enable channelwise attention through the squeeze-excitation operation. As such, high-order intramodal feature interactions are explored during the encoder procedure.

After that, the embedded attended features from all the above layers $\hat{f}$ are combined and sent into LSTMs during the decoder procedure. The output hidden state h_t of the LSTMs together with the final attended visual feature $g^{(n+1)}$ are then sent into an m-linear decoder block to explore multimodal feature interactions, which is later used for word prediction.

After pretraining with such an encoder–decoder architecture for some epochs, RL is used to boost the model's performances, where we employ the SCST algorithm and use a hybrid weighted reward for optimization. To generate more readable descriptions, we use a repetition penalty module during sentence generation to increase the readability of the generated reports. We present our network design and implementation details in Section III. The frequently used symbols are included and explained in Nomenclature.

A. Visual Encoder CNN

In our work, a ResNet-101-based network pretrained on ImageNet [26] is employed to extract the global and regional visual features of the patient's multiview (frontal and LL) CRX images. We resize our input images to 224 × 224 to keep them consistent with our pretrained ResNet-based CNN encoder. Then, the regional features $f \in \mathbb{R}^{2048\times196}$ (reshaped from 2048 × 14 × 14) are extracted from the last convolutional layer of ResNet-101, which represents 196 subregions. A global average pooling is applied to the extracted regional features to obtain the global features f_{global}. After that, both global and regional features are concatenated according to different views before feeding into the next layers. In this article, we choose to generate the *Impression* and the *Findings* sections. To learn high-order feature interactions both spatially and channelwise, we introduce an m-linear-attention-based encoder, and detailed information is as follows.

1) M-Linear Attention: Although works in medical report generation have frequent uses of the attention mechanisms,

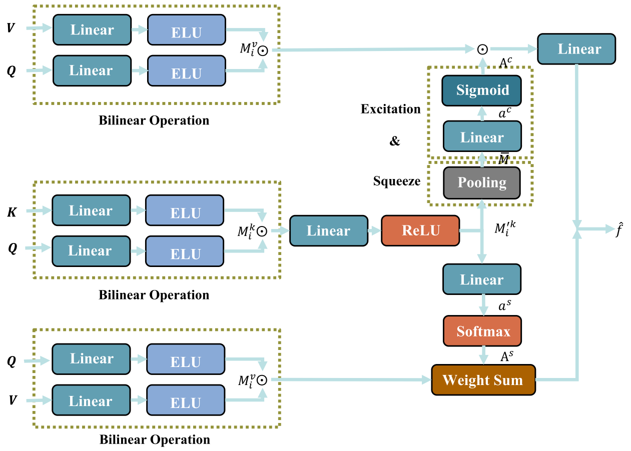


Fig. 2. Architecture of our proposed M-linear attention, which is used in encoder and decoder to learn high-order and multidimensional feature interactions.

there are no works exploring the high-order interactions of features. The existing attention mechanisms tend to focus on low-order spatial or semantic attention, ignoring the high-order feature reasoning, thus resulting in the weakness of locating the prominent abnormalities. Therefore, bilinear pooling is imported to our attention mechanism, which is first used for fine-grained image classification, and has recently been applied to multimodal feature fusion in visual question answering, aiming at exploring multimodal interactions. We thus use it in our medical report generation to explore high-order intramodal interactions in medical image features and intermodal interactions between the radiography features and the respective reports. Furthermore, due to the specificity of data from different domains, a similar mechanism may have different effects on different areas. We are thus motivated to explore a different feature fusion pattern apart from the usually adopted cascaded spatial and channelwise attention mechanism.

To enhance the visual features obtained above, we employ a stack of m-linear attention blocks, which are capable of catching high-order feature interactions. Taking the combined regional features f as the initial input keys $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^N$ and the values $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$, and the global features f_{global} as the initial input query $\mathbf{Q}$, we attend the input features with our m-linear attention as in Fig. 2.

First, we do a low-rank bilinear pooling [27] on both the input query and keys, and the query and values to get the joint bilinear representation of query-key $\mathbf{M}_i^k$ and query-value $\mathbf{M}_i^v$, which encodes the second-order feature interactions of query-key and query-value

$$\mathbf{M}_i^k = \sigma(W_k \mathbf{k}_i) \odot \sigma\left(W_q^k \mathbf{Q}\right) \tag{1}$$

$$\mathbf{M}_i^v = \sigma(W_v \mathbf{v}_i) \odot \sigma\left(W_q^v \mathbf{Q}\right) \tag{2}$$

where W_k, W_v, W_q^k, and W_q^v are embedding matrices, $\sigma(\cdot)$ denotes an exponential linear unit (ELU), and $\odot$ represents elementwise multiplication. Next, we transform the query-key representations $\mathbf{M}_i^k$ into $\mathbf{M}_i'^k$ with an embedding layer

$$\mathbf{M}_i'^k = \delta\left(W_m^k \mathbf{M}_i^k\right) \tag{3}$$

where W_m^k is an embedding matrix and $\delta(\cdot)$ denotes a rectified linear unit (ReLU). Then, both spatial and channelwise bilinear attention distributions are obtained according to the transformed query-key representations $\mathbf{M}_i'^k$.

Specifically, we perform spatial attention via another embedding layer to obtain the spatial attention weights a^s, and normalize it with a softmax layer to get the spatial attention distribution A^s

$$a_i^s = W_s \mathbf{M}_i'^k \tag{4}$$

$$A^s = \text{softmax}\left(\mathbf{a}^s\right) \tag{5}$$

where W_s is an embedding matrix. Meanwhile, the channelwise attention is achieved via a squeeze and excitation operation [28], in which the squeeze performs a global average pooling to obtain a global channelwise descriptor $\bar{\mathbf{M}}$, and the excitation produces the channelwise attention distribution A^c with the gating mechanism via a fully connected (FC) layer using sigmoid as the activation function

$$\bar{\mathbf{M}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{M}_i'^k \tag{6}$$

$$\mathbf{a}^c = W_e \bar{\mathbf{M}} \tag{7}$$

$$A^c = \text{sigmoid}\left(\mathbf{a^c}\right) \tag{8}$$

where W_e is an embedding matrix. Finally, we generate the m-linear-attended features $\hat{f}$ by computing the combination of spatial and channelwise weighted bilinear features

$$\begin{aligned} \hat{f} &= \text{Attention}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) \\ &= \text{Concat}\left(W_c\left(A^c \odot \mathbf{M_i^v}\right), \sum_{i=1}^{N} A^s \mathbf{M}_i^v\right) \end{aligned} \tag{9}$$

where W_c is an embedding matrix and $\odot$ represents elementwise multiplication.

2) M-Linear Encoder Block: To extend to higher order intramodal feature interactions, our visual encoder is composed of a stack of m-linear attention blocks. As shown in Fig. 1, each block has been marked with a gray box. Specifically, taking the calculation of the nth block as an example, we take the previous output m-linear attended feature $\hat{f}^{(n-1)}$ as input query $\mathbf{Q}$, along with the current keys $\mathbf{K}^{(n-1)} = \{\mathbf{k}_i^{(n-1)}\}_{i=1}^N$ and values $\mathbf{V}^{(n-1)} = \{\mathbf{v}_i^{(n-1)}\}_{i=1}^N$, which is updated by concatenating with the new attended feature $\hat{f}^{(n)}$ via residual connection and layer normalization

$$\hat{f}^{(n)} = \text{Attention}\left(\mathbf{K}^{(n-1)}, \mathbf{V}^{(n-1)}, \hat{f}^{(n-1)}\right) \tag{10}$$

$$\mathbf{k}_i^{(n)} = \text{LN}(\delta(W_n^k[\hat{f}^{(n)}, \mathbf{k}_i^{(n-1)}]) + \mathbf{k}_i^{(n-1)}) \tag{11}$$

$$\mathbf{v}_i^{(n)} = \text{LN}(\delta(W_n^v[\hat{f}^{(n)}, \mathbf{v}_i^{(n-1)}]) + \mathbf{v}_i^{(n-1)}) \tag{12}$$

where W_n^v and W_n^k are embedding matrices, $\mathbf{v_i^{(n)}}$ and $\mathbf{k_i^{(n)}}$ represents the ith key and value, $\delta(\cdot)$ denotes an ReLU unit, and LN denotes layer normalization. We repeat this process four times to achieve eighth-order feature interactions.

B. Sentence Decoder RNN

In Fig. 1, the sentence decoder uses LSTMs to take the combination of different order attended region-level visual features $\hat{f}^{(1)}, \ldots, \hat{f}^{(n+1)}$ and the global image feature f_{global} as input $\tilde{f}$, and generates the report word by word

$$\tilde{f} = W_f[\hat{f}^{(0)}, \hat{f}^{(1)}, \ldots, \hat{f}^{(n+1)}] \tag{13}$$

$$h_t = \text{LSTM}(\tilde{f}; w_t, h_{t-1}, c_{t-1}) \tag{14}$$

where W_f is an embedding matrix, $\hat{f}^{(0)} = f_{\text{global}}$, w_t is the current input word, h_{t-1} is the previous LSTM hidden state, and c_{t-1} is the previous context vector.

