Attribute Prototype-Guided Iterative Scene Graph for Explainable Radiology Report Generation

Ke Zhang, *Member, IEEE*, Yan Yang, *Member, IEEE*, Jun Yu, *Senior Member, IEEE*, Jianping Fan, Hanliang Jiang, Qingming Huang, *Fellow, IEEE*, and Weidong Han

***Abstract*— The potential of automated radiology report generation in alleviating the time-consuming tasks of radiologists is increasingly being recognized in medical practice. Existing report generation methods have evolved from using image-level features to the latest approach of utilizing anatomical regions, significantly enhancing interpretability. However, directly and simplistically using region features for report generation compromises the capability of relation reasoning and overlooks the common attributes potentially shared across regions. To address these limitations, we propose a novel region-based Attribute Prototype-guided Iterative Scene Graph generation framework (AP-ISG) for report generation, utilizing scene graph generation as an auxiliary task to further enhance interpretability and relational reasoning capability. The core components of AP-ISG are the Iterative Scene Graph Generation (ISGG) module and the Attribute Prototype-guided Learning (APL) module. Specifically, ISSG employs an autoregressive scheme for structural edge reasoning and a contextualization mechanism for relational reasoning. APL enhances intra-prototype matching and reduces inter-prototype semantic overlap in the visual space to fully model the potential attribute commonalities among regions. Extensive experiments on the MIMIC-CXR with Chest ImaGenome datasets demonstrate the superiority of AP-ISG across multiple metrics.**

Manuscript received 9 April 2024; revised 11 June 2024; accepted 3 July 2024. Date of publication 8 July 2024; date of current version 2 December 2024. This work was supported by the National Natural Science Foundation of China under Grant 62125201, Grant 62020106007, and Grant 62176230. *(Corresponding authors: Weidong Han; Jun Yu.)*

Ke Zhang and Yan Yang 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: ke.zhang@hdu.edu.cn; yangyan@hdu.edu.cn).

Jun Yu is with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China, and also with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yujun@hit.edu.cn).

Jianping Fan is with the AI Laboratory, Lenovo Research, Beijing 100094, China (e-mail: jfan1@Lenovo.com).

Hanliang Jiang is with the Regional Medical Center, National Institute of Respiratory Diseases, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou 310016, China (e-mail: aock@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).

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).

Digital Object Identifier 10.1109/TMI.2024.3424505

***Index Terms*— Radiology report generation, scene graph generation, prototype learning, interpretability.**

I. Introduction

RADIOGRAPHIC X-ray chest imaging, one of the most common medical imaging modalities, efficiently detects basic lesions in organs like the lungs, mediastinum, and heart with low radiation exposure [1], [2], [3]. In routine practice, radiologists examine these images to provide detailed descriptions of normal or abnormal findings in various anatomical regions. However, manually diagnosing the increasing volume of radiographic images is time-consuming and burdensome. Therefore, automatic report generation has gained increasing attention due to its potential to optimize diagnostic processes [4], [5]. Automatic radiology report generation is significantly more challenging than natural image captioning, as the reports contain 4-8 times more sentences and require detailed descriptions of more complex semantic relationships and their corresponding regions within the image. Inspired by the success in image captioning, recent studies usually employ Transformer [6] as the backbone due to its long-range dependency modeling capabilities through self-attention mechanisms. Unlike previous methods that only utilize global image features for report generation, Tanida et al. [7] is the first to propose visually grounding report sentences to local anatomical regions, significantly enhancing report interpretability. However, direct and simplistic input of regional features for report generation fails to fully exploit the potential of regions in enhancing the generation process. This not only results in suboptimal reasoning capability for semantic relationships within images but also overlooks potential attribute commonalities among regions.

Differing from straightforward modeling of regional features, we propose a novel region-based Attribute Prototype-guided Iterative Scene Graph generation framework (AP-ISG) for report generation, leveraging scene graph generation as an auxiliary task to enhance the relational reasoning capabilities. Particularly, integrating scene graph generation into report generation is unprecedented in prior research. Additionally, we introduce an Attribute Prototype-Guided Learning approach to fully explore potential attribute commonalities among regions. Specifically, this involves using intra-prototype

matching to establish tighter connections between region pairs and attributes, and inter-prototype regularization to minimize semantic overlap among attribute prototypes. Moreover, the discovered common attributes between regions also serve as supplementary nodes in scene graph generation to enrich its content. Extensive experiments demonstrate that our AP-ISG achieves optimal performance across multiple metrics. The main contributions of our work are summarized as:

- We propose a scene graph augmented region-based report generation method, utilizing scene graph generation as an auxiliary task to enhance interpretability and relational reasoning capabilities. To the best of our knowledge, this is the first framework to combine scene graph generation into report generation process.
- We introduce an Iterative Scene Graph Generation module to reason semantic relationships in images for report generation assistance, using autoregressive generation for structural reasoning of connected edges and a contextualization mechanism for relational reasoning.
- We propose an Attribute Prototype-Guided Learning module to fully explore potential attribute commonalities among region pairs by enhancing intra-prototype matching and reducing inter-prototype semantic overlap in the visual space.
- Extensive quantitative comparisons across various metrics and in-depth qualitative experiments demonstrate that our AP-ISG outperforms other state-of-the-art report generation models.

II. Related Works

A. Medical Report Generation

In the field of automated medical report generation, architectures can be categorized into RNN-based and Transformer-based approaches [8]. RNN-based methods have limitations in adequately modeling dependencies in long sequences, leading to the increasing predominance of Transformer architectures. R2Gen [9] was the first to adapt the Transformer for medical report generation, introducing the concept of a memory-driven transformer. Building upon [9], R2GenCMN [10] further incorporated a shared memory to capture cross-modal alignment details. PPKED [11] emulated the workflow of radiologists by integrating report retrieval with medical prior knowledge. Inspired by curriculum learning, Nooralahzadeh et al. [12] proposed a two-phase report generation process as initial rough generation followed by detailed refinement. Liu et al. [13] contrasted the current image with normal images to extract differential information. AlignTransformer [14] learned multi-granularity representations by aligning visual regions with predicted disease labels. To improve inconsistent report generation, Miura et al. [15] applied reinforcement learning to optimize rewards for consistency. DeltaNet [16] uses a step-by-step retrieval of historical reports and images for conditional report generation, while Yang et al. [17] integrates an automatic knowledge base and multimodal semantic alignment to facilitate report generation. Recognizing the highly structured nature and the extreme data imbalance of medical reports, ITA [18] proposed a task-aware framework turning image-level reports into structure-level descriptions. Subsequently, Nicolson et al. [19] evaluated numerous encoder-decoder combinations, ultimately selecting the best checkpoint combination for warm starting. RAMT-U [8] focused on the issue of data scarcity, taking the lead in applying semi-supervised learning to report generation with a relation-aware mean teacher framework while UAR [20] further introduced a "unify, align, and refine" approach to learn multi-level cross-modal alignment. Additionally, KiUT [21] presented a Knowledge-injected U-Transformer for multi-level visual representations, adaptively extracting contextually and clinically relevant information for word prediction. METransformer [22] incorporated multiple learnable expert tokens into the transformer encoder and decoder, further developing a metrics-based expert voting strategy. Diverging from methods focusing on specific cross-modal alignment or report optimization, RGRG [7] abandoned the traditional reliance on global image features, proposing a truly anatomy region-guided approach for interpretable report generation. Our AP-ISG builds on [7] and introduces scene graph generation into report generation for the first time, further exploring attribute commonalities between regions and enhancing the capability for interpretable relation reasoning.

