Severity-aware Radiology Report Generation: Knowledge Graph Expansion and Momentum-Guided Classification Enhancement

1st Chunyan Yu
College of Computer and Data Science
Fuzhou University
Fuzhou, China
therica@fzu.edu.cn

2nd Jiannan You
College of Computer and Data Science
Fuzhou University
Fuzhou, China
231020008@fzu.edu.cn

3rd Shengbiao Huang*
College of Computer and Data Science
Fuzhou University
Fuzhou, China
231027119@fzu.edu.cn

4th Zejie Yan
College of Computer and Data Science
Fuzhou University
Fuzhou, China
231020050@fzu.edu.cn

5th Wanjian Xu
College of Computer and Data Science
Fuzhou University
Fuzhou, China
231027205@fzu.edu.cn

6th Zexi Lin
College of Computer and Data Science
Fuzhou University
Fuzhou, China
231027036@fzu.edu.cn

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356157

***Abstract*—Radiology report generation aims to provide comprehensive clinical descriptions and ease radiologists' workloads. Previous research has explored using knowledge graphs and auxiliary classification tasks to enhance the model's ability to generate accurate reports. However, due to the lack of information in the knowledge graphs or insufficient class label information, these methods fail to provide models with clinical severity information about the same disease at different stages of development, resulting in less accurate reports. To address this issue, we propose a Severity-Guided Radiology Report Generation method (SR2Gen), which guides the model in identifying internal severity variations of the disease from both explicit and implicit dimensions. Specifically, SR2Gen includes two innovative modules: a Knowledge Enhancement Module (KEM) and a Disease Severity-Aware Module (DSAM). First, KEM explicitly guides the report generation model by constructing a knowledge graph containing disease severity information as prior knowledge. Secondly, DSAM enhances the severity-aware classifier using pseudo-labels generated through momentum distillation and further incorporates an adaptive disease severity learning method, implicitly guiding the model to learn disease progression. Extensive experiments and analyses on IU X-Ray and MIMIC-CXR datasets demonstrate that SR2Gen outperforms previous state-of-the-art methods.**

***Index Terms*—Medical Report Generation, Disease Severity, Knowledge Graph, Momentum Distillation**

I. Introduction

Medical image analysis plays a crucial role in disease detection [1]. In clinical practice, radiologists analyze medical images to detect lesions and assess their severity, which is time-consuming, labor-intensive, and error-prone. To address this challenge, research on automating the generation of medical image reports has garnered increasing attention [2]. In recent years, data-driven neural networks have been widely applied to generate descriptive text for given images (e.g. image captioning tasks) [3]. Compared to image captioning tasks, medical image report generation faces the core challenge of significant visual-to-text data discrepancies [4], [5]. Medical images are often highly similar due to imaging techniques and tissue characteristics, yet key features indicating abnormalities can be subtle and difficult to detect, further complicated by a lack of labeled data. Additionally, there is a large imbalance in the number of samples for common versus rare diseases, making it challenging for models to accurately describe rare lesions.

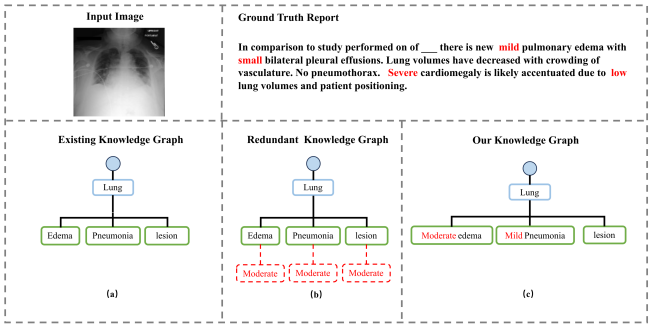

Fig. 1. An example from the MIMIC-CXR dataset and a comparison of knowledge graph construction methods.

To address these challenges, some researchers have proposed more suitable radiology report generation (RRG) methods using the encoder-decoder framework [6]–[11]. Some of these methods enhance the model's performance by integrating medical knowledge into the report generation system. For example, MKG [12] extracts seven organs/tissues and twenty disease descriptions from medical reports as graph nodes, linking disease-related keywords for the same organ

* is Corresponding author.

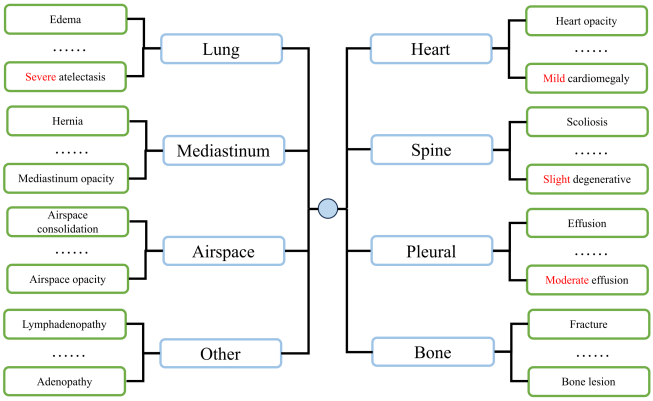

Fig. 2. The knowledge graph we proposed. This graph includes the "Normal" category and eight disease categories, covering seven organ categories and one "Other" category. Each category is further subdivided into its corresponding specific diseases. The severity of the diseases is highlighted in red font.

