Adapter-Enhanced Hierarchical Cross-Modal Pre-Training for Lightweight Medical Report Generation

Ting Yu, *Member, IEEE*, Wangwen Lu, Yan Yang, *Member, IEEE*, Weidong Han, Qingming Huang, *Fellow, IEEE*, Jun Yu, *Senior Member, IEEE*, and Ke Zhang, *Member, IEEE*

***Abstract*—Automatic medical report generation is an emerging field that aims to transform medical images into descriptive, clinically relevant narratives, potentially reducing the workload for radiologists significantly. Despite substantial progress, the increasing model parameter size and corresponding marginal performance gains have limited further development and application. To address this challenge, we introduce an Adapter-enhanced Hierarchical cross-modal Pre-training (AHP) strategy for lightweight medical report generation. This approach significantly reduces the pre-trained model's parameter size while maintaining superior report generation performance through our proposed spatial adapters. To further address the issue of inadequate representation of visual space details, we employ a convolutional stem combined with hierarchical injectors and extractors, fully integrating with traditional Vision Transformers to achieve more comprehensive visual representations. Additionally, our cross-modal pre-training model effectively handles the inherent complex visual-textual relationships in medical imaging. Extensive experiments on multiple datasets, including IU X-Ray, MIMIC-CXR, and bladder pathology, demonstrate our model's exceptional generalization and transfer performance in downstream medical report generation tasks, highlighting AHP's potential in significantly reducing model parameters while enhancing report generation accuracy and efficiency.**

***Index Terms*—Cross-modal pre-training, lightweight model, medical report generation, multi-task learning.**

I. Introduction

MEDICAL imaging plays a crucial role in the routine diagnosis and treatment of patients. However, healthcare professionals encounter significant challenges when analyzing the increasing volume of medical images and drafting diagnostic reports based on clinical observations. These reports, which typically include detailed descriptions of both normal and abnormal findings, demand extensive knowledge and experience. Despite their expertise, doctors find this task time-consuming and prone to errors, especially as the number of images continues to grow. To alleviate the burden on healthcare providers and enhance clinical efficiency, the implementation of automated medical report generation is essential. This technology not only simplifies the diagnostic process and reduces doctors' workload but also improves the consistency and accuracy of reports, ultimately promoting better patient care and outcomes.

Recently, automated medical report generation has garnered significant interest and achieved substantial advancements through the integration of deep learning and healthcare [1], [2], [3], [4], [5], [6]. For medical cross-modal tasks, it is crucial for models to capture and describe both the rare but significant details and the overall imaging characteristics. Following the workflow of radiologists, when given a medical image, they first inspect the global region to identify anomalies and then scrutinize smaller, easily overlooked local areas before drafting a precise medical report based on medical knowledge and experience. However, traditional pre-trained models typically use a standalone Transformer Encoder to encode images, as shown in Fig. 1(a). The traditional Vision Transformer (ViT) [7] utilizes a self-attention mechanism to compute global dependencies, which is highly effective for learning long-range dependencies and global context [8]. However, it tends to overlook local structural information, and the global nature of ViT's positional encoding is significantly different from the local receptive fields of CNNs, which often leads to insufficient representation of spatial details [9]. Previous approaches [10], [11], [12], [13] have

Received 2 July 2024; revised 22 January 2025; accepted 24 January 2025. Date of publication 28 January 2025; date of current version 4 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62125201, Grant 62020106007, Grant 62002314, and Grant 62406093, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY23F020005 and Grant LQ24F020032. *(Corresponding author: Ke Zhang.)*

Ting Yu and Wangwen Lu are with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: yut@hznu.edu.cn; luwangwen@stu.hznu.edu.cn).

Yan Yang and Ke Zhang are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yangyan@hdu.edu.cn; ke.zhang@hdu.edu.cn).

Weidong Han is with the Department of Colorectal Medical Oncology, Zhejiang Cancer Hospital, Hangzhou 310022, China, and also with the College of Mathematical Medicine, Zhejiang Normal University, Jinhua 321017, China (e-mail: hanwd@zju.edu.cn).

Qingming Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qmhuang@ucas.ac.cn).

Jun Yu is with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: yujun@hit.edu.cn).

Our code is available on the project page: https://github.com/OpenMICG/AHP.

Digital Object Identifier 10.1109/JBHI.2025.3535699

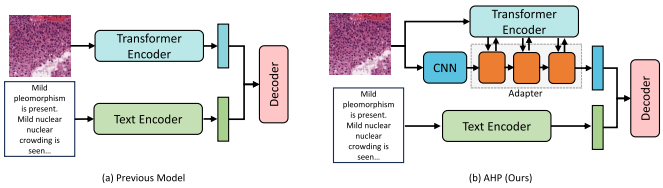


Fig. 1. Previous model vs. our AHP: (a) The previous model utilizes a single-branch transformer encoder exclusively for visual feature extraction. (b) We propose a convolution-augmented spatial adapter hierarchically integrated with the transformer encoder to effectively merge global and local representations, producing more accurate medical image reports.

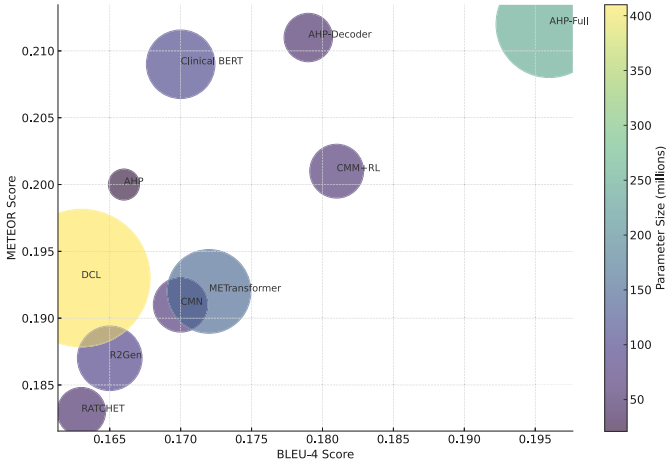


Fig. 2. Comparative analysis of parameter size versus generative performance across various models on the IU X-Ray dataset.

shown that the interaction between CNNs and transformers can more effectively extract image features. To more adequately extract and combine spatial details present in images, we introduce CNNs to extract local visual features and integrate them hierarchically with the multi-scale features from Vision Transformers (ViT) through our proposed adapter, as illustrated in Fig. 1(b). This hierarchical fusion process enriches feature representation, enabling more accurate medical report generation.

Furthermore, despite significant advancements in automatic medical report generation and cross-modal pre-training models, the increasing size of model parameters coupled with minimal performance gains has hindered further development and application of medical report generation. Especially with smaller downstream datasets, large pre-trained models often overfit. To address this issue, we propose a lightweight solution: An Adapter-enhanced Hierarchical cross-modal Pre-training model (AHP) for lightweight medical report generation. Specifically, we conduct joint cross-modal pre-training on multiple datasets including ROCO [14], MedICaT [15], and PMC-OA [16] to learn transferable latent knowledge and universal cross-modal representations. In downstream tasks on three imaging and pathology datasets, we can only fine-tune the proposed lightweight spatial adapter to achieve report generation results comparable to, or even better than, state-of-the-art methods. Compared with publicly available methods in terms of parameter count and performance, as depicted in Fig. 2, our proposed AHP achieves even better performance with a much smaller parameter size than the baseline.

Our main contributions can be summarized as follows:

- We propose an Adapter-enhanced Hierarchical cross-modal Pre-training Model for lightweight medical report generation, which employs adapters to enhance the extraction of spatially hierarchical visual features for medical report generation and utilizes multiple cross-modal matching tasks to better align image and text modalities.
- To overcome the issue of insufficient representation of spatial details in traditional ViT architectures, we propose using a convolutional backbone to extract local spatial features, which are then fused with multi-scale global features through hierarchical injectors and extractors, achieving a richer and more comprehensive feature representation.
- Extensive experiments and analyses on the IU X-Ray, MIMIC-CXR, and Bladder Pathology datasets have validated the efficacy and superior performance of our method. Additionally, our approach has demonstrated the capability to generate reports on par with or better than state-of-the-art methods with significantly fewer parameters on downstream datasets.

II. Related Works

A. Medical Image and Report Pre-Training

In recent years, the integration of vision and language pre-training has been advanced for medical applications through various models. ConVIRT [17] employed a domain-agnostic, bidirectional contrastive learning strategy, enhancing the general applicability across different medical scenarios. GLoRIA [18] improved understanding by aligning image sub-regions with corresponding words in radiology reports. Multi-Granularity Cross-modal Alignment (MGCA) [19] leveraged semantic correspondences at multiple levels—pathological region, instance, and disease—to refine medical visual representations. PPIOR [20] focused on fine-grained local alignment and cross-modality conditional reconstruction, targeting detailed visual and clinical linguistic features in medical images and reports. Knowledge-enhanced Auto Diagnosis (KAD) [21] incorporated medical domain knowledge into pre-training, specifically with chest X-rays and associated reports. Inspired by these strategies, we aim to enhance the alignment between vision and language, thereby improving the efficacy of model pre-training.

B. Image Captioning

In the realm of deep learning, image captioning is a pivotal task that automates the generation of descriptive textual captions for images, integrating computer vision and natural language processing (NLP) technologies. This task represents a crucial research direction in artificial intelligence. Recent advancements have seen a surge in methodologies that have significantly pushed the boundaries of the state-of-the-art [22], [23], [24], [25], [26], [27], [28]. Image captioning leverages natural language to describe images, aiming to produce meaningful and

grammatically correct text based on a comprehensive understanding of the visual content. The generated captions provide a high-level summary, facilitating automatic image annotation and support for individuals with visual impairments.

