Multi-Grained Radiology Report Generation With Sentence-Level Image-Language Contrastive Learning

Aohan Liu, Yuchen Guo, Jun-Hai Yong, and Feng Xu

***Abstract*— The automatic generation of accurate radiology reports is of great clinical importance and has drawn growing research interest. However, it is still a challenging task due to the imbalance between normal and abnormal descriptions and the multi-sentence and multi-topic nature of radiology reports. These features result in significant challenges to generating accurate descriptions for medical images, especially the important abnormal findings. Previous methods to tackle these problems rely heavily on extra manual annotations, which are expensive to acquire. We propose a multi-grained report generation framework incorporating sentence-level image-sentence contrastive learning, which does not require any extra labeling but effectively learns knowledge from the image-report pairs. We first introduce contrastive learning as an auxiliary task for image feature learning. Different from previous contrastive methods, we exploit the multi-topic nature of imaging reports and perform fine-grained contrastive learning by extracting sentence topics and contents and contrasting between sentence contents and refined image contents guided by sentence topics. This forces the model to learn distinct abnormal image features for each specific topic. During generation, we use two decoders to first generate coarse sentence topics and then the fine-grained text of each sentence. We directly supervise the intermediate topics using sentence topics learned by our contrastive objective. This strengthens the generation constraint and enables independent fine-tuning of the decoders using reinforcement learning, which further boosts model performance. Experiments on two large-scale datasets MIMIC-CXR and IU-Xray demonstrate that our approach outperforms existing state-of-the-art methods, evaluated by both language generation metrics and clinical accuracy.**

***Index Terms*— Medical report generation, contrastive learning, multi-grained.**

Manuscript received 5 December 2023; revised 17 February 2024; accepted 26 February 2024. Date of publication 5 March 2024; date of current version 1 July 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0704000 and Grant 2023YFC3305600, and in part by the National Natural Science Foundation of China under Grant 62021002. *(Corresponding authors: Yuchen Guo; Jun-Hai Yong; Feng Xu.)*

Aohan Liu, Jun-Hai Yong, and Feng Xu are with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: liuaohan123@163.com; yongjh@tsinghua.edu.cn; xufeng2003@gmail.com).

Yuchen Guo is with BNRist, Tsinghua University, Beijing 100084, China (e-mail: yuchen.w.guo@gmail.com).

Digital Object Identifier 10.1109/TMI.2024.3372638

I. Introduction

RADIOLOGY report generation aims to generate coherent and accurate textual descriptions for radiology images automatically. It is an interdisciplinary research field, with a technical relationship with image captioning [1], [2] in artificial intelligence, and clinical value for helping radiologists to produce accurate imaging reports with less workload [3], [4]. Due to its technical and practical importance, report generation has been recognized as a frontier research topic, and attracted considerable research interest recently [3], [5], [6], [7], [8], [9].

Report generation has a problem formulation similar to image captioning [10], [11], [12], i.e., generating texts from input images. However, it has two unique features and challenges compared to conventional image captioning. First, image captioning has an unbiased preference for all concepts, while report generation cares more about abnormal findings [13] which can indicate various diseases. In medical practice, as an all-around description of medical images, the imaging reports usually contain overwhelming normal descriptions, while the important abnormal descriptions are less (Fig.1). This scarcity and imbalance greatly hinder models from learning abnormal features in images when trained on conventional text generation objectives. Second, image captioning usually has short text targets [1], while report generation needs to generate long paragraphs consisting of multiple sentences with various topics. Conventional text generation approaches [14], [15] where texts of different sentences in the whole paragraph are predicted as a single sequence do not conform to this feature and hinder the model from describing each specific topic in a fine-grained way.

To address the above challenges, many approaches have been proposed previously. To prevent the model from being dominated by the normal text descriptions, previous methods mainly resort to extra labels or knowledge to force the model to learn abnormal features by performing auxiliary tasks such as classification, or injecting the prior knowledge into the model architecture. However, the acquirement of the labels and the construction of knowledge graphs is burdensome and not well generalizable to new modalities or datasets. To generate long and multi-sentence reports, hierarchical recurrent neural networks (HRNN) [3], [7], [8], [16] and Transformers have been used to either prevent the decoder from gradient