to the root node, strengthening associations between organs and abnormalities through graph convolution. The knowledge graph constructed by this method [12] has been widely adopted in subsequent studies [13]–[16]. However, these methods still have limitations. As shown in Fig. 1, many words describing disease severity, such as "mild" and "small", appear in real reports. However, as shown in Fig. 1(a), existing knowledge graphs have significant limitations in modeling the severity dimension of diseases, failing to capture severity-related descriptors that frequently appear in reports. As shown in Fig. 1(b), directly adding severity modifiers as graph nodes to the existing knowledge graph oversimplifies the process, resulting in redundancy and making it impossible to ensure the most suitable knowledge scope. In contrast, as shown in Fig. 1(c), we extract severity modifiers for each organ-disease pair from real reports, avoiding redundancy while ensuring the accuracy of prior knowledge. In addition to incorporating prior knowledge, recent studies [17]–[23] have optimized feature representations through joint training of auxiliary tasks. As noted in [20], existing methods have failed to fully exploit diagnostic information in medical images, and the bias in disease distribution has weakened the model's ability to describe rare conditions, diminishing the clinical value of generated reports.

Inspired by these challenges, we propose Severity-Guided Radiology Report Generation (SR2Gen), a novel framework based on knowledge graph expansion and auxiliary classification task optimization. Our framework is composed of two modules: a Knowledge Enhancement Module (KEM) and a Disease Severity-Aware Module (DSAM). These modules are designed to inject disease severity information into the model from both explicit and implicit dimensions. KEM accurately integrates disease severity information into the knowledge graph and uses a graph convolution mechanism to guide the model's focus on disease severity information. Building on the knowledge enriched by KEM, DSAM enhances the Disease Severity-Aware Classifier (DSAC) through momentum distillation, and it also incorporates Adaptive Disease Severity Learning (ADSL), which uses entropy as a dynamic adjustment factor to optimize classification loss, enabling the model to focus more on learning disease severity.

Our main contributions are as follows:

- We propose a more comprehensive knowledge graph that not only contains organ-disease pair information but also covers disease severity, revealing the disease progression.
- We introduce DSAM, which optimizes traditional auxiliary classification tasks through momentum distillation, and addresses the problem of poor learning of minority disease severity information by combining ADSL.
- Extensive experiments on the benchmark datasets IU X-Ray [24] and MIMIC-CXR [25] demonstrate the superiority of the proposed method.

II. Related Work

The goal of RRG is to produce clinical descriptive text for radiological images. Inspired by natural image captioning, most RRG models employ an encoder-decoder architecture to generate reports [14], [21]. However, due to the high visual similarity of radiology images, detecting subtle abnormalities is much more challenging than with natural images. Many methods have been proposed to address this issue, mainly falling into two categories: incorporating knowledge graphs and setting auxiliary tasks.

A. Knowledge Graph

Medical knowledge is crucial for report generation. To incorporate medical knowledge, researchers have experimented with knowledge graphs. For example, KERP [26] employed an encode module that transforms visual features into a structured abnormality graph by incorporating prior medical knowledge. However, the abnormality graph fails to distinguish between organs and diseases. Zhang et al. [12] used a universal graph of seven organs/tissues and twenty finding entities, where entities associated with the same organ were connected. DCL [15] retrieves organ-disease entities during the generation process as a supplement to the pre-built organ-disease graph. Additionally, DCG [16] separately constructed disease-free and disease-specific nodes within the knowledge graphs to enable model to consciously focus on abnormal information and mitigate the impact of excessively common diseases on RRG. None of the above methods consider the importance of disease severity in report generation. Compared with the above methods, as shown in Fig. 2, We propose integrating this crucial information about disease severity into the knowledge graph to help the model generate higher-quality reports.

B. Auxiliary Task Learning

The goal of auxiliary task learning is to enhance the model's ability to extract discriminative features through collaborative learning of multiple related tasks, thereby improving overall performance in the main task. In radiology report generation, common auxiliary tasks include contrastive learning, lesion segmentation, matching medical images with reports, and