The success of the Transformer architecture [29] across both vision and NLP domains has inspired numerous innovations in image captioning techniques. The hybrid model [22] employed both bottom-up and top-down attention mechanisms to explore interrelations among objects and key regions within an image, enhancing the extraction of visual features. Vinyals et al. [23] utilized a deep recurrent network for generating image descriptions. Further, Herdade et al. [30] introduced the Object Relation Transformer, which enriched the encoder-decoder framework by incorporating spatial relationships between detected objects through geometric attention. The X-Linear Attention Networks (X-LAN) [28] introduced X-Linear attention blocks to represent visual features and capture complex interrelations among image regions using higher-order interactions and bilinear pooling. The Meshed-Memory Transformer [27] integrated a multi-level representation of visual features by embedding contextual knowledge across image regions. Additionally, the Dual-Level Collaborative Transformer (DLCT) [31] combined grid-level and region-level features to enhance captioning performance. The Semantic-Conditional Diffusion Networks (SCD-Net) [32] and SmallCap [33] further extended capabilities by integrating semantic priors and leveraging diffusion models for visual-language alignment and linguistic coherence. Despite their success in generic image captioning, these models face challenges when adapted to medical report generation due to the existence of the domain gap.

C. Medical Report Generation

Medical reports, characterized by detailed sentences that interpret medical images, represent a specialized advancement from general image captioning. The generation of medical reports requires not only longer textual outputs but also a higher level of detail and accuracy in descriptions, crucial for clinical relevance and utility. Significant strides have been made in medical report generation. HRGR-Agent [34] employed a hybrid of retrieval-based and learning-based methods with reinforcement learning to produce structured, diverse, and accurate medical reports. Liu et al. [35] generated chest X-ray reports by first predicting topics from images, then crafting sentences from these topics, optimized for readability and clinical accuracy through reinforcement learning.

Recent advancements have leveraged the Transformer model [29]. R2Gen [2] enhanced clinical automation with a memory-driven Transformer, incorporating relational memory and memory-driven conditional layer normalization within the decoder. Cross-modal memory networks (CMN) [1] recorded image-text alignments in shared memory to improve radiology report generation by enhancing cross-modal interactions. RATCHET [36], an end-to-end CNN-RNN-based medical transformer, extracted features from chest radiographs to generate precise medical reports, aiming to enhance clinical workflow efficiency. METransformer [3] used a multi-expert token system within its architecture to mimic multi-specialist consultations, thereby improving image analysis and report generation.

There is a growing trend towards incorporating supplementary knowledge into report generation. Zhang et al. [37] focused on disease-specific graphs and relationships to enhance radiology report generation, introducing a tailored evaluation metric. The PPKED [38] mimicked radiologist diagnostic patterns through modules that explored and distilled knowledge. KERP [39] utilized a Knowledge-driven Encode, Retrieve, Paraphrase framework with a Graph Transformer to generate accurate and coherent reports by learning structured abnormality graphs and adapting text templates to specific cases. Distinct from previous methods, RAMT [40] addressed the issue of scarce labeled data by proposing a semi-supervised hybrid feature consistency learning approach to fully exploit the commonalities of diseases. The DCL [4] enhanced automatic radiology report generation by using a dynamic knowledge graph and contrastive learning to adaptively refine and utilize medical knowledge. Lastly, the Knowledge-injected U-Transformer (KiUT) [41] integrated multi-modal information and clinical knowledge through a novel U-connection schema and knowledge distillation techniques to enhance radiology report generation. In this paper, we introduce a novel adapter-enhanced hierarchical pre-training model for medical report generation, leveraging pre-trained implicit knowledge to significantly improve spatial feature extraction from medical images and report fluency and coherence.

III. Methodology

The goal of medical report generation is to automatically generate a target description report $R = \{r_1, r_2, \ldots, r_n\}$ for a given medical image I, where n represents the number of tokens in the report. To enhance the generalization of pre-trained models in medical report generation and reduce the parameter count when transferred to downstream datasets, we propose a novel lightweight Adapter-enhanced Hierarchical cross-modal Pre-training model, AHP. Our model employs adapters to enhance the extraction of spatially hierarchical visual features for medical report generation and utilizes multiple cross-modal matching tasks to better align image and text modalities. Fig. 3 illustrates the overall framework of the model, which consists of a visual encoder, spatial adapter, report encoder, image-report encoder, and report decoder. In this section, we will introduce each of these components in detail.

A. Visual Encoder

As depicted in Fig. 3(a), our visual encoder is a standard Vision Transformer (ViT), consisting of a patch embedding module and N blocks, each containing L Transformer encoder layers. The input medical image is initially fed into the patch embedding module, where the image is divided into non-overlapping patches of size 16×16. Each patch is then flattened and augmented with position embeddings to incorporate spatial information. These patches are subsequently passed through N blocks to obtain the visual features of this branch. Throughout this process, the resolution of the features is reduced to 1/16 of

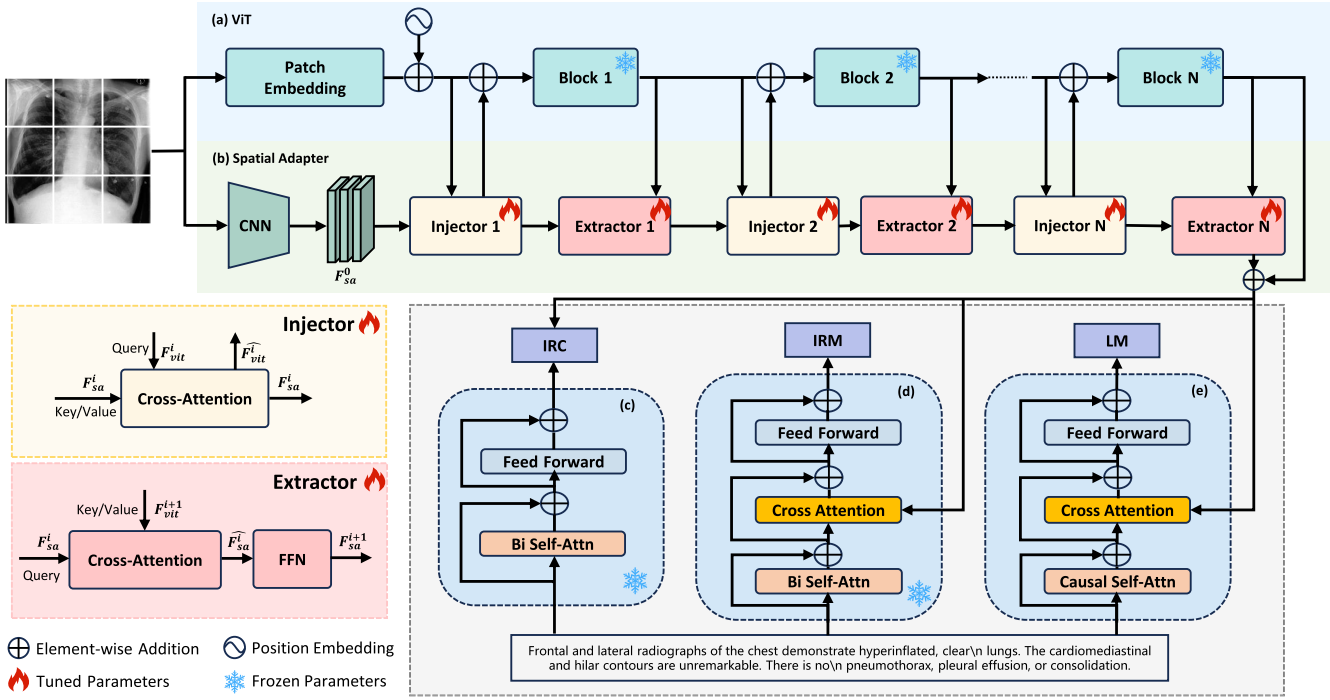


Fig. 3. Overall architecture of our proposed AHP. (a) ViT, which acts as a visual encoder and includes N layers. (b) Spatial Adapter, which consists of CNN, Injector, and Extractor. (c) Unimodal report encoder, to encode the report for Image-Report Contrastive (IRC) learning. (d) Image-Report encoder, to encode image-report pairs for Image-Report Matching (IRM). (e) Report decoder, to generate reports given images in an autoregressive manner constrained by Language Modeling loss (LM).

the original image, and a [CLS] token is introduced to represent the global features of the original image.

B. Spatial Adapter

To enable the visual encoder to integrate spatial prior information while extracting hierarchical features to assist in report generation, we propose the spatial adapter module. As shown in Fig. 3(b), this module comprises three key components: (1) a CNN backbone, responsible for capturing spatial prior features from the input image; (2) spatial injectors, which integrate spatial prior information into the ViT; and (3) feature extractors, which extract hierarchical features from the multi-scale outputs of the ViT.