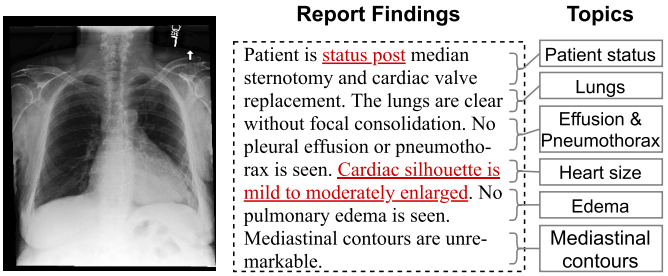

Fig. 1. An example of medical imaging reports in MIMIC-CXR. Abnormal findings are highlighted. Topics of sentences are estimated by us for better illustration instead of being predefined.

vanishing or directly exploit the long-range dependencies of words in the report. However, neither of these methods actually conforms to the multi-topic nature of radiology reports as they do not have direct supervision of the generated topics.

In this work, we propose a multi-grained report generation (MRG) approach incorporating a novel sentence-level image-language contrastive learning (SILC) method. We first introduce SILC, a fine-grained image-language contrastive learning objective to enable better learning of abnormal image features without requiring extra labels. SILC treats the report as a set of semantically independent sentences and learns a topic feature and a content feature for each sentence. The sentence topic is used to extract topic-guided image content from the image features, which is supervised contrastively between the sentence content. SILC encourages the learning of fine-grained contrastive relationships between image and report, which enables the model to capture subtle yet informative patterns in radiology images. We then develop the MRG model based on SILC. The model first generates the coarse-grained topics of sentences that the report should describe given the input image. Then it generates the fine-grained texts of each sentence based on the topics and the image. Unlike HRNN which applies sequence generation loss only on the final texts, MRG directly supervises the intermediate topics using the ground truth sentence topics learned by SILC. In particular, to enable better topic generation, we utilize vector quantization (VQ) [17], [18] to map the topic features to discrete feature tokens so that the model generates topics in a finite space without outliers. Furthermore, MRG enables a two-stage finetuning of the model using reinforcement learning [19], which substantially improves model performance. Our novelty lies in the utilization of the multi-topic nature of radiology reports in the framework by introducing SILC and MRG. To the best of our knowledge, we are the first to explicitly learn and supervise the sentence topics without manual definition or extra supervision. We evaluate our approach on two large-scale datasets - MIMIC-CXR [20] and IU-Xray [21] of chest X-Ray exams. Experimental results demonstrate the superior performance of our method, which surpasses existing state-of-the-art methods on both conventional language generation metrics and clinical accuracy [13]. Ablation study and qualitative analysis demonstrate the effectiveness of our design.

In summary, the contributions of our work are as follows:

1) We propose a sentence-level image-language contrastive learning (SILC) approach, which forces the model to learn fine-grained image features for each type of sentence topic.
2) We propose a multi-grained report generation approach based on SILC, which first generates coarse-grained sentence topics to talk about and then generates the fine-grained texts of each sentence, conforming to the multi-topic nature of radiology reports.
3) We perform extensive experiments on two large-scale datasets, MIMIC-CXR and IU-Xray, and show that our approach outperforms previous state-of-the-art approaches on both natural language generation metrics and clinical accuracy.

II. Related Works

A. Image Captioning

Image captioning aims at generating brief textual descriptions for given images, and has attracted extensive research interests in recent years [1], [10], [11], [12]. Image captioning approaches are mostly based on the encoder-decoder architecture, where an image encoder first captures image features which are then fed to a text decoder to generate the captions. Different encoder and decoder architectures including CNN, RNN, and Transformer [22] have been explored. The use of various training approaches including reinforcement learning [19] and adversarial training [23], has also been shown effective in the image captioning task.

B. Medical Imaging Report Generation

The automatic generation of clinically accurate reports for medical images has attracted extensive research interest in recent years. Existing approaches are usually derived from image captioning methods, but focus on the unique features of medical images and reports. Report generation approaches could be mainly divided into two categories. The first category uses only image-report pairs for training, which are easy to acquire. These works aim at learning image features and report patterns from the training data more effectively. Contrastive learning [8] and attention mechanism [5] are used to help the model capture distinctive image features. Hierarchical decoding [3], retrieval-based methods [24], and memory mechanism [6] are used to better generate long, regular, and patterned reports. More recent works utilize reinforcement learning to boost model performance [8], [25]. Methods in the other category use extra domain-specific data during training. Some works use tags or disease labels as extra supervisions by performing classification [3], [26], feature clustering [7] or report generation fine-tuning [27], [28]. Auxiliary knowledge has also been used to help the report generation task [29]. Recently, works that use large-scale datasets and disease labels for model pretraining using objectives such as BERT and classification have also shown promising results for report generation by performing transfer learning [30], [31]. While the utilization of additional data improves model performance, the acquirement of these data is expensive, burdensome, and sometimes infeasible. Our work belongs to the first category, using only images and reports for training.