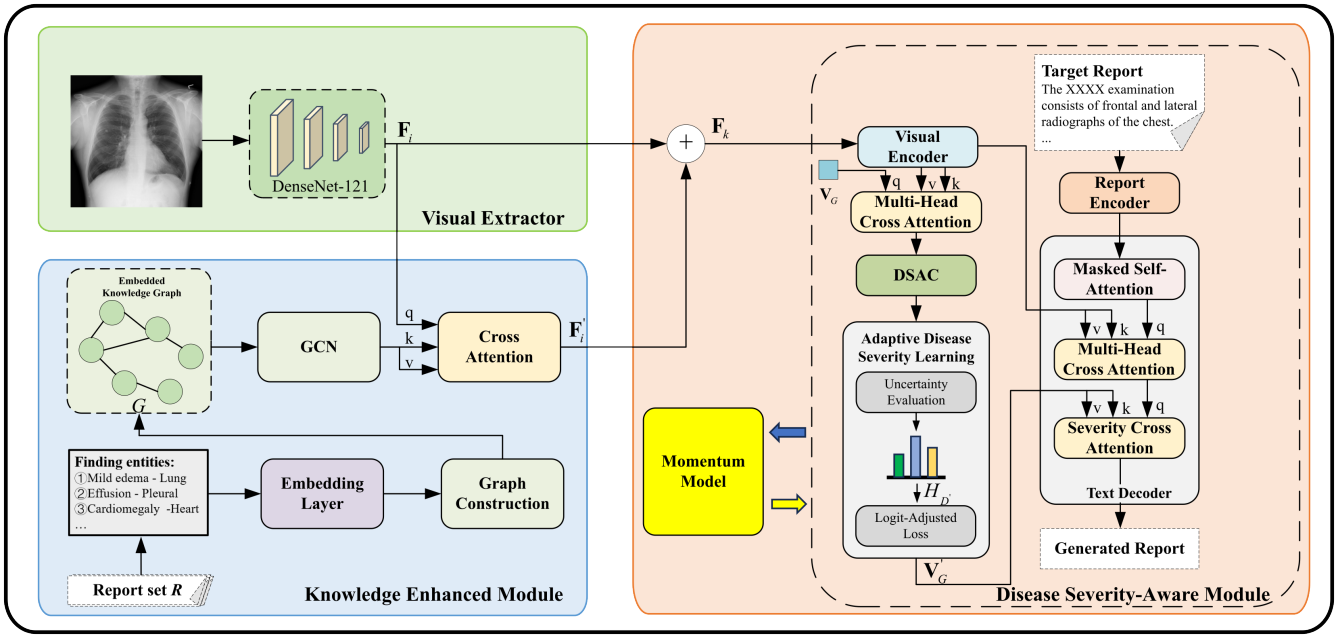

Fig. 3. The overall architecture of the SR2Gen model we proposed. The architecture consists of a visual feature extractor, a knowledge enhancement module, and a disease severity-aware module. We also maintain a momentum version of the disease severity-aware module.

disease classification. Yan et al. [27] introduced weakly supervised contrastive learning as an auxiliary task. This method learns a semantically rich representation to improve the quality of report generation. Meanwhile, lesion segmentation serves as another auxiliary task, providing more detailed disease location information and guiding the model's attention to these critical areas of the body [28], [29].

Furthermore, researchers have explored image-text matching to obtain fine-grained, aligned image-text representations [15], [30]. Notably, disease classification is a commonly used auxiliary task that allows models to distinguish between the presence or absence of abnormalities and acquire discriminative features [20], [21], [31]. All these studies aim to generate more accurate reports by integrating different auxiliary tasks.

Existing methods primarily focus on coarse-grained disease presence identification, whereas our framework innovatively extends auxiliary classification to fine-grained disease severity assessment. We use DSAM and ADSL to guide the model in capturing the underlying patterns of disease progression, generating clinically graded descriptive reports that accurately reflect the different stages of disease development.

III. Methodology

In this section, we introduce the detailed implementation of the proposed SR2Gen. An overview of SR2Gen architecture is presented in Fig. 3, which consists of two main modules: KEM and DSAM. The latter is further divided into two submodules: DSAC and ADSL. We will first briefly describe the RRG task, followed by an introduction to the two main modules we have proposed. Finally, we will present the components related to the decoder and the overall loss function.

A. Task Definition

The RRG task aims to produce a tokenized sequence $Y = \{y_1, ..., y_t, ..., y_T\}$, which describes the input radiology image I. Here, y_i represents a subword token from predefined vocabulary H, T is the sequence length in tokens. The process of generating the tokenized sequence can be expressed as:

$$P(Y|I) = \prod_{t=1}^{T} P(y_t|y_1, ..., y_{t-1}, I). \tag{1}$$

Typically, the model is trained by maximizing the conditional probability $p(Y|I)$ through minimizing the negative log-likelihood of Y given I, as expressed in the equation:

$$\theta^* = \arg_\theta \max \sum_{t=1}^{T} \log p(y_i|y_1, ..., y_{t-1}, I; \theta), \tag{2}$$

where θ are the parameters of the model. In this paper, we adopt the standard image captioning structure, including an image encoder and an auto-regressive text decoder, while also introducing a severity-aware mechanism to enhance the clinical relevance of generated reports.

Fig. 3 illustrates the overall architecture of SR2Gen. After extracting visual features, we enhance the model with disease severity knowledge through a knowledge graph and auxiliary classification tasks. Specifically, we propose two modules: KEM and DSAM, which inject disease severity information into the model from both explicit knowledge injection and implicit feature learning perspectives. This multi-faceted approach facilitates effective knowledge transfer and utilization, significantly improving the model's ability to discern disease severity. Finally, we use a text decoder to generate the reports.

B. Knowledge Enhancement Module

Introducing knowledge graph as structured prior knowledge has been shown to significantly improve model's clinical semantic understanding in RRG. By embedding disease severity information in the graph, the model better captures disease progression. As shown in Fig. 1(c), we propose a small, task-focused knowledge graph incorporating severity information while minimizing redundant data.