For the spatial injector, this module integrates the feature map generated by the CNN backbone into the original ViT branch. Specifically, we designate the initial feature map F_{sa}^0 derived from the CNN stem as both the key and value, while the feature F_{vit}^i obtained from the original ViT branch serves as the query. The spatial injection process is detailed as follows.

$$\hat{F_{vit}^i} = F_{vit}^i + \lambda^i SparseAttn(LN(F_{vit}^i), LN(F_{sa}^i)). \quad (1)$$

Here, i signifies the index of the ViT block. $LN(\cdot)$ denotes the Layer Normalization operation [42], and $SparseAttn(\cdot)$ represents the Sparse Attention. Additionally, we introduce a learnable vector λ^i to regulate the balance between the output of the attention layer and the input feature F_{vit}^i, initialized to zero. This aims to gradually inject spatial information into the ViT to maintain training stability.

The resulting $\hat{F_{vit}^i}$ is then input into the encoder layers of the i-th block, yielding the output feature F_{vit}^{i+1}. Subsequently, we employ a feature extractor composed of a cross-attention layer and a feed-forward network (FFN) to facilitate further interaction between the hidden layers of the original ViT branch and the spatial injector. Specifically, in the cross-attention layer, F_{vit}^{i+1} is used as the key and value input, while F_{sa}^i is used as the query input. The $\hat{F_{sa}^i}$ is then processed through the FFN to obtain F_{sa}^{i+1}. This process can be described as follows:

$$\hat{F_{sa}^i} = F_{sa}^i + SparseAttn(LN(F_{sa}^i),\ LN(F_{vit}^{i+1})), \quad (2)$$

$$F_{sa}^{i+1} = \hat{F_{sa}^i} + FFN(LN(\hat{F_{sa}^i})). \quad (3)$$

Notably, the resulting spatial feature F_{sa}^{i+1} is subsequently used as the input for the following spatial injector.

Ultimately, following N interaction rounds between the ViT branch and the spatial adapter branch, the composite visual features fed into the report module consist of F_{sa}^N, which contains spatial detail information, and F_{vit}^N, which incorporates global relevance information. These features undergo normalization to yield the final visual feature representation F_v.

$$F_v = Norm(F_{sa}^N + F_{vit}^N). \quad (4)$$

C. Report Encoder

In this study, we utilize a unimodal report encoder and an image-report encoder to process the target report, each trained using distinct loss functions: image-report contrastive loss and

image-report matching loss, respectively, detailed further in subsequent sections. The unimodal report encoder, illustrated in Fig. 3(c), is based on the BERT framework [43], featuring transformer blocks that combine bi-directional self-attention layers with feed-forward networks. A `[CLS]` token is prefixed to the input to encapsulate the sentence's global features. Diverging from this, the image-report encoder (Fig. 3(d)) incorporates a cross-attention mechanism to enhance the integration of visual data, optimizing the encoder for multimodal information processing.

D. Decoder

To facilitate the generation of diagnostic reports from medical images, we implement an autoregressive model for medical report generation. This model incorporates a 12-layer Transformer decoder, detailed in Fig. 3(e), to construct the reports. The decoder employs cross-attention mechanisms to integrate multimodal embeddings and initiates the generation process with a sequence start token `[CLS]`. A sequence end token `[EOS]` is appended to the output of the decoder, signaling the completion of the report generation.

E. Training Objectives

Compared to traditional captioning datasets, the collection of medical images and their corresponding reports is costly and time-consuming, resulting in relatively small datasets for medical report generation training. Constrained by the limited size of these medical datasets, we adopt cross-modal pre-training with image-report pairs to extract commonalities from the medical report dataset. Specifically, we introduce three distinct pre-training objectives that effectively enhance the representations of both vision and text while strengthening the alignment between image and text modalities, thus improving the model's generalization on downstream datasets. The three pre-training loss functions are detailed as follows.

1) Image-Report Contrastive Loss (IRC): IRC aligns the representations of medical images with their corresponding reports by leveraging the `[CLS]` tokens from both modalities. To address the issue of weak associations in the training data, we adopt the ALBEF methodology [44] to utilize a momentum model that generates pseudo-targets, which are then applied in the base model's contrastive learning. The similarity between two `[CLS]` representations is calculated as follows:

$$s(I,R)=g_v(v_{cls})^{\intercal}g'_w(w'_{cls}),\quad s(R,I)=g_w(w_{cls})^{\intercal}g'_v(v'_{cls}), \tag{5}$$

where g_v and g_w denote linear transformations that map the `[CLS]` token into low-dimensional spaces. $g'_v(v'_{cls})$ and $g'_w(w'_{cls})$ represent the feature representations from the momentum visual encoder and momentum text encoder, respectively. Specifically, IRC encourages the model to bring positive image-report pairs closer in the representation space, while distancing negative pairs. Two momentum queues are utilized to store the most relevant M representations of image-report pairs $\{I_m, R_m\}, m \in [1, M]$, derived from the momentum-driven image and text encoders. The image-to-report similarity $f_m^{i2r}(I)$ and report-to-image similarity $f_m^{r2i}(R)$ are calculated as follows:

$$f_m^{i2r}(I)=\frac{exp(s(I,R_m)/\tau)}{\sum_{m=1}^{M}exp(s(I,R_m)/\tau)}, \tag{6}$$

$$f_m^{r2i}(R)=\frac{exp(s(R,I_m)/\tau)}{\sum_{m=1}^{M}exp(s(R,I_m)/\tau)}, \tag{7}$$

where τ is a learnable temperature parameter. The image-report contrastive loss is defined by calculating the cross-entropy loss between $f^{i2r}(I)$, $y^{r2i}(R)$, and their respective ground truths as follows:

$$L_{IRC}=\frac{1}{2}(CE(y^{i2r}(I),f^{i2r}(I))+CE(y^{r2i}(R),f^{r2i}(R))). \tag{8}$$

Here, $CE(\cdot)$ denotes the cross-entropy loss, and $y^{i2r}(I)$ and $y^{r2i}(R)$ represent the ground-truth one-hot similarities, where negative pairs are assigned a probability of 0 and the positive pair a probability of 1.

2) Image-Report Matching Loss (IRM): IRM is employed to determine whether a given image-report pair is positive (matching) or negative (non-matching). A multimodal encoder that integrates image and report features through a cross-attention mechanism is used, producing multimodal embedded features. The softmax function is then applied to predict the two-class probability, p^{irm}. The IRM is defined as follows:

$$L_{IRM}=CE(y^{irm},p^{irm}(I,R)), \tag{9}$$

where y^{irm} denotes a binary label vector, encoded as a two-dimensional one-hot vector, which indicates the ground-truth label.

3) Language Modeling Loss (LM): LM engages the report decoder to generate reports corresponding to provided medical images. This optimization utilizes cross-entropy loss, guiding the model through autoregressive training to maximize the likelihood of the generated report. Following the decoder architecture illustrated in Fig. 3(e), the report token $\hat{r_t}$ at time step t is predicted by inputting the preceding report tokens $r_{1:t-1} = \{r_1, \ldots, r_{t-1}\}$ into the causal self-attention, cross-attention, and feed-forward layers, as described below:

$$h_{att}=LN(CMSA(r_{1:t-1})+r_{1:t-1}), \tag{10}$$

$$h_{ca}=LN(CA(h_{att},f_I)+h_{att}), \tag{11}$$

$$\hat{r_t}=LN(FFN(h_{ca})+h_{ca}), \tag{12}$$

where $CMSA(\cdot)$ and $CA(\cdot)$ respectively denote Causal Multi-head Self-Attention and Cross-Attention. A `[EOS]` token is employed to signal the end of the report. The entire autoregressive generation process during the inference stage can be expressed as follows:

$$p(\hat{r}|I)=\prod_{t=1}^{n}p(\hat{r_t}|r_{1:\hat{t}-1},I) \tag{13}$$

where $r_{1:t-1}$ is the input report token in time step t.

In our model training, we aim to minimize the negative log-likelihood of $P(R|I)$ given the image features and the target report. The training process of the language model is facilitated

by calculating the cross-entropy loss between the predicted token $\hat{r_t}$ and the ground truth report r_t as follows:

$$L_{LM} = -\sum_{t=1}^{n} \log\ p(r_t|r_1, \ldots, r_{t-1}, I). \tag{14}$$

Our overall objective function is the combination of the above three pre-training loss functions:

$$L_{all} = L_{IRC} + L_{IRM} + L_{LM}. \tag{15}$$

IV. Experiments

A. Datasets

Our research on medical report generation involved validation on three distinct downstream datasets, each with unique characteristics. Comprehensive descriptions of these datasets are provided below. There is no overlap between the samples in the training, validation, and test sets across the three datasets.

1) IU X-Ray: The IU X-Ray dataset [45], established by Indiana University, serves as a widely recognized benchmark for evaluating radiology report generation methods. It comprises 7,470 chest X-ray images and 3,955 corresponding radiology reports. The dataset is typically segmented into training, validation, and test sets with proportions of 70%, 10%, and 20%, respectively, resulting in 2,069 training samples with an average report length of 31.77, 296 validation samples with an average report length of 31.12, and 590 testing samples with an average report length of 28.23. Notably, patient overlap is prevented across these sets. In the image preprocessing pipeline, the input images are resized to a fixed resolution of 224×224. Additionally, normalization is applied to each channel using a mean value of 0.483 and a standard deviation of 0.235, as referenced in the R2Gen [2]. These operations ensure that the input images are well-suited for the model, enhancing training stability and convergence. Each sample includes both a frontal and a lateral image corresponding to the same report. The images are processed separately through the Visual Encoder, then concatenated and passed through a linear layer with dimensions 768×2 to 768, enabling them to adapt to the model effectively. Reports Preprocessing steps include tokenization, conversion to lowercase, setting a maximum report length of 60, and replacing infrequently occurring words (less than three times) with $'<\texttt{unk}>'$ token. A corresponding vocabulary index is also constructed to map words to indices and vice versa. This index is used to encode words into indices compatible with the model and to reconstruct reports from the indices. The vocabulary of the dataset covers over 99.0% of the words in the corpus.

