scientific reports

Check for updates

OPEN

Multi-view contrastive learning and symptom extraction insights for medical report generation

Qi Bai[1,2,7], Xiaodi Zou[1,3,7], Ahmad Alhaskawi[1], Yanzhao Dong[1], Haiying Zhou[1], Sohaib Hasan Abdullah Ezzi[4], Vishnu Goutham Kota[5], Mohamed Hasan Hasan AbdullaAbdulla[5], Sahar Ahmed Abdalbary[6], Xianliang Hu[2]✉ & Hui Lu[1]✉

The task of generating medical reports automatically is of paramount importance in modern healthcare, offering a substantial reduction in the workload of radiologists and accelerating the processes of clinical diagnosis and treatment. Current challenges include handling limited sample sizes and interpreting intricate multi-modal and multi-view medical data. In order to improve the accuracy and efficiency for radiologists, we conducted this investigation. This study aims to present a novel methodology for medical report generation that leverages Multi-View Contrastive Learning (MVCL) applied to MRI data, combined with a Symptom Consultant (SC) for extracting medical insights, to improve the quality and efficiency of automated medical report generation. We introduce an advanced MVCL framework that maximizes the potential of multi-view MRI data to enhance visual feature extraction. Alongside, the SC component is employed to distill critical medical insights from symptom descriptions. These components are integrated within a transformer decoder architecture, which is then applied to the Deep Wrist dataset for model training and evaluation. Our experimental analysis on the Deep Wrist dataset reveals that our proposed integration of MVCL and SC significantly outperforms the baseline model in terms of accuracy and relevance of the generated medical reports. The results indicate that our approach is particularly effective in capturing and utilizing the complex information inherent in multi-modal and multi-view medical datasets. The combination of MVCL and SC constitutes a powerful approach to medical report generation, addressing the existing challenges in the field. The demonstrated superiority of our model over traditional methods holds promise for substantial improvements in clinical diagnosis and automated report generation, indicating a significant stride forward in medical technology

Keywords Medical report generation, Transformer, Multi-view contrastive learning, Symptom consultant

The concept of medical report generation holds immense significance in the healthcare landscape. These reports serve as critical documents that record and communicate essential patient information and play a pivotal role in enhancing patient care and clinical decision-making. Medical reports provide a comprehensive overview of a patient's medical history, current condition, diagnostic findings, and treatment plans[1]. This wealth of information is invaluable for healthcare providers as it enables them to make informed decisions, monitor a patient's progress, and ensure that the most appropriate care is delivered. Medical reports facilitate effective communication and collaboration among healthcare teams. They serve as a common reference point, ensuring that all team members are on the same page regarding a patient's status and treatment plan. This seamless flow of

[1]Department of Orthopedics, The First Affiliated Hospital, Zhejiang University, #79 Qingchun Road, Hangzhou, Zhejiang Province 310003, People's Republic of China. [2]School of Mathematical Sciences, Zhejiang University, # 866 Yuhangtang Road, Hangzhou, Zhejiang Province 310058, People's Republic of China. [3]Department of Orthopedics, The Second Affiliated Hospital of Zhejiang Chinese Medical University, Xinhua Hospital of Zhejiang Province, Hangzhou, Zhejiang Province 310003, People's Republic of China. [4]Department of Orthopedics, Third Xiangya Hospital, Central South University, #138 Tongzi Po RoadHunan Province, Changsha 410013, People's Republic of China. [5]Zhejiang University School of Medicine, #866 Yuhangtang Road, Hangzhou, Zhejiang Province 3100058, People's Republic of China. [6]Department of Orthopedic Physical Therapy, Faculty of Physical Therapy, Nahda University in Beni Suef, Beni Suef, Egypt. [7]Qi Bai and Xiaodi Zou these have contributed equally to this work. ✉email: xlhu@zju.edu.cn; huilu@zju.edu.cn

information ultimately contributes to better-coordinated care and improved patient outcomes. Beyond clinical applications, medical reports also have a crucial role in research, quality control, and legal documentation[2]. Researchers rely on the data within these reports to conduct studies, analyze trends, and advance medical knowledge. Healthcare institutions use them to monitor and improve the quality of care provided, identifying areas where enhancements are needed. In legal contexts, medical reports serve as essential documents in malpractice claims or insurance disputes[3]. They provide an objective record of a patient's condition and the care provided, aiding in the resolution of disputes and ensuring the protection of patients' rights.

TFCC injury refers to various forms of damage to the Triangular Fibrocartilage Complex within the wrist joint area, including tears, wear, or other forms of injury. This can lead to wrist joint pain, discomfort, and functional impairments for patients[4,5]. The significance of magnetic resonance imaging (MRI) in diagnosing TFCC injuries is crucial, as it provides high-resolution images that accurately depict the wrist joint structures, including the TFCC, therefore, aids healthcare professionals in precisely diagnosing the type and extent of the injury[6]. The importance of medical report generation lies in the process of documenting and conveying this crucial information to the medical team and patients. These reports underscore the critical nature of the diagnosis, providing a diagnostic foundation for patients and facilitating the development of effective treatment and rehabilitation plans. Thus, the significance of medical report generation supports effective medical practices and patient care.

However, writing high-quality medical reports can be challenging. Medical reports demand clarity, conciseness, comprehensiveness, and logical coherence. Even the composition of a report by a seasoned radiologist necessitates a profound degree of concentration and consumes a considerable amount of time. In recent years, due to the rapid advancement of image captioning technology[7–9], the pursuit of automated medical report generation has gained momentum, aiming to alleviate the burdensome workload of radiologists while concurrently reducing time, mistakes, and costs.

Similar to image caption, the majority of medical report generation models employ an encoding–decoding framework originally harnessed in machine translation[10]. The encoder extracts semantic features from medical images, while the decoder formulates reports aligned with the medical images, relying on the semantic features. In CNN-RNN model[11], the network in network[12] encoder undertakes multi-label classification of the input image, while the RNN decoder transmutes the encoder embeddings into descriptions of medical images. To address the limitation inherent in the CNN-RNN model, which struggles with the generation of long sequences, a hierarchical RNN framework[13] has been introduced. This hierarchical RNN incorporates two hierarchical of Long Short-Term Memory(LSTM) units[14] : the sentence LSTM is responsible for generating topics and the word LSTM composes sentences that harmonize with these topics. However, a hierarchical decoder produces redundant sentences due to the omission of considerations for contextual coherence and to overcome this issue, a multi-modal CNN-LSTM[15] is proposed, which combines the preceding sentence and the input image to produce an attention input. This input serves as a guiding principle for generating the present sentence, thereby ensuring the coherence between sentences.