1) Severity Knowledge Graph Construction: As shown in Fig. 2, we construct a knowledge graph incorporating disease severity, reflecting the relationships between organs and diseases and how severity levels manifest in different organ-disease pairs. It should be noted that the knowledge graph we proposed has been expanded according to [12] and [16]. Our knowledge graph contains two types of nodes: organ nodes and disease nodes. Organ nodes represent different organs (e.g. "Lung", "Heart", "Bone") and disease nodes formed by combining severity modifiers with disease names (e.g. "Severe atelectasis", "Moderate effusion"). Furthermore, the knowledge graph also includes a "Normal" node, representing the absence of disease. The disease nodes are linked to their corresponding organ nodes, with severity modifiers providing more accurate clinical information. This approach reduces confusion caused by similar severity levels in different pairs of organ-diseases. During graph construction, we filter severity modifiers based on their frequency in the training set for each disease, generating organ-disease pairs with severity information, which are then encoded as nodes in the graph.

2) Knowledge Enhancement: To accurately extract the knowledge relevant to the input image I_{input} from the knowledge graph, we adopt a retrieval-augmented approach that retrieves a similar image I_{sim}, and uses its corresponding report to activate nodes in the graph. Specifically, for I_{input}, we retrieve top-k similar images from a pre-built index set $D = \{Avg(I_i)\}_{i=1}^{M}$, where $Avg(\cdot)$ denotes global average pooling over features extracted by DenseNet-121 [32], and M represents the total number of training images. We compute cosine similarities between $Avg(I_{input})$ and D to identify the top-k similar images. The reports of the top-k images are combined to form a candidate report set $R = [Y_1, Y_2, ..., Y_k]$, from which organ-disease entities are extracted.

As illustrated in Fig. 2, Throughout the training set, the organ-disease entities are designated as M' nodes $V = \{v_1, v_2, ..., v_{M'}\}$ in the disease severity graph $G = \{E, V\}$, where E and V represent the edges and nodes, respectively. Node relationships include "exists" and "does not exist". Each node $v \in V$ is encoded by ClinicalBERT [33]. The [CLS] embedding from the final layer of ClinicalBERT is used as the node embedding $\mathbf{F}_v = \{f_{v_i} \in \mathbb{R}^d\}_{i=1}^{M'}$. If a predefined organ-disease pair appears in the report set R, the corresponding edge $e_j \in E$ in G is activated. For example, if "Heart-Mild cardiomegaly" appears, the edge between "Heart" and "Mild cardiomegaly" is activated. After constructing the graph G, we use a Graph Convolutional Network (GCN) to aggregate disease-related features and update node embeddings $\mathbf{F}'_v \in \mathbb{R}^{M' \times d}$. Multi-Head Cross Attention (MHCA) is then applied to enhance the fine-grained image patch embeddings $\mathbf{F}_i$:

$$\mathbf{F}'_i = MHCA\left(\mathbf{F}'_v, \mathbf{F}_i\right), \tag{3}$$

and the final output is obtained by summing $\mathbf{F}_i$ and $\mathbf{F}'_i$:

$$\mathbf{F}_k = \mathbf{F}'_i + \mathbf{F}_i, \tag{4}$$

where $\mathbf{F}_k \in \mathbb{R}^{N \times d}$ combines visual features and disease severity information, aiding the text decoder in report generation.

C. Disease Severity-Aware Module

Incorporating auxiliary classification tasks into report generation improves the model's ability to discriminate diseases. However, traditional classification tasks typically use hard labels that only indicate disease presence, neglecting severity. To address this, inspired by the work of [21] and leveraging prior knowledge from the knowledge graph, we further introduce momentum distillation to generate pseudo-labels, enabling the model to implicitly learn disease severity information. Meanwhile, we propose ADSL to address the issue that class imbalance among diseases leads to insufficient learning of severity information for minority diseases.

1) Enhancement of Classification Task: To enhance the classification task's ability to identify disease severity, we define a set of good classification labels and avoid using the graph nodes as labels, as they contain severity information that may bias the model's learning. After defining the labels, we use a topic embedding mechanism to represent various labels in a continuous vector space, denoted as $\mathbf{V}_G$. Then, we apply MHCA to allow $\mathbf{V}_G$ to learn severity information from $\mathbf{F}_k$:

$$\mathbf{V}'_G, \alpha_{cls} = ADSL\left(DSAC\left(MHCA\left(\mathbf{F}_k, \mathbf{V}_G\right)\right)\right), \tag{5}$$

where $ADSL$ denotes adaptive disease severity learning, which will be introduced later, $\mathbf{V}'_G$ represents the features that contain the classification results and will be utilized in subsequent processes to further assist the decoder in generating precise descriptions of disease severity, while α_{cls} serves as the attention scores, functioning as the probability distribution used for classification loss calculation. Then, we use momentum distillation to generate pseudo-labels and calculate the classification loss. During the training process, we maintain a momentum version of the DSAM with the same structure as the original model, essentially a continuously evolving teacher model. The update process of the momentum model at time step t combines its historical parameters θ_m^{t-1} (influenced by the momentum factor $\alpha \in [0, 1)$) with the current parameters $\theta_{original}^{t-1}$ of the original model:

$$\theta_m^t = \alpha\theta_m^{t-1} + (1-\alpha)\theta_{original}^{t-1}. \tag{6}$$

In DSAC, we denoted P_{cls} and P_{cls_m} as the parameters of classifier $C(\cdot)$ and momentum-based classifier $C_m(\cdot)$, respectively. After updating P_{cls_m}, we pass visual representations through $C_m(\cdot)$ to construct pseudo-labels using the categorical truth label L_g and obtained α'_{cls}, we define the following distillation-base classifier loss function:

$$P_\theta^{target} = \alpha_{dis} L_{gt} + (1-\alpha_{dis})\text{softmax}(\alpha'_{cls}), \tag{7}$$

TABLE I
THE STATISTICS OF IU X-RAY AND MIMIC-CXR.