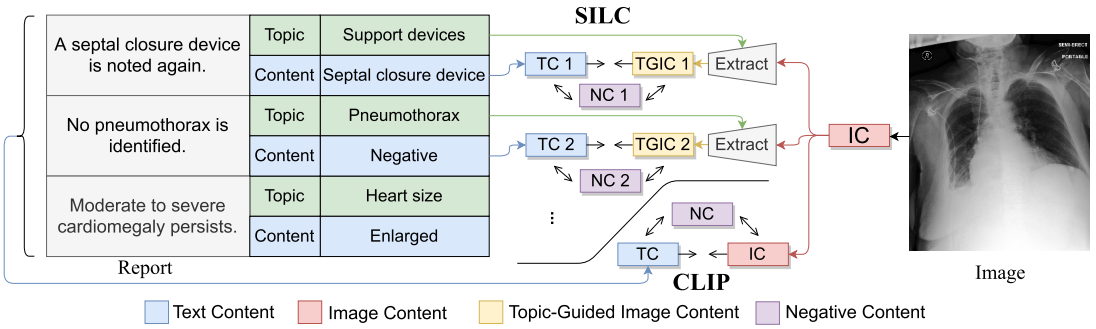

Fig. 2. Conventional and sentence-level contrastive learning. Conventional image-text contrastive learning (CLIP) views the whole report as a whole, while we perform fine-grained contrastive learning between images and semantically independent sentences. Note that the topics and contents are estimated by us just for better illustration. Our method does not use these topics or content definitions.

C. Image-Language Contrastive Learning

Image-language contrastive learning is widely used in weakly-supervised model pretraining. In general, texts and images are fed into encoders, and the output features are supervised by contrastive objectives. It has been shown as an effective method to learn image and text encoders from large-scale datasets [32] and also produces promising results in medical image pretraining [33]. It is also utilized in report generation approaches to help the models learn better image and text features [7]. However, these works treat the report as a whole despite the fact that it is composed of multiple sentences with different topics and contents. In [34], a sentence is randomly selected from the report as the text sample, which forces the image encoder to learn distinctive features for each sentence. However, as different sentences have different topics and contents, it is hard to constrain that all the sentence features in a report are close to the image feature. In [35], a refined contrastive learning method is proposed for pretraining based on the attention mechanism, but it processes word pieces independently, omitting the high-level semantics of sentences. MedCLIP [36] enables contrastive learning between sentences and images but relies heavily on extra image and text labels. Different from these works, we propose a sentence-level image-language contrastive learning approach, which treats each sentence separately and enables the encoder to learn fine-grained image and text features for each topic.

D. Hierarchical Decoding

The hierarchical decoding mechanism has been adopted in previous works to generate long paragraphs with multiple sentences. It generally consists of a decoder that first generates sentence features and a second decoder that then generates the words by further decoding the sentence features. In [37], hierarchical decoding is utilized to generate long paragraphs in order to deal with the gradient vanishing problem in recurrent neural networks. Due to the multi-sentence nature of imaging reports, it is also frequently adopted in the report generation task [3], [8], [16]. This hierarchical decoding schedule is to some extent similar to the multi-sentence nature of radiology reports. However, none of these methods have direct supervision on the generation of the sentence features aside from whether to stop generating new sentences. Moreover, these works are all based on RNN, where the sentence feature sequence could be easily generated recurrently during the training process. The utilization of hierarchical decoding on the more powerful Transformer architecture [22] is more difficult. As Transformer is typically trained after a parallel forward pass using an attention mask unlike the recurrent fashion of RNN, sentence feature sequence cannot be generated without additional intermediate supervision during training. In this work, we propose a multi-grained report generation method, which is essentially different from the hierarchical RNNs as we explicitly supervise the intermediate sentence features generated by the first decoder. This also enables the use of the Transformer architecture and two-stage reinforcement fine-tuning on the two decoders independently, which further improves model performance.