1) M-Linear Decoder Block: To exploit high-order intermodal interactions between visual features and semantic features, we also introduce the m-linear attention into the decoder. The block has been marked with a gray box in Fig. 1. Specifically, we take the output hidden state h_t of the LSTM as the input query $\mathbf{Q}$ of an m-linear attention block, and the final output attended visual feature $g^{(n+1)}$ from the visual encoder as keys $\mathbf{K}$ and values $\mathbf{V}$

$$\begin{aligned}\hat{f}_d &= \text{Attention}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) \\ &= \text{Attention}\big(g^{(n+1)}, g^{(n+1)}, h_t\big).\end{aligned} \tag{15}$$

After getting the m-linear-attended features $\hat{f}_d$, we compute the current context vector c_t with a residual connection and a gated linear unit (GLU), followed by an FC layer:

$$c_t = W_L(\epsilon(h_t + W_d(h_t + \hat{f}_d))) \tag{16}$$

where W_L and W_d are embedding matrices and $\epsilon(\cdot)$ denotes a GLU unit. Finally, the context vector c_t is sent into a softmax layer to predict the next word w_{t+1}

$$w_{t+1} = \text{Softmax}(c_t). \tag{17}$$

C. Repetition Penalty Module

Different from traditional image captioning tasks, which aim at generating a single sentence, our task requires the generation of paragraphed reports, which consist of hundreds of words. Inevitably, this leads to an increase in generation difficulty, thus resulting in repeated terms. Furthermore, the employed metrics (i.e., BLEU-1), which focus on the matching of words, also aggravate this problem. Though our proposed m-linear attention alleviates the problem to some extent by improving the capacity of feature extraction, we also propose an adaptive repetition penalty module that constrains the probabilities of words resulting in repeated trigrams to decrease the probability of repeated terms generated in the reports.

Although Melas-Kyriazi et al. [10] also discovered this problem and proposed to add a penalty when generating words, its penalty is a preset static hyperparameter. Since the more the occurrence of repetition terms, the less likely the sentence is to be correct and readable, to address the above-mentioned problems, we propose to assign an exponential weight to the generated repeated words. Specifically, we update the log-probability of the output word by subtracting a value proportional to the number of times the trigram has been generated

$$p_w = p_w - (1 - e^{-n_w}) \tag{18}$$

where p_w is the log-probability of the word w, and n_w is the number of times that the trigram has generated the word w. We employ this update mechanism during our greedy search process in SCST to generate more diverse paragraphs, thus avoiding the repetition problem.

D. Reinforcement Learning for Comprehensive Optimization

The RL algorithm commonly used in existing medical report generation [5], [6] is the SCST algorithm [19], which directly optimizes the automatic natural language generation metrics. Specifically, SCST adopts a policy gradient method to optimize a nondifferentiable metric such as CIDEr. To normalize the reward and reduce the variance during training, SCST utilizes the REINFORCE algorithm with a baseline, which is obtained from the inference reward by greedy search. The goal is to minimize the negative expected reward. The final gradient of the optimization object is

$$\frac{\partial L(\theta)}{\partial s_t} = (r(w^s) - r(\hat{w}))\,\nabla_\theta \log p_\theta(w^s|x) \tag{19}$$

where $w^s = (w_1^s, \ldots, w_T^s)$ is a Monte-Carlo sample from our generation model p_θ, $r(w^s)$ is the current reward, and $r(\hat{w})$ is the reward obtained by the inference algorithm.

Since most of the existing RL-based methods take CIDEr [29] as the reward, it inevitably leads to the failure of optimization on other metrics. Furthermore, the existing works did not reasonably explain the use of the CIDEr-based reward; although the success of the CIDEr-based reward was proven in the task of image captioning, it has not been verified in the area of medical report generation. Intuitively, we consider verifying the effectiveness of all the metrics and utilizing them to achieve an overall optimization. Since each metric focuses on different aspects, and some metrics are more important in our task (e.g., compared with BLEU, METEOR, and ROUGE-L consider not only precision but also recall), simply applying the same weight to each metric is not reasonable. We thus attach different weights to different metrics to achieve a weighted optimization. Though Liu et al. [30] also proposed to adopt a mixture of the metrics, they did not give any explanation for the combination nor provide any solution to find the suitable weights. Therefore, besides proposing a different weighted hybrid reward (DWHR), we also develop a solution to search for the local optimal weights.

In our work, the seven most frequently used natural language generation metrics are adopted as the rewards, and a DWHR is proposed to achieve an overall promotion. The gradient of our optimization object is

$$\frac{\partial L(\theta)}{\partial s_t} = \left(\sum_{i=0}^{7} \lambda_i r(w^s) - \sum_{i=0}^{7} \lambda_i r(\hat{w})\right) \nabla_\theta \log p_\theta(w^s|x) \tag{20}$$

where λ_i is the weight of the corresponding metric.

Generally, the search for the optimal weight is time consuming. The complexity of grid search is determined by the size of the searched hyperparameter space. Given the hyperparameter space is Θ, each Θ_i corresponds to a hyperparameter (i.e., the weight of an evaluation metric here), $i = 1, 2, \ldots, m$, i.e., having m hyperparameters (weights) in total, and each hyperparameter has n alternative values, the size of the hyperparameter space is

$$\Theta = \Theta_1 \times \Theta_2 \cdots \times \Theta_m. \tag{21}$$

When conducting the grid search, each possible value combination in the hyperparameter space has to be tried, so the model needs to be trained and evaluated the same number of times as the size of the hyperparameter space. So the computational complexity of grid search is exponential to the number of candidate values n, i.e., $O(n^m)$.

To reduce the complexity of the search method, we propose a greedy search-based solution, called the local optimal weight search algorithm, whose pseudocodes are shown in Algorithm 1 and the detailed processes and complexity analysis are as follows. Given the same settings as grid search (i.e., the number of weights for the evaluation metrics is m and each weight has n alternative values), the local optimal weight search algorithm first initializes the values of all weights to be 1, and obtains the corresponding performance, so the complexity for the initialization is $O(1)$.

Algorithm 1 Local Optimal Weight Search

```
Input: $A$: a set of weights of all the metrics
  $m$: the number of hyperparameter
  $n$: the number of candidate values for hyperparameter
  $Score$: model of compute the NLP metric scores
Output: $A$: a set of local optimal weight values of all the
  metrics
  Initialize all $A_i = 1$
  insert $Score(A)$ to $Map(key = A)$
  while There exists at least one $A_i$ whose local optimal value
  has not been searched do
    for each $i \in [1, m]$ and the local optimal value of $A_i$ has
    not been searched do
      $A_i \leftarrow A_i + 1$
      if $key = A$ does not exit in the $Map$ then
        $S_i \leftarrow Score(A)$
        Insert $S_i$ to $Map(key = A)$
      else
        $S_i \leftarrow Map(key = A)$
      end if
    end for
    find out most influenced $A_i$ based on $S$
    set other element (except the fixed ones) of $A$ to be 1
    set $A_i \leftarrow 1$, $S'_1 \leftarrow Map(key = A)$
    set $A_i \leftarrow 2$, $S'_2 \leftarrow Map(key = A)$
    for  each $j \in [3, n]$ do
      $A_i \leftarrow j$
      if $key = A$ does not exit in the $Map$ then
        $S'_j \leftarrow Score(A)$
        Insert $S'_j$ to $Map(key = A)$
      else
        $S'_j \leftarrow Map(key = A)$
      end if
    end for
    find out the local optimal value of $A_i$ based on $S$
    fix the value of $A_i$ to the optimal one in the remaining
    iterations
  end while
```

Then we conduct iterative operations to add 1 to the value of solely a weight each time and also obtain the corresponding performance; after m times of iterations, we select the metric with the highest performance improvement after its weight is added by 1 as the *most influential metric*; so the complexity of selecting the most influential metric at this step (i.e., the first for loop in Algorithm 1) is $O(m)$. Furthermore, using A_i to denote the weight of the most influential metric, we aim to find the local optimal value of A_i by increasing the value of A_i one by one from 3 (because the performances of $A_i = 1, 2$ have already been obtained) to n; therefore, the complexity of finding the optimal value for A_i (i.e., the second for loop in Algorithm 1) is $O(n - 2)$. Consequently, the complexity of all operations for finding the local optimal value for the first weight is $O(m + (n - 2))$.

After that, we fix the local optimal value for A_i and then repeat the above operations to find the local optimal value for the most influential weight among the remaining $m - 1$ weights. Here, it is easy to derive that the complexity of finding the most influential weight among $m - 1$ weights is $O(m - 1)$ while the complexity of determining its corresponding local optimal value is still $O(n - 2)$. So the complexity of all operations for finding the second local optimal weight value is $O((m - 1) + (n - 2))$.