2) MIMIC-CXR: The MIMIC-CXR dataset [46], currently the largest radiology dataset, contains 377,110 chest X-ray images and 227,835 radiology reports sourced from 64,588 patients who underwent examination at Beth Israel Deaconess Medical Center, covering the period from 2011 to 2016. According to official partitioning guidelines, the dataset includes 270,790 training samples with an average report length of 47.88, 2,130 validation samples with an average report length of 48.00, and 3,858 test samples with an average report length of 60.83. Similar to the IU-Xray dataset, we perform the following steps on the MIMIC-CXR dataset: input images are resized to 224×224 resolution, and each channel is normalized with a mean of 0.483 and a standard deviation of 0.235. The preprocessing of reports involves tokenization, conversion to lowercase, setting a maximum report length of 100, and substituting words with frequencies less than 10 with $'<\texttt{unk}>'$ token. Additionally, special characters, extra spaces, and numbering are removed to simplify and standardize the text. The vocabulary encompasses approximately 4,000 words, and the reports provide detailed clinical findings for each image.

3) Bladder Pathology: The Bladder Pathology dataset [47], consists of 4,253 bladder pathology images derived from 221 slides of non-invasive high-grade (HG) and low-grade (LG) papillary urothelial carcinoma. The associated reports primarily describe five critical morphological visual features for classifying urothelial carcinoma: nuclear pleomorphism, cell crowding, cell polarity, mitosis, and nucleoli prominence. We follow the official data partitioning method, excluding samples labeled as "insufficient information" and retaining those labeled as "normal", "LG papillary urothelial carcinoma", and "HG papillary urothelial carcinoma". This partition results in 2,076 training samples with an average report length of 35.56 and 1,734 testing samples with an average report length of 36.06, with each image associated with 5 reports describing similar content. Similar to the previous two datasets, images are resized to a resolution of 224×224, and each channel is normalized. Report preprocessing includes tokenization, conversion to lowercase, and replacing words occurring less than three times with $'<\texttt{unk}>'$ token. The final vocabulary consists of 113 words.

B. Implementation Details

Our model is pre-trained on eight NVIDIA GeForce RTX 4090 GPUs over 30 epochs with PyTorch [60]. During this phase, the visual transformer's initial parameters are sourced from ViT [7] pre-trained on ImageNet [61], while the text transformer initializes with BERTbase [43] parameters. The ViT is configured with 4 blocks, each comprising 3 encoder layers. The model dimension is set at $d = 768$. We employ the AdamW [62] optimizer, with a weight decay of 0.05 and an initial learning rate of $1e-4$, decaying to $1e-5$ following a cosine schedule. For pre-training, we utilize three datasets: PMC-OA [16], Radiology Objects in Context (ROCO) [14], and MedICaT [15]. PMC-OA includes 1.6 million biomedical image-caption pairs. ROCO provides 81,825 radiology-related image-caption pairs and 6,127 out-of-class pairs, though only the radiology pairs were used. MedICaT contains 217,060 figures from 131,410 open-access papers, with additional annotations for complex figures, ensuring no overlap with ROCO. The pre-training dataset encompasses diverse medical imaging data such as CT scans, ultrasound images, X-rays, MRIs, and more, with a maximum caption length of 60 to enhance the learning of general medical knowledge.

For the downstream task of medical report generation, we fine-tune the model on four NVIDIA GeForce RTX 4090 GPUs across 30 epochs, continuing to use the AdamW optimizer with

TABLE I
PERFORMANCE COMPARISONS OF OUR PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE IU X-RAY DATASET WITH RESPECT TO NATURAL LANGUAGE GENERATION (NLG) METRICS

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Trainable Params
Show-Tell [29]	0.243	0.130	0.108	0.078	0.157	0.307	0.197	-
Transformer [33]	0.372	0.251	0.147	0.136	0.168	0.317	0.310	-
M^2Transformer [31]	0.402	0.284	0.168	0.143	0.170	0.328	0.332	-
CA [50]	0.492	0.314	0.222	0.169	0.193	0.381	-	-
R2Gen [2]	0.470	0.304	0.219	0.165	0.187	0.371	-	90.8M
CMN [1]	0.475	0.309	0.222	0.170	0.191	0.375	-	64.8M
RATCHET [38]	0.452	0.292	0.211	0.163	0.183	0.356	0.603	51M
AlignTransformer [5]	0.484	0.313	0.225	0.173	0.204	0.379	-	-
CMM+RL [51]	0.494	0.321	0.235	0.181	0.201	0.384	-	63M
METransformer [3]	0.483	0.322	0.228	0.172	0.192	0.380	0.435	152M
MedEPT-GPT2 [52]	0.468	-	-	0.167	-	0.364	0.397	-
MedEPT-OPT [52]	0.477	-	-	0.171	-	0.373	0.416	-
R2GenGPT(Shallow) [53]	0.466	0.301	0.211	0.156	0.202	0.370	0.405	4.2M
R2GenGPT(Delta) [53]	0.470	0.299	0.213	0.162	0.211	0.369	0.419	5.0M
R2GenGPT(Deep) [53]	0.488	0.316	0.228	0.173	0.211	0.377	0.438	90.9M
PPKED [40]	0.483	0.315	0.224	0.168	0.190	0.376	0.351	-
KERP [41]	0.482	0.325	0.226	0.162	-	0.339	0.280	
DCL [4]	-	-	-	0.163	0.193	0.383	0.586	410M
MGSK [6]	0.496	0.327	0.238	0.178	-	0.381	0.382	-
Clinical BERT [54]	0.495	0.330	0.231	0.170	0.209	0.376	0.432	102M
AHP	0.478	0.308	0.221	0.166	0.200	0.373	0.607	21M
AHP-Decoder	0.483	0.320	0.233	0.179	0.211	0.378	**0.684**	51M
AHP-Full	**0.502**	**0.338**	**0.250**	**0.196**	**0.212**	**0.388**	0.670	247M

The best results are highlighted in bold. BLEU-n denotes the BLEU score calculated using up to n-grams.

a weight decay of 0.05 and an initial learning rate of $1e-4$, which decays to $1e-5$ following a cosine schedule. Images are consistently cropped to a resolution of 224×224, as in pre-training.

To evaluate model performance, we employ natural language generation (NLG) metrics common in medical report generation research. These include BLEU-n [63], which measures the proportion of matching n-grams between the generated and reference reports, ROUGE-L [64], which assesses the longest common subsequence between the two reports, METEOR [65], which considers word-level and phrase-level similarities along with word order, and CIDEr [66], which uses a TF-IDF weighting scheme to evaluate the significance of various n-grams in captioning tasks. We also utilize BERTScore [67], a text evaluation method based on semantic similarity. These metrics collectively provide a comprehensive assessment of the text generation quality. In addition, we employ clinical metrics, including SembScore [68], which leverages semantic embedding techniques to calculate semantic similarity between reports from a medical domain perspective. RadCliQ-v1 [69] aims to more accurately predict the number of errors in radiology reports, thereby providing stronger alignment with radiologists' ratings. RadGraph [70] analyzes the consistency of key entities (e.g., anatomical structures and lesion types) and their relationships in the generated reports using information extraction techniques. Finally, RaTEScore [71] comprehensively evaluates the reports in terms of content completeness, clinical relevance, and linguistic fluency.

C. Performance Comparison

In this study, we compare our method against various established report generation models across three downstream datasets. We categorize our proposed AHP into three variants: **AHP**, **AHP-Decoder**, and **AHP-Full**, corresponding to fine-tuning only the adapter parameters, both the adapter and report decoder parameters, and all parameters, respectively. For image captioning, our benchmarks include Show-Tell [23], the Transformer [29], and the Meshed-Memory Transformer (M^2 Transformer) [27]. In the domain of medical report generation, we compare against models such as CA [48], R2Gen [2], CMN [1], AlignTrans [5], CMM+RL [49], METransformer [3], PPKED [38], DCL [4], MGSK [6], KERP [39], and Clinical BERT [52]. Notably, models like PPKED [38], DCL [4], MGSK [6], KiUT [41], KERP [39], and Clinical BERT [52] utilize clinical graph knowledge and pre-training to enhance performance by incorporating domain-specific knowledge. And also LLM-based models such as R2GenGPT [51], MedEPT [50], and BiomedGPT_IU [58], etc.

For the IU X-Ray dataset, as depicted in Table I, our proposed AHP achieves performance comparable to state-of-the-art (SOTA) methods with only 21 M trainable parameters. Specifically, our 21 M parameter AHP outperforms the 410 M parameter DCL model in BLEU-4 and METEOR scores and surpasses all SOTA methods in the CIDEr metric, indicating superior capture of critical disease information. Compared to methods like R2GenGPT and MedEPT, which also use parameter-efficient tuning, AHP achieves better performance with the same or even smaller parameter settings. Compared to AHP, both our AHP-Decoder and AHP-Full models show further performance improvements, with AHP-Full exceeding the performance of SOTA methods on nearly all metrics.