III. Method

In this section, we first introduce our sentence-level image-language contrastive learning (SILC) in III-A, then we introduce our multi-grained report generation approach based on SILC in III-B. Finally, we describe how the two approaches are trained and fine-tuned in III-C.

A. Sentence-Level Image-Language Contrastive Learning

1) Basic Idea: Our SILC approach enables fine-grained image and text feature learning by performing contrastive learning for each sentence topic. As is shown in Fig.2, a medical imaging report is usually composed of multiple sentences, each with a topic that it focuses on describing and the corresponding content. Conventional image-language contrastive learning treats these sentences as a whole and encodes the entire report as a text feature, which is constrained to be close to the paired image feature and distant from unpaired image or text features. As the contents of different topics are mixed in a single feature, the model may achieve a relatively low loss by encoding the contents of the easiest and distinctive topics in the feature and learning the corresponding image features, while topics related to subtle and detailed image features are not learned well. In contrast, we propose to perform fine-grained contrastive learning for each semantically independent sentence to force the model to learn image features for each topic. Specifically, the topic of each sentence

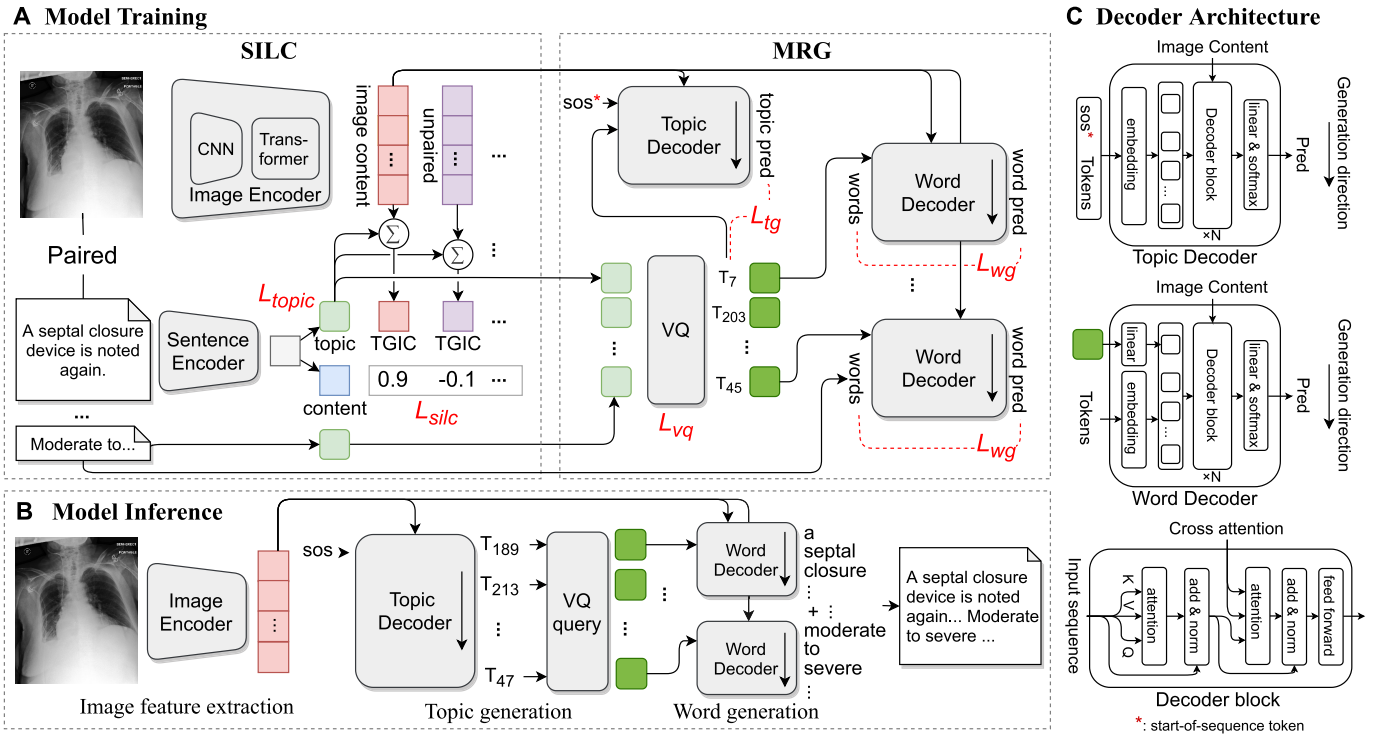

Fig. 3. Illustration of our full approach. **A**, the training of the model with two components, sentence-level image language contrastive learning and multi-grained report generation. **B**, model inference. **C**, the detailed architecture of the decoders.

is used to extract the corresponding image content, namely topic-guided image content (TGIC). The TGIC is supervised contrastively with the text content, using the TGIC of the paired image as the positive sample and the TGICs of other images with the same topic as negative samples. Instead of manually defining the basic topics and assigning each sentence a topic, our SILC approach automatically learns the feature representation of sentence topics as well as image and sentence contents from the image-report pairs in the dataset, as is shown in Fig.3.