Consequently, it is also easy to derive that the complexities for finding the third, fourth, $\ldots$, and the last local optimal weight values are $O((m - 2) + (n - 2))$, $O((m - 3) + (n - 2)), \ldots,$ and $O(1 + (n - 2))$, respectively. Finally, by adding the above results together, we have the final complexity of the proposed local optimal weight search algorithm as follows, which is thus linear to the number of candidate values n:

$$O(1) + O(m + (n-2)) + O((m-1) + (n-2)) + O(m-2 + (n-2)) + \cdots + O(1 + (n-2)) \tag{22}$$

$$= O\left(1 + \sum_{m}^{1} K + m(n-2)\right). \tag{23}$$

As previously stated, the complexity of grid search is $O(n^m)$ while that of our proposed algorithm is $O(1 + \sum_m^1 K + m(n-2))$; since there are seven metrics in our work, we have $m = 7$, the resulting complexity of grid search is $O(n^7)$ (exponential) while that of our proposed algorithm is $O(15 + 7n)$ (linear). Consequently, by proposing the local optimal weight search algorithm, we are able to reduce the complexity of weight search from exponential to linear.

IV. Experiments

A. Datasets

To evaluate the performance of our proposed HReMRG-MR, extensive experiments have been conducted on the following two public CXR image datasets.

1) IU X-Ray: IU X-Ray [31] is a set of CXR images with paired diagnostic reports collected by Indiana University. It contains 7470 medical images and 3955 corresponding reports, which mainly consist of two sections: *Impression* and *Findings*. As some images or reports are missing, we filter out the unpaired data and use the remaining 6222 images including both frontal and LL views, and 3111 paired reports. We take *Impression* and *Findings* as our generation object and preprocess the texts by converting all words to lowercase. After that, we tokenize the reports, resulting in 2068 unique words in total. As most words occur rarely, we exclude tokens appearing less than five times and get 776 tokens.

2) MIMIC-CXR: MIMIC-CXR [32] is the largest publicly available dataset of chest medical images with free-text radiology reports, which consists of 377 110 images along with 227 827 free-text reports. Similarly, we remove the unpaired data and obtain 76 724 remaining pairs of images combined of posteroanterior (PA) or anteroposterior (AP) view and LL view with their reports, which also consist of both *Impression* and *Findings*. We then apply the same preprocessing and keep tokens with more than five occurrences, ending up with 2991 tokens overall.

Finally, for both datasets, we randomly partition them by patients into training, validation, and testing sets with a ratio of 7:1:2 and make sure that there is no overlap between the sets. Words excluded are replaced by the token of "UNK"

B. Evaluation Metrics

To evaluate the performance of our proposed model, we use several different automatic language generation metrics including BLEU [33], METEOR [34], ROUGE-L [35], and

TABLE I

EXPERIMENTAL RESULTS OF HREMRG-MR AND THE STATE-OF-THE-ART BASELINES ON IU X-RAY (TOP PART) AND MIMIC-CXR (BOTTOM PART). ALL RESULTS ARE OBTAINED BY OUR REIMPLEMENTATION. THE BEST RESULTS ARE BOLD AND THE SECOND BEST ONES ARE UNDERLINED

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	top-down [22]	0.279	0.178	0.119	0.079	0.206	0.144	0.334	1.339
	MRMA [3]	0.382	0.252	0.173	0.120	0.325	0.163	0.309	1.724
	RTMIC [5]	0.3356	0.2236	0.15	0.1003	0.2779	0.1664	0.3371	1.5909
	X-LAN [23]	0.3782	0.2636	0.1858	0.1314	0.4508	0.1735	0.3490	1.8004
	HReMRG-MR	**0.4399**	**0.3059**	**0.2139**	**0.1490**	**0.5239**	**0.1971**	**0.3810**	**2.2107**
MIMIC-CXR	top-down [22]	0.233	0.159	0.119	0.093	0.359	0.134	0.319	1.416
	MRMA [3]	0.361	0.244	0.182	0.141	0.324	0.157	0.330	1.739
	RTMIC [5]	0.3654	0.2448	0.1772	0.134	0.3575	0.1521	0.3208	1.7518
	X-LAN [23]	0.362	0.2579	0.1801	0.126	0.365	0.169	0.3402	1.8002
	HReMRG-MR	**0.4806**	**0.3431**	**0.2555**	**0.1921**	**0.3715**	**0.2070**	**0.3802**	**2.2301**

CIDEr [29]. In particular, BLEU is a popular machine translation evaluation metric, which aims at measuring the ratio of correct matches of n-grams between generated sequences and the ground truth. To address some defects in BLEU, METEOR further considers the accuracy and recall based on the whole corpus. ROUGE-L is a metric for summary evaluation by measuring the longest common sequence between two sequences based on recall. Differently, CIDEr was designed especially for image description, which is most suitable for our task. It computes the term frequency-inverse document frequency (TF-IDF) weight of n-grams to obtain the similarity between the candidate sequences and the reference sequences.

C. Baselines

We compare our method with the four state-of-the-art image captioning and medical report generation models: 1) our reimplementation of the top-down model [22], which is a classic encoder–decoder-based model for image captioning employing a conventional attention mechanism that calculates the contribution of regional features to the generated texts; 2) MRMA [3], an encoder–decoder-based model specially designed for medical report generation, where reports are generated sentence by sentence in a recurrent way to get long paragraphs; 3) RTMIC, which is a state-of-the-art medical report generation method based on RL; it enhances the capacity of the generation model by RL, and increases the clinical accuracy by a transformer; and 4) X-LAN [23], which is an image captioning model employing x-linear attention and improving it by RL with CIDEr as the reward; since image captioning is similar to our task to some extent, we also take this model as our baseline. For our implemented methods, we use the same visual features and train/val/test split on both datasets.

D. Implementation Details

We use ResNet-101 pretrained on ImageNet to extract the region-level features, which are from the last convolutional layer. As both views of medical images are sent into the model simultaneously, two original 2048-D region features are concatenated to get a 4096-D vector, which is later transformed into the visual embedding with the size of 1024. Next, four stacks of m-linear attention blocks are used to explore the high-order intramodal interactions. In decoding, we set the hidden layer size and word embedding dimension to 1024. As there exists a fixed upper limit to the length of the generated report, we set the max generation length of IU X-Ray and MIMIC-CXR to 114 and 184, respectively.

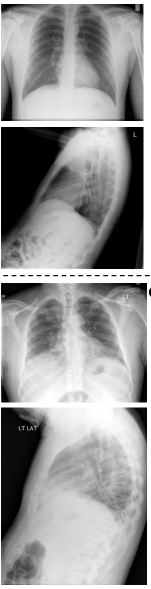
In training, we first pretrain the model with cross-entropy loss for 60 epochs with a batch size of 8 using eight NVIDIA RTX 2080 GPUs. Setting the base learning rate to 0.0001, we utilize the Noam decay strategy with 10 000 warmup steps and use Adam [36] as the optimizer. In the RL stage, we set the DWHR obtained via our search solution as our training reward. The search space k is set to 5. We set the base learning rate to 0.00001 and decay using Cosine Annealing with a period of 15 epochs. We also set the maximum iteration to 60 epochs and the batch size to 3. We use beam search with a beam size of 2 for training.

E. Main Results

Table I shows the experimental results of the proposed HReMRG-MR and four baselines in terms of seven natural language generation metrics and the sum of them, where all baselines are reimplemented by ourselves. In addition, Fig. 3 exhibits some examples of reports generated by these models.

Generally, our proposed HReMRG-MR greatly outperforms all state-of-the-art baselines in all metrics in Table I, and Fig. 3 shows that HReMRG-MR also generates more coherent and accurate reports than the baselines (with more matching terms). Specifically, in Table I, we underline the best performing baselines on each metric. For the four BLEU-n word matching metrics, on IU X-Ray, HReMRG-MR is 15.15%, 16.05%, 28.63%, and 13.39% higher than the best-performing baseline, respectively; on MIMIC-CXR, HReMRG-MR is better than the best-performing baseline by 31.53%, 33.04%, 40.38%, and 36.24%, respectively. The improvements are due to the following reasons: 1) RTMIC conducts RL using a single-metric-reward, so it cannot achieve the overall optimization of the model; 2) as shown in Fig. 3, MRMA does not perform well in long text generation because it does not use RL; and 3) X-LAN cannot fuse channel attention and spatial attention well to identify suspicious regions of images because it only uses low-order single-dimensional attention. For CIDEr, METEOR and ROUGE metrics, X-LAN performs best among the baselines, but HReMRG-MR still outperforms it by 16.22%, 13.6%, and 9.17%, respectively, on IU X-Ray, while by 1.78%, 22.49%, and 11.76% on MIMIC-CXR. This is because X-LAN's low-order single-dimensional attention cannot explore the high-order interactions of visual and semantic features, resulting in low accuracy in multimodal reasoning. HReMRG-MR also achieved the best results in overall score, which are 22.79% and 23.88% higher than the second place on IU X-Ray and MIMIC-CXR datasets, proving its significant advantages.