For the MIMIC-CXR dataset, as shown in Table II, our AHP-Decoder model remains highly competitive with only 51 M trainable parameters. Relative to methods like Clinical BERT that utilize clinical graph knowledge and pre-training, AHP achieves comparable performance with only one-fifth the parameters, while AHP-Decoder significantly surpasses their performance with half the parameters. Compared to the 63 M parameter CMM+RL method, our AHP-Decoder increases the

TABLE II
PERFORMANCE COMPARISONS OF OUR PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE MIMIC-CXR DATASET WITH RESPECT TO NATURAL LANGUAGE GENERATION (NLG) METRICS

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Trainable Params
Show-Tell [29]	0.308	0.190	0.125	0.088	0.122	0.256	0.096	-
Transformer [33]	0.316	0.199	0.140	0.092	0.129	0.267	0.134	-
M^2Transformer [31]	0.332	0.210	0.142	0.101	0.134	0.264	0.142	-
CA [50]	0.350	0.219	0.152	0.109	0.151	0.283	-	-
R2Gen [2]	0.353	0.218	0.145	0.103	0.142	0.277	-	90.8M
CMN [1]	0.353	0.218	0.148	0.106	0.142	0.278	-	64.8M
RATCHET [38]	0.326	0.205	0.139	0.099	0.136	0.280	0.140	51M
AlignTransformer [5]	0.378	0.235	0.156	0.112	**0.158**	0.283	-	-
CMM+RL [51]	0.381	0.232	0.155	0.109	0.151	0.287	-	63M
METransformer [3]	0.386	**0.250**	0.169	0.124	0.152	**0.291**	**0.362**	152M
RaDialog-INS [55]	0.340	-	-	0.097	0.136	0.270	-	-
RaDialog-RG [55]	0.346	-	-	0.095	0.140	0.271	-	-
MedEPT-GPT2 [52]	0.362	-	-	0.107	-	0.272	0.287	-
MedEPT-OPT [52]	0.368	-	-	0.113	-	0.282	0.338	-
PPKED [40]	0.360	0.224	0.149	0.106	0.149	0.284	0.237	-
DCL [4]	-	-	-	0.109	0.150	0.284	0.281	410M
MGSK [6]	0.363	0.228	0.156	0.115	-	0.284	0.203	-
Clinical BERT [54]	0.383	0.230	0.151	0.106	0.144	0.275	0.167	102M
AHP	0.349	0.220	0.150	0.109	0.136	0.280	0.155	21M
AHP-Decoder	0.384	0.240	0.163	0.119	0.148	0.284	0.182	51M
AHP-Full	**0.400**	**0.250**	**0.172**	**0.126**	0.154	0.285	0.169	247M

TABLE III
PERFORMANCE COMPARISONS OF OUR PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON CLINICAL METRICS AND LEXICAL METRICS

Dataset	Model	Years	Clinical Metrics				Lexical Metrics	Parameters
			1/RadCliQ-v1 ↑	SembScore ↑	RadGraph ↑	RaTEScore ↑	BERTScore ↑	
IU X-Ray	R2Gen [2]	2020	1.400	0.530	0.318	0.615	0.491	90.8M
	CMN [1]	2021	1.406	0.552	0.295	0.603	0.497	64.8M
	RATCHET [38]	2021	1.264	0.531	0.321	0.614	0.446	51M
	RaDialog [55]	2023	1.086	0.544	0.205	0.586	0.444	-
	RGRG [56]	2023	1.174	**0.602**	0.229	0.620	0.437	254.1M
	RadFM [57]	2023	1.187	0.566	0.230	0.627	0.459	14B
	LLM-CXR [58]	2024	0.486	0.057	0.023	0.280	0.186	3B
	GPT4V [59]	2024	0.708	0.405	0.146	0.517	0.274	-
	BiomedGPT_IU [60]	2024	0.956	0.522	0.213	0.543	0.375	182M
	AHP	2024	1.414	0.547	0.300	0.629	0.500	21M
	AHP-Decoder	2024	1.431	0.566	0.285	0.625	0.504	51M
	AHP-Full	2024	**1.626**	0.581	**0.338**	**0.653**	**0.507**	247M
MIMIC-CXR	R2Gen [2]	2020	0.679	0.331	0.176	0.478	0.396	90.8M
	CMN [1]	2021	0.679	0.342	0.174	0.488	0.399	64.8M
	RATCHET [38]	2021	0.809	0.338	0.162	0.485	0.392	51M
	RadFM [57]	2023	0.650	0.259	0.109	0.450	0.313	14B
	VLCI [61]	2023	0.680	0.305	0.140	0.450	0.304	69.41M
	RGRG [56]	2023	0.755	0.344	0.168	**0.491**	0.348	254.1M
	LLM-CXR [58]	2024	0.516	0.156	0.046	0.341	0.181	3B
	GPT4V [59]	2024	0.558	0.214	0.084	0.423	0.207	-
	BiomedGPT_IU [60]	2024	0.544	0.224	0.059	0.360	0.192	182M
	AHP	2024	0.785	0.322	0.163	0.482	0.395	21M
	AHP-Decoder	2024	0.812	0.334	0.178	0.487	0.403	51M
	AHP-Full	2024	**0.824**	**0.349**	**0.180**	**0.491**	**0.404**	247M

BLEU-4 score from 0.109 to 0.119. Although slightly trailing in some other metrics, the higher performance of CMM+RL is contingent upon using a reinforcement learning approach with a significantly larger search space, substantially increasing model training time under similar parameter conditions. Similarly, the multi-expert Transformer architecture proposed by METransformer improves performance but significantly increases the computational costs of training and inference, placing high demands on hardware resources. Compared to methods like RaDialog and MedEPT, which also employ parameter-efficient tuning, AHP still delivers more competitive predicted reports at the text matching level. This highlights the effectiveness of the spatial adapter in achieving parameter-efficient fine-tuning.

In addition to commonly used NLP metrics, we also employ clinical metrics and BERTScore to conduct comprehensive comparative experiments between the latest methods and AHP from different perspectives, as shown in Table III. The results reveal that, compared to other lightweight domain-specific models and generative large models, our proposed AHP-Full achieves a significantly superior SOTA performance across almost all clinical metrics. Meanwhile, AHP has also demonstrated highly competitive results. This validates that AHP not only leads in the textual aspects of report generation but also excels in clinical relevance, semantic structural consistency, and content completeness.

For the Bladder Pathology dataset, as illustrated in Table IV, our AHP leads SOTA methods significantly with only 21 M parameters, specifically improving the BLEU-4 metric by 25.1% and 13.8% compared to CMM+RL and R2Gen, respectively, and performing comparably on other metrics with SOTA methods that have many more parameters. This demonstrates that our AHP approach is particularly effective on smaller datasets, maintaining or even enhancing performance while significantly reducing parameter count. Furthermore, comparative analysis of our fully fine-tuned AHP-Full reveals that simply increasing parameter count does not significantly enhance performance on

TABLE IV
PERFORMANCE COMPARISONS OF OUR PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON THE BLADDER PATHOLOGY DATASET WITH RESPECT TO NATURAL LANGUAGE GENERATION (NLG) METRICS

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Trainable Params
R2Gen [2]	0.580	0.425	0.337	0.276	0.284	**0.538**	1.191	90.8M
CMN [1]	0.548	0.401	0.332	0.270	0.279	0.535	1.069	64.8M
CMM+RL [51]	-	-	-	0.251	0.268	0.517	0.992	63M
RATCHET [38]	0.567	0.420	0.332	0.273	0.290	0.530	1.147	51M
AHP	0.604	0.457	0.372	0.314	0.290	0.524	1.195	21M
AHP-Decoder	**0.610**	**0.459**	**0.373**	**0.316**	0.289	0.524	1.198	51M
AHP-Full	0.605	0.454	0.369	0.312	**0.292**	0.522	**1.206**	247M

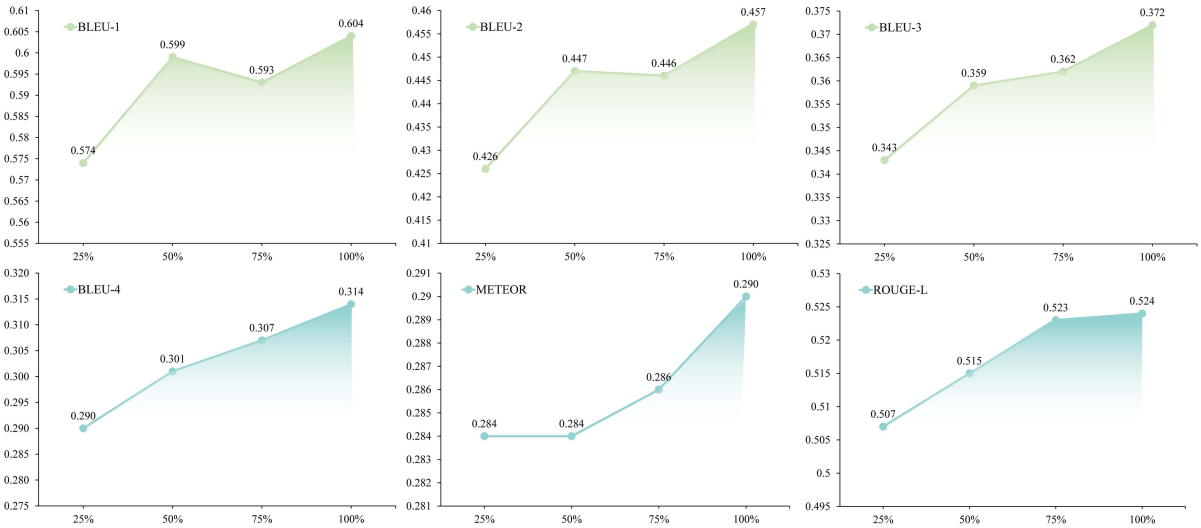


Fig. 4. Multi-sample experiments on the Bladder Pathology dataset using different data proportions for fine-tuning with our proposed AHP.

smaller datasets and may increase the risk of overfitting. To further validate the performance improvement of our proposed AHP on the Bladder Pathology dataset at a finer granularity, we conducted multi-sample experiments using data proportions of 25%, 50%, 75%, and 100%. As shown in Fig. 4, the changes in six evaluation metrics are reported. It can be observed that, as the data proportion increases, all metrics show an overall upward trend. Notably, the BLEU-4 score exhibits a nearly linear increase, indicating that larger dataset sizes contribute to improved AHP performance. Additionally, using only 25% of the data leads to a noticeable performance decline, while the performance stabilizes with a moderate reduction in data volume. This suggests that AHP is not highly sensitive to small fluctuations in dataset size, and model performance can be further enhanced through techniques such as data cleaning.