B. Scene Graph Generation

Scene graph generation was first introduced by [23] with the aim of constructing graph-based representations that capture the rich semantic structure of scenes by modeling objects and relationships among them. Most scene graph generation methods follow a typical workflow: object detection, followed by pairwise interaction modeling to generate relationship triplets representing the edges of the scene graph. In recent years, various models have been proposed from different perspectives to address the scene graph generation task. Early approaches [24] attempted to detect entities and relationships using separate networks, often overlooking the wealth of contextual information. Subsequently, [25] demonstrated that context information can significantly improve relationship prediction and introduced an iterative message-passing mechanism to refine the representations of entities and relationships. Motifs [26] further emphasized the importance of context information among objects and employed BiLSTM to encode extensive contextual information between entities and edges. Additionally, to mitigate the influence of noise, VCTree [27] and Graph R-CNN [28] designed sparse structures to enhance the model's ability for contextual modeling. RU-Net [29] further employed graph regularization to suppress spurious connections between nodes and enhance the diversity of relationship prediction through rank maximization. Scene graph generation has also been widely applied in captioning tasks. Li and Jiang [30] proposed using scene graphs to structurally model visual features and semantic knowledge, while the ASG2Caption [31] model utilized abstract scene graph structure to represent user intention with fine granularity and guide the caption generation process. SGAE [32] was proposed to incorporate the language inductive bias into the encoder-decoder image captioning framework for more

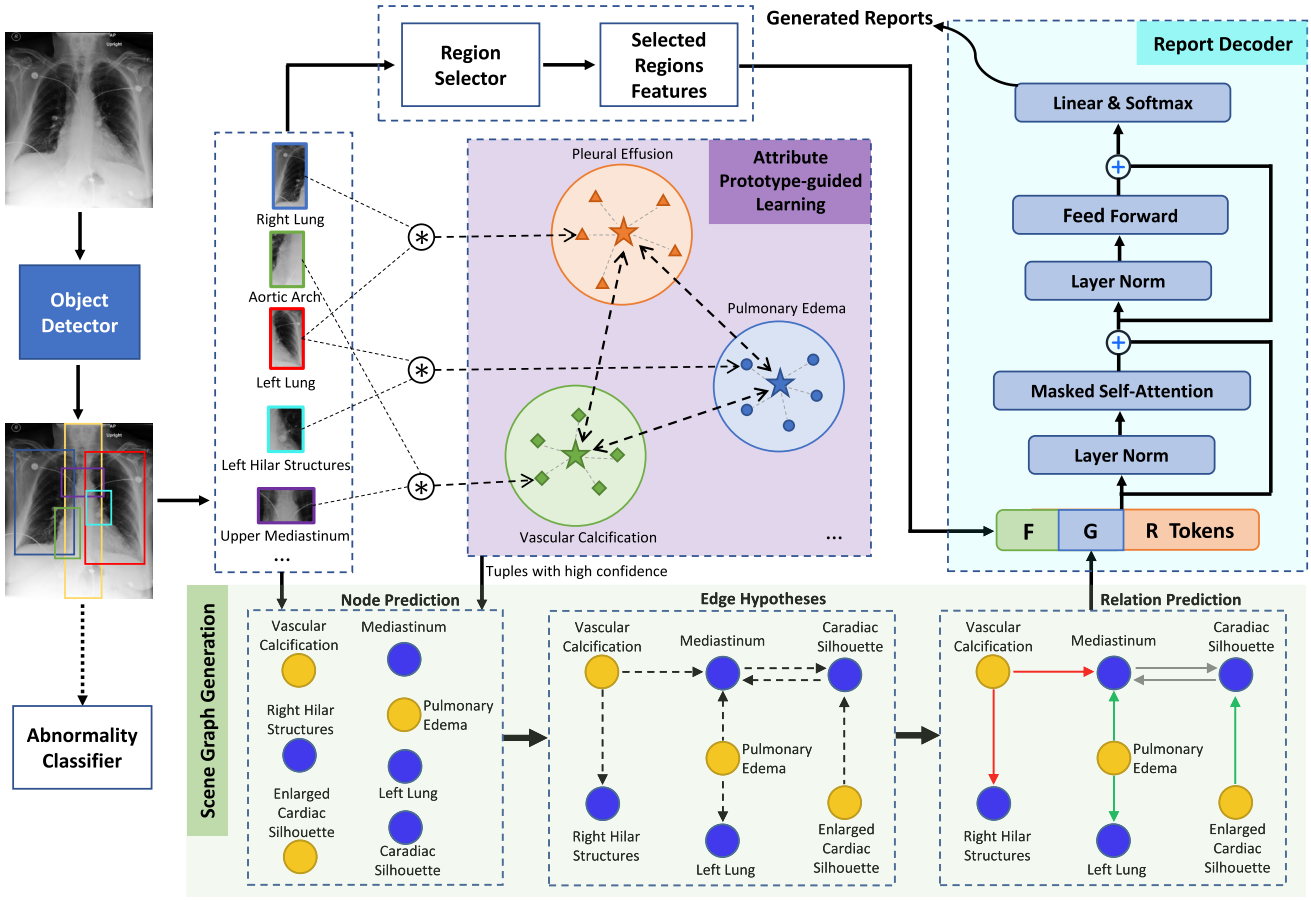


Fig. 1. Overview of our proposed AP-ISG framework. The Attribute Prototype-guided Learning (APL) module takes fused anatomical regions output by Object Detector as inputs and outputs triplets formed from attribute prototypes exceeding a certain matching threshold. The Iterative Scene Graph Generation (ISSG) module takes both the anatomical regions and the triplets outputted by the APL module as input, generating the scene graph through three distinct stages. ⊛ denotes the feature fusion process between regions. Pentagrams in APL denote attribute prototypes.

human-like captions. Furthermore, Lu and Gao [33] proposed the utilization of scene graphs for guidance and interaction to address the issue of hallucination in text generation. Our approach assists in scene graph generation by improving intra-prototype matching and reducing the semantic overlap between inter-prototypes in the visual space, thereby facilitating a more effective exploration of attribute commonalities.

III. Method

In this section, we elucidate the technical details of our proposed method. As illustrated in Fig. 1, the entire framework comprises components, including the Object Detector, Region Selector, Abnormality Classifier, Attribute Prototype-guided Learning combined with Iterative Scene Graph Generation, and the Report Generation Decoder. Subsequent subsections will provide a comprehensive exposition of these modules.

A. Object Detection

We use a pre-trained Faster R-CNN as the object detector for experiments, and extract region proposals from image features with Region Proposal Network. We follow the settings of [7] to extract anatomical region categories and region features $f \in \mathcal{R}^{29\times1024}$. We further define the i-th bounding box coordinate and the predicted class labels as $b_i \in [0, 1]^4$ and $l_i \in \mathcal{C}$.

As depicted in Fig. 1, the Region Selector and Abnormality Classifier utilize the region features f outputted by the Object Detector, and separately employ binary classification to discern whether a report sentence should be generated for the current region proposal or if the current region proposal is abnormal. Specifically, during the inference process, we initially employ two independent Multi-Layer Perceptrons (MLPs) to reduce the dimensionality of region features $f \in \mathcal{R}^{29\times1024}$ to 1, yielding 29 logit values. Each logit is then compared with a threshold (0.269 in probability space) to determine whether this region is selected for sentence generation or to assess if the region is abnormal. Throughout the training phase, we compute the Binary Cross-Entropy (BCE) loss only for the locations corresponding to regions detected by the object detector.

B. Attribute Prototype-Guided Learning

To address the issue of insufficient extraction of potential attribute commonalities between regions, we propose an Attribute Prototype-guided Learning (APL) scheme to establish a more compact intra-prototype matching between region pairs and their corresponding attributes, while utilizing inter-prototype regularization to reduce the semantic overlap between attribute prototypes. Firstly, we use predefined

attributes from Chest ImaGenome dataset [34] as prototypes, e.g., diseases, anatomical findings, and tubesandlines, etc. Then, we employ Glove [35] to obtain the class label's word embedding by processing class-specific semantic prototypes, and its prototype embedding can be defined as E_P^j. Each prototype represents a specific category of attribute related to abnormality. We use linear transformation to transform prototype embedding from semantic space to visual space $p_j = W_p \cdot E_P^j$, W_p represents learnable spatial mapping parameters. We assume that there are potential attribute commonalities between image regions. In order to sufficiently extract attribute commonalities between two regions, we first apply the Hadamard product to fuse two region features, and the fused feature vector is represented as $C_{i,k} = f_i \circledast f_k$. We hope to select a prototype p_u that matches $C_{i,k}$ as the most similar and representative attribute to represent the commonality of attributes between region i and region k. Cosine similarity is applied to select the closest prototype as follows:

$$u = \arg\max_j \frac{C_{i,k} \cdot p_j}{||C_{i,k}|| \cdot ||p_j||}. \tag{1}$$

If the value of $\frac{C_{i,k} \cdot p_j}{||C_{i,k}|| \cdot ||p_j||}$ is greater than the pre-defined similarity threshold τ, the triplets (f_i, f_k, p_u) will be delivered as auxiliary information for the scene graph generation in the next subsection. Otherwise, they will be left blank.

1) Intra-Prototype Matching: In order to effectively assist in matching commonalities between region pairs and prototypes, we design loss functions from the perspectives of cosine similarity and Euclidean distance. Firstly, we need to increase the cosine similarity between the regional commonality and its corresponding prototype, and the cosine similarity loss function is defined as:

$$\mathcal{L}_{Intra}^{cos} = -\log \frac{\exp(\langle C_{i,k}, p_u \rangle / T)}{\sum_{j=1}^{m} \exp(\langle C_{i,k}, p_j \rangle / T)}, \tag{2}$$

where T is a learnable temperature hyperparameter. Due to the fact that cosine similarity only measures the spatial angle similarity between vectors, we additionally introduce Euclidean distance to constrain its spatial distance $\|C_{i,k} - p_u\|_2^2$, it encourages narrowing the Euclidean distance between regional commonalities and its corresponding prototype, and widening the Euclidean distance with other prototypes. To implement this, we have constructed a Triplet loss as:

$$\mathcal{L}_{Intra}^{euc} = \max(0, \|C_{i,k} - p_u\|_2^2 - \frac{1}{m-1} \sum_{\substack{j=1 \cup \\ j \neq u}}^{m} \|C_{i,k} - p_j\|_2^2 + \lambda_1), \tag{3}$$

where λ_1 is the hyperparameter to adjust the Euclidean distance margin.