Furthermore, some additional observations are as follows. MRMA generally outperforms top-down in almost all metrics because it adopts an attention-guided recurrent text generation method to better generate long texts. Similarly, RTMIC also achieves better results than top-down due to the use of RL. MRMA and RTMIC obtain similar scores in terms of overall performance, proving that the attention-guided long text



Ground Truth	Top Down	MRMA	RTMIC	XLAN	Ours (HReMRG-MR)
no acute **cardiopulmonary** findings, **cardiomediastinal silhouette pulmonary vasculature** are within normal limits, lungs are **clear**, no **pneumothorax** or **pleural effusion**, no acute **osseous** findings.	no acute **cardiopulmonary** process, the lungs are clear without focal consolidation. no **pleural effusion** or **pneumothorax** is seen.	no acute **cardiopulmonary** abnormality, the heart size is normal. Lungs are clear, no focal airspace consolidation **pleural effusion** or **pneumothorax**, no acute bony abnormality.	no acute **cardiopulmonary** abnormality . heart size and mediastinal contours appear within normal limits **pulmonary vasculature** is normal . lungs effusion no focal consolidation . no **pneumothorax** or **pleural effusion**	no evidence of acute **cardiopulmonary** findings, the heart size and mediastinal contours are within normal limits. The lungs are **clear** of focal airspace disease. No **pleural effusion** or **pneumothorax**, there are no acute bony findings.	no acute **cardiopulmonary** findings, the **cardiomediastinal silhouette** and **pulmonary vasculature** are within normal limits. the lungs are **clear**. There is no focal consolidation. no **pneumothorax** or **pleural effusion**, no acute bony findings.
Ground Truth negative for acute **cardiopulmonary** disease, heart size borderline **enlarged,** stable **cardiomediastinal silhouette**, no **pneumothorax** or large **pleural effusion**, no focal **airspace** disease, low **lung volumes**, **nodular densities** consistent with chronic **granulomatous** disease, mild **degenerative disease** of the **thoracic spine**	**Top Down** no acute **cardiopulmonary** process the lungs are clear without focal consolidation no **pleural effusion** or **pneumothorax** is seen .the cardiac and **Mediastinal silhouettes** are stable	**MRMA** no acute **cardiopulmonary** abnormality . the heart size and pulmonary vascularity appear within normal limits . Pulmonary vasculature is normal. There is no **pneumothorax** or **pleural effusion**	**RTMIC** no acute **cardiopulmonary** abnormality . heart size and **mediastinal** contours appear within normal limits . pulmonary vasculature is normal . lungs effusion no focal consolidation . no **pneumothorax** or **pleural effusion**	**XLAN** low **lung volumes** . no evidence of active disease . the heart size and pulmonary vascularity appear within normal limits . the lungs are clear of focal **airspace** disease . there is no **pleural effusion** or **pneumothorax**. there is no acute bony abnormality	**Ours (HReMRG-MR)** low **lung volumes** . no acute **cardiopulmonary** disease . the heart size and pulmonary xxxx are within normal limits . the lungs are clear . there is no focal **airspace** consolidation . no **pleural effusion** or **pneumothorax** . **degenerative** changes of the **thoracic spine.**

Fig. 3. Examples of reports generated by Top Down, MRMA, RTMIC, XLAN, and HReMRG-MR. Matching medical keywords are bold.

TABLE II

ABLATION STUDIES, WHERE HREMRG INDICATES THE REINFORCED MEDICAL REPORT GENERATION MODEL WITH DWHR, X INDICATES X-LINEAR ATTENTION, M INDICATES M-LINEAR ATTENTION, AND R INDICATES REPETITION PENALTY. THE BEST RESULTS ARE BOLD AND THE SECOND-BEST ONES ARE UNDERLINED

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	HReMRG	0.3491	0.2245	0.1551	0.1139	0.4489	0.1511	0.2925	1.7351
	HReMRG-X	**0.4399**	0.3081	0.2135	0.1466	0.4378	0.1942	0.374	2.1141
	HReMRG-XR	0.4374	**0.3096**	**0.2177**	**0.1544**	0.4589	0.1926	0.3733	2.1439
	HReMRG-M	0.4322	0.3034	0.2113	0.1462	0.4929	0.1945	0.3795	2.1599
	HReMRG-MR	**0.4399**	0.3059	0.2139	0.1490	**0.5239**	**0.1971**	**0.3810**	**2.2107**
MIMIC-CXR	HReMRG	0.3084	0.2131	0.1611	0.1252	0.3164	0.1461	0.3383	1.6086
	HReMRG-X	**0.4849**	0.3429	0.2548	0.1897	0.3423	**0.2092**	0.3742	2.1980
	HReMRG-XR	0.4760	0.3385	0.2533	0.1908	0.3541	0.2059	**0.3910**	2.2096
	HReMRG-M	0.4821	0.3428	**0.2558**	0.1920	0.3529	0.2081	0.3752	2.2089
	HReMRG-MR	0.4806	**0.3431**	0.2555	**0.1921**	**0.3715**	0.2070	0.3802	**2.2301**

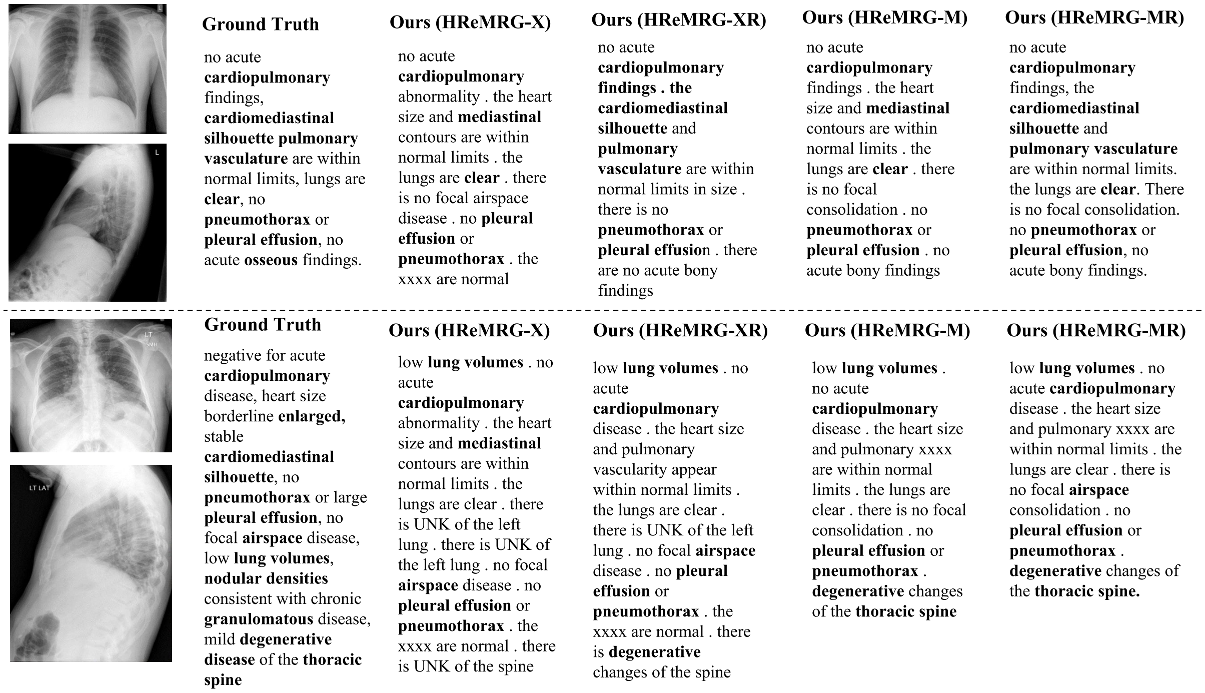
generation solution and RL are both effective in improving the quality of generated reports. As for X-LAN, the finding that it outperforms RTMIC proves the effectiveness of x-linear attention because they both use RL; furthermore, it also outperforms MRMA, proving that the utilization of a suitable attention mechanism and the employment of RL is more effective than the recurrent text generation method. Therefore, our work is proposed based on these observations.

Our HReMRG-MR uses RL methods for long text generation, and designs hybrid reward to further improve the overall performance of the model, then introduces a high-order multidimensional attention mechanism to help the model perform multimodal reasoning. Thanks to these technical contributions, we make up for the shortcomings of existing methods and obtain superior results than existing models. It is worth mentioning that these deficiencies we have solved generally exist in most medical report generation methods. The incomplete optimization makes it difficult to generate long texts, and the lack of high-level multidimensional interaction results in a lack of understanding of features, which all cause performance limitations of existing methods. With the improvement of these technical issues, our method is believed to be able to achieve better results in most other clinical practice situations.

F. Ablation Studies

In this section, we report on a series of ablation experiments, using HReMRG-MR and four incrementally implemented intermediate models to show the effectiveness of using the proposed m-linear attention modules and the adaptive repetition penalty in our work. Specifically: 1) we first use only DWHR to construct a reinforced medical report generation model, called HReMRG, which does not use an attention mechanism nor repetition penalty; 2) we then incorporate the x-linear attention (an SOTA attention mechanism) and m-linear attention mechanisms into HReMRG to obtain HReMRG-X and HReMRG-M, respectively; and 3) finally, we incorporate the proposed adaptive repetition penalty into HReMRG-X and HReMRG-M, respectively, and obtain the HReMRG-XR and HReMRG-MR models.

Table II shows the results of the above five models. By comparing the results of HReMRG to those of HReMRG-M, we find that, by additionally introducing the m-linear attention modules into the model, HReMRG-M greatly outperforms HReMRG in both the overall score and each of the seven evaluation metrics. Consequently, this greatly supports our theoretical argument that m-linear attention can help the model better learn high-order and multidimensional features and also the intermodal dependencies between text and image, which thus improves the model's performances. Furthermore, by introducing an additional repetition penalty mechanism on HReMRG-M, HReMRG-MR achieved the highest overall scores, indicating that the proposed repetition penalty can greatly improve the fluency of generating reports.