D. Ablation Studies

In this section, we conducted ablation studies on our proposed AHP using the IU X-Ray dataset to assess the contribution of each component in our model.

1) Contribution of Training Objectives: To explore the contribution of different training objectives to medical report generation, we investigate two additional training objectives—Image-Text Contrastive Loss (ITC) and Image-Text Matching Loss (ITM)—beyond the standard Language Modeling Loss (LM). We assess the effectiveness of these objectives in enhancing report generation capabilities.

As illustrated in Table V, both ITC and ITM significantly contribute to the model's performance. Notably, ITC appears to offer greater benefits by aligning text and image information more effectively, thereby enhancing feature representations. This alignment is particularly effective between words and image patches, as reflected by improved BLEU-1 scores, suggesting more accurate word generation. Conversely, while ITM may lead to lower BLEU-1 scores than Variant #1, possibly due to the complexity of matching tasks in longer medical reports, it substantially enhances report continuity. This improvement is evident from the higher BLEU-4 scores with ITM compared to those with LM+ITC (Variant #2). In conclusion, the incorporation of contrastive learning significantly influences the quality of medical report generation, with each training objective contributing uniquely to different aspects of the generated reports.

2) Contribution of Components on the Spatial Adapter: To evaluate the impact of each component within the adapter, we sequentially integrate the CNN, Injector, and Extractor into the baseline model. As depicted in Table VI, the inclusion of spatial features via CNN in Variant #2 results in a notable performance enhancement over BASE, evidenced by increases of 0.007 in BLEU-4, 0.012 in ROUGE-L, 0.007 in METEOR, and 0.083 in CIDEr. This enhancement underscores the significance of local information in the precise generation of medical reports. Further,

TABLE V
ABLATION STUDIES OF TRAINING OBJECTIVES IN IU X-RAY DATASET

#	Training Loss	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
1	LM	0.470	0.308	0.224	0.173	0.201	0.365	0.620
2	LM+ITC	0.493	0.326	0.235	0.180	0.211	0.381	0.640
3	LM+ITM	0.464	0.311	0.232	0.185	0.202	0.370	0.637
4	LM+ITC+ITM	**0.502**	**0.338**	**0.250**	**0.196**	**0.212**	**0.388**	**0.670**

TABLE VI
ABLATION STUDIES OF KEY COMPONENTS ON THE SPATIAL ADAPTER IN IU X-RAY

#	Component	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
1	BASE	0.471	0.309	0.222	0.168	0.200	0.368	0.486
2	+CNN	0.484	0.320	0.229	0.175	0.207	0.380	0.569
3	+Injector	0.491	0.321	0.235	0.183	0.207	0.378	0.639
4	+Extractor	**0.502**	**0.338**	**0.250**	**0.196**	**0.212**	**0.388**	**0.670**

TABLE VII
ABLATION STUDIES OF USING DIFFERENT PRE-TRAINING DATASETS ON IU X-RAY

#	ROCO	MedICaT	PMC-OA	BLEU-4	METEOR	ROUGE-L
1				0.175	0.192	0.378
2	✓			0.179	0.201	0.380
3		✓		0.178	0.208	0.381
4			✓	0.188	**0.214**	0.383
5	✓	✓		0.184	0.210	0.387
6	✓	✓	✓	**0.196**	0.212	**0.388**

TABLE VIII
ABLATION STUDIES OF STRUCTURALLY SIMILAR MODELS ON IU X-RAY

Model	Params	BLEU-4	METEOR	ROUGE-L
BLIP-2 w/ FlanT5$_{XL}$	103M	0.125	0.179	0.331
BLIP-2 w/ OPT$_{2.7B}$	104M	0.112	0.187	0.306
BLIP-2 w/ OPT$_{6.7B}$	108M	0.152	0.190	0.320
BLIP	129M	0.160	0.193	0.333
AHP	21M	0.166	0.200	0.373
AHP-Decoder	51M	**0.179**	**0.211**	**0.378**

TABLE IX
COMPARISON OF TIME AND COMPUTATIONAL EFFICIENCY DURING FINE-TUNING ON THE MIMIC-CXR DATASET

Methods	Training(h)	Inference(s)	GFLOPs	Memory(GB)
DCL [4]	2.24±0.010	627.98±68.37	38.86	2.39
BLIP [73]	1.155±0.003	689.47±27.20	27.53	1.60
AHP	1.015±0.008	560.01±44.20	29.87	1.72
w/o CNN	0.913±0.006	510.47±78.17	28.75	1.71
w/o Injector	0.909±0.005	506.15±7.42	29.22	1.70
w/o Extractor	0.822±0.002	538.52±55.08	28.98	1.67

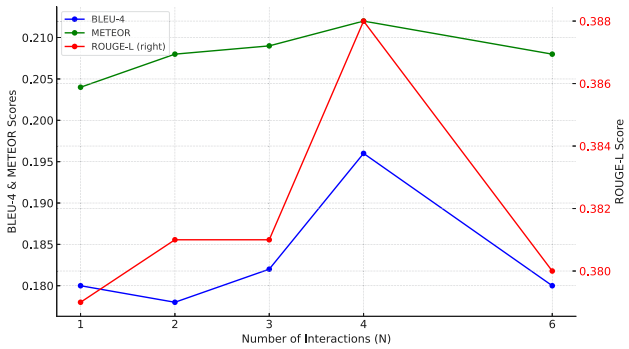


Fig. 5. Hyperparameter analysis of the number of interactions N in IU X-Ray.

the introduction of the Injector in Variant #3 and the subsequent addition of the Extractor in Variant #4 lead to substantial gains across all evaluated metrics, highlighting their combined effectiveness in promoting the integration and utilization of local and global information. Collectively, these findings demonstrate that the three components significantly enhance the capability of the model in medical report generation.

3) Impact of Pre-Training Dataset Selection: To evaluate the impact of different pre-training datasets (ROCO [14], MedICaT [15], PMC-OA [16]), we perform an ablation study focusing on the pre-training phase. Table VII presents our findings that PMC-OA is essential for pre-training, primarily because it provides a significantly larger data volume compared to ROCO and MedICaT, highlighting the importance of data volume in enhancing model performance. Despite the greater total volume of data in PMC-OA, performance measured by the ROUGE-L metric is inferior to that achieved when pre-training with ROCO and MedICaT combined. We attribute this to the content of our test dataset, which consists of radiology images, whereas ROCO and MedICaT predominantly include radiology images, but PMC-OA encompasses a broader array of medical data types, leading to diminished performance in metrics.

4) Comparison of Structurally Similar Models: In Table VIII, we present a comparative analysis of our AHP with various structurally similar models, including BLIP-2 [73]. In the case of BLIP-2, we utilized different large-scale language models such as OPT [74] and FlanT5 [75], while training solely the Q-former module and maintaining the rest of the modules in a frozen state. Specifically for FlanT5, the instruction was configured for "medical report generation." Notably, in the domain of medical report generation, BLIP-2 exhibited inferior performance compared to its predecessor BLIP, despite the employment of large language models. We attribute this underperformance to the substantial discrepancies between general image-text pairings and specialized medical image-report associations, which likely impeded the model's effectiveness.