Dataset	Split	#Images	#Reports	#Patients	Avg. Len.
IU X-Ray [24]	Train	5,212	2,780	2,780	38.29
	Val	720	402	402	36.58
	Test	1,534	800	800	37.63
MIMIC-CXR [25]	Train	368,960	222,758	64,586	53.00
	Val	2,991	1,808	500	53.05
	Test	5,159	3,269	293	66.40

$$\mathcal{L}_{cls} = KL(P_\theta^{target} \| \alpha_{cls}), \tag{8}$$

where $\alpha_{dis} \in [0, 1)$ is the distillation coefficient, and we use Kullback-Leibler divergence to minimize the difference between the targeted probability distribution P_θ^{target} and α_{cls}.

2) Adaptive Disease Severity Learning: The imbalanced distribution of diseases leads to uneven learning of disease severity. To address this, we propose ADSL, an algorithm that automatically adjusts learning objectives for different diseases based on the entropy values. To balance the learning across diseases, we introduce the logit-adjusted loss [34], which encourages a large relative margin between logits of rare versus dominant labels. Specifically, given a specific disease, the logit-adjusted loss with respect to the positive label can be expressed as follows:

$$\mathcal{L}'_{D'}(y = P, f(x)) = -\log \frac{e^{f_y(x) + \log \pi_{D'}}}{\sum_{y' \neq P} e^{f_{y'}(x)} + \left(e^{f_y(x) + \log \pi_{D'}}\right)}, \tag{9}$$

where $f_y(x)$ is the logit of class y, and $\pi_{D'}$ is the prior probability of disease D'. However, Jin et al. [20] mentions that the fixed class distribution which is used for logit adjustment is unable to reflect the learning dynamics of diseases. This is because these diseases have both diverse distributions and different learning difficulties. To solve this problem, we propose to use entropy, which can represent uncertainty, as an adjustment factor to dynamically adjust the logit:

$$H_{D'} = -\pi_{D'} \log \pi_{D'}, \tag{10}$$

$$\mathcal{L}_{D'}(y = P, f(x)) = -\log \frac{e^{f_y(x) + \log(\pi_{D'} - H_{D'})}}{\sum_{y' \neq P} e^{f_{y'}(x)} + \left(e^{f_y(x) + \log(\pi_{D'} - H_{D'})}\right)}, \tag{11}$$

where $H_{D'}$ is the entropy of D'. The loss against non-positive labels remains the same as standard cross-entropy loss $\mathcal{L}_{CE}$. Through this entropy value, ASDL dynamically adjusts the weight of each class. During the training process, by dynamically adjusting the logit based on entropy, the model can more accurately focus on the categories that need learning, improving severity recognition. Meanwhile, we regularize the decoder using momentum distillation to reduce its focus on frequently occurring words and enhance its attention to important, less frequent words that represent disease severity. This process can be expressed by the following formula: $\mathcal{L}_{MD} = KL(O_d \parallel O_{d_m})$, where O_d is the output of decoder, O_{d_m} is the output of momentum-based decoder. The loss function of DSAM is composed of DSAC and MD, with ADSL serving as enhancement to DSAC for better handling class imbalance, and is formulated as follows:

$$\mathcal{L}_{DSAM} = \mathcal{L}_{DSAC} + \mathcal{L}_{MD}. \tag{12}$$

D. Decoder and Loss Function

After obtaining the $\mathbf{V}'_G$ which encodes the knowledge about disease severity in Eqn. 5, we use an n-layer transformer as the decoder to generate the final radiology reports. Formally, let $R' = \{r_0, r_1, \ldots, r_t\}$ represent a report which consists of t words that has been generated. The decoder takes the report R' as input and predicts the next word auto-regressively, conditional on the $\mathbf{F}_k$ in Eqn. 4 and $\mathbf{V}'_G$ in Eqn. 5:

$$\begin{aligned} \mathbf{h}_t^v &= MHCA\left(\mathbf{h}'_t, \mathbf{F}_k, \mathbf{F}_k\right), \\ \mathbf{h}_t^{cls} &= MHCA\left(\mathbf{h}_t^v, \mathbf{V}'_G, \mathbf{V}'_G\right), \end{aligned} \tag{13}$$

where $\mathbf{h}'_t$ is obtained by applying a masked multi-head attention mechanism to the previously generated word $\mathbf{h}_t$. Finally, the $\mathbf{h}_t^{cls}$ is passed to a Feed-Forward Network [35] and linear layer to predict the next word:

$$\mathbf{h}_{t+1} = \text{softmax}\left(FFN\left(\mathbf{h}_t^{cls}\right)\mathbf{W}_p + \mathbf{b}_p\right), \tag{14}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the learnable parameters. Given the ground truth report $Y^* = \{y_1^*, \ldots, y_2^*, \ldots, y_T^*\}$, we use cross entropy loss to optimize the report generation task:

$$\mathcal{L}_{CE}(\theta) = -\sum_{i=1}^{T} \log\left(p_\theta\left(y_i^* \mid y_{1:i-1}^*\right)\right), \tag{15}$$

the total training loss of our model is

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{DSAM} + \lambda_2 \cdot \mathcal{L}_{CE}, \tag{16}$$

where λ_1 and λ_2 are loss weighting hyper-parameters.

IV. EXPERIMENTS

A. Datasets, Metrics and Settings

We evaluate the effectiveness of our SR2Gen on two established benchmarks for report generation: IU X-Ray [24] and MIMIC-CXR [25].

Datasets. Consistent with the data splits used in previous works [4], [11], [22], we partition the IU X-Ray dataset into training (70%), validation (10%), and test (20%) sets, excluding samples that lack both image views. For MIMIC-CXR, we use its official split [25]. The patient distribution across the training, validation, and test sets is non-overlapping. Table I presents the statistics of these two datasets.

Metrics. For radiology report generation, we use three widely adopted evaluation metrics: BLEU [36], ROUGE-L [37], and METROR [38]. BLEU assesses the precision of n-grams (typically up to 4-grams) in machine-generated text compared to reference texts. ROUGE-L focuses on recall by identifying the longest common subsequences between generated and reference texts. METEOR combines unigram precision and recall, considering alignment, stemming, and synonym matching to enhance its evaluation.

TABLE II
THE PERFORMANCES OF OUR PROPOSED SR2GEN COMPARED WITH OTHER STATE-OF-THE-ART METHODS ON IU X-RAY AND MIMIC-CXR DATASET. THE HIGHEST SCORES ARE HIGHLIGHTED IN BOLD, THE SECOND-HIGHEST SCORES ARE INDICATED WITH AN UNDERLINE.

Datasets	Methods	BL-1	BL-2	BL-3	BL-4	METEOR	RG-L
IU X-Ray	R2Gen [11]	0.47	0.304	0.219	0.165	0.187	0.371
	PPKED [14]	0.483	0.315	0.224	0.168	0.19	0.376
	XPRONET [22]	0.525	0.357	0.262	0.199	0.22	0.411
	DCL [15]	-	-	-	0.163	0.193	0.383
	EKAGen [23]	0.526	0.361	0.267	0.203	0.214	0.404
	DCG [16]	0.514	0.33	0.241	0.186	0.211	0.401
	GMoD [21]	**0.53**	0.363	0.267	0.203	**0.217**	0.418
	Ours	0.528	**0.38**	**0.282**	**0.214**	0.212	**0.42**
MIMIC-CXR	R2Gen [11]	0.353	0.218	0.145	0.103	0.142	0.27
	PPKED [14]	0.36	0.224	0.149	0.106	0.149	0.284
	XPRONET [22]	0.344	0.215	0.146	0.105	0.138	0.279
	DCL [15]	-	-	-	0.109	0.15	0.284
	EKAGen [23]	**0.419**	0.258	0.17	0.119	0.157	0.287
	DCG [16]	0.397	0.258	0.166	0.126	0.162	0.295
	GMoD [21]	0.398	0.251	0.172	0.124	**0.166**	0.286
	Ours	0.395	**0.26**	**0.175**	**0.128**	0.158	**0.299**

Settings. In this work, we use the pre-trained DenseNet-121 to extract image features. For the encoder-decoder backbone, it is complemented by an 6-layer image encoder, a 6-layer text encoder and a 15-layer report decoder for report generation. For the definition of classification labels, we extract the top 100 most frequently occurring symptom phrases, combined with the 14 types of lung diseases extracted from CheXpert [39], to form the labels for the classification task. We used an Adam [40] optimizer with a weight decay of 5e-5, and set the learning rate to 1e-4. For knowledge graph generation, we followed the processing method in DCG [16] to extract predefined lists of organs and diseases. Subsequently, pre-trained ClinicalBERT is used to extract finding entities as node embedding. We used 3-layer GCN to aggregate the disease-related features through the knowledge graph. The momentum coefficient was set to 0.995, and the distillation coefficient was set to 0.995 for both datasets. The model was trained on the RTX 4090 GPU with a batch size of 64 and 30 epochs.

B. Quantitative Results

In Table II, we compare our SR2Gen with several state-of-the-art radiology report generation systems across two benchmark datasets. R2Gen [11] has been widely used as a baseline RRG model in recent years. PPKED [14], DCL [15], and GMoD [21] aim to integrate medical knowledge into typical RRG baseline models. XPRONET [22] and EKAGen [23] use auxiliary task learning to enhance feature extraction and improve performance. Since we used the same setup, we directly refer to the results from the original papers. As shown in Table II, our SR2Gen achieves state-of-the-art performance on most evaluation metrics, with a 1.7% improvement in the BLEU-2 score and a 1.1% improvement in the BLEU-4 score on the IU X-Ray dataset. SR2Gen scores slightly lower than GMoD and EKAGen on a few metrics, which may be due to GMoD's graph-driven classification being more favorable for generating short texts, and EKAGen's aggregated discriminative attention maps leveraging weak supervision signals to generate discriminative regions while reducing background influence. Furthermore, our results indicate that incorporating prior medical knowledge and severity guidance can better integrate diagnostic information, improving report generation.