Recently, self-attention mechanisms[16] have attracted a lot of attention in the field of natural language processing, and transformer architecture has showcased remarkable achievements in natural image captioning tasks[7,17]. These findings have enhanced a wide range of transformer-based medical report generation models. In RATCHET model[18], the DenseNet[19] assumes the role of the encoder to extract salient features from medical images, while transformer serves as the decoder for generating the pertinent medical report. Lovelace et al.[20] employs DenseNet to extract spatial image features, subsequently inputting them into the encoder of a transformer as a sequential input. Chen et al.[21] introduced the concept of a memory-driven transformer, wherein memory is integrated into the decoder of a transformer to archive information stemming from the generation process. Srinivasan et al.[22] employed a dual Transformer architecture in conjunction with cross-attention mechanisms to segregate and address both normal and abnormal cases. Lee et al.[23] integrates the local and global features, which have been extracted using the global–local visual extractor, into the decoder in the cross encoder-decoder transformer. Muksimova et al.[24] developed a hybrid architecture that integrates EfficientNetV2 for local feature extraction and Vision Transformers for capturing global dependencies, achieving superior performance in cervical cancer diagnosis. Tu et al.[25] proposed Med-PaLM M, a generalist biomedical AI system capable of processing multiple types of biomedical data including MRI and other medical imaging modalities using a single set of model weights. Schmidt et al.[26] proposed using ChatGPT to automatically simplify radiology reports, demonstrating that AI-generated summaries enhance patient comprehension of MRI findings while maintaining factual accuracy. However, they emphasize that such tools should complement, rather than replace, physician–patient discussions. Hamamci et al.[27] proposed CT2Rep, the first framework for automated radiology report generation for 3D medical imaging, specifically targeting chest CT volumes, leveraging a novel auto-regressive causal transformer and relational memory.

Contrastive learning aims to enhance feature representation by contrasting positive and negative pairs. Inspired by the recent achievements of contrastive learning, certain studies have incorporated it into report generation tasks[28,29]. Gao et al.[30] introduced a straightforward contrastive sentence embedding framework to produce more robust sentence embeddings. Liu et al.[31] proposed a contrastive attention model designed to compare the reference image with normal images to identify abnormal regions. Zhang et al.[32] pre-trained a medical image encoder using the bidirectional contrastive objective to improve visual representations. Wu et al.[33] enhanced visual and textual representations through contrastive learning techniques, facilitating the generation of medical reports using recursive networks.

Leveraging machine learning for medical report generation presents several formidable challenges and intricacies, particularly within the intricate and sensitive healthcare landscape[34]. When embarking on the utilization of machine learning for the generation of medical reports, a multitude of intricate aspects come into play, each carrying its own set of considerations and challenges. Foremost among these is the fundamental

concern of data quality. Machine learning models rely heavily on extensive volumes of high-quality medical data for training. However, the realm of medical data presents unique challenges, including issues related to its reliability, comprehensiveness, and precision. These challenges encompass erroneous diagnoses, incomplete medical records, and data gaps, all of which can have profound consequences on the reliability and accuracy of the generated medical reports[35,36]. Moreover, the medical field itself is highly specialized, encompassing numerous distinct domains and disease categories, each replete with its own specialized terminology, standards, and diagnostic methodologies. Crafting a universal machine learning model capable of seamlessly generating medical reports across these diverse domains proves to be a formidable and complex undertaking, often requiring extensive domain-specific expertise[37]. One of the critical considerations in the adoption of machine learning for medical report generation is the imperative for human–machine collaboration. Physicians and healthcare professionals must play an integral role in developing and validating machine-generated reports to ensure their precision, clinical relevance, and adherence to medical standards. Establishing an effective and efficient process for this collaboration can be intricate, demanding not only technical integration but also cultural and procedural alignment between the realms of healthcare and artificial intelligence[38,39]. Privacy and security concerns loom large when dealing with medical data. The sensitive nature of patient information necessitates stringent safeguards. Applying machine learning to the generation of medical reports inherently involves accessing and processing vast quantities of patient data, underscoring the need for meticulous data security measures and strict compliance with regulatory standards and privacy laws[40]. Furthermore, interpretability challenges come to the forefront. Machine learning models often operate as black-box systems, making it challenging for healthcare professionals to understand their decision-making processes. In the medical context, it is paramount that physicians can grasp the report generation process and be capable of elucidating and corroborating the conclusions produced by machines. Achieving interpretability stands as a substantial challenge, especially when clinical decisions are at stake[41,42]. Additionally, standardization and regulatory compliance pose significant hurdles. Medical reports must adhere to specific standards and regulatory frameworks to ensure their quality and comprehensibility. Developing machine learning models that can reliably generate reports in alignment with these standards may necessitate extensive engineering and standardization efforts, further complicating the adoption of this technology. Finally, legal and ethical considerations come into play. The adoption of machine learning for medical report generation introduces a host of complex issues, including questions of liability attribution, the legal standing of reports generated by machines, and the ethical responsibilities of healthcare providers. These matters demand meticulous examination within the broader context of legal and ethical frameworks, as the integration of artificial intelligence into healthcare raises new and uncharted ethical and legal territory[43–45].

The majority of the aforementioned models pertain to chest X-rays[46,47], and the data for each model includes one or two medical images, along with their corresponding reports and tags or observations extracted from these reports. In the actual diagnostic process, physicians depend on multi-modal and multi-view information in plane and sequence. Patients articulate symptoms correlated with their ailment, while the medical images acquired by physicians are typically more intricate and comprehensive in scope. Despite the commendable performance of current medical report generation models, they hinge upon substantial volumes of data for their efficacy.

In this paper, we propose a transformer-based framework to address the challenge of medical report generation with a limited number of samples, bearing in mind that the data is multi-modal and multi-view. In detail, multi-view contrastive learning (MVCL) and symptom consultant (SC) are used to enhance medical report generation. The MVCL leverages the multi-view MRI data, reducing the gap between similar views and accentuating distinctions among disparate views; this, in turn, enhances the quality of the visual features derived from the visual extractor. The SC functions as a way for extracting medical insights from symptoms. Subsequently, these insights are harnessed to guide the layer normalization process within the transformer decoder. As a result, MVCL and SC gleans medical information from multi-modal and multi-view diagnostic data, thereby facilitating the transformer ingenerating medical reports. Experimental results substantiate the validity and effectiveness of our approach. To summarize, the contributions of this paper are four-fold:

1. We propose a novel transformer-based framework to generate medical report using multi-modal and multi-view data.
2. We propose multi-view contrastive learning to improve the quality of the visual features extracted from multi-view MRI data.
3. We propose symptom consultant to apprehend the medical insights within symptoms and to provide guidance to the layer normalization process in decoder of transformer.
4. Comprehensive experiments have been conducted, and the outcomes demonstrate that our proposed models surpass the baseline models in performance.

Materials

Patients' data

From June 2020 to June 2023, 111 patients with wrist joint disorders who attended the First Affiliated Hospital of Zhejiang University School of Medicine were enrolled. TFCC injury was considered as the primary diagnosis during initial outpatient consultations. Subsequently, all patients underwent wrist joint 3.0T MRI examinations. Both outpatient medical records and MRI reports were comprehensive.