2) Model Architecture: We assume that the content of an image could be represented as N features each corresponding to a certain basic topic or concept, while N is a hyper-parameter. A sentence in the report normally talks about one or several (or more specifically, a linear combination) of these basic topics. Thus, the topic of each sentence could be represented as a non-negative vector with a sum of 1, indicating the extent it is related to each basic topic.

Specifically, as shown in Fig.3, for an input image we first extract the image content $V = \{v_1, v_2, \ldots, v_N\}, v_i \in \mathbb{R}^D$ using an image encoder, where D is the feature dimension. The image encoder is composed of a convolutional neural network (CNN) followed by a Transformer encoder. The CNN outputs a 7×7 feature map where each feature corresponds to a certain spatial location. However, because of the difference in patient position, the feature in each location does not always correspond to the same content for different images. Thus, we use a Transformer encoder to further encode the flattened features into N image content features $\{v_1, v_2, \ldots, v_N\}$, each representing the image content for a certain topic.

For a sentence, a Transformer-based sentence encoder encodes it into a feature S, and then predicts a topic feature $T = softmax(l_1(S)) \in \mathbb{R}^N$ and a content feature $C = l_2(S) \in \mathbb{R}^D$, where l_1 and l_2 represent two linear projections. The topic feature T serves as a soft representation of the actual sentence topic. It contains N non-negative elements that sum up to 1, representing how much the sentence topic is related to the N basic topics. For example, if v_1 in V corresponds to topic *lung volume* and v_2 corresponds to *opacity*, then it is rational that the sentence "*the lung volumes are low accompanied by parenchymal opacity*" has a topic similar to $[0.5, 0.5, 0, 0, \ldots]$. Note that all the image contents, sentence topics, and sentence contents are automatically learned from the image-report pairs without any pre-definition or extra labeling.

3) Objectives: Here we discuss how to train the sentence topic feature T, content feature C, and the image contents V using contrastive objectives.

To perform sentence-level contrastive learning, we need to define the similarity between an image and a sentence. For an image I and a sentence S, we use the sentence topic T^S as the weight to extract topic-guided image content (TGIC) from image content V^I using a weighted summation:

$$TGIC(I, S) = \sum_{i=1}^{N} T_i^S v_i^I \tag{1}$$

The TGIC represents the image content corresponding to the certain topic of the sentence. The similarity between the image and the sentence is defined as the cosine similarity between the sentence content C^S and the TGIC:

$$sim(I, S) = cos(C^S, TGIC(I, S)) \tag{2}$$

During training, we compute the similarities between all images and all sentences in the same batch. We constrain that paired sentences and images are similar and unpaired ones are dissimilar. Formally, the contrastive loss for a batch of size B is based on the InfoNCE loss [38] and represented as:

$$L_{silc} = \frac{-1}{B} \sum_{i=1}^{B} \sum_{S \in R_i} \log \frac{e^{sim(I_i, S)\tau}}{\sum_{k=1}^{B} e^{sim(I_k, S)\tau}} \tag{3}$$

where R_i and I_i denote the i-th report and image, S denotes a sentence in the i-th report, and τ is a trainable parameter.

Furthermore, we notice that different sentences in a report normally describe different topics. Thus, we apply an auxiliary loss on the topic features to constrain their difference. For a report containing m sentences, the auxiliary topic loss is defined as:

$$L_{topic} = \frac{-1}{m} \sum_{i=1}^{m} \log \frac{1}{\sum_{k=1}^{m} e^{cos(T_i, T_k)}} \tag{4}$$

where T_i is the topic feature of the i-th sentence. This loss is also based on the InfoNCE loss, treating a sentence itself as the positive sample and all other sentences in the same report as negative samples.