Ground Truth	Ours (HReMRG-X)	Ours (HReMRG-XR)	Ours (HReMRG-M)	Ours (HReMRG-MR)
no acute **cardiopulmonary** findings, **cardiomediastinal silhouette pulmonary vasculature** are within normal limits, lungs are **clear**, no **pneumothorax** or **pleural effusion**, no acute **osseous** findings.	no acute **cardiopulmonary** abnormality . the heart size and **mediastinal** contours are within normal limits . the lungs are **clear** . there is no focal airspace disease . no **pleural effusion** or **pneumothorax** . the xxxx are normal	no acute **cardiopulmonary findings . the cardiomediastinal silhouette** and **pulmonary vasculature** are within normal limits in size . there is no **pneumothorax** or **pleural effusio**n . there are no acute bony findings	no acute **cardiopulmonary** findings . the heart size and **mediastinal** contours are within normal limits . the lungs are **clear** . there is no focal consolidation . no **pneumothorax** or **pleural effusion** . no acute bony findings	no acute **cardiopulmonary** findings, the **cardiomediastinal silhouette** and **pulmonary vasculature** are within normal limits. the lungs are **clear**. There is no focal consolidation. no **pneumothorax** or **pleural effusion**, no acute bony findings.
Ground Truth	Ours (HReMRG-X)	Ours (HReMRG-XR)	Ours (HReMRG-M)	Ours (HReMRG-MR)
negative for acute **cardiopulmonary** disease, heart size borderline **enlarged,** stable **cardiomediastinal silhouette**, no **pneumothorax** or large **pleural effusion**, no focal **airspace** disease, low **lung volumes**, **nodular densities** consistent with chronic **granulomatous** disease, mild **degenerative disease** of the **thoracic spine**	low **lung volumes** . no acute **cardiopulmonary** abnormality . the heart size and **mediastinal** contours are within normal limits . the lungs are clear . there is UNK of the left lung . there is UNK of the left lung . no focal **airspace** disease . no **pleural effusion** or **pneumothorax** . the xxxx are normal . there is UNK of the spine	low **lung volumes** . no acute **cardiopulmonary** disease . the heart size and pulmonary vascularity appear within normal limits . the lungs are clear . there is UNK of the left lung . no focal **airspace** disease . no **pleural effusion** or **pneumothorax** . the xxxx are normal . there is **degenerative** changes of the spine	low **lung volumes** . no acute **cardiopulmonary** disease . the heart size and pulmonary xxxx are within normal limits . the lungs are clear . there is no focal consolidation . no **pleural effusion** or **pneumothorax** . **degenerative** changes of the **thoracic spine**	low **lung volumes** . no acute **cardiopulmonary** disease . the heart size and pulmonary xxxx are within normal limits . the lungs are clear . there is no focal **airspace** consolidation . no **pleural effusion** or **pneumothorax** . **degenerative** changes of the **thoracic spine.**

Fig. 4. Examples of reports generated by HReMRG, HReMRG-X, HReMRG-XR, HReMRG-M, and HReMRG-MR. Matching medical keywords are bold.

The superiority of m-linear attention modules can be demonstrated by comparing the results of HReMRG-M (resp., HReMRG-MR) with those of HReMRG-X (resp., HReMRG-XR), where a state-of-the-art attention baseline, x-linear attention, is used. As shown in Table II HReMRG-M (resp., HReMRG-MR) outperforms HReMRG-X (resp., HReMRG-XR) in terms of most evaluation metrics on both datasets. Specifically, the overall scores of HReMRG-M (resp., HReMRG-MR) are better than those of HReMRG-X (resp., HReMRG-XR) on both datasets, which thus proves that m-linear attention is generally superior to x-linear in the medical report generation tasks. Furthermore, HReMRG-M and HReMRG-MR greatly outperform HReMRG-X and HReMRG-XR, respectively, in CIDEr on both datasets (e.g., on IU X-Ray, HReMRG-MR has a 14.16% improvement with respect to HReMRG-XR, while HReMRG-M has a 12.59% improvement with respect to HReMRG-X); since CIDEr incorporates the TF-IDF of words in the entire evaluation corpus and gives higher weight to informative phrases, the great performances of HReMRG-M and HReMRG-MR in CIDEr proves that, comparing to x-linear attention, m-linear attention can help the model generate informative report containing more meaningful medical terms. Similar superior performances of m-linear attention are also observed in terms of METEOR and ROUGE-L, proving that using m-linear attention can result in better recall and sentence-level matching. All the above findings demonstrate the superiority of m-linear attention with respect to the SOTA attention baseline, x-linear attention, and also support our argument on the advantages of m-linear attention.

The visualized examples in Fig. 4 also exhibit the superiority of m-linear attention: First, the descriptions generated by HReMRG-M and HReMRG-MR contain significantly fewer meaningless words (e.g., “UNK, xxxx”) than those generated by HReMRG-X and HReMRG-XR. Second, HReMRG-M (resp., HReMRG-MR) matches more keywords (especially those for abnormalities) than HReMRG-X (resp., HReMRG-XR). Third, HReMRG-M (resp., HReMRG-MR) can describe the suspicious areas more accurately, e.g., on the MIMIC dataset, HReMRG-M (resp., HReMRG-MR) can accurately state “thoracic spine” while HReMRG-X (resp., HReMRG-XR) only states “spine.”

The superiority of repetition penalty modules can be demonstrated by comparing the results of HReMRG-X (resp., HReMRG-M) with those of HReMRG-XR (resp., HReMRG-MR). In Table II, HReMRG-XR (resp., HReMRG-MR) achieves much better performances than HReMRG-X (resp., HreMRG-M) in terms of most metrics, especially on BLEU-4 and CIDEr. This is because, with the repetition penalty, the generated texts tend to have more diverse phrases and avoid repetitive words with little information. Thus, the values of BLEU-4, which evaluate the matching of four-gram phrases, increase significantly, and the values of CIDEr, which measures the matching of informative phrases, also rise greatly. Similarly, the example report of HReMRG-MR in Fig. 3 is also more accurate and readable than that generated by HReMRG-M. These findings prove that with the proposed repetition penalty, medical report generation models can achieve much better performances. We thus employ m-linear attention and repetition penalty in our final model HReMRG-MR.

G. Comparison Between Single-Reward and Hybrid-Reward

Almost all existing RL-based medical report generation methods employ CIDEr as the reward, which was proven to be the best reward in image captioning, while this has not been demonstrated in our task. Table III shows the experimental results of our HReMRG-X model using different rewards on the basis of X-LAN. Assuming that all natural language evaluation metrics are equally important, we take the sum of them as the final evaluation metric, denoted Score. Taking each metric as the reward separately, we can find that it usually has the highest value when the corresponding metric

TABLE III

RESULTS OF USING DIFFERENT REWARDS IN RL, WHERE SWHR DENOTES SAME WEIGHTED HYBRID REWARD, AND DWHR DENOTES DIFFERENT WEIGHTED HYBRID REWARD. THE BEST RESULTS ARE BOLD AND THE SECOND-BEST RESULTS ARE UNDERLINED

Dataset	Reward	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	BLEU-1	0.4420	0.2968	0.2038	0.1397	0.3423	0.1887	0.3597	1.9732
	BLEU-2	0.4225	0.3039	0.2128	0.1491	0.4292	0.1874	0.3633	2.0683
	BLEU-3	0.4034	0.2926	0.2100	**0.1510**	0.4348	0.1832	0.3630	2.0436
	BLEU-4	0.3762	0.2772	0.2041	0.1488	0.3944	0.1670	0.3241	1.8917
	CIDEr	0.3782	0.2636	0.1858	0.1314	**0.4508**	0.1735	0.3490	1.8004
	METEOR	0.4377	0.2850	0.2000	0.1400	0.2021	0.1915	0.3442	1.9499
	ROUGE-L	0.3632	0.2717	0.1931	0.1357	0.4147	0.1729	0.3716	1.9311
	SWHR	0.4160	0.2952	0.2101	0.1497	0.4507	0.1870	0.3650	2.0741
	DWHR	**0.4399**	**0.3081**	**0.2135**	0.1466	0.4378	**0.1942**	**0.374**	**2.1141**
MIMIC-CXR	BLEU-1	**0.5002**	0.3205	0.2239	0.1630	0.2674	0.2072	0.3632	2.0454
	BLEU-2	0.4801	0.3415	0.2523	0.1870	0.3476	0.2066	0.1698	2.1850
	BLEU-3	0.4713	0.3388	0.258	0.1942	0.3456	0.2036	0.1303	2.1800
	BLEU-4	0.4613	0.3373	**0.2601**	**0.2033**	0.3494	0.1998	0.381	2.1922
	CIDEr	0.362	0.2579	0.1801	0.126	0.365	0.169	0.3402	1.8002
	METEOR	0.3972	0.2835	0.2148	0.1658	0.1166	**0.2183**	0.3596	1.7558
	ROUGE-L	0.4150	0.2924	0.2185	0.1648	0.3323	0.1822	**0.3927**	1.9979
	SWHR	0.4699	0.3339	0.2504	0.1910	**0.3753**	0.2043	0.3707	2.1956
	DWHR	0.4849	**0.3429**	0.2548	0.1897	0.3423	0.2092	0.3742	**2.1980**