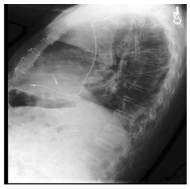
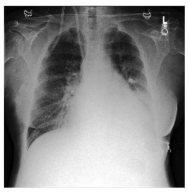
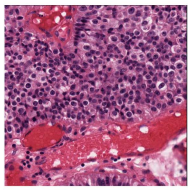
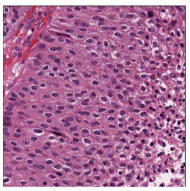
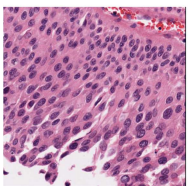
	Input Images	Reference Reports	R2Gen	CMN	Our AHP
(a)		the patient is status post median sternotomy and cabg . left-sided dual-chamber pacemaker is noted with leads terminating in the right atrium and right ventricle . there is mild enlargement of the cardiac silhouette which is stable . the aorta remains tortuous . there is mild pulmonary edema and a small right pleural effusion . previously noted left pleural effusion is not clearly seen on the current study . patchy ill-defined opacity in the right base persists and is likely due to atelectasis though infection cannot be excluded . there is no pneumothorax . no acute osseous abnormalities are present.	dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle . no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are stable with the aorta calcified and tortuous and the cardiac silhouette top normal to mildly enlarged . no overt pulmonary edema is seen .	left-sided dual-chamber pacemaker device is noted with leads terminating in the right atrium and right ventricle . mild enlargement of the cardiac silhouette is unchanged . the aorta is diffusely calcified . mediastinal and hilar contours are similar . pulmonary vasculature is not engorged . lungs are hyperinflated with emphysematous changes again noted most pronounced in the lung apices . no focal consolidation pleural effusion or pneumothorax is present . mild degenerative changes are noted in the thoracic spine .	dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle . patient is status post median sternotomy and cabg . there is a small left pleural effusion with overlying atelectasis . left base retrocardiac opacity most likely represents combination of pleural effusion and atelectasis however underlying consolidation is difficult to exclude . the right lung is clear . the cardiac silhouette remains mildly enlarged . the aortic knob is calcified .
(b)		there has been interval placement of a large-bore dual-lumen right central venous catheter distal aspect not well seen but likely terminating at the cavoatrial junctionproximal right atrium . the cardiac silhouette is mildly enlarged . there is a left base opacity likely represents combination of pleural effusion and atelectasis . there is a moderate pulmonary vascular congestion . no pneumothorax seen .	the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable .	right internal jugular central venous catheter tip terminates in the proximal right atrium . moderate enlargement of cardiac silhouette is re- demonstrated . the mediastinal contour is unchanged . there is mild pulmonary vascular congestion . patchy opacities in the lung bases likely reflect areas of atelectasis . no large pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .	there has been interval placement of a right internal jugular central venous catheter which terminates at the cavoatrial junction without evidence of pneumothorax . there are low lung volumes . left mid to lower lung linear atelectasisscarring is seen . there is slight blunting of the left costophrenic angle which may be due to a trace pleural effusion . the cardiac silhouette is top-normal to mildly enlarged . mediastinal contours are unremarkable . no overt pulmonary edema is seen .
(c)		swan-ganz catheter has been removed and a right-sided port-a-cath is noted with tip in the lower svc . consolidative opacity within the right lower lobe is concerning for pneumonia . there is elevation of the right hemidiaphragm with lateralization of the diaphragmatic peak suggesting a subpulmonic effusion . the cardiac silhouette size is top normal . there is mild prominence of the pulmonary vascular markings . no left-sided pleural effusion is seen and there is no pneumothorax . there are no acute osseous abnormalities .	frontal and lateral views of the chest were obtained . right-sided port-a-cath is again seen terminating in the low svc . there are low lung volumes . there is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema versus atypical infection . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are stable .	right-sided port-a-cath tip terminates at the junction of the svc and right atrium . lung volumes are low . heart size is mildly enlarged . the aorta is tortuous . there is crowding of the bronchovascular structures without overt pulmonary edema . patchy opacities in the lung bases likely reflect areas of atelectasis . no pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .	right-sided port-a-cath tip terminates at the junction of the svc and right atrium . heart size is mildly enlarged . mediastinal and hilar contours are unremarkable . pulmonary vasculature is not engorged . elevation of the right hemidiaphragm is chronic . patchy opacities in the lung bases likely reflect areas of atelectasis . no focal consolidation pleural effusion or pneumothorax is present . there are no acute osseous abnormalities .
(d)		mild pleomorphism is present . nuclei are moderately crowded together . polarity is partially lost . mitosis is infrequent throughout the tissue . prominent nucleoli are easily identified in low magnification scanning . high grade .	mild pleomorphism and cytologic atypia is present . the nuclei are crowded to a mild degree . architecturally no overt lack of polarity is observed . there are infrequent mitotic figures throughout the tissue . visible and prominent nucleoli are not observed or exceedingly rare . low grade .	mild pleomorphism is present . nuclei are mildly crowded together . insufficient information . insufficient information . the nuclei have inconspicuous nucleoli . low grade .	nuclear features show moderate pleomorphism . mild crowding of the nuclei can be seen . architecturally the cells show complete lack of polarity toward the surface urothelium . mitosis is infrequent throughout the tissue . the nucleoli of nuclei are prominent . high grade .
(e)		mild pleomorphism and cytologic atypia is present . pictured nuclei exhibit normal crowding . the urothelium shows some degree of loss of polarity . mitosis is rare throughout the tissue . visible and prominent nucleoli are not observed or exceedingly rare . low grade .	mild pleomorphism is present . nuclei are mildly crowded together . there is no full-thickness lack of polarity observed . mitosis appears to be rare . the nuclei have inconspicuous nucleoli . low grade .	slight variability in nuclear size shape and outline consistent with mild pleomorphism . mild nuclear crowding is seen . polarity is partially lost . mitosis is infrequent . the nuclei have inconspicuous nucleoli . low grade .	mild pleomorphism and cytologic atypia is present . there is a mild degree of crowding . polarity is not completely lost toward the surface urothelium . mitosis is rare throughout the tissue . visible and prominent nucleoli are not observed or exceedingly rare . low grade .
(f)		nuclear features show moderate pleomorphism . mild nuclear crowding is seen . architecturally no overt lack of polarity is observed . mitosis is rare throughout the tissue . visible and prominent nucleoli are not observed or exceedingly rare . low grade .	slight variability in nuclear size shape and outline consistent with mild pleomorphism . there is a mild degree of crowding . polarity is not completely lost toward the surface urothelium . mitosis is rare throughout the tissue . nucleoli is absent to inconspicuous . low grade .	mild pleomorphism and cytologic atypia is present . mild nuclear crowding is seen . polarity is partially lost . mitotic figures including the atypical forms are frequently seen in all levels of the urothelium . the nuclei have inconspicuous nucleoli . high grade .	mild pleomorphism is present . nuclei are mildly crowded together . the urothelium shows some degree of loss of polarity . mitosis is rare throughout the tissue . prominent nucleoli are easily identified in low magnification scanning . low grade .

Fig. 6. Illustrations of reports from ground truth, R2Gen, CMN, and our AHP for six samples from the MIMIC-CXR and Bladder Pathology datasets. Content closely aligned with the reference report is highlighted in blue, while inaccuracies are marked in red.

Our observations from the BLIP series variants indicate that our AHP achieves state-of-the-art results in report generation using only one-fifth the parameter count of comparable models. Further fine-tuning of the report decoder on this framework could enhance performance significantly. Remarkably, even with these adjustments, the AHP-Decoder requires only half the parameter volume compared to other architectures.

5) Computational Efficiency Analysis: We further present a comprehensive analysis of the training and inference speed with memory consumption and computational cost of the AHP under different configurations. We quantify the computational demand on a per-sample basis using gigaflops (GFLOPs) and represent memory consumption in gigabytes. Additionally, we also provide the total training hours for the complete training

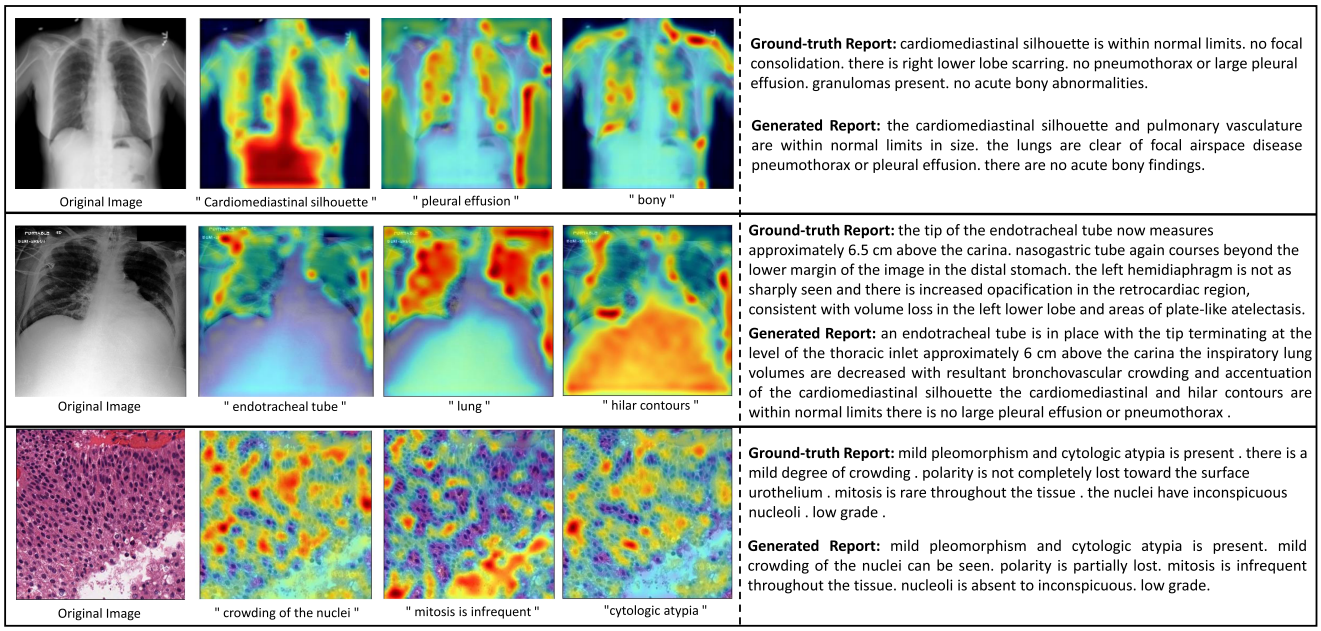


Fig. 7. Examples of attention map visualizations for key medical terms in the generated reports are provided, where more vibrant colors indicate greater attention to specific regions.

on the MIMIC-CXR dataset, as well as the total inference time for generating all reports in one round. We compare our model against the baseline BLIP model and a variant, DCL, which is also based on BLIP. As observed from Table IX, our model demonstrates superior efficiency in both GFLOPs and memory consumption compared to DCL. Notably, memory consumption, as a measure of GPU memory usage, includes the base model's memory requirements indiscriminately, making it less flexible in evaluating the model's performance. Our proposed AHP increases computational cost and memory consumption only marginally compared to BLIP, yet significantly reduces both training and inference time. This highlights the effectiveness of using the Spatial Adapter in AHP, enabling the model to leverage task-specific hints with minimal additional training parameters while maintaining throughput and computational efficiency.