Description of the dataset

The Deep Wrist dataset utilized in this paper was gathered from patients diagnosed with TFCC conditions. Each sample within the dataset comprises the symptoms, multiple sagittal, coronal, and transverse MRI scans

of the wrist, and the corresponding medical report associated with images. The MRI images have been carefully chosen by experienced professionals, and all images that are not relevant to the disease have been excluded or filtered out. Furthermore, the medical reports have been authored by experts who have analyzed the MRIs as their basis. The complete dataset comprises a total of 111 samples, and each of these samples is composed of a triplet containing a symptom description, associated MRIs (from different views), and a corresponding medical report. An illustrative example within the dataset is presented in Fig. 1. From left to right, the images portray the coronal, sagittal, and transverse views of the wrist. Adjacently, the accompanying report delineates the symptoms and corresponding images on the right-hand side.

Data preparation

In medical report generation tasks, the quality of medical images plays a pivotal role in shaping the quality of the generated reports. We conducted preprocessing on the dataset to enhance image quality, encompassing resizing, cropping, and applying Contrast Limited Adaptive Histogram Equalization (CLAHE).

The original MRIs exhibit variations in size. To ensure uniformity in the input image dimensions, we initially employed a bilinear interpolation technique to resize the images to a standardized 512×512 size. Following the standardization of image dimensions, it became apparent that the disease-related regions in the MRIs are predominantly situated at the center of the image. This central region is encompassed by extraneous black background and certain parameter-related information from the MRIs. To enable the model to harness the information contained within the image more effectively, we maintain the central focus and apply a crop operation, resulting in an image size of 336×336. This cropping process eliminates superfluous information surrounding the MRIs and a substantial portion of the black background while preserving the principal content pertinent to report generation. Furthermore, it's noteworthy that MRIs often exhibit noise and an absence of contrast, necessitating enhancements to augment image quality and legibility, and to address this, contrast enhancement techniques are frequently employed to ameliorate MRI quality. Among these techniques, CLAHE[48] is a commonly utilized method for reducing noise and heightening contrast in medical images. Unlike traditional histogram equalization that processes the entire image uniformly, CLAHE divides the image into small regions and enhances the contrast in each region separately, which helps preserve important details while preventing over-amplification of noise in medical images. In the final stages of data preprocessing, CLAHE is utilized to enhance MRI quality. Within this procedure, the context area size is consistently set at 24×24, leading to the segmentation of each input image into 196 non-overlapping blocks. Moreover, a clipping limit of 0.03 is applied, effectively enhancing the contrast of the MRIs.

In summary, the preprocessing workflow initiates by standardizing the image to a uniform 512×512 size. Subsequently, a central cropping operation is performed, yielding a 336×336 image, thereby concentrating on significant regions. Finally, CLAHE is harnessed to augment image contrast, enhancing the visibility of anatomical structures. The preprocessed MRIs are shown in Fig. 2.

Methods

Architecture overview

The task of medical report generation involves the transformation of images into textual narratives. We maintain consistency with established methodologies[8,9,21] and employ a sequence-to-sequence model. The comprehensive structure of our proposed model is depicted in Fig. 3. The entirety of the model can be categorized into three distinct components: the visual extractor, the encoder, and the decoder. Input data comprises symptom descriptions and MRIs captured from various views, while the output consists of corresponding medical reports.

The visual extractor is a pre-trained convolutional neural network responsible for extracting visual feature sequences $X = \{x_1, x_2, \cdots, x_S\}, x_i \in R^d$ from the input MRIs. Each MRI corresponds to a distinct visual feature sequence X. Here, S represents the sequence length of the visual features, and d signifies the dimensionality of the visual feature vector. This process can be represented as follows:

Fig. 1. Example in dataset. The images displayed from left to right represent the coronal, sagittal, and transverse views. On the right side, the content consists of symptoms and medical report information.

Fig. 2. Preprocessed MRIs. The MRIs have undergone resizing, cropping, and Contrast-Limited Adaptive Histogram Equalization (CLAHE) preprocessing.

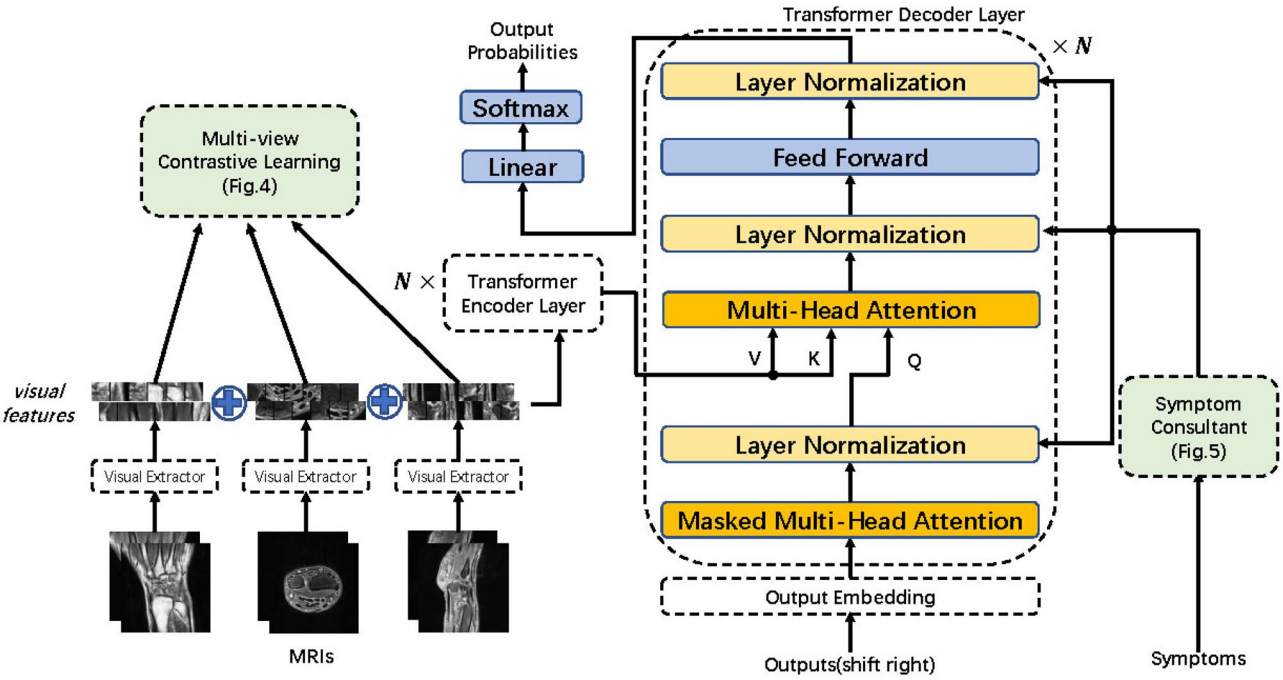

Fig. 3. The overview of our proposed model.

$$\{x_1, x_2, \cdots, x_S\} = f_v\,(\mathrm{MRI})$$

where x_i symbolizes the visual extractor. In this context, we employ ResNet101[49] as the visual extractor and integrate multi-view contrastive learning (MVCL) into the visual extraction process, which is introduced later.