TABLE IV

PROCESSES AND RESULTS OF THE PROPOSED LOCAL OPTIMAL WEIGHT SEARCH ALGORITHM USING THREE METRICS TO FORM THE HYBRID REWARD ON IU X-RAY. THE FINAL SCORE CORRESPONDING TO THE LOCAL OPTIMAL WEIGHTS IS BOLD, AND THE UNDERLINED SCORES ARE THE ONES USED TO COMPARE AND SELECT THE FIRST MOST INFLUENTIAL METRIC

Dataset	Reward	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	0:0:0:1:1:1:0	0.4210	0.2968	0.2074	0.1457	0.4114	0.1868	0.3624	2.0314
	0:0:0:2:1:1:0	0.4054	0.2936	0.2135	0.1551	0.3737	0.1832	0.3534	1.9779
	0:0:0:1:2:1:0	0.3972	0.2871	0.2068	0.1483	0.3968	0.1835	0.3599	1.9795
	0:0:0:1:1:2:0	0.4202	0.2958	0.2108	0.1517	0.448	0.1888	0.3679	**2.0833**
	0:0:0:1:1:3:0	0.4048	0.2871	0.2025	0.1428	0.447	0.1841	0.3594	2.0277
	0:0:0:2:1:2:0	0.4202	0.2958	0.2108	0.1517	0.448	0.1879	0.3647	2.0792
	0:0:0:1:2:2:0	0.4028	0.2917	0.2106	0.1513	0.4339	0.1828	0.3564	2.0296
	0:0:0:3:1:2:0	0.4235	0.2997	0.2124	0.1498	0.3959	0.1892	0.3638	2.0343
	0:0:0:1:2:2:0	0.4028	0.2917	0.2106	0.1513	0.4339	0.1828	0.3564	2.0296
	0:0:0:1:3:2:0	0.4215	0.2921	0.2081	0.1488	0.4335	0.1853	0.3651	2.0545

is used as a reward. Consequently, we believe the mixture of them will achieve improved results on all the metrics, which is verified in Table III, where SWHR achieves higher overall scores compared with single-reward-based models. Furthermore, since different metrics act in different roles, we are motivated to search for the local optimal weight for the hybrid reward. As shown in Table III, the results of the DWHR further improve the performances in almost all the metrics and achieve the highest final scores. Examples in Fig. 3 also proves that our HReMRG-X (with the weighted hybrid reward) generates a more accurate report (i.e., has more correctly matched terms) than that generated by XLAN (using only CIDEr as the reward). Therefore, we employ the DWHR in our research.

H. Effectiveness of the Proposed Search Solution

To show the effectiveness of our proposed search solution, additional experiments are conducted to quantitatively compare the complexities of our proposed method with the conventional grid search. Since the complexity of grid search is exponential, due to page limit, it is impossible to list all the combinations (i.e., BLEU-4, CIDEr, and METEOR) for all the seven metrics, so three of them are selected here for demonstration and the number of alternative values for the weights is limited to 3. Table IV shows the processes and results of our proposed local optimal weight search algorithm and Table V shows the processes and results of the grid search.

Specifically, in Table IV, we first initialize all weights to 1, and then add 1 to the weights of BLEU-4, CIDEr, and METEOR one by one. By comparing the result of comprehensive metric Score, we can find that the most influential metric in this case is METEOR. Subsequently, we increase the weight of METEOR by 1 and also obtain the corresponding Score. Then by comparing the values of Score of the three combinations (i.e., the ones underlined in Table IV), we find 2 is the local optimal weight of METEOR. After that, we fix the weight of METEOR to 2, and repeat the above operations to find the most influential indicators in BLEU-4 and CIDEr one by one according to Algorithm 1 to find their corresponding local optimal weights. Finally, we have the local optimal weights of BLEU-4, CIDEr, and METEOR to be 1:1:2.

By comparing the results and processes of these two tables, we can find that the proposed method achieves the same optimal search results along with the original search, while greatly reducing the search complexity from $O(n^3)$ (here $n = 3$) to $O(10)$. Please also note that our algorithm is an approximation solution that finds the local optimal results which are not guaranteed to always be equal to the global optimal results of grid search; however, an approximation solution does not mean "bad solution," it is a trade-off between accuracy and efficiency, i.e., sacrificing tiny accuracy in exchange for significant improvements in efficiency (the complexity is reduced from exponential to linear using our proposed local optimal weight search algorithm) and making the intractable work become possible. In addition, we also find that the best result obtained using three metrics to construct the weighted hybrid reward is worse than that of using seven metrics, which also demonstrates the necessity of adopting

TABLE V

PROCESSES AND RESULTS OF THE CONVENTIONAL GRID SEARCH ALGORITHM USING THREE METRICS TO FORM THE HYBRID REWARD ON IU X-RAY. THE SCORE CORRESPONDING TO THE GLOBAL OPTIMAL WEIGHTS ARE BOLD

Dataset	Reward	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	0:0:0:1:1:1:0	0.4210	0.2968	0.2074	0.1457	0.4114	0.1868	0.3624	2.0314
	0:0:0:2:1:1:0	0.4054	0.2936	0.2135	0.1551	0.3737	0.1832	0.3534	1.9779
	0:0:0:1:2:1:0	0.3972	0.2871	0.2068	0.1483	0.3968	0.1835	0.3599	1.9795
	0:0:0:1:1:2:0	0.4202	0.2958	0.2108	0.1517	0.448	0.1888	0.3679	**2.0833**
	0:0:0:2:1:2:0	0.4202	0.2958	0.2108	0.1517	0.448	0.1879	0.3647	2.0792
	0:0:0:1:2:2:0	0.4028	0.2917	0.2106	0.1513	0.4339	0.1828	0.3564	2.0296
	0:0:0:3:1:2:0	0.4235	0.2997	0.2124	0.1498	0.3959	0.1892	0.3638	2.0343
	0:0:0:1:3:2:0	0.4215	0.2921	0.2081	0.1488	0.4335	0.1853	0.3651	2.0545
	0:0:0:1:3:1:0	0.3828	0.2781	0.2002	0.1433	0.4573	0.1771	0.3597	1.9984
	0:0:0:3:1:1:0	0.4058	0.2956	0.2139	0.1538	0.3695	0.1785	0.3212	1.9383
	0:0:0:2:2:1:0	0.3952	0.292	0.2116	0.1511	0.3993	0.1792	0.3613	1.9897
	0:0:0:2:3:1:0	0.3988	0.2924	0.211	0.1508	0.3694	0.1784	0.3499	1.9507
	0:0:0:1:1:3:0	0.4048	0.2871	0.2025	0.1428	0.447	0.1841	0.3594	2.0277
	0:0:0:3:2:1:0	0.4002	0.2916	0.2102	0.1499	0.4215	0.1818	0.3664	2.0216
	0:0:0:3:3:1:0	0.3858	0.2773	0.2009	0.1442	0.4196	0.1753	0.3573	1.9604
	0:0:0:2:1:3:0	0.4269	0.3019	0.2131	0.1507	0.4287	0.1894	0.3616	2.0723
	0:0:0:1:2:3:0	0.4233	0.2950	0.2062	0.1443	0.4428	0.1895	0.3652	2.0662
	0:0:0:3:1:3:0	0.4014	0.286	0.2059	0.1483	0.474	0.1828	0.3584	2.0567
	0:0:0:2:2:3:0	0.3134	0.2971	0.2134	0.1534	0.3633	0.1836	0.3333	1.9575
	0:0:0:2:3:3:0	0.4048	0.2842	0.2019	0.1446	0.4427	0.1843	0.3591	2.0216
	0:0:0:1:3:3:0	0.3983	0.2854	0.2022	0.1443	0.452	0.1805	0.3573	2.02
	0:0:0:3:2:3:0	0.4215	0.296	0.2067	0.145	0.4289	0.1881	0.3607	2.047
	0:0:0:3:3:3:0	0.3998	0.2827	0.2077	0.1436	0.4195	0.1824	0.3569	1.9857
	0:0:0:2:2:2:0	0.4043	0.2998	0.217	0.157	0.4298	0.1829	0.3564	2.0472
	0:0:0:2:3:2:0	0.4037	0.2902	0.2069	0.1496	0.4246	0.1818	0.3564	2.0132
	0:0:0:3:2:2:0	0.3958	0.2844	0.2065	0.1490	0.4749	0.1766	0.3589	2.046
	0:0:0:3:3:2:0	0.41	0.2963	0.21	0.1479	0.3906	0.1853	0.3587	1.9988

TABLE VI

RESULTS OF COMPARING THE PROPOSED M-LINEAR ATTENTION WITH THE STATE-OF-THE-ART ATTENTION MECHANISMS. THE BEST RESULTS ARE BOLD AND THE SECOND-BEST ONES ARE UNDERLINED