C. Ablation Study

In order to comprehensively investigate the contributions made by our proposed DSAC, KEM, ADSL, and MD, major results are shown in Table III.

Effect of DSAC. Our DSAC learns disease severity information from pseudo-labels generated through momentum learning. Compared to the baseline model with setting (a), DSAC significantly improves performance, with a 9.4% increase in BLEU-1 and a 5.9% increase in BLEU-4. This highlights the effectiveness of introducing the momentum-based auxiliary classification task, which helps the model acquire more discriminative features and integrate disease severity knowledge.

Effect of KEM. Next we use a knowledge graph with disease severity information to enhance visual representations. The performance of setting (b) shows that introducing the knowledge graph further improves performance over setting (a). This underscores the importance of prior medical knowledge, as the complexity of medical images and the gap between image and text modalities can severely reduce the representational power of features.

Effect of ADSL and MD. Additionally to address the issue of poor learning of disease severity due to data bias in classification tasks, we adopted ADSL. As shown in setting (c), ADSL effectively alleviates the above issue. Learning disease severity information improved both the BLEU-2 and BLEU-3 scores by 0.6%. Furthermore, as shown in setting (d), momentum distillation-based decoder regularization slightly boosts the model's performance.

TABLE III

THE ABLATION STUDY OF SR2GEN ON THE IU X-RAY DATASET. THE BASELINE REPRESENTS THE SIMPLEST ENCODER-DECODER STRUCTURE WE HAVE IMPLEMENTED. KEM DENOTES THE KNOWLEDGE-ENHANCED MODULE, DSAC REPRESENTS THE DISEASE SEVERITY-AWARE CLASSIFIER, ADSL INDICATES THE ADAPTIVE DISEASE SEVERITY LEARNING AND MD REPRESENTS THE MOMENTUM DISTILLATION CONSTRAINT ON THE TEXT DECODER.

Settings	DSAC	KEM	ADSL	MD	BL-1	BL-2	BL-3	BL-4	METEOR	RG-L
baseline					0.421	0.273	0.19	0.142	0.173	0.321
(a)	✓				0.515	0.362	0.267	0.201	0.205	0.41
(b)	✓	✓			0.52	0.373	0.275	0.21	0.211	0.411
(c)	✓	✓	✓		0.523	0.379	0.281	0.213	**0.212**	0.413
SR2Gen	✓	✓	✓	✓	**0.528**	**0.38**	**0.282**	**0.214**	**0.212**	**0.420**

Image	Ground Truth	Retrieved Pairs	Baseline	Ours
 IU-Xray	There is moderate cardiomegaly. There are bilateral interstitial opacities, increased since the previous exam. No focal airspace consolidation, pleural effusions or pneumothorax. No acute bony abnormalities.	cardiomegaly-heart moderate cardiomegaly-heart no edema-lung no effusion-pleural no pneumothorax-pleural no airspace consolidation-airspace	Lungs are clear. ~~Heart and mediastinal contours are normal~~. There is no pneumothorax or pleural effusion.	Moderate cardiomegaly. The lungs are clear. There is no focal airspace disease. No pneumothorax or pleural effusion.
 MIMIC-CXR	Frontal and lateral views of the chest. Severe cardiomegaly has increased since ___ with right and left atrial enlargement, consistent with right heart decompensation. Lung volumes are low with a possibly small left pleural effusion. No focal consolidation or pneumothorax. A left subclavian vascular stent is new since the prior exam.	low volume-lung small effusion-pleural no pneumothorax-pleural cardiomegaly-heart no consolidation-lung no edema-lung normal-airspace ~~mild enlarge-heart~~	Pa and lateral views of the chest were obtained. There is no focal consolidation or pneumothorax. The lung fields are clear. The ~~cardiomediastinal silhouette is normal~~. There is ~~no pleural abnormality~~.	Frontal and lateral views of the chest were obtained. The lung volumes are low which may suggest small pleural effusion. ~~The size of the cardiac silhouette is likely mildly enlarged~~. The mediastinal and hilar contours are unremarkable. There is no pneumothorax. The lungs are well expanded without focal consolidation concerning for pneumonia.

Fig. 4. Illustrations of reports from ground truth, ours and baseline and retrieved disease-organ pairs from two datasets. Different colors are used to highlight various medical entities for better visualization.