The encoder utilized here mirrors the architecture of the Transformer[16] encoder. Its primary function is to transform the visual feature sequence X into a sequence of hidden states denoted as H:

$$H = \{h_1, h_2, \cdots, h_S\} = f_e\,(X)$$

with f_e representing the encoder, h_i is the hidden state corresponding to visual feature x_i.

The decoder leverages the identical architecture as the R2Gen[21] model, with the sole distinction being the integration of the Symptom Consultant (SC) to extract medical insights from the symptoms and provide guidance for the layer normalization within the decoder. The decoding procedure can be articulated as follows:

$$y_t = f_d\,(H, y_1, \cdots, y_{t-1}, f_s\,(\mathrm{D}))$$

where D signifies the symptom embedding, f_s represents the SC, f_d denotes the decoder, and y_i stands for the generated token. Meanwhile, $Y = \{y_1, y_2, \cdots, y_T\}$ represents the target textual sequence, T is the length of medical report.

Multi-view contrastive learning

Contrastive learning is a self-supervised learning approach that trains models to recognize similar and dissimilar pairs of samples by maximizing agreement between different views of the same data while minimizing agreement between views of different data. The MRIs within the Deep Wrist dataset encompass multi-views, including coronal, sagittal, and transverse, each consisting of numerous individual MRI scans. To extract a more comprehensive set of medical insights from these multi-view MRIs and to accentuate the distinctiveness of information derived from various views, we employ contrastive learning techniques to bolster the capabilities of the feature extractor. Specifically, our approach treats MRI scans from the same view as positive pairs while scans from different views are treated as negative pairs, enabling the model to learn distinctive features that capture the unique characteristics of each anatomical view.

The structure for multi-view contrastive learning is illustrated in Fig. 4. The input consists of visual feature sequences derived from different views. Randomly, we choose two latent representations from each view, denoted as $X_{c1}, X_{c2}, X_{s1}, X_{s2}, X_{t1}, X_{t2}$. Following two fully connected layers, six latent representations $l_{c1}, l_{c2}, l_{s1}, l_{s2}, l_{t1}, l_{t2}$ are produced. The computational procedure can be elucidated through the following equation:

$$l_k = W_2 f_r(W_1 X_k + b_1) + b_2.$$

In this equation, W and b denote the weights and biases of the fully connected layer, f_r represents the Rectified Linear Unit (ReLU) activation function, X_k signifies the visual feature, and l_k corresponds to the associated latent representation. Subsequently, the contrastive learning technique is used to minimize the disparity between representations derived from identical views while concurrently accentuating the distinctions among representations arising from diverse views. To achieve this objective, we employ a triplet loss function[50] that operates on anchor, positive, and negative samples. For any given anchor sample l_1 from a specific view, we define a positive sample l_2 from the same view and a negative sample l_3 from a different view. The triplet loss is formulated as:

$$L_m(l_1, l_2, l_3) = \max(d(l_1, l_2) - d(l_1, l_3) + \mathrm{m}, 0)$$

where $d(\cdot)$ is the distance function and m is a hyperparameter that enforces a minimum distance between positive and negative pairs. This loss function encourages the network to learn view-specific features by pulling representations of the same view closer while pushing representations of different views apart in the embedding space.

Symptom consultant

Differing from other datasets, the Deep Wrist dataset incorporates symptom descriptions. To enable more effective utilization of symptoms in guiding the generation of medical reports, we introduce the concept of symptom consultant module. SC extracts medical insights from symptom embeddings and employs them to guide the layer normalization process within the transformer. Layer normalization is a technique designed to normalize the inputs across the features. For a given input x, layer normalization computes the mean μ and standard deviation σ across the feature dimension, and the normalized output is then scaled and shifted using learnable parameters γ and β:

$$y = \gamma \cdot \frac{x - \mu}{\sigma} + \beta.$$

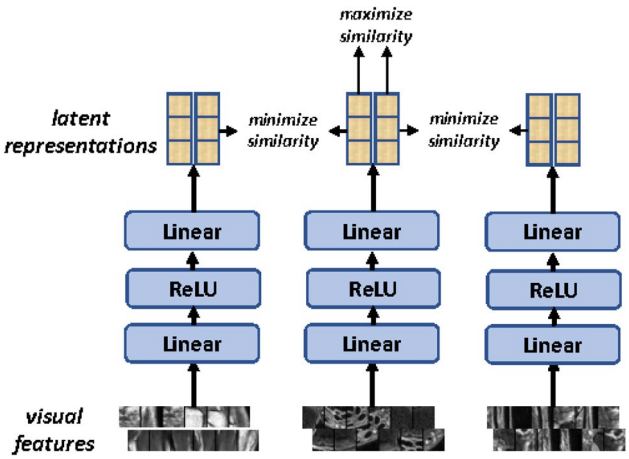

Fig. 4. Multi-view contrastive learning.

Instead of directly learning γ and β parameters in layer normalization, we propose to modulate these parameters based on the medical insights f_{sc} extracted from the data. Specifically, we compute the adaptive changes $\Delta\gamma$ and $\Delta\beta$ as functions of f_{sc}:

$$\Delta\gamma = W_\gamma f_{sc} + b_\gamma, \Delta\beta = W_\beta f_{sc} + b_\beta.$$

where W_γ, W_β are learnable weight matrices and b_γ, b_β are bias terms. The final layer normalization output is then computed as:

$$y = (\gamma + \Delta\gamma) \cdot \frac{x - \mu}{\sigma} + (\beta + \Delta\beta).$$

The configuration of the symptom consultant module is depicted in Fig. 5. We employ a multi-head self-attention mechanism to derive the attention weight matrix A_w and outcomes att from the symptom embeddings:

$$A_w = softmax\left(\frac{QK^T}{\sqrt{d}}\right), att = A_w V$$

where Q, K, V are the query, key and value obtained via three linear transformations, d is the dimension of symptom embeddings. Upon averaging the attention weights, we derive a weight vector. Then, the positions of the top k largest values in this vector are identified, and they are used as indices to extract the corresponding features f_{sc} from att. These extracted features constitute the medical insight for guiding the layer normalization, which is the same as R2Gen[21], thereby facilitating the report generation process.

Loss function

The loss function of the medical report generation model is composed of the summation of two distinct components. The first part is comprised of the multi-view contrastive loss function, and the second part is constituted by the report generation loss function.