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE-L	Score
IU X-Ray	HReMRG	0.3491	0.2245	0.1551	0.1139	0.4489	0.1511	0.2925	1.7351
	HReMRG-Att2in [19]	0.4071	0.2755	0.1925	0.1329	0.3636	0.1884	0.3652	1.9252
	HReMRG-AdaAtt [24]	0.3534	0.2453	0.174	0.1213	0.3217	0.1658	0.3472	1.7287
	HReMRG-AdaAttMO [24]	0.3836	0.2479	0.1761	0.1289	0.3154	0.1683	0.3309	1.7511
	HReMRG-X [23]	**0.4399**	**0.3081**	**0.2135**	**0.1466**	0.4378	0.1942	0.374	2.1141
	HReMRG-M	0.4322	0.3034	0.2113	0.1462	**0.4929**	**0.1945**	**0.3795**	**2.1599**
MIMIC-CXR	HReMRG	0.3084	0.2131	0.1611	0.1252	0.3164	0.1461	0.3383	1.6086
	HReMRG-Att2in2	0.3664	0.2666	0.2045	0.1585	0.3401	0.1705	**0.3875**	1.8941
	HReMRG-AdaAtt	0.4173	0.3004	0.2287	0.1734	0.3101	0.1834	0.3838	1.9972
	HReMRG-AdaAttMO	0.2886	0.1936	0.1441	0.1125	**0.4008**	0.1486	0.3319	1.6201
	HReMRG-X	**0.4849**	**0.3429**	0.2548	0.1897	0.3423	**0.2092**	0.3742	2.1980
	HReMRG-M	0.4821	0.3428	**0.2558**	**0.1920**	0.3529	0.2081	0.3752	**2.2089**

all the evaluation metrics as the reward in the medical report generation task.

I. Effectiveness of Different Attention

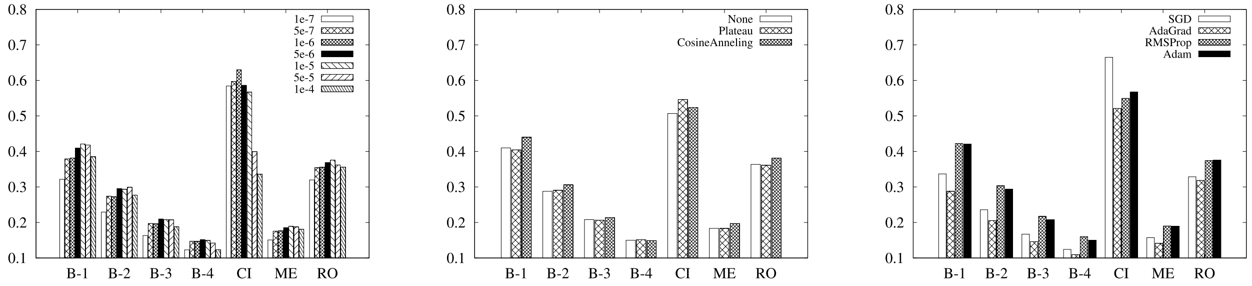
The superiority of m-linear can be further demonstrated by comparing it with other state-of-the-art attention mechanisms besides x-linear attention. Therefore, in this section, we use the reinforced medical report generation model with solely the DWHR, i.e., HReMRG, as the backbone, and incorporate it with the state-of-the-art attention baselines, Att2in [19], AdaAtt [24], and AdaAttMO [24], respectively, to obtain the models HReMRG-Att2in, HReMRG-AdaAtt, and HReMRG-AdaAttMO. Table VI shows the additional results of these models as well as those of HReMRG, HReMRG-X, and HReMRG-M.

As shown in Table VI, using m-linear attention (i.e., HReMRG-M) always not only achieves the highest overall score but also achieves the best or second best results in nearly all metrics. The superiority of m-linear comes from the following reason: Att2in2, AdaAtt, and its improved version AdaAttMO learn single-dimension and low-order feature interactions only, so the features learned by them are less accurate and comprehensive than those learned by m-linear attention with the multidimensional high-order attention mechanism. Consequently, the superiority of m-linear attention is further demonstrated, i.e., by fusing spatial and channelwise attention, stacking multiple blocks to learn high-order features, and learning intermodal dependencies between text and images, m-linear can overcome the problem of the existing attention baselines and achieve better performances in medical report generation tasks.

J. Influence of Different Hyperparameters

In this section, We further evaluate the influence of learning rate, strategies of learning rate, and optimizers. To obtain the optimal parameters of training, we train HReMRG-MR model on IU X-Ray for 100 epochs with a batch size of two.

To obtain the optimized initial learning rate, we compare the performances using different initial learning rates ranging from $1e^{-4}$ to $1e^{-7}$. Here, we employ no learning decay strategies and use the Adam optimizer. According to Fig. 5, the overall performance of $1e^{-5}$ outperforms the others, so we apply it as the initial learning rate.



(a)

(b)

(c)

Fig. 5. Results of varying hyperparameters, where B, CI, ME, and RO denote BLEU, CIDEr, METEOR, and ROUGE-L, respectively. (a) Different initial learning rate. (b) Different learning rate decay strategies. (c) Different optimizers.

After setting the initial learning rate to be $1e^{-5}$, we further investigate the influence of the strategy of learning rate decay when training with RL by comparing the performances using *Plateau* strategy, *Cosine Annealing* strategy, and that of not using learning rate decay strategies. Experimental results are presented in Fig. 5. For *Plateau* strategy, the learning rate is reduced by a factor of 0.8, once learning stagnates for 3 epochs. For the Cosine Annealing strategy, the cosine function is used as the learning rate annealing function, the maximum number of iterations epoch is set to be 10, and the minimum learning rate is $4e^{-8}$. From Fig. 5, we can observe that the performance of *Plateau* is about the same as *None*, and most evaluation values using Cosine Annealing are better than those using the Plateau decay strategy. Therefore, we apply the Cosine Annealing learning rate decay strategy during RL in our work.

Furthermore, we investigate the influence of optimizers by comparing SGD, AdaGrad, RMSProp, and Adam. Here, we set the initial learning rate to be $1e^{-5}$ and the learning rate decay strategy to be Cosine Annealing. In Fig. 5, we can observe that Adam and RMSProp perform best on the whole. Since the CIDEr score of Adam outperforms that of RMSProp a lot and CIDEr pays more attention to the important terms, we choose Adam in our experiments.

V. Discussion and Future Work

A. Application Scope of Proposed Algorithm

In this section, we further discuss the application scope of our work in clinical practices to show that our work can achieve superior performances not only in the scenarios tested in our experiments but also in the majority of situations of real-world applications of medical report generation models. First, all the baselines used in our work are the state-of-the-art methods in medical image generation tasks; so by showing that our work is superior to all these state-of-the-art baselines on the two widely used benchmark datasets in our experimental studies, it is reasonable to say that our work will also be superior to other existing medical report generation methods. Second, two benchmark datasets, IU X-Ray and MIMIC-CXR are used in our experimental studies, these two datasets are highly representative datasets and widely used in all the state-of-the-art medical report generation works [3], [5], because their samples are obtained from the real-world clinical practices, and they have diverse dataset sizes, i.e., thousands of samples for IU X-Ray and hundreds of thousands of samples for MIMIC-CXR, which make them able to cover the majority of the real-world application of medical report generation models (indeed, the medical report generation model is designed to reduce the workload of radiologists in medium or large medical institutions by generating the report automatically, where the volumes of medical images are normally between thousands and hundreds of thousands). Therefore, it is widely believed by researches working on medical report generation that, if the proposed new model can outperform the SOTA baselines on these two datasets, they are guaranteed to be able to also achieve superior performances in (maybe not all but) the majority scenarios of real-world applications of medical report generation models. Since the superiority of our work on these two benchmark datasets has been shown and proved in the experimental studies, it is reasonable to say that the proposed method can also have similar superiority in most other real-world clinical situations. Finally, we have clearly stated and analyzed the problems of the existing SOTA medical report generation methods, and also analyzed the technical difficulties in solving this problem and how our proposed model deal with these technical difficulties and overcome all the problems in Section I. Therefore, due to the inherent technical advantages, it is natural for our work to achieve superior performances than the existing SOTA methods. Overall, with all the above analysis, it is reasonable to say that our proposed work has fantastic application scope and great value in clinical practices.

B. Social Impact for Proposed Algorithm

This model can be widely used in medical report generation to effectively reduce the workload of doctors and improve the efficiency and accuracy of report generation. Clinically, this judgment on medical images is usually performed by radiologists. A radiologist is often required to provide numerous reports of relevant medical findings to support his judgment. However, this kind of interpretation of medical images requires a lot of clinical experience. Due to the current shortage of clinical radiologists, especially the shortage of experienced experts, and because the interpretation of medical images only relies on the professional skills and experience of doctors, there are subjective analysis biases that easily lead to miscalculations. In addition, due to the large number of patients and the shortage of experienced radiologists, a radiologist may perform dozens or even hundreds of medical imaging examinations every day and then write corresponding reports, which puts a huge workload on radiologists and affects the efficiency and accuracy of their film reading. Since the global epidemic of the COVID-19, medical imaging techniques, including computed tomography (CT) or CXR, have been largely used to facilitate the diagnosis. Due to the complexity of the condition, the large number of patients, and the serious shortage of medical resources, the automatic generation of medical reports has attracted more attention, which has a significant positive effect on helping the rapid diagnosis of COVID-19 [13]. Therefore, it turns out that being able to automatically generate high-quality medical imaging reports is of great research interest. This greatly reduces the workload of doctors and saves valuable time and money for patients' treatment.