2) Inter-Prototype Regularization: In the experiment, it is observed that prototypes are prone to semantic overlap between representations in visual space. In order to reduce the similarity between attribute prototypes, we further propose an inter-prototype regularization approach. Similar to the previously mentioned matching method, we firstly normalize the prototype vectors $P = [p_1, p_2, \ldots, p_m]$ and calculate the similarity matrix M between prototypes by:

$$s_{ij}^m \in M = P \cdot P^\top, M \in \mathcal{R}^{m \times m}, \tag{4}$$

where m is the overall number of prototypes, s_{ij}^m represents the cosine similarity between prototype p_i and p_j. To minimize similarity, it is preferable for s_{ij}^m to be as small as possible. Consequently, the cosine similarity loss function can be formulated as:

$$\mathcal{L}_{Inter}^{cos} = (\sum_{i=1}^{m} \sum_{j=1}^{m} (s_{ij}^m)^2)^{\frac{1}{2}}. \tag{5}$$

Merely reducing cosine similarity is not enough to achieve sufficient discrimination between prototypes. We also increase the Euclidean distance between prototypes, which can be defined as $d_{inter} = \frac{2}{m \times (m+1)} \sum_{i=1}^{m} \sum_{j=i}^{m} \|p_i - p_j\|_2^2$, the Euclidean distance loss function is expressed as:

$$\mathcal{L}_{Inter}^{euc} = \max(0, -d_{inter} + \lambda_2), \tag{6}$$

where λ_2 is the hyperparameter that controls the distance margin.

The overall loss function of attribute prototype-guided learning contains the above-mentioned four components:

$$\mathcal{L}_{APL} = \mathcal{L}_{Intra}^{cos} + \mathcal{L}_{Intra}^{euc} + \mathcal{L}_{Inter}^{cos} + \mathcal{L}_{Inter}^{euc}. \tag{7}$$

C. Iterative Scene Graph Generation

In order to enhance the intermediate reasoning and interpretability of the overall report generation framework, we additionally introduce scene graph generation as an auxiliary task. We implement an iterative approach to generate scene graph, which consists of three stages: 1) node prediction, 2) edge hypotheses, and 3) relationship prediction. Firstly, predict the entity nodes may appear in the graph with their labels, then iteratively sample the connected edges between nodes and output the final adjacency matrix consisting of adjacency lists for each node. Finally, predict the relation labels of directed edges based on the adjacency lists and nodes features, and output the final directed scene graph as one of the inputs for the following report decoding process.

1) Node Prediction: Firstly, the regions predicted in our Object Detector are used as the location nodes. Unlike traditional scene graph generation, the detected location nodes are not equivalent to the whole entity nodes in our scene graphs. In addition to the location nodes, the Chest ImaGenome dataset [34] contains anatomicalfinding nodes, tubesanlines nodes, disease nodes, and other attribute nodes. Collectively, these are referred to as attribute nodes. We define node objects as $O = \{o_1, \ldots, o_n, o_{n+1}, \ldots, o_{n+m}\}$, where the first n nodes denotes location nodes while the subsequent m nodes are designated as attribute nodes. The bounding boxes, visual region features, and predicted class label corresponding to the i-th location node are defined as b_i, f_i, and l_i. Therefore, location node o_i can be formulated as $o_i = (f_i, b_i, l_i), i \in [1, n]$. For attribute nodes, we define $o_j = s_j, j \in [1, m]$, where s_j represents the semantic features corresponding to

Algorithm 1 Algorithm of the Iterative Decoding Process of Generating the Adjacency Lists for Edge Hypotheses

Input: $O = \{o_1, \ldots, o_n, o_{n+1}, \ldots, o_{n+m}\}$
Output: $G = \hat{A} = \{A_i\}$
$G \leftarrow \varnothing$ ▷ *Initialize graph*
$\hat{A} \leftarrow \varnothing$ ▷ *Initialize adjacency matrix*
for each node o_i in $O^{1:n+m}$ **do**
  if $i > n$ & $e_i = 0$ **then**
    break ▷ *Skip unselected attribute nodes*
  else
    $e_i \leftarrow [o_{1:i}; A_{1:i}]$ ▷ *Preceding edges for decoding*
    $e'_i \leftarrow e_i + PositionEncoding(e_i)$
    $h_t \leftarrow LN(e'_i + MLP(e'_i))$ ▷ *Linear projection*
    $c_t \leftarrow LN(h_t + MHA(h_t))$ ▷ *Transformer decode*
    $\hat{c_t} \leftarrow MLP(c_t)$ ▷ *Feature mapping*
    $A_i \leftarrow Sample(\sigma(MLP(\hat{c_t})))$ ▷ *Sample o_i's edge*
    $\hat{l_i} \leftarrow Softmax(MLP(\hat{c_t}))$ ▷ *Auxiliary label*
    $\hat{A} \leftarrow \hat{A} \cup \{A_i\}$ ▷ *Aggregate adjacency list*
  end if
end for
return $G \leftarrow \{O, \hat{A}\}$

the attribute label of o_j, generated by an embedding layer initialized with pre-trained word embeddings, such as Glove. Therefore, each node can be uniformly represented as:

$$o_i = \begin{cases} (f_i, b_i, l_i), & i \in [1, n] \\ s_{i-n}, & i \in [n+1, n+m]. \end{cases} \tag{8}$$

We project the node features o_i onto a shared visual-semantic space and link this to a Multi-layer Perceptron (MLP) to identify the presence of node objects in the graph. This process ultimately results in the output of a binary mask vector $\hat{B}$, representing the occurrence of each node. On this basis, according to the auxiliary triplet (f_i, f_k, p_u) output from the APL module in Subsection III-B, we incorporate the attribute corresponding to prototype p_u into set $\hat{B}$, explicitly setting $\hat{B_{n+u}} = 1$. The training loss function of node prediction is formalized using binary cross-entropy between the predicted probability of occurrence $\hat{B_i}$ and target mask B_i:

$$\mathcal{L}_{NP} = \sum_{i}^{n+m} -[B_i \log p(\hat{B_i}) + (1 - B_i) \log(1 - p(\hat{B_i}))]. \tag{9}$$

2) Edge Hypotheses: In order to reduce the computational cost of relationship prediction in the graph, we adopt an iterative sampling method to gradually predict the connectivity between nodes before relationship prediction, expecting to output a simple directed graph structure. We use Transformer to model the iterative decoding process of the adjacency list A_i corresponding to node o_i, and the simplified pseudocode process is shown in Algorithm 1. Firstly, the graph G and the adjacency matrix $\hat{A}$ are initialized. During the autoregressive decoding process, conditioned on $o_{1:i}$ and $A_{1:i}$, the task is to predict the adjacency list A_i for each node o_i. The final decoding output size of adjacency matrix $\hat{A} = \{A_1, \ldots, A_{n+m}\}$ is $(n+m) \times (n+m)$, and each element in matrix $\hat{A}$ is a binary value determined by threshold ϵ. If $A_{ij} = 1$, it indicates a connection between nodes i and j, with the direction of the edge from i to j. We define the objective function as maximizing the conditional probability of graph G condition on image I, formulated as follows:

$$P(G \mid I) = P(\hat{A} \mid I) = \prod_{i=1}^{n} p(A_i \mid A_{1:i}, f_{1:i}, l_{1:i}, b_{1:i}) \cdot \prod_{i=n+1}^{n+m} p(A_i \mid A_{1:i}, f_{1:n}, l_{1:n}, b_{1:n}, s_{1:i-n}), \tag{10}$$

where the overall conditional probability is decomposed into the joint probability for each node o_i on its corresponding adjacency list.