The contrastive loss function[50] is defined as follows:

$$L_m(l_1, l_2, l_3) = \max(d(l_1, l_2) - d(l_1, l_3) + \mathrm{m}, 0)$$

wherein $d(\cdot, \cdot)$ represents a function that quantifies the dissimilarity between two latent representations, with l_1 and l_2 originating from the same views, and l_3 stemming from differing views. Additionally, m denotes a hyperparameter, signifying the minimum acceptable separation between distinct views. Each view engenders a contrastive loss in relation to the other two views, resulting in a total of six components within the loss function.

The loss function associated with report generation is the cross-entropy loss function, which measures the dissimilarity between the predicted word distribution and the ground truth distribution. Specifically, for a sequence of T tokens, the loss is computed as: $L_g(\theta) = -\sum_{i=1}^{T} \log(p_\theta(y_i|y_{1:i-1}))$.

where $y_i \in Y$ is the generated target, p_θ represents the probability of generating the current word, and θ encompasses the parameters within the entire generation model. This cross-entropy loss serves multiple crucial purposes in training the medical report generation model. By penalizing incorrect word predictions and rewarding accurate ones, it drives the model to not only assign higher probabilities to the correct medical terms, but also establish meaningful connections between visual features and their corresponding textual descriptions.

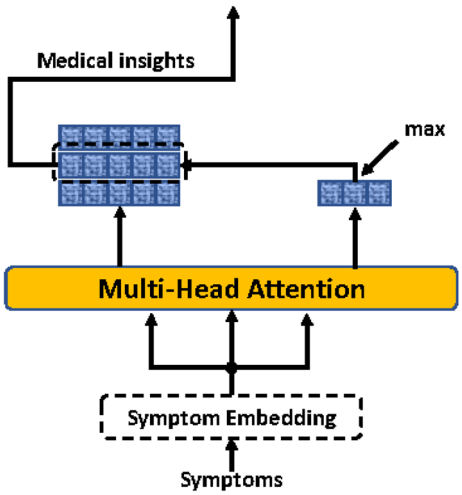

Fig. 5. Symptom consultant.

Through the training process, as this loss is backpropagated through the decoder, it simultaneously optimizes the model's language modeling capabilities and visual-textual alignment, enabling the generation of contextually appropriate medical terminology while maintaining proper grammatical structure in the output reports.

Experiment

Experiment settings

All experiments are conducted using the Deep Wrist dataset, which comprises 111 reports and 2366 MRIs. To make full use of the limited dataset while ensuring reliable evaluation, we partition the dataset into training, validation and testing sets with a 7:1:2 ratio. This split ratio is chosen to maximize the training samples while maintaining a sufficient validation set for model monitoring and testing set for final evaluation. During the training process, we monitor the model's performance on the validation set and implement early stopping strategy to prevent overfitting. Specifically, the training process is terminated when the model's performance on the validation set shows no improvement for 30 epochs, and the model weights that achieve the best performance on the validation set are saved as the final model. The test set is strictly reserved for the final performance evaluation. The embedding dimension for both reports and symptoms are set to 512, and the number of heads for all multi-head attention layers in the model is 8, with an output dimension of 512. The encoder and decoder each consist of three layers. In the MVCL module, two MRIs are randomly selected from each view to compute contrast, while in the SC module, the three locations with the highest attention values are chosen. Model training employs the ADAM optimizer, with a learning rate of $5e-5$ for the visual extractor and $1e-4$ for all other components. Finally, a random sampling strategy is employed for generating words from the obtained distribution.

To assess model performance, natural language generation metrics specifically relevant to medical report evaluation were employed. BLEU[51] assesses the precision and accuracy of generated reports through n-gram matching, which is crucial for capturing the standardized medical terminology and conventional expressions commonly used in radiology reports. BLEU-1, BLEU-2, BLEU-3, and BLEU-4 were utilized to evaluate different levels of phrase matching. Meteor[52] evaluates text generation quality by considering n-gram matching and incorporating syntactic and semantic information, which helps assess the medical accuracy and completeness of the generated reports.. Rouge-L[53] employs the longest common subsequence to measure the similarity between the generated and human reference reports, particularly effective in capturing the sequential nature of medical findings and diagnoses. CIDEr[54] evaluates the quality of generated reports by considering the consistency between reports and their lexical diversity, which is essential for maintaining both the standardization and necessary variation in medical reporting.

Results and analyses

To evaluate the performance of our proposed modules, MVCL and SC, we conducted ablation experiments using the Deep Wrist dataset, with the R2Gen model serving as the baseline. Additionally, we explored the impact of summing and concatenating multi-view visual features. The results are detailed in Table 1. "+SC" indicates the inclusion of the SC module, "+MVCL" signifies the integration of the MVCL module, and "Full" designates the inclusion of both modules. The "1" and "2" following MVCL and Full denote the approaches of summing and concatenating multi-view visual features, respectively.

The following phenomena are reflected in the experimental data. Firstly, the inclusion of the SC module results in a performance improvement surpassing the baseline, affirming the efficacy of the SC module. This observation underscores the role of medical insights extracted from symptom descriptions in enhancing report generation. Secondly, the model with the MVCL module performed comparably to the baseline, indicating that while MVCL provides alternative feature extraction pathways, its standalone impact on performance improvement is limited under current experimental conditions. Thirdly, the full model combining both modules demonstrate measurable performance gains, suggesting that the synergistic impact of both modules can be effectively leveraged when operating in conjunction. This outcome underscores the utility of utilizing multi-modal and multi-view data in report generation.

Moreover, we observed that the multi-view visual feature fusion approach influences the performance. A comparison of results using the summation method and the concatenation method is presented in Fig. 6. When exclusively employing the MVCL module, the experimental outcomes indicate comparable performance between feature summation and concatenation methods. However, when both the SC and MVCL modules are utilized concurrently, the summation approach demonstrates superior effectiveness compared to concatenation. This phenomenon suggests that the interaction between visual features and symptom descriptions requires careful

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	Rouge-L	CIDEr
Baseline	0.2333	0.1468	0.0981	0.0613	0.1699	0.2382	0.1551
+SC	0.3234	0.2205	0.1603	0.1200	0.2131	0.2780	0.3581
+MVCL1	0.2212	0.1485	0.1056	0.0771	0.1561	0.2257	0.1243
Full1	**0.4183**	**0.2933**	**0.2090**	**0.1496**	**0.2497**	**0.3394**	**0.5329**
+MVCL2	0.2016	0.1322	0.0939	0.0649	0.1702	0.2387	0.2711
Full2	0.3382	0.2490	0.1867	0.1435	0.2187	0.3307	0.5223

Table 1. The performance of baseline and our full model on the test sets of Deep Wrist dataset.