C. Limitations and Future Work

Similar to the existing medical report generation solutions, we directly use the medical images as the inputs of our model; since the abnormal regions normally only take up a small portion of the medical images, even with the attention mechanism, it is still challenging for the model to accurately focus its learning on the most important abnormal areas. Therefore, to further improve the model's performances, an interesting future research direction is to first rank the importance of different medical images [37] or different regions within the given medical image [11], and then assign different priorities or weights for the images or regions according to their importance during the model's learning process.

Both our methods and the SOTA baselines train the medical report generation model from scratch. However, recent studies have proved that using large-scale datasets that are relevant to the learning task to pretrain the model is beneficial. Since the existing works have proved that the semantic information in the related area is helpful for learning better image captioning [12] or medical image generation [38] models, inspired by the recent advances of introducing large-scale multimodal pretraining models [39], a promising future research direction to improve our work is to first conduct multimodal pretraining on the medical related datasets and then use the resulting pretrained model in the subsequent medical reports generation model learning.

Finally, although the reward of our proposed model has taken into account multiple evaluation measurements, all of them focus on evaluating the quality of the generated report solely. However, in the context of medical report generation, we should consider not only the quality of the text but also whether the matching between the text and the image is accurate. Therefore, similar to [40] that develops a new evaluation metric to measure the matching qualities between natural images and the generated image captions, future research works can be conducted to develop a new image-report matching quality metric to better evaluate the intermodal matching quality and also use it as an additional reward in our RL process to enhance the model's performances.

VI. Conclusion

We have proposed a hybrid reinforced medical report generation method, HReMRG-MR, which used a DWHR to optimize the generated medical reports, and a search solution was developed to obtain the local optimal weights for the hybrid reward. We have also proposed the m-linear attention, which stacked bilinear pooling operations to explore high-order feature interactions for intramodal and intermodal reasoning in medical report generation. An adaptive repetition penalty was finally proposed to generate more readable and coherent reports. We have conducted extensive experiments on two publicly available datasets, which have demonstrated the superior performances of our proposed HReMRG-MR in medical report generation tasks.

References

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.

[2] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 2577–2586.

[3] Y. Xue et al., "Multimodal recurrent model with attention for automated radiology report generation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2018, pp. 457–466.

[4] Y. Keneshloo, T. Shi, N. Ramakrishnan, and C. K. Reddy, "Deep reinforcement learning for sequence-to-sequence models," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2469–2489, Jul. 2020, doi: 10.1109/TNNLS.2019.2929141.

[5] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Proc. Int. Workshop Mach. Learn. Med. Ima.*, Oct. 2019, pp. 673–680.

[6] G. Liu et al., "Clinically accurate chest X-ray report generation," in *Proc. Mach. Learn. Heal. Conf.*, Apr. 2019, pp. 249–269.

[7] M. Yang, W. Huang, W. Tu, Q. Qu, Y. Shen, and K. Lei, "Multitask learning and reinforcement learning for personalized dialog generation: An empirical study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 49–62, Jan. 2021, doi: 10.1109/TNNLS.2020.2975035.

[8] A. Chaturvedi and U. Garain, "Mimic and fool: A task-agnostic adversarial attack," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1801–1808, Apr. 2021, doi: 10.1109/TNNLS.2020.2984972.

[9] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019, doi: 10.1109/TNNLS.2018.2851077.

[10] L. Melas-Kyriazi, A. Rush, and G. Han, "Training for diversity in image paragraph captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 757–761.

[11] Z. Shao, J. Han, D. Marnerides, and K. Debattista, "Region-object relation-aware dense captioning via transformer," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 11, 2022, doi: 10.1109/TNNLS.2022.3152990.

[12] J. Zhang, Z. Fang, H. Sun, and Z. Wang, "Adaptive semantic-enhanced transformer for image captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 29, 2022, doi: 10.1109/TNNLS.2022.3185320.

[13] G. Liu et al., "Medical-VLBERT: Medical visual language BERT for COVID-19 CT report generation with alternate learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3786–3797, Sep. 2021, doi: 10.1109/TNNLS.2021.3099165.

[14] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13748–13757.

[15] I. Najdenkoska, X. Zhen, M. Worring, and L. Shao, "Uncertainty-aware report generation for chest X-rays by variational topic inference," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102603, doi: 10.1016/j.media.2022.102603.

[16] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1439–1449.

[17] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 1537–1547.

[18] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6570–6580.

[19] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.

[20] Z. Xu et al., "ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution," *Neurocomputing*, vol. 500, pp. 177–190, Aug. 2022, doi: 10.1016/j.neucom.2022.05.053.

[21] L. Sun, W. Wang, J. Li, and J. Lin, "Study on medical image report generation based on improved encoding-decoding method," in *Proc. Int. Conf. Intell. Comput.*, Jul. 2019, pp. 686–696.

[22] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[23] Y. Pan, T. Yao, Y. Li, and T. Mei, “X-linear attention networks for image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10968–10977.

[24] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3242–3250.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[27] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” in *Proc. 32nd Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 1571–1581.

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[29] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[30] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Improved image captioning via policy gradient optimization of SPIDEr,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 873–881.

[31] D. Demner-Fushman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.

[32] A. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci. Data*, vol. 6, no. 1, pp. 317–325, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[34] A. Lavie and A. Agarwal, “Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 65–72.

[35] C.-Y. Lin and E. Hovy, “Manual and automatic evaluation of summaries,” in *Proc. ACL Workshop Autom. Summarization*, Jun. 2002, pp. 74–81.

[36] D. P. Kingma and J. Ba, “ADAM: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 13–28.

[37] F. Liu, S. Ge, and X. Wu, “Competence-based multimodal curriculum learning for medical report generation,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2021, pp. 3001–3012.

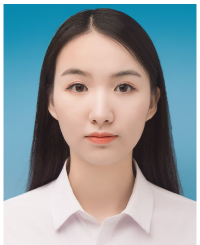
[38] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, “AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation,” in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, Mar. 2021, pp. 72–82.

[39] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. ICML*, Jul. 2021, pp. 8748–8763.

[40] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, and M. Bansal, “Fine-grained image captioning with CLIP reward,” in *Proc. Findings Assoc. Comput. Linguistics (NAACL)*, 2022, pp. 517–527.

Zhenghua Xu received the M.Phil. degree in computer science from The University of Melbourne, Melbourne, VIC, Australia, in 2012, and the D.Phil. degree in computer science from the University of Oxford, Oxford, U.K., in 2018.

From 2017 to 2018, he was a Research Associate with the Department of Computer Science, University of Oxford. He is currently a Professor with the Hebei University of Technology, Tianjin, China. He has authored more than 40 papers in top AI or database conferences and journals, e.g., NeurIPS, AAAI, IJCAI, IEEE TRANSACTIONS ON MEDICAL IMAGING (TMI), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), and *Medical Image Analysis*. His research interests include intelligent medical image analysis, deep learning, reinforcement learning, federated learning, and computer vision.

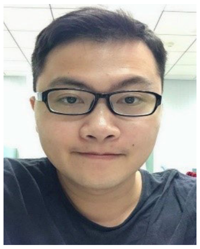


Wenting Xu received the B.Eng. degree in biomedical engineering from Southern Medical University, Guangzhou, China, in 2019. She is currently pursuing the master’s degree with the Hebei University of Technology, Tianjin, China.

Her research interests include medical report generation using deep learning methods.

Ruizhi Wang received the B.Eng. degree in automation from Qufu Normal University, Jining, China, in 2021. She is currently pursuing the master’s degree with the Hebei University of Technology, Tianjin, China.

Her research interests include medical report generation using deep learning methods.



Junyang Chen (Member, IEEE) received the Ph.D. degree in computer and information science from the University of Macau, Macau, China, in 2020.

He is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include graph neural networks, text mining, deep learning, and recommender systems.

Chang Qi received the B.Eng. degree in biomedical engineering from Southern Medical University, Guangzhou, China, in 2018, and the master’s degree in biomedical engineering from the Hebei University of Technology, Tianjin, China, in 2021.

Her research interests include medical image generation using deep learning.

Thomas Lukasiewicz is currently a Professor with the Institute of Logic and Computation, Vienna University of Technology (TU Vienna), Vienna, Austria, and the Department of Computer Science, University of Oxford, Oxford, U.K. He also holds an AXA Chair Grant on “Explainable Artificial Intelligence in Healthcare” and a Turing Fellowship at the Alan Turing Institute, London, U.K., which is the U.K.’s National Institute for Data Science and Artificial Intelligence. His research interests include artificial intelligence and machine learning.

Dr. Lukasiewicz has been a fellow of the European Association for Artificial Intelligence (EurAI) since 2020. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019.