The self-regressive decoding process is supervisedly trained using adjacency loss $\mathcal{L}_A$ and node label loss $\mathcal{L}_l$. The adjacency loss $\mathcal{L}_A$ calculates the binary cross-entropy loss between the predicted adjacency matrix and the ground truth A', which is designed to capture dependencies within the graph structure effectively.

$$\mathcal{L}_A = \frac{1}{(n+m)^2} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} -[A'_{ij} \log(\hat{A_{ij}}) + (1 - A'_{ij}) \log(1 - \hat{A_{ij}})]. \tag{11}$$

The node label loss $\mathcal{L}_l$ models the semantic consistency between the predicted auxiliary node labels $\hat{l_i}$ from Algorithm 1 and the node labels l_i predicted by the object detector. It incorporates the foundational semantics from l_i into the decoding process. It is computed using the cross-entropy loss function as follows:

$$\mathcal{L}_l = -\frac{1}{n} \sum_{i=1}^{n} l_i \log p(\hat{l_i}). \tag{12}$$

Combining two loss functions by:

$$\mathcal{L}_{EH} = \lambda \mathcal{L}_A + (1 - \lambda) \mathcal{L}_l, \tag{13}$$

where λ serves as a trade-off hyperparameter for balancing $\mathcal{L}_A$ and $\mathcal{L}_l$.

Algorithm 2 Algorithm of Relation Predicting for Edges

Input: $G = \hat{A} = \{A_i\}$, $O = \{o_i\}$
Output: $\hat{r} = \{\hat{r_k}\}$
$\hat{r} \leftarrow \varnothing$ ▷ *Initialize relation*
for each edge A_{ij} in $\hat{A}$ **do**
  if $A_{ij} = 0$ **then**
    break ▷ *Skip disconnected edges*
  else
    $\hat{h_i} \leftarrow ReLU(W_{hi} \cdot o_i)$ ▷ *Linear projection*
    $\hat{h_j} \leftarrow ReLU(W_{hj} \cdot o_j)$ ▷ *Linear projection*
    $\mathcal{E}_{ij} \leftarrow ReLU(W_{\mathcal{E}}[h_i; h_j])$ ▷ *Aggregate nodes*
    $\mathcal{E}'_{ij} \leftarrow LN(\mathcal{E}_{ij} + MHA(\mathcal{E}_{ij}, \mathcal{E}_{ij}, \mathcal{E}_{ij}))$
    $\hat{\mathcal{E}_{ij}} \leftarrow LN(\mathcal{E}'_{ij} + MHA(\mathcal{E}'_{ij}, f_v, f_v))$ ▷ *Integrate f_v*
    $r_{ij} \leftarrow Softmax(MLP(\hat{\mathcal{E}_{ij}}))$ ▷ *Relation label*
    $\hat{r} \leftarrow \hat{r} \cup \{r_{ij}\}$
  end if
end for
return $\hat{r}$

3) Relation Prediction: The final stage in generating the scene graph is relation prediction, which involves predicting the relation labels for the corresponding connected edges in the previously predicted adjacency matrix $\hat{A}$. The simplified pseudocode process is presented in Algorithm 2. Given the binary adjacency matrix $\hat{A}$, the target is to classify the relation label r_{ij} between nodes o_i and o_j. The prediction process can be formalized as maximizing the contextualization relation probability $P(r_{ij}|o_i, o_j, f_v)$, where f_v represents the visual global contextualized features extracted by a pre-trained ResNet. The specific composition of the node o has been described in Equation 8. We employ a Transformer encoder-decoder scheme to model the process of extracting the probability distribution of relations, with the key component being Multi-Head Attention (MHA) formalized as:

$$MHA(Q, K, V) = [Softmax(\frac{QW_i^Q \cdot (KW_i^K)^\top}{\sqrt{d}})VW_i^V]W^M, \tag{14}$$

where W_i^Q, W_i^K, and W_i^V are learnable projection parameters while W^M is multi-head aggregation parameter.

The initial phase of our modeling process involves applying a linear projection to map node features into a shared visual-semantic space, yielding intermediate node embeddings $\hat{h}_i$. We then aggregate these node embeddings $\hat{h}_i$ and $\hat{h}_j$ to derive the edge embedding $\mathcal{E}_{ij}$, which is followed by the application of self-attention layers and the updating of edge embeddings. Further refinement of the edge representation is achieved by integrating f_v. To address the long-tail distribution of relation categories, a weighted cross-entropy loss function is employed for training.

$$\mathcal{L}_{RP} = -\sum_{k}\sum_{c=1}^{\mathcal{C}_\mathcal{R}} w_c \cdot r_{kc} \log(\hat{r_{kc}}), \tag{15}$$

where $\hat{r_{kc}}$ denotes the prediction probability of the k-th relation belonging to category $c \in \mathcal{C}_\mathcal{R}$ while r_{kc} represents target relation category, and w_c is the weight set as the reciprocal of the frequency of relation c's occurrence with normalization.

Overall training loss function of iterative scene graph generation is an integration of the above three stages:

$$\mathcal{L}_{ISGG} = \mathcal{L}_{NP} + \mathcal{L}_{EH} + \mathcal{L}_{RP}. \tag{16}$$

D. Report Decoder

Following the practice in [7], for the report generation decoder, we utilize a GPT-2 Medium model with 335M parameters, fine-tuned on PubMed abstracts [36]. To enhance the report generation process by incorporating multi-modal information during decoding, we inject scene graph features G and selected visual region features F into the self-attention layer with report tokens $R = \{y_1, \ldots, y_T\}$. This integration facilitates the extraction of diverse mutual information, which can be formalized as follows:

$$SA(F, G, R) = Softmax\left(RW_q \begin{bmatrix} FW_k^F \\ GW_k^G \\ RW_k^R \end{bmatrix}^\top\right) \begin{bmatrix} FW_v^F \\ GW_v^G \\ RW_v^R \end{bmatrix}, \tag{17}$$

where W_k^F, W_k^G and W_v^F, W_v^G represents the newly added learnable parameters for key and value projection, and $SA(\cdot)$ denotes the Self-Attention mechanism.

Typically, the report generation objective can be formalized as minimizing a cross-entropy loss by comparing the predicted token with ground truth:

$$\mathcal{L}_R = -\sum_{t=1}^{T} \log p(y_t \mid y_{1:t-1}, F, G). \tag{18}$$

The overall training objective amalgamates three distinct loss components: attribute prototype-guided learning, iterative scene graph generation, and report generation. This integrated objective can be succinctly formulated as follows:

$$\mathcal{L}_{all} = \alpha\mathcal{L}_{APL} + \beta\mathcal{L}_{ISGG} + \mathcal{L}_R, \tag{19}$$

where the hyperparameters α and β serve to balance the three loss terms.

IV. Experiments and Results

A. Dataset and Evaluation Metrics

We use the Chest ImaGenome v1.0.0 [34] dataset to train and evaluate our proposed AP-ISG, which is constructed from the MIMIC-CXR [37] dataset. The original MIMIC-CXR dataset consists of 377,110 chest X-ray images with 227,835 reports. The Chest ImaGenome dataset contains pairs of chest X-ray images and corresponding reports along with automatically extracted 242,072 scene graphs. We follow [7] to use the official split as 166,512 for training, 23,952 for validation, and 47,389 for testing. For more dataset details, please refer to Appendix-A.

We evaluate AP-ISG from two levels of metrics. On report level, we adopt the widely-used Natural Language Generation (NLG) metrics to calculate the matching degree between generated reports and ground truth reports, including BLEU [38], METEOR [39], ROUGE-L [40], and CIDEr [41]. In order to comprehensively measure the diagnostic accuracy of the generated reports, we additionally adopt Clinical Efficacy (CE) metrics to directly compare the presence of prominent clinical observations. Following [7], the CE metrics consist of two parts. The first part involves computing the micro-average of accuracy, precision, recall, and F1 scores over 5 observations for: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. The second part is the example-based precision, recall, and F1 scores calculated over 14 observations by comparing the classifications for each generated report and corresponding reference report. On scene graph level, we follow frequently used evaluation criteria to evaluate the scene graph generation performance of AP-ISG on three sub-tasks, including Predicate Classification (PredCls), Scene Graph Classification (SGCls), and Scene Graph Detection (SGDet). Mean Recall@100 (mR@100) is taken as the primary evaluation metric.