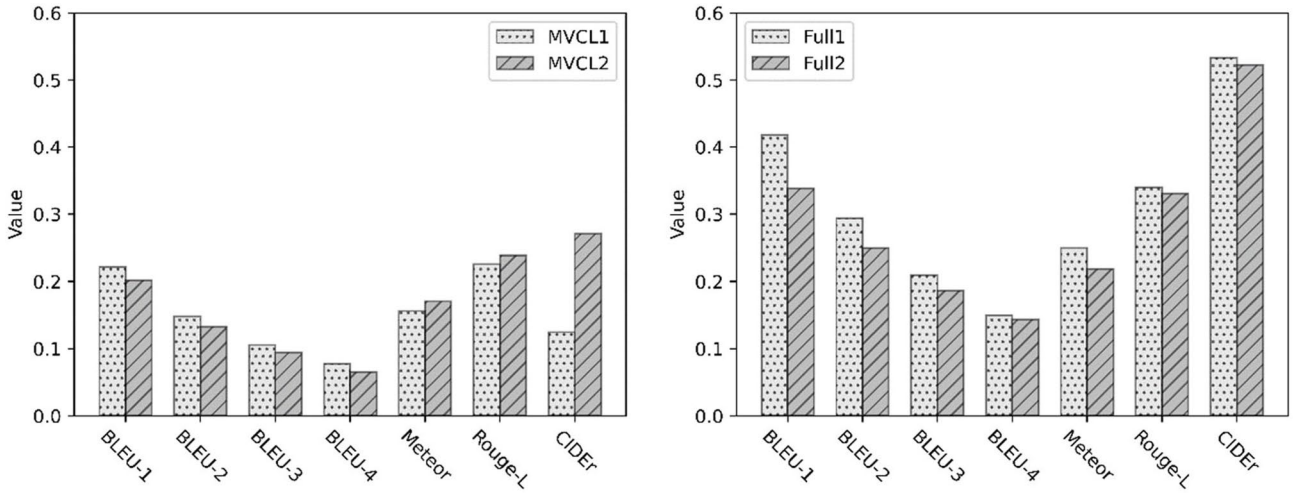

Fig. 6. Comparison of the performance of different strategies for multi-view visual feature fusion.

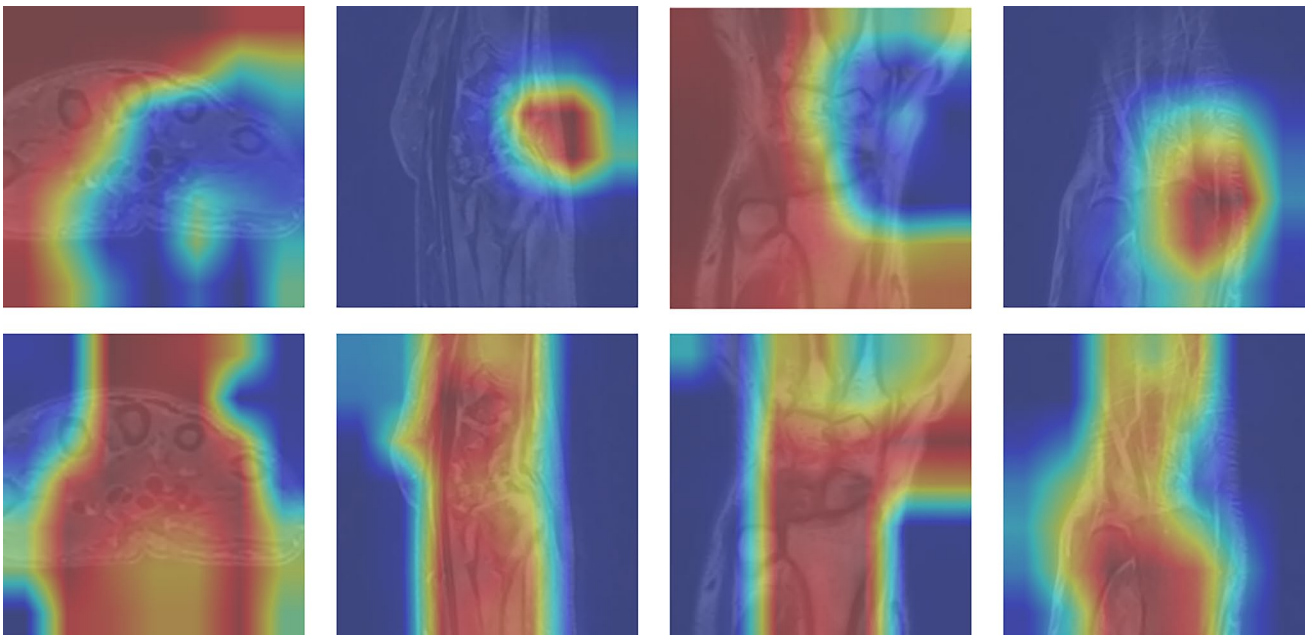

Fig. 7. Contrastive visualization of attention maps in MRI report generation with and without symptom consultation.

architectural consideration. By using the summation method to maintain a shorter visual feature sequence length of 49 instead of 147, we observe that the compressed feature representation works synergistically with the SC module to better distill clinically relevant information. The shorter sequence length in summation helps prevent attention dilution while still preserving sufficient medical insight from symptom descriptions, ultimately enabling the generation of higher-quality reports through this more focused feature representation.

Figure 7 presents a comparative visualization of attention heatmaps across four distinct MRI cases under two experimental conditions: the upper images depict baseline attention distributions without symptom consultant, while the lower images demonstrate attention patterns symptom consultant. The incorporation of symptom consultant critically enhances the model's ability to localize clinically relevant regions. When the symptom consultant is disabled, the attention maps exhibit a scattered distribution, lacking clear focus on pathological areas. In contrast, integrating symptom sharpens the model's attention, concentrating it on lesion-specific regions. This demonstrates that symptom-consultant not only aligns with radiologists' diagnostic reasoning but also ensures that the generated reports prioritize medically salient observations. The visualizations confirm that symptom acts as a semantic bridge between image features and textual descriptions, enabling precise and context-aware report generation (Fig. 7).

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor	Rouge-L	CIDEr
ST	0.1116	0.0939	0,9781	0.0645	0.1722	0.3329	0.2547
SAT	0.0938	0.0698	0.0557	0.0434	0.1498	0.2768	0.2510
COATT	0.1036	0.0767	0.0594	0.0456	0.1451	0.1160	0.1354
R2Gen	0.2333	0.1468	0.0981	0.0613	0.1699	0.2382	0.1551
Proposed	**0.4183**	**0.2933**	**0.2090**	**0.1496**	**0.2497**	**0.3394**	**0.5329**

Table 2. The performance of different models on the test sets of Deep Wrist dataset.