B. Multi-Grained Report Generation

Based on SILC, we develop a multi-grained report generation approach aiming to generate accurate reports for medical images. It is composed of a vector quantization (VQ) module [17], a topic decoder, and a word decoder, as shown in Fig.3. The VQ module learns a discrete space for the sentence topic features. The topic decoder generates sentence topic indexes in the VQ space based on the image content, which are translated to topic features by the VQ module. The word decoder generates the words of each sentence based on the sentence topic feature and image content.

1) VQ Module: We use a VQ module to map continuous sentence topic features to a discrete set of features to fit the generation task of the topic decoder, similar to VQGAN [18]. For an input feature v, the VQ module finds its nearest embedding feature e_v in the embedding space E and outputs the feature and the corresponding index. However, the loss function we use differs from that in the original VQ-VAE. We only push the vectors in the embedding to the input vectors but remove the commitment loss that pushes the input vectors to the embedding, and we do not back-propagate the gradients of the output vectors to the input. This is because the topic features are trained by the SILC objective and we want to perform one-way supervision from SILC to report generation but not in the reverse direction. We also push each unused embedding feature e to its nearest input feature v_e to make more embedding features used. The overall VQ loss is defined as:

$$L_{vq} = \sum_{v \in input} \|sg(v) - e_v\| + \sum_{e \in E} \|sg(v_e) - e\| \tag{5}$$

where $sg()$ denotes the stop-gradient operation.

Note that it is also possible to train a generation model without the VQ module and let the topic decoder generate sequences of continuous feature vectors rather than token indexes. However, this could lead to the prediction of outlier topic features and result in extremely poor generation performance.

2) Topic Decoder: Given an image, a Transformer-based topic decoder first generates the topics that the report should talk about. The image contents are used as the source of cross-attention and the topic decoder generates the topic indexes in an auto-regressive manner. For a report whose sentence topics are mapped by the VQ module to an index sequence $\mathcal{T} = t_1, t_2, \ldots t_n$, and the image content V, the topic decoder is optimized by a cross-entropy loss:

$$L_{tg} = -\sum_{i=1}^{n} \log(P(t_i | t_{1:i-1}; V; \theta_{TD})) \tag{6}$$

where θ_{TD} denotes the parameters of the topic decoder.

3) Word Decoder: A transformer-based word decoder generates the words for each sentence given the sentence topic and the image contents. The sentence topic is provided by either the sentence encoder during training or the topic decoder during inference. The topic feature quantized by the VQ module is mapped to the same dimension as the hidden size of the word decoder and fed at the first input position to the word decoder without token embedding. The word decoder then generates the word sequence auto-regressively using the image contents as cross-attention. Formally, for an input topic feature T, the ground truth word sequence $\mathcal{W} = w_1, w_2, \ldots w_m$, and the image content V, the loss is defined as:

$$L_{wg} = -\sum_{i=1}^{m} \log(P(w_i | T, w_{1:i-1}; V; \theta_{WD})) \tag{7}$$

where θ_{WD} denotes the parameters of the word decoder.

During inference, the topic decoder first generates a sequence of sentence topic indexes based on the image. The VQ module extracts the topic features given the topic indexes, which are then fed to the word decoder independently to generate the words of the sentences. The sentences generated by the word decoder are concatenated sequentially to acquire the final report.

C. Model Training

Our SILC model and report generation model are jointly trained in an end-to-end manner optimized by the total loss:

$$L = \lambda_{silc} L_{silc} + \lambda_{topic} L_{topic} + \lambda_{vq} L_{vq} + \lambda_{sg} L_{sg} + \lambda_{wg} L_{wg} \tag{8}$$

where λ are the weights to balance the loss terms.

Since each decoder in MRG has its own target sequence independently, we propose to finetune the two decoders separately using reinforcement learning after joint training. Specifically, we first train the topic decoder and then train the word decoder using the self-critical sequence training method [19]. When training each decoder, all other model parameters are frozen. We use the recommended CIDEr-D score [39] as the reward during fine-tuning, which is shown to be the most effective in [19]. However, when training the

topic decoder, the target is the topic feature index sequence, which is only about 6 to 7 tokens long. We find that using the CIDEr-D score, which is by default based on up to 4-grams, is too strict for the topic decoder. Thus, we resort to a CIDEr-D score with 1-gram when training the topic decoder. The word decoder is trained using the original CIDEr-D score.

IV. Experiments

A. Datasets

We evaluate our method on two datasets: MIMIC-CXR and IU-Xray.