B. Implementation Details

To maintain the original aspect ratio, all images are padded to a square shape based on the longer edge and then resized

TABLE I

PERFORMANCE COMPARISONS OF OUR PROPOSED METHOD WITH STATE-OF-THE-ART METHODS ON MIMIC-CXR DATASET WITH RESPECT TO NATURAL LANGUAGE GENERATION (NLG) METRICS. THE BEST AND THE SECOND-BEST VALUES ARE HIGHLIGHTED IN BOLD. † DENOTES OUR IMPLEMENTATIONS ON THE OFFICIAL SPLIT OF CHEST IMAGENOME DATASET

Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Parameters
R2Gen [9]	2020	0.353	0.218	0.145	0.103	0.142	0.277	-	90.8M
R2GenCMN [10]	2021	0.353	0.218	0.148	0.106	0.142	0.278	-	64.8M
PPKED [11]	2021	0.360	0.224	0.149	0.106	0.149	0.284	0.237	-
$\mathcal{M}^2$ TR. PROGRESSIVE [12]	2021	0.378	0.232	0.154	0.107	0.145	0.272	-	-
Contrastive Attention [13]	2021	0.350	0.219	0.152	0.109	0.151	0.283	-	-
AlignTransformer [14]	2021	0.378	0.235	0.156	0.112	0.158	0.283	-	-
ITA [18]	2022	**0.395**	**0.253**	0.170	0.121	0.147	0.284	-	-
CvT-212DistilGPT2 [19]	2022	0.392	0.245	0.169	0.124	0.153	0.285	0.361	-
RAMT-U [8]	2023	0.362	0.229	0.157	0.113	0.153	0.284	-	168.5M
UAR [20]	2023	0.363	0.229	0.158	0.107	0.157	**0.289**	0.246	-
KiUT [21]	2023	**0.393**	0.243	0.159	0.113	0.160	0.285	-	-
METransformer [22]	2023	0.386	0.250	0.169	0.124	0.152	**0.291**	0.362	152.0M
R2Gen† [9]	2020	0.331	0.205	0.128	0.092	0.136	0.262	0.287	90.8M
R2GenCMN† [10]	2021	0.338	0.209	0.131	0.097	0.138	0.263	0.301	64.8M
RAMT-U† [8]	2023	0.345	0.217	0.141	0.105	0.142	0.276	0.264	168.5M
RGRG [7]	2023	0.373	0.249	**0.175**	**0.126**	**0.168**	0.264	**0.495**	254.1M
AP-ISG	2023	0.391	**0.258**	**0.182**	**0.129**	**0.175**	0.282	**0.526**	287.3M

to 512×512, followed by normalization to a standard normal distribution. For the reports, we follow [7] not to employ lowercasing for fair comparisons. The findings section is used as the target report, with the removal of any excess spaces, special characters, and reports with blank findings.

During the training process, we directly employ the pre-trained checkpoints from [7] for the object detector, region selector, and abnormality classifier components. Subsequently, the entire model is trained in an end-to-end fashion, encompassing attribute prototype-guided learning, iterative scene graph generation, and report decoder. We train our framework on a single NVIDIA GeForce RTX 4090 with PyTorch 1.12.1. AdamW [42] is applied as the optimizer with a weight decay of 1e-2, reducing the learning rate by a factor of 0.5, and early stopping is included. The initial learning rate is set to 5e-5 with a batch size of 2. Batch sizes are accumulated to 64 before gradient updating. In particular, we set hyperparameters λ, λ_1, λ_2, α and β as 0.7, 1, 7, 2, 5, which are determined in the Ablation Studies.

For inference, we employ beam search with beam size as 3 and apply BERTScore [43] with a threshold of 0.9 to remove similar sentences, in which only highly similar sentences are deduplicated. The code of our framework is available at https://github.com/giantke/AP-ISG.

C. Comparisons With State-of-the-Art

For radiology report generation, we conduct performance comparisons with state-of-the-art (SOTA) methods: R2Gen [9], R2GenCMN [10], PPKED [11], $\mathcal{M}^2$ TR. PROGRESSIVE [12], Contrastive Attention [13], AlignTransformer [14], $\mathcal{M}^2$ Trans w/ NLL [15], $\mathcal{M}^2$ Trans w/ NLL+BS+fc_E [15], $\mathcal{M}^2$ Trans w/ NLL+BS+fc_EN [15], ITA [18], CvT-21 2DistilGPT2 [19], RAMT-U [8], RGRG [7], UAR [20], KiUT [21], and METransformer [22].

1) NLG Metrics: As shown in Table I, we employ a double line to divide approaches into two sections based on different dataset splits. The section above adopts the official split of MIMIC-CXR following [9], while the section below follows [7] for the official division of the Chest ImaGenome dataset. In comparison with existing methodologies, the traditional CNN-Transformer-based R2Gen framework and its variants exhibit inferior performance. This limitation stems from their reliance on a conventionally small-parameter CNN as the visual extractor, which inadequately addresses the significant semantic disparity between image and text modalities. In contrast, most methods that incorporate cross-modal alignment have demonstrated enhanced performance over R2Gen. Despite utilizing fewer training images than half of those in the official MIMIC-CXR dataset split, our approach almost achieves optimal results compared to state-of-the-art methods across various metrics. Specifically, in terms of METEOR and BLEU-4 metrics, our proposed AP-ISG achieves an improvement of 4.2% and 2.4% over the second-best method, respectively. This highlights the refined capability of AP-ISG in accurately localizing within the visual space and generating more precise extended sentences. The elevated CIDEr score further validates this advancement, which also focuses on the matching degree of key information. Regarding the suboptimal performance of our AP-ISG in achieving the SOTA ROUGE-L metric, we hypothesize that this is attributable to the relatively simplistic design of the Region Selector, which consequently selects more regions than necessary for sentence generation.

We further present the model size for our method and the available baselines. Owing to the strong supervision employed in object detection, it is evident that the model size of RGRG substantially exceeds that of other baselines not utilizing region information. On the basis of RGRG, our method achieves a desirable enhancement in performance with only a minimal increase in parameter quantity.

2) CE Metrices: To address the limitations of NLG Metrics in measuring the clinical accuracy of generated reports, we also conduct comparative experiments using CE Metrics.

TABLE II
PERFORMANCE COMPARISONS WITH CLINICAL EFFICACY (CE) METRICS ON MIMIC-CXR DATASET, EVALUATED MICRO-AVERAGED OVER FIVE OBSERVATIONS (MIC-5) AND EXAMPLE-BASED AVERAGED OVER 14 OBSERVATIONS (EX-14). RL REPRESENTS REINFORCEMENT LEARNING. THE 'W/O APL' IS THE ABBREVIATION OF 'WITHOUT ATTRIBUTE PROTOTYPE-GUIDED LEARNING (APL)'

Method	RL	Year	P_{mic-5}	R_{mic-5}	$F_{1,mic-5}$	P_{ex-14}	R_{ex-14}	$F_{1,ex-14}$
R2Gen [9]	✗	2020	0.412	0.298	0.346	0.331	0.224	0.228
$\mathcal{M}^2$ Trans w/ NLL [15]	✗	2021	0.489	0.411	0.447	-	-	-
$\mathcal{M}^2$ Trans w/ NLL+BS+fc_E [15]	✓	2021	0.463	**0.732**	**0.567**	-	-	-
$\mathcal{M}^2$ Trans w/ NLL+BS+fc_{EN} [15]	✓	2021	**0.503**	0.651	**0.567**	-	-	-
R2GenCMN [10]	✗	2021	-	-	-	0.334	0.275	0.278
Contrastive Attention [13]	✗	2021	-	-	-	0.352	0.298	0.303
$\mathcal{M}^2$ TR. PROGRESSIVE [12]	✗	2021	-	-	-	0.240	0.428	0.308
CvT-212DistilGPT2 [19]	✗	2022	-	-	-	0.359	0.412	0.384
RGRG [7]	✗	2023	0.491	0.617	0.547	0.461	**0.475**	0.447
AP-ISG w/o APL	✗	2023	0.502	0.643	0.559	**0.467**	0.474	**0.449**
AP-ISG	✗	2023	**0.518**	**0.695**	**0.582**	**0.486**	**0.493**	**0.462**

TABLE III
PERFORMANCE COMPARISONS OF SCENE GRAPH GENERATION ON MIMIC-CXR DATASET WITH RESPECT TO MEAN RECALL@100 (MR@100) METRIC. † DENOTES OUR IMPLEMENTATIONS USING THEIR OFFICIAL CODE

Method	PredCls mR@100	SGCls mR@100	SGDet mR@100
VCTree† [27]	12.3	6.8	5.6
RU-Net† [29]	17.7	**10.2**	7.5
AP-ISG	**18.3**	**10.2**	**8.8**

TABLE IV
ABLATION STUDY OF OUR PROPOSED AP-ISG ON MIMIC-CXR DATASET, INCLUDING APL AND ISGG MODULE

Module		NLG Metrics				
APL	ISGG	BLEU-1	BLEU-4	METEOR	ROUGE-L	FLOPs
✗	✗	0.373	0.124	0.167	0.263	6.9×10^{11}
✗	✓	0.380	0.127	0.173	0.275	7.2×10^{11}
✓	✗	0.378	0.124	0.171	0.269	6.9×10^{11}
✓	✓	**0.391**	**0.129**	**0.175**	**0.282**	7.2×10^{11}

As illustrated in Table II, our AP-ISG surpasses the SOTA approaches across all CE metrics by a significant margin. Specifically, AP-ISG outperforms the second-best method by 2.6% and 3.6% on $F_{1,mic-5}$ and $F_{1,ex-14}$ scores, respectively. Even compared to methods optimized with reinforcement learning (RL), our approach remains highly competitive, showing a 2.6% improvement on the F1 metric. This demonstrates that AP-ISG more effectively enhances the consistency between image regions and corresponding sentences while improving the hit rate of clinical category descriptive phrases in the generated sentences.