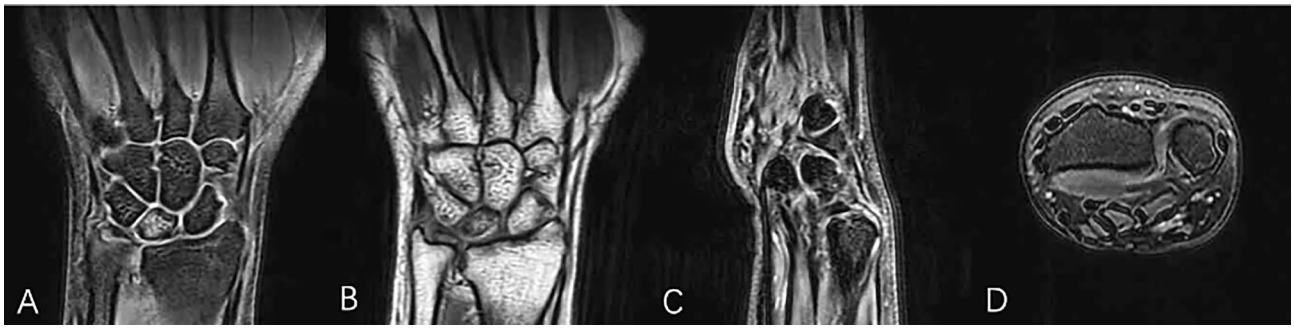

Fig. 8. The standard report indicates lunate bone edema with suspected avascular necrosis and increased median nerve signal, suggesting possible carpal tunnel syndrome.

To validate the effectiveness of our proposed model, we conducted a comprehensive comparison with several state-of-the-art models, including ST[2], SAT[3], COATT[55], and R2Gen[15]. The experimental results, as shown in Table 2, demonstrate that the proposed model achieves the best performance across all evaluation metrics, including BLEU (1–4), METEOR, ROUGE-L, and CIDEr, with scores of 0.4183, 0.2933, 0.2090, 0.1496, 0.2497, 0.3394, and 0.5329, respectively, indicating its superior ability to generate captions that are both semantically aligned and diverse. SAT and COATT exhibit significantly lower performance compared to the Proposed model, highlighting their substantial challenges in managing longer sequences and maintaining semantic relevance. ST and R2Gen show relatively better performance when compared to SAT and COATT, indicating their weaker recall and semantic alignment. Overall, our proposed model consistent dominance across all metrics underscores its effectiveness in image captioning, while the comparative analysis reveals the varying strengths and weaknesses of the other models.

Case studies

To further investigate the quality and readability of generated reports, we performed qualitative analysis on a case studies. For the right wrist case, the standard report describes lunate bone edema with suspected avascular necrosis and increased median nerve signal, suggesting possible carpal tunnel syndrome. Other models identified a triangular fibrocartilage complex (TFCC) injury and minor joint effusion but failed to provide comprehensive details regarding neural involvement. In contrast, the proposed method not only accurately identifies median nerve thickening and effusion in the distal radioulnar joint space but also demonstrates superior recognition of TFCC abnormalities. This highlights the model's enhanced capability in detecting soft tissue injuries crucial for diagnosing wrist instability and related conditions (Fig. 8).

Conclusions

In this paper, we propose a transformer-based framework to generate medical reports. The data employed for medical report generation is multi-modal, encompassing MRI scans and symptom descriptions. Furthermore, the MRI data utilized consists of multi-view images, which include coronal, sagittal, and transverse views. The multi-view contrastive learning module makes full use of multi-view MRI data to extract valuable information from distinct views. Meanwhile, the symptom consultant leverages symptom information to extract medical insight, guiding the layer normalizations within decoder. The synergy between these two modules harnesses the potential of both multi-modal and multi-view data, collaboratively enhancing the quality of medical report generation. Experimental results on the Deep Wrist dataset verify the effectiveness of the MVCL and SC modules. In the future, the combination of these two methods has the potential to play a significant role in the early detection and treatment of TFCC injuries, improving the quality of patient care, and providing enhanced support for medical decision-making. These approaches can have broader applications in the field of medical imaging and report generation, potentially benefiting the diagnosis and treatment of various medical conditions. Their integration of advanced technologies like deep learning and computer vision, as well as the incorporation of extensive clinical data and symptom information, holds promise for enhancing medical care and decision support across a wide range of medical conditions and scenarios.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 11 April 2025; Accepted: 29 April 2025
Published online: 23 May 2025