E. Hyperparameter Analysis of N

In this study, the symbol N denotes both the number of interactions between the ViT branch and the spatial adapter branch, and the number of blocks within the ViT. To maintain a constant total of 12 layers, the product of the number of blocks N and the number of layers per block is kept fixed. To evaluate the impact of different numbers of interactions, we experimented with N values from the set $\{1, 2, 3, 4, 6\}$. As shown in Fig. 5, it indicates that increasing the number of interactions generally enhances performance. However, setting $N = 2$ causes a reduction in BLEU-4 and CIDEr scores. This outcome suggests that two interactions may prioritize accuracy at the expense of report continuity, affecting the fluidity of longer reports. Moreover, increasing the number of interactions beyond a certain threshold does not consistently yield performance gains. Consequently, we select $N = 4$ as the optimal number of interactions for our model, balancing performance and efficiency.

F. Case Studies and Visualization

In Fig. 6, we present report generation examples for chest X-ray and bladder pathology images, comparing our results with those from models released by R2Gen and CMN. As illustrated in Fig. 6(a), (b), and (c), our model effectively generates reports identifying abnormalities in chest X-ray images, unlike the comparative models which either omit or inaccurately describe these abnormalities. For bladder pathology images, as shown in Fig. 6(d), (e), and (f), our model produces reports closely resembling the reference reports. For instance, in Fig. 6(a), our model accurately generates phrases such as "dual-lead left-sided pacemaker" in "right atrium and right ventricle," and notes "the cardiac silhouette remains mildly enlarged" and "patient is status post median sternotomy and cagb." However, there are discrepancies, such as in Fig. 6(c) where our model incorrectly reports "heart size is mildly enlarged" whereas the reference states "the cardiac silhouette size is top normal." This indicates that while our model is proficient in localizing specific anatomical features, accurately assessing the severity of minor abnormalities remains a challenge.

To improve interpretability, further evaluation involves visualizing image-text attention maps for randomly selected images from the IU-Xray, MIMIC-CXR, and Bladder Pathology test sets. As depicted in Fig. 7, these maps demonstrate effective alignment of our model's focus areas with the indicated abnormalities, where attention intensities range from blue (low) to red (high). Notably, our model precisely locates the "endotracheal tube" in the second example. These results underscore

our model's enhanced capability in medical report generation and improved image-report alignment. Nevertheless, in conditions like pleural effusion and hilar boundaries, while accurate localization is achieved, some inaccuracies in lung focus are observed. This might be attributed to the broad scope of attention required, leading to minor errors in depicting edge details.

V. Conclusion

In this paper, we focus on reducing model size and propose a lightweight solution: an Adapter-enhanced Hierarchical cross-modal Pre-training model (AHP) for medical report generation. The AHP utilizes spatial adapters to augment the extraction of spatially hierarchical visual features and employs multiple cross-modal matching tasks to enhance the alignment between image and text modalities. To further address the issue of inadequate spatial detail representation in traditional ViT architectures, we integrate a convolutional backbone to extract local spatial features, which are then combined with multi-scale global features through hierarchical injectors and extractors, thus providing a richer and more comprehensive feature representation. Extensive experiments and analyses on the IU X-Ray, MIMIC-CXR, and Bladder Pathology datasets demonstrate that our AHP generates reports comparable to or surpassing state-of-the-art methods with significantly fewer parameters. Although the effectiveness of AHP has been validated on a broad spectrum of medical imaging and pathological datasets, certain modalities, such as PET and endoscopy, have not been fully explored. Future work will focus on extending its application to these areas for comprehensive coverage. Furthermore, we plan to integrate deeper medical knowledge, combining explicit and implicit knowledge to enhance the interpretability of our framework, thereby further advancing our capabilities in intelligent healthcare systems.

References

[1] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5904–5914.

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

[3] Z. Wang, L. Liu, L. Wang, and L. Zhou, "MetransFormer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11558–11567.

[4] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 3334–3343.

[5] 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. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 72–82.

[6] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, 2022, Art. no. 102510.

[7] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–22.

[8] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 68–80.

[9] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[10] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[11] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[12] W. Wang et al., "PVT V2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.

[13] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

[14] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (ROCO): A multimodal image dataset," in *Proc. Intravascular Imag. Comput. Assist. Stenting Large-Scale Annotation Biomed. Data Expert Label Synth.*, 2018, pp. 180–189.

[15] S. Subramanian et al., "MedICaT: A dataset of medical images, captions, and textual references," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 2112–2120.

[16] W. Lin et al., "PMC-Clip: Contrastive language-image pre-training using biomedical documents," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2023, pp. 525–536.

[17] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proc. Mach. Learn. Healthcare Conf.*, 2022, pp. 2–25.

[18] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3942–3951.

[19] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 33536–33549.

[20] P. Cheng, L. Lin, J. Lyu, Y. Huang, W. Luo, and X. Tang, "Prior: Prototype representation joint learning from medical images and reports," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21361–21371.

[21] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 4542.

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

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[24] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[25] 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.*, 2017, pp. 375–383.

[26] J. Ji et al., "Improving image captioning by leveraging intra-and inter-layer global representation in transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1655–1663.

[27] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10575–10584.

[28] 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.*, 2020, pp. 10971–10980.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[30] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 11137–11147.

[31] Y. Luo et al., "Dual-level collaborative transformer for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2286–2293.

[32] J. Luo et al., "Semantic-conditional diffusion networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23359–23368.

[33] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, "Smallcap: Lightweight image captioning prompted with retrieval augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2840–2849.

[34] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31, pp. 1537–1547.
[35] G. Liu et al., "Clinically accurate chest X-ray report generation," in *Proc. Mach. Learn. Healthcare Conf.*, 2019, pp. 249–269.
[36] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: Medical transformer for chest X-ray diagnosis and reporting," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, vol. 12907, pp. 293–303.
[37] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12910–12917.
[38] 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.*, 2021, pp. 13753–13762.
[39] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6666–6673.
[40] K. Zhang et al., "Semi-supervised medical report generation via graph-guided hybrid feature consistency," *IEEE Trans. Multimedia*, vol. 26, pp. 904–915, 2024.
[41] Z. Huang, X. Zhang, and S. Zhang, "KIUT: Knowledge-injected u-transformer for radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19809–19818.
[42] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[43] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[44] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 9694–9705.
[45] 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, 2016.
[46] A. E. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, 2019, Art. no. 317.
[47] Z. Zhang et al., "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 236–245, 2019.
[48] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest X-ray report generation," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 269–280.
[49] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Proc. Findings Assoc. Comput. Linguistics*, Dublin, Ireland, May 2022, pp. 448–458.
[50] Q. Li, "Harnessing the power of pre-trained vision-language models for efficient medical report generation," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, 2023, pp. 1308–1317.
[51] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," *Meta- Radiol.*, vol. 1, no. 3, 2023, Art. no. 100033.
[52] B. Yan and M. Pei, "Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 2982–2990.
[53] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher, "Radialog: A large vision-language model for radiology report generation and conversational assistance," 2023, *arXiv:2311.18681*.
[54] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7433–7442.
[55] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology," 2023, *arXiv:2308.02463*.
[56] S. Lee, W. J. Kim, J. Chang, and J. C. Ye, "LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation," in *Proc. 12th Int. Conf. Learn. Representations*, 2024.
[57] Z. Yang et al., "The dawn of LMMs: Preliminary explorations with GPT-4V (ision)," 2023, *arXiv:2309.17421*.
[58] K. Zhang et al., "A generalist vision–language foundation model for diverse biomedical tasks," *Nature Med.*, vol. 30, pp. 3129–3141, 2024.
[59] W. Chen et al., "Cross-modal causal intervention for medical report generation," 2023, *arXiv:2303.09117*.
[60] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.
[61] 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.*, 2009, pp. 248–255.
[62] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations*, 2019, pp. 1–19.
[63] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[64] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out, Post-Conf. Workshop*, Jul. 2004, pp. 74–81.
[65] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. Second Workshop Stat. Mach. Transl.*, 2007, pp. 228–231.
[66] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[67] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.
[68] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "ChexBERT: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," 2020, *arXiv:2004.09167*.
[69] F. Yu et al., "Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, 2023, Art. no. 100802.
[70] S. Jain et al., "Radgraph: Extracting clinical entities and relations from radiology reports," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 1)*, 2021.
[71] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Ratescore: A metric for radiology report generation," in *Proc. 2024 Conf. Emp. Methods Natural Lang. Process.*, 2024, pp. 15004–15019.
[72] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.
[73] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
[74] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.
[75] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.