D. Qualitative Analysis

Case Study. As shown in Fig. 4, to further explore the effectiveness of our method, we conducted a qualitative analysis on the IU X-Ray [24] and MIMIC-CXR [25] datasets. The analysis includes the disease-organ pairs retrieved from the datasets, as well as the real reports, the reports generated by our model, and the reports generated by the baseline model. In these two examples, we use different colors to highlight keywords related to organs and diseases for clear distinction, with keywords related to disease severity highlighted in red boxes. From these two examples, it can be seen that the reports generated by our model are highly consistent with the real reports. Phrases such as "moderate cardiomegaly" and "small pleural effusion" accurately describe the severity of the disease. This indicates that our model can effectively focus on lesions in radiographic images and provide appropriate descriptions. In contrast, the reports generated by the baseline model fail to describe the severity of the disease and may even misrepresent the lesion.

Error Analysis. The red strikethrough parts in Fig. 4 represent errors in the generated reports. In the example from MIMIC-CXR, although our model successfully detected cardiomegaly, its severity was incorrectly judged as "mild". We observed that this occurred because the retrieved disease-organ pair contained erroneous information: "mild cardiomegaly". Using more accurate retrieval methods could help alleviate this issue. Additionally, our model identified "small pleural effusion on the left side" as simply "small pleural effusion", failing to describe the disease's location. To address this issue, auxiliary segmentation tasks could be introduced in model to enhance its ability to extract location-specific features.

V. CONCLUSION

In this paper, we propose a novel SR2Gen method for generating radiology reports that focus on disease severity. Our model explicitly learns the relationship between visual representations and disease severity by incorporating an expanded knowledge graph, and implicitly learns feature representations related to disease severity through the Disease Severity-Aware Module. Extensive experiments and analysis on the IU X-Ray and MIMIC-CXR datasets validate the effectiveness of our SR2Gen method, with results showing that learning disease severity improves the accuracy of generated radiology reports.

ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China under Grant 2023YFB2904000 and the Natural Science Foundation of the Fujian Province, China, under Grant 2022J01574.

REFERENCES

[1] S. Wang and R. M. Summers, "Machine learning and radiology," *Medical image analysis*, pp. 933–951, 2012.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[3] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[4] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, 2018.

[5] M. Li, Y. Chen, R. Ji, X. Wu, Y. Wong, and Y. Yang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *World Wide Web*, pp. 253–270, 2023.

[6] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 753–13 762.

[7] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 457–466.

[8] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *International conference on medical image computing and computer-assisted intervention*, 2019, pp. 721–729.

[9] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[10] B. Kang, Y. Xiong, J. Jiao, Y. Zhang, X. Jia, and J. Li, "Bridging the gap: Cross-modal knowledge driven network for radiology report generation," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2023, pp. 1202–1209.

[11] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *arXiv preprint arXiv:2010.16056*, 2020.

[12] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, pp. 12 910–12 917.

[13] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Medical image analysis*, p. 102510, 2022.

[14] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 753–13 762.

[15] M. Li, Y. Chen, R. Ji, X. Wu, Y. Wong, and Y. Yang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3334–3343.

[16] X. Liang, Z. Hu, H. Zhang, E. P. Xing, and D. Xu, "Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1233–1242.

[17] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[18] Z. Wang, J. Liu, G. Li, and W. Zuo, "Automated radiographic report generation purely on transformer: A multicriteria supervised approach," *IEEE Transactions on Medical Imaging*, pp. 2803–2813, 2022.

[19] B. Yan and M. Pei, "Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2982–2990.

[20] H. Jin, C. Che, F. Liu, F. Yin, K. McKeown, D. Rubin, E. Xing, and W. Xiong, "Promptmrg: Diagnosis-driven prompts for medical report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 2966–2974.

[21] Z. Xiang, Y. Chen, R. Ji, X. Wu, Y. Wong, and Y. Yang, "Gmod: Graph-driven momentum distillation framework with active perception of disease severity for radiology report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024, pp. 123–133.

[22] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *European Conference on Computer Vision*, 2022, pp. 563–579.

[23] S. Bu, F. Liu, S. Ge, X. Wu, W. Fan, and Y. Zou, "Instance-level expert knowledge and aggregate discriminative attention for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 534–12 543.

[24] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, pp. 304–310, 2016.

[25] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[26] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 6666–6673.

[27] A. Yan, Y. He, B. Shao, L. Li, C. Lin, J. Xu, C. Zhang, J. Xiao, Y. Xu, E. Xing *et al.*, "Weakly supervised contrastive learning for chest x-ray report generation," *arXiv preprint arXiv:2109.12242*, 2021.

[28] S. Bu, F. Liu, S. Ge, X. Wu, W. Fan, and Y. Zou, "Instance-level expert knowledge and aggregate discriminative attention for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 534–12 543.

[29] T. Gu, X. Wu, F. Liu, S. Ge, W. Fan, and Y. Zou, "Complex organ mask guided radiology report generation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1971–1980.

[30] B. Yan and M. Pei, "Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2982–2990.

[31] J. Wang, A. Bhalerao, and Y. He, "Camanet: class activation map guided attention network for radiology report generation," *IEEE Journal of Biomedical and Health Informatics*, pp. 2199–2210, 2024.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[33] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[34] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," *arXiv preprint arXiv:2007.07314*, 2020.

[35] G. Bebis and M. Georgiopoulos, "Feed-forward neural networks," *Ieee Potentials*, pp. 27–31, 1994.

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[37] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[38] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[39] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 590–597.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.