References

1. Bali, A., Bali, D., Iyer, N. & Iyer, M. Management of medical records: Facts and figures for surgeons. *J. Maxillofac. Oral. Surg.* **10**, 199–202. https://doi.org/10.1007/s12663-011-0219-8 (2011).
2. Usher, J. L. The importance of medical records in the home health field. *Caring* **6**, 92–94 (1987).
3. Struik, M. H. L. et al. The preferences of users of electronic medical records in hospitals: Quantifying the relative importance of barriers and facilitators of an innovation. *Implement Sci.* **9**, 69. https://doi.org/10.1186/1748-5908-9-69 (2014).
4. Treiser, M. D., Crawford, K. & Iorio, M. L. TFCC injuries: Meta-analysis and comparison of diagnostic imaging modalities. *J. Wrist. Surg.* **7**, 267–272. https://doi.org/10.1055/s-0038-1629911 (2018).
5. Jawed, A., Ansari, M. T. & Gupta, V. TFCC injuries: How we treat?. *J. Clin. Orthopaed. Trauma* **11**, 570–579. https://doi.org/10.1016/j.jcot.2020.06.001 (2020).
6. Mb, Z. MR imaging of ligaments and triangular fibrocartilage complex of the wrist. *Radiol. Clin. North Am.* https://doi.org/10.1016/j.rcl.2006.04.010 (2006).
7. Herdade, S., Kappeler, A., Boakye, K. Soares, J. Image captioning: Transforming objects into words. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2019).
8. Vinyals, O., Toshev, A., Bengio, S. Erhan, D. Show and tell: A neural image caption generator. pp 3156–3164 (2015).
9. Xu, K., Ba, J. Kiros, R., et al. Show, Attend and tell: neural image caption generation with visual attention (2016).
10. Cho, K., van Merrienboer, B. Gulcehre, C., et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on empirical methods in natural language processing (EMNLP 2014) (2014).
11. Shin, H-C., Roberts, K., Lu, L. et al. Learning to read chest X-Rays: recurrent neural cascade model for automated image annotation. pp 2497–2506 (2016).
12. Lin, M., Chen, Q. Yan, S. Network in network (2014).
13. Krause, J., Johnson, J., Krishna, R. Fei-Fei, L. A hierarchical approach for generating descriptive image paragraphs (2017)
14. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
15. Xue, Y. et al. Multimodal recurrent model with attention for automated radiology report generation. In *Medical image computing and computer assisted intervention – MICCAI 2018* (eds Frangi, A. F., Schnabel, J. A., Davatzikos, C. et al.) 457–466 (Springer International Publishing, 2018).
16. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. In: Advances in neural information processing systems. Curran Associates, Inc. (2017).
17. Yu, J., Li, J., Yu, Z. & Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **30**, 4467–4480. https://doi.org/10.1109/TCSVT.2019.2947482 (2020).
18. Hou, B., Kaissis, G., Summers, R. M. & Kainz, B. RATCHET: Medical transformer for chest X-ray diagnosis and reporting. In *Medical image computing and computer assisted intervention – MICCAI 2021* (eds De Bruijne, M., Cattin, P. C., Cotin, S. et al.) 293–303 (Springer International Publishing, 2021).
19. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q. Densely Connected Convolutional Networks. pp 4700–4708 (2017).
20. Lovelace, J., Mortazavi, B. Learning to generate clinically coherent chest X-Ray reports. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1235–1243 (2020).
21. Chen, Z., Song, Y., Chang, T.-H., Wan, X. Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 1439–1449 (2020).
22. Srinivasan, P., Thapar, D., Bhavsar, A. & Nigam, A. Hierarchical X-Ray report generation via pathology tags and multi head attention. In *Computer Vision – ACCV 2020* (eds Ishikawa, H. et al.) 600–616 (Springer International Publishing, 2021).
23. Lee, H. et al. Cross encoder-decoder transformer with global-local visual extractor for medical image captioning. *Sensors* **22**, 1429. https://doi.org/10.3390/s22041429 (2022).
24. Muksimova, S. et al. Novelty classification model use in reinforcement learning for cervical cancer. *Cancers* **16**, 3782. https://doi.org/10.3390/cancers16223782 (2024).
25. Tu, T., Azizi, S., Driess, D., et al. Towards Generalist Biomedical AI. NEJM AI 1:AIoa2300138. https://doi.org/10.1056/AIoa2300138.(2024).
26. Schmidt, S. et al. Simplifying radiologic reports with natural language processing: A novel approach using ChatGPT in enhancing patient understanding of MRI results. *Arch. Orthop. Trauma Surg.* **144**, 611–618. https://doi.org/10.1007/s00402-023-05113-4 (2024).
27. Hamamci, I. E., Er, S. & Menze, B. CT2Rep: Automated radiology report generation for 3D medical imaging. In *medical image computing and computer assisted intervention MICCAI 2024* (eds Linguraru, M. G., Dou, Q., Feragen, A. et al.) 476–486 (Springer, 2024).
28. Dai, B., Lin, D. Contrastive learning for image captioning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017).
29. Radford, A., Kim, J.W., Hallacy, C., et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp 8748–8763 (2021).
30. Gao, T., Yao, X., Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 6894–6910 (2021).
31. Liu, F., Yin, C., Wu, X. et al. Contrastive attention for automatic chest X-ray report generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, Online, pp 269–280 (2021).
32. Zhang, Y., Jiang, H., Miura, Y. et al. Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference. PMLR, pp 2–25 (2022).
33. Wu, X., Li, J., Wang, J. & Qian, Q. Multimodal contrastive learning for radiology report generation. *J. Ambient. Intell. Human Comput.* **14**, 11185–11194. https://doi.org/10.1007/s12652-022-04398-4 (2023).
34. Pang, T., Li, P. & Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. *Biomed. Eng. Online* **22**, 48. https://doi.org/10.1186/s12938-023-01113-y (2023).
35. Park, D. J. et al. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci. Rep.* **11**, 7567. https://doi.org/10.1038/s41598-021-87171-5 (2021).
36. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29. https://doi.org/10.1038/s41591-018-0316-z (2019).

37. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **20**, e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4 (2019).
38. Reverberi, C. et al. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **12**, 14952. https://doi.org/10.1038/s41598-022-18751-2 (2022).
39. Henry, K. E. et al. Human–machine teaming is key to AI adoption: Clinicians' experiences with a deployed machine learning system. *npj Digit. Med.* **5**, 1–6. https://doi.org/10.1038/s41746-022-00597-7 (2022).
40. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311. https://doi.org/10.1038/s42256-020-0186-1 (2020).
41. Teng, Q. et al. A survey on the interpretability of deep learning in medical diagnosis. *Multimed. Syst.* **28**, 2335–2355. https://doi.org/10.1007/s00530-022-00960-4 (2022).
42. Kolyshkina, I. & Simoff, S. Interpretability of machine learning solutions in public healthcare: The CRISP-ML approach. *Front. Big Data* **4**, 660206 (2021).
43. Da Silva, M. et al. Legal concerns in health-related artificial intelligence: A scoping review protocol. *Syst. Rev.* **11**, 123. https://doi.org/10.1186/s13643-022-01939-y (2022).
44. Currie, G. & Hawk, K. E. Ethical and legal challenges of artificial intelligence in nuclear medicine. *Semin. Nucl. Med.* **51**, 120–125. https://doi.org/10.1053/j.semnuclmed.2020.08.001 (2021).
45. Cath, C. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosop. Trans. Royal Soc. A Math. Phys. Eng. Sci.* **376**, 20180080. https://doi.org/10.1098/rsta.2018.0080 (2018).
46. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310. https://doi.org/10.1093/jamia/ocv080 (2016).
47. Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R. et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs (2019).
48. Pizer, S. M. et al. Adaptive histogram equalization and its variations. *Comput. Vision Graph. Image Process.* **39**, 355–368 (1987).
49. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. pp 770–778 (2016).
50. Schroff, F., Kalenichenko, D., Philbin, J. FaceNet: A unified embedding for face recognition and clustering. pp 815–823 (2015).
51. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318 (2002).
52. Denkowski, M., Lavie, A. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Edinburgh, Scotland, pp 85–91 (2011)
53. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp 74–81 (2004).
54. Vedantam, R., Lawrence Zitnick, C., Parikh, D. CIDEr: Consensus-based image description evaluation. pp 4566–4575 (2015).
55. Jing, B., Xie, P., Xing, E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp 2577–2586 (2018).

Acknowledgements
Not applicable

Author contributions
H.L. and X.H. conceptualized the study. X.Z. curated the data, performed formal analysis, and acquired funding. Y.D. conducted the investigation. H.Z. developed the methodology. S.H.A.E. handled the software development. V.G.K. supervised the project. M.H.A.H.A. validated the findings. S.A.A. created the visualizations. Q.B., A.A. and X.Z. wrote the original draft and reviewed and edited the manuscript. All authors reviewed the manuscript.

Funding
The study was funded byAlibaba Youth Studio Project. The funding body had no role in the design of the study; in collection, analysis, and interpretation of data; and in drafting the manuscript.

Declarations

Competing interest
The authors declare no competing interests.

Ethical approval
This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of First Affiliated Hospital, Zhejiang University School of Medicine (Approval Number:2024-0399). Informed consent was obtained from all participants involved in the study.

Consent for publication
Written informed consent was obtained from the patient for the publication of clinical details and clinical images. Upon request, a copy of the consent form is available for review by the Editor of this journal.

Additional information
Correspondence and requests for materials should be addressed to X.H. or H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.