3) Scene Graph Metrics: To validate the performance of scene graph generation in our proposed method, we additionally adapt VCTree [27] and RU-Net [29] to the MIMIC-CXR dataset using their official code in Table III. Our approach either surpasses or matches the performance of the current SOTA method RU-Net across three sub-tasks. This underscores the efficacy of our iterative scheme and attribute prototype-guided learning in enhancing structural reasoning and contextual learning within the scene graphs.

TABLE V
ABLATION STUDY ON WHETHER ADDITIONAL SCENE GRAPH INFORMATION (DENOTED AS G) IS INTEGRATED INTO THE VISUAL FEATURES (F) AND REPORTS (R) WITHIN OUR AP-ISG FRAMEWORK PRIOR TO REPORT GENERATION

G	BLEU-1	BLEU-4	METEOR	ROUGE-L
✗	0.378	0.124	0.172	0.270
✓	**0.391**	**0.129**	**0.175**	**0.282**

D. Ablation Studies

To validate the efficacy of each component proposed in our study, we conduct ablation experiments on the MIMIC-CXR dataset, with results reported in Table IV. We set the baseline by omitting the APL and ISGG modules, which is almost identical to RGRG. The other variants include adding APL alone (defined as '+APL'), adding ISGG alone (defined as '+ISGG'), and fully integrated our AP-ISG. Initially, both '+APL' and '+ISGG' enhance the baseline's performance across all four NLG metrics, demonstrating that the augmented relation reasoning capability of '+ISGG' and the additional exploration of potential attribute commonalities between regions by '+APL' both contribute to generating more accurate and comprehensive reports. Furthermore, our complete AP-ISG achieves the best generative performance, indicating that the components in our approach synergistically enhance each other, optimizing the report generation process. We have also included experiments on FLOPs, further demonstrating that the computational cost introduced by our designed APL and ISGG modules is exceedingly limited, while yielding significant performance improvements. Concurrently, we conduct an ablation study on the incorporation of additional scene graph information in the report decoder, as shown in Table V, the experimental results validate the necessity of integrating scene graph information prior to report generation.

We further perform an ablation study to investigate the influence of different hyperparameters λ_1, λ_2 and λ. Since λ_1 and λ_2 are introduced in the APL module, they are a prerequisite for λ in the ISGG module. Thus we firstly fix λ to 0.5 and investigate the other two. As illustrated in Table VI, the optimal results are achieved with $\lambda_1 = 1$ and $\lambda_2 = 7$. Subsequent refinement of λ to 0.7 further optimized

TABLE VI
ABLATION STUDY ON HYPERPARAMETERS λ_1, λ_2 AND λ WITH RESPECT TO MEAN RECALL@100 (MR@100) METRIC

λ_1	λ_2	λ	PredCls mR@100	SGCls mR@100	SGDet mR@100
3	1	0.5	14.2	7.5	6.6
1	3	0.5	16.7	8.9	7.3
1	7	0.5	17.5	9.4	7.8
1	12	0.5	16.9	9.1	7.5
1	7	0.7	**18.3**	**10.2**	**8.8**
1	7	0.9	17.8	9.6	8.3
1	7	0.3	16.5	8.8	7.1

TABLE VII
ABLATION STUDY ON HYPERPARAMETERS α AND β IN EQN.19

α	β	BLEU-4	ROUGE-L	CIDEr
1	1	0.121	0.273	0.405
2	5	**0.129**	**0.282**	**0.526**
2	10	0.124	0.278	0.475
5	2	0.117	0.270	0.361

TABLE VIII
GENERALIZABILITY VALIDATION OF OUR METHOD ACROSS DIFFERENT BACKBONE ARCHITECTURES. THE ABBREVIATION '*w/o*' STANDS FOR 'WITHOUT'

Backbone	BLEU-4	METEOR	ROUGE-L
Faster RCNN + GPT-2	0.129	0.175	0.282
Fast RCNN + Transformer	0.115	0.156	0.271
w/o APL	0.113	0.152	0.265
w/o ISGG	0.110	0.149	0.262
w/o APL,ISGG	0.109	0.142	0.258

performance, highlighting the crucial role of precise adjacency list accuracy during the edge hypotheses phase for superior model effectiveness.

Furthermore, we conduct an additional ablation study to explore the impact of crucial hyperparameters α and β on balancing the overall objective. As depicted in Table VII, our model demonstrates enhanced performance with α and β set to 2 and 5, respectively. This demonstrates that when the model employs attribute prototypes as an auxiliary and focuses more on the primary task of scene graph generation, it will achieve superior overall performance.

To validate the generalizability of our method across different backbone architectures, additional ablation experiments are conducted using the 'Fast R-CNN + Transformer' architecture combination. Specifically, we try to replace the object detector from Faster R-CNN to Fast R-CNN and the report generator from GPT-2 to Transformer. The experimental results, presented in Table VIII, reveal that the proposed APL and ISGG approaches continue to deliver significant performance enhancements, even with the substitution of the entire backbone architecture. This demonstrates the generalizability of our method on other architectures.

E. Case Studies

1) Report-Level: In this study, we compare the quality of report generation between the baseline R2GenCMN, the current SOTA method RGRG, and our approach. We can observe from the upper part of Fig. 2 that our proposed AP-ISG generates reports with more comprehensive and accurate coverage of diagnosis, whereas the comparison methods miss several key descriptions. Specifically, for common and easily identifiable diagnoses like lung opacity and heart size, all methods accurately recognize them. However, for anatomical areas with frequent overlap and complexity, e.g., acute cardiopulmonary process and vascular calcification, only AP-ISG provides diagnoses essentially identical to the reference reports. This indicates that our method maintains stable predictive accuracy in more challenging anatomical regions. Nevertheless, all methods exhibit errors or omissions in diagnosing osseous structures. We speculate that the observed discrepancies may arise from the limited sample size for this category and the inherent complexity in interpreting osseous structures in X-ray images.

In addition, we visualize the three stages of scene graph generation in AP-ISG, depicted in the lower part of Fig. 2. As can be observed in the Relation Prediction stage, vascular calcification is accurately associated with anatomical areas including the aortic arch, mediastinum, and hilar structure. This demonstrates the robust predictive capability of our iterative scene graph generation module among complex regional relationships, enhancing the structural localization and reasoning ability in report generation.

2) Sentence-Level: As illustrated in Fig. 3, we conduct a qualitative comparison of the generative prediction performance of RGRG and our proposed AP-ISG at sentence-level corresponding to anatomical regions. Specifically, for the regions of Cardiac silhouette and Spine, both methods provide predictions closely matching the reference sentences. However, for the diagnosis of bilateral lung, the region bounding boxes of RGRG show a relatively larger deviation and notably miss detailed descriptions such as hyperinflation, biapical scarring, and focal consolidation. In contrast, AP-ISG is still able to thoroughly extract diagnostic details within the corresponding areas. This indicates that our approach, bolstered by the auxiliary task of scene graph generation, possesses a more robust capability in capturing detailed descriptive information of specific regions.

F. Qualitative Visualization

1) Effect of APL: To validate the capability of the APL module in producing compact intra-prototype and distinguishable inter-prototype representations more vividly, we manually select the top-15 attribute prototypes with the highest frequency of occurrence in a sample. Taking advantage of the t-SNE technique, we visualize the feature distribution between these attribute prototypes and regional entity pairs, as depicted in Fig. 4. Upon contrasting subfigures 4(a) and 4(b), it is discernible that the prototype feature distribution in subfigure 4(a) demonstrates significant intra-class variance and substantial intra-class overlap without APL. In subfigure 4(b), however, with the application of the APL module, the features of entity pairs within each prototype appear considerably more compact. Furthermore, the distinction between prototypes is enhanced, thereby reducing the likelihood of semantic overlap. This visually explicates the intrinsic reason behind our