MIMIC-CXR [20] is a large-scale dataset of chest X-Ray exams consisting of 473,057 images and 227,835 reports from 63,478 patients. Each report may contain different sections including *Indication*, *Preamble*, *Comparison*, *Findings*, and *Impression*, etc. Each report corresponds to a single or multiple X-Ray images. We adopt the official data split with 222,758 training reports, 1,808 validation reports, and 3,269 test reports. There is no overlap of patients between the training, validation, and test sets. For a fair comparison, we follow the experimental settings of previous works [6]. Specifically, we use single X-Ray images as input because about 35% of the reports only correspond to single images, and we generate *Findings* of the reports as it is the most informative section. Reports without a *Findings* section are removed, resulting in a training set of 152,173 reports and 270,790 images, a validation set of 1,196 reports and 2,130 images, and a test set of 2,347 reports and 3,858 images. We tokenize the words in the *Findings* of the reports and remove words that appear less than 10 times. The maximum length of each report is set to 100 tokens following previous works.

IU-Xray [21] is a widely-used chest X-Ray dataset collected from the Indiana Network consisting of 7,470 images and 3,955 reports. Each report has four sections - *Indication*, *Comparison*, *Findings* and *Impression*, and corresponds to a frontal view and/or a lateral view X-Ray image. We follow the experimental settings of previous works [6], [24], [40]. Specifically, we use the two image views to generate the *Findings* section of the reports. Reports without a *Findings* section or two image views are removed, resulting in a total of 2,955 reports. We then split these reports by a ratio of $7:1:2$ into the training, validation, and test set, resulting in a training set of 2,069 reports, a validation set of 296 reports, and a test set of 590 reports. As the reports in IU-Xray all correspond to different patients, there is no overlap of patients between the training, validation, and test sets. We tokenize the words of the *Findings* section and remove tokens that appear less than 3 times. We crop each report to a maximum of 60 tokens following previous works.

Note that we manage the input image views differently on the two datasets. These settings are directly adopted from previous works that we compare with to ensure a fair comparison. As we use single images to generate reports on MIMIC-CXR, we view each image-report pair as a sample. If a report corresponds to multiple images, each image would be used to generate a report independently, and all the generated reports would be used for evaluation. Differently on IU-Xray, we view each study-report pair as a sample and each report would only be generated once.

B. Implementation Details

On both datasets, we resize the input image to the size of 224×224 pixels. For the image encoder, we use a ResNet-101 [45] pretrained on ImageNet [46] as the backbone which predicts a 7×7 feature map for each input image. For IU-Xray we average-fuse the feature maps of the two images. The feature map is then flattened and fed into a 3-layer Transformer encoder that generates 49 image content features (i.e., $N = 49$). The sentence encoder is a 6-layer Transformer encoder. Both the topic decoder and the word decoder are Transformer decoders with 6 layers for MIMIC-CXR and 3 layers for IU-Xray. All the Transformers we use have 8 attention heads and a hidden size of 512, and are trained from scratch. We set the embedding size of the VQ module to 1,024. We use an Adam optimizer with a learning rate of $3e-5$ and $\beta = (0.9, 0.98)$. The term weight λ_{silc} is set to 0.3 while all other term weights are set to 1. We train the model for 8 epochs and 300 epochs on MIMIC-CXR and IU-Xray respectively, and the model with the best BLEU-4 score on the validation set is saved for final evaluation. On MIMIC-CXR, we finetune the topic decoder for 5 epochs and the word decoder for 1 epoch using reinforcement learning with a learning rate of $1e-6$. We do not perform any model finetuning on IU-Xray as the dataset is relatively small and the model converges quickly. We use batch sizes of 32 and 8 for model training and RL finetuning respectively. During inference, we use the greedy decoding strategy due to its high efficiency. We have also tried the beam search strategy but see no significant improvement. We use a Tesla V-100 GPU with 32GB memory for model training and evaluation.

C. Evaluation Metrics

We use both conventional natural language generation (NLG) metrics and clinical accuracy to evaluate our model. The NLG metrics include BLEU [47], METEOR [48], ROUGE-L [49] and CIDEr-D [39]. These metrics are evaluated using the standard evaluation protocol.[1] We also evaluate the clinical accuracy of the generated reports using the CheXpert tool [13]. Specifically, for each report, CheXpert automatically labels the existence of 14 critical findings as 1 (confidently present), 0 (confidently absent), -1 (uncertainly present), and *blank* (not mentioned). Following previous works,