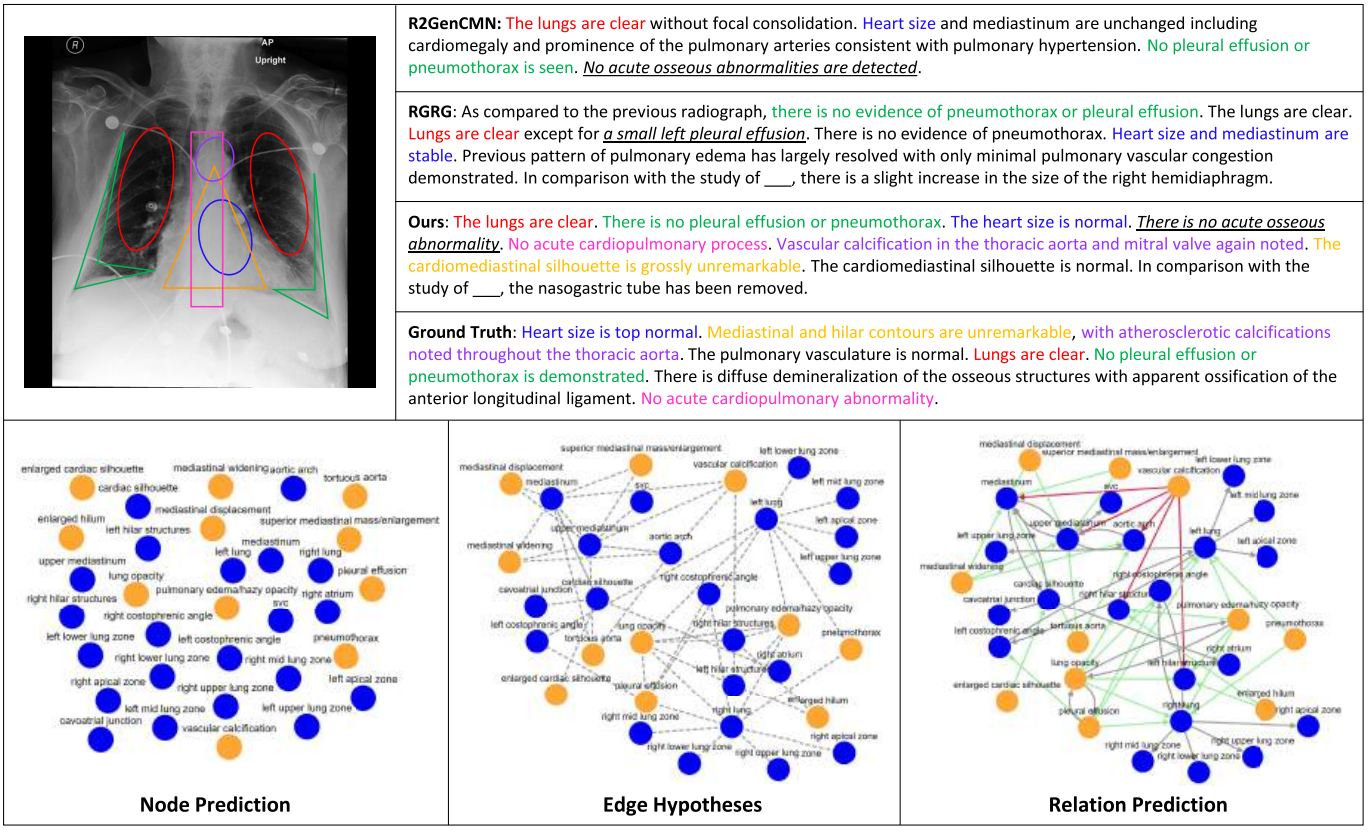


Fig. 2. Illustrations of reports from R2GenCMN, RGRG, ours, and ground truth for one sample from MIMIC-CXR. For better visualization, different colors are utilized to highlight various corresponding matching medical information, and the corresponding bounding box locations which are manually annotated. Incorrect descriptions are marked with italics and underline. The figure below displays a visualization of our AP-ISG during three stages of scene graph generation, wherein red arrows denote affirmation, green arrows indicate negation, and grey arrows represent other relationships, including positional and subordinate associations.

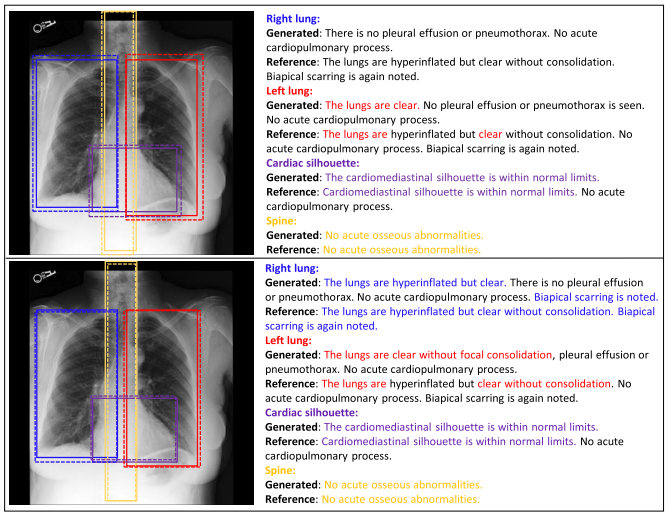
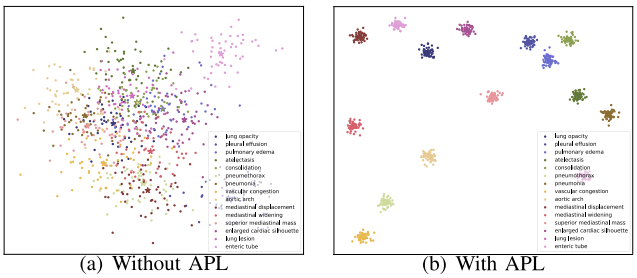


Fig. 3. Qualitative sentence-level comparison results between RGRG (upper part) and our AP-ISG (lower part). The generated sentences are annotated with colors consistent with the corresponding regions, highlighting the parts that are in agreement with the ground truth.



(a) Without APL (b) With APL

Fig. 4. Comparison of t-SNE visualization on attribute prototypes and region pairs between without and with the Attribute Prototype-guided Learning (APL) module. Pentagrams denote distinct prototypes.

APL module's effectiveness in thoroughly mining the latent common attributes among regional entities.

2) Interpretability Analysis: To demonstrate the interpretability of our framework on report generation, we conduct a visual analysis using Grad-CAM [44] as illustrated in Fig. 5. The CAM maps highlight areas with high attention values, which exhibit good spatial consistency with the corresponding locations of the diagnosis provided by professional pathologists. Specifically, our proposed AP-ISG is capable of accurately focusing on significant locations across various anatomical structures, such as the bilateral lungs (a), heart (b), right lower lung (c), and central aorta (c). This indicates that our proposed method possesses robust disease region localization capabilities, which can assist the framework in filtering out a substantial amount of redundant and irrelevant image patches, thereby aiding in precise report generation.

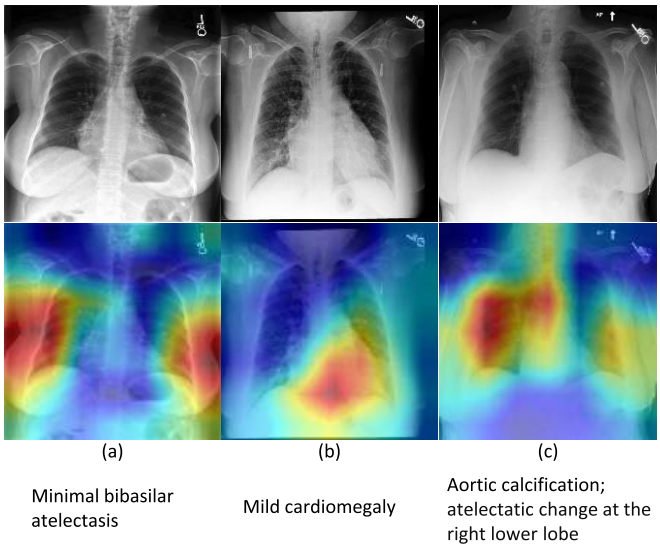


Fig. 5. Visualization of Grad-CAMs for interpretability with ground truth diagnoses.

V. Conclusion

In this work, we propose a novel region-based Attribute Prototype-guided Iterative Scene Graph generation framework (AP-ISG) for report generation, utilizing scene graph generation as an auxiliary task to enhance interpretability and relational reasoning capabilities. The iterative scene graph generation module employs an autoregressive scheme for structural edge reasoning and a contextualization mechanism for relational reasoning. APL enhances intra-prototype matching and reduces inter-prototype semantic overlap in the visual space to fully model the potential attribute commonalities among regions. Extensive comparative experiments on the MIMIC-CXR dataset across multiple metrics demonstrate that AP-ISG achieves state-of-the-art results for radiology report generation. Despite achieving optimal performance in report generation, AP-ISG still exhibits some limitations. Firstly, ISGG creates additional scene graphs to assist in report generation, enhancing the reasoning capability for diagnoses. However, this design may lead to a slight decrease in inference speed. This issue could be addressed by designing a stage-wise local location encoding scheme. Secondly, APL is designed to unearth hidden attribute commonalities across regions, enabling the model to generate more comprehensive and diversified scene graphs. However, this mining process may also introduce additional erroneous noise, leading to inaccurate reasoning in the final scene graph. Nevertheless, the proposed APL has been proven to bring positive gains to both report and scene graph generation tasks. In the future, we will extend AP-ISG to more medical diagnostic tasks and further enhance its reasoning speed and noise control capabilities.

Appendix

A. Dataset Details

The Chest ImaGenome dataset [34] is derived through a streamlined methodology referenced in [45] and [46], employing a text pipeline [47] that segments radiology reports to selectively extract sentences from findings and impressions. This approach involves the application of a CXR-specific lexicon, pre-compiled by radiologists using a concept expansion engine [48], to identify and contextualize CXR-related entities, differentiating their negated or affirmed states. Ambiguities in entity interpretation, such as the differentiation between types of 'collapse', are resolved through sentence-level filtering. The natural language parser SpaCy then maps CXR attributes to corresponding anatomical locations within sentences. Utilizing a radiologist-constructed CXR ontology, the system corrects misalignments in attribute-to-anatomy assignments, culminating in the aggregation of attributes at the exam level into a radiology knowledge graph. This graph delineates the relationships between anatomical sites and their documented CXR attributes for each report.

The attributes we use can be mainly classified into the following categories.

- 'anatomical finding': describes observations on anatomies with subjectivity in phrase grouping for label extraction.
- 'disease': diagnostic-level descriptions requiring patient information beyond images, highly subjective to the radiologist's inference.
- 'technical assessment': concerns image quality issues impacting radiologic interpretation.
- 'tubes and lines': medical support devices needing placement issue reporting.
- 'devices': medical devices with less emphasis on placement issues.
- 'texture': highly non-specific attributes (e.g., opacity, lucency, interstitial, airspace) for initial objective image descriptions by radiologists.

For more detailed content, please refer to the source dataset link: https://physionet.org/content/chest-imagenome/1.0.0/.

TABLE IX
Performance Comparisons Between RGRG, AP-ISG and AP-*ISG**. $\star$ Denotes That AP-ISG Is Trained From Scratch, Without Applying the Checkpoint From RGRG. B-4, M, and R-L Denote BLEU-4, METEOR, and ROUGE-L, Respectively

Method	NLG Metrics			CE Metrics		
	B-4	M	R-L	P_{mic-5}	R_{mic-5}	$F_{1,mic-5}$
RGRG	0.126	0.168	0.264	0.491	0.617	0.547
AP-ISG	0.129	0.175	0.282	0.518	0.695	0.582
AP-ISG*	0.128	0.175	0.279	0.523	0.686	0.580

B. Train From Scratch

Although our proposed AP-ISG demonstrates superior performance over RGRG across multiple metrics, some enhancements may be attributable to the pre-trained checkpoints from RGRG. To elucidate the distinct performance contributions of the proposed APL and ISGG more clearly, experiments have been also conducted without leveraging RGRG's pre-trained checkpoints. As shown in Table IX, the performance of AP-ISG* and AP-ISG across all metrics is nearly identical, corroborating the pivotal role of our approach in enhancing the final performance while also negating the impact of pre-trained checkpoints.

TABLE X
PERFORMANCE COMPARISONS OF MEDICAL REPORT GENERATION ON IU X-RAY DATASET WITHOUT FINE-TUNING. † DENOTES THE RESULTS CITED FROM [50]

Model	Year	B-4	M	$P_{ex\text{-}14}$	$R_{ex\text{-}14}$	$F_{1,ex\text{-}14}$
R2Gen† [9]	2020	0.059	0.131	0.141	0.136	0.136
CVT2Dis.† [19]	2022	**0.082**	0.147	0.174	0.172	0.168
M2KT† [17]	2023	0.078	**0.153**	0.153	0.145	0.145
DCL† [51]	2023	0.074	0.152	0.168	0.167	0.162
AP-ISG	2023	0.076	0.148	**0.192**	**0.194**	**0.189**

C. Test Without Fine-Tuning

To further validate the generalizability of our proposed method in scenarios without fine-tuning, we have conducted comparative tests with the current SOTA methods on the IU X-Ray dataset [49]. Following [50], all models were trained on the MIMIC-CXR dataset and then directly tested on the complete IU X-Ray dataset to ensure fairness. As shown in Table X, the experimental results indicate that AP-ISG performs comparably to SOTA methods on NLG metrics such as BLEU-4 and METEOR. Moreover, AP-ISG significantly outperforms current SOTA methods in accuracy metrics, demonstrating that our proposed AP-ISG maintains excellent generalizability even without fine-tuning.

REFERENCES

[1] Y. Han, G. Holste, Y. Ding, A. Tewfik, Y. Peng, and Z. Wang, "Radiomics-guided global-local transformer for weakly supervised pathology localization in chest X-rays," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 750–761, Mar. 2023.

[2] F. Wang et al., "Lumbar bone mineral density estimation from chest X-ray images: Anatomy-aware attentive multi-ROI modeling," *IEEE Trans. Med. Imag.*, vol. 42, no. 1, pp. 257–267, Jan. 2023.

[3] N. Gaggion, L. Mansilla, C. Mosquera, D. H. Milone, and E. Ferrante, "Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest X-ray analysis," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 546–556, Feb. 2023.

[4] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed abnormality graph embedding for clinically accurate X-ray report generation," *IEEE Trans. Med. Imag.*, vol. 42, no. 8, pp. 2211–2222, Aug. 2023.

[5] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated radiographic report generation purely on transformer: A multicriteria supervised approach," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2803–2813, Oct. 2022.

[6] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[7] 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. (CVPR)*, Jun. 2023, pp. 7433–7442.

[8] K. Zhang et al., "Semi-supervised medical report generation via graph-guided hybrid feature consistency," *IEEE Trans. Multimedia*, vol. 26, pp. 904–915, 2023.

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

[10] 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.

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

[12] F. Nooralahzadeh, N. Perez Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2824–2832.

[13] 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.

[14] 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 *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, M. de Bruijne et al., Eds., Cham, Switzerland: Springer, 2021, pp. 72–82.

[15] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 5288–5304.

[16] X. Wu et al., "DeltaNet: Conditional medical report generation for COVID-19 diagnosis," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 2952–2961.

[17] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102798.

[18] L. Wang, M. Ning, D. Lu, D. Wei, Y. Zheng, and J. Chen, "An inclusive task-aware framework for radiology report generation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Cham, Switzerland: Springer, 2022, pp. 568–577.

[19] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artif. Intell. Med.*, vol. 144, Oct. 2023, Art. no. 102633.

[20] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, and Y. Zou, "Unify, align and refine: Multi-level semantic alignment for radiology report generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2863–2874.

[21] 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. (CVPR)*, Jun. 2023, pp. 19809–19818.

[22] 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. (CVPR)*, Jun. 2023, pp. 11558–11567.

[23] J. Johnson et al., "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3668–3678.

[24] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 852–869.

[25] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5410–5419.

[26] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[27] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6619–6628.

[28] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 670–685.

[29] X. Lin, C. Ding, J. Zhang, Y. Zhan, and D. Tao, "RU-Net: Regularized unrolling network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19457–19466.

[30] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2117–2130, Aug. 2019.

[31] X. Yang, H. Zhang, and J. Cai, "Auto-encoding and distilling scene graphs for image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2313–2327, May 2022.

[32] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9962–9971.

[33] X. Lu and Y. Gao, "Guide and interact: Scene-graph based generation and control of video captions," *Multimedia Syst.*, vol. 29, no. 2, pp. 797–809, Apr. 2023.

[34] J. T. Wu et al., "Chest imagenome dataset for clinical reasoning," in *Proc. 34th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–14.

[35] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[36] Y. Papanikolaou and A. Pierleoni, "DARE: Data augmented relation extraction with GPT-2," 2020, *arXiv:2004.13845*.
[37] A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.
[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[39] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.
[41] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.
[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–43.
[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[45] J. T. Wu et al., "Ai accelerated human-in-the-loop structuring of radiology reports," in *Proc. AMIA Annu. Symp.*, 2020, p. 1305.
[46] J. Wu et al., "Automatic bounding box annotation of chest X-ray data for localization of abnormalities," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 799–803.
[47] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 590–597.
[48] A. Coden, D. Gruhl, N. Lewis, M. Tanenblatt, and J. Terdiman, "SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora," in *Proc. IEEE 2nd Int. Conf. Healthcare Informat., Imag. Syst. Biol.*, Sep. 2012, pp. 33–39.
[49] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.
[50] H. Jin, H. Che, Y. Lin, and H. Chen, "PromptMRG: Diagnosis-driven prompts for medical report generation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 3, pp. 2607–2615.
[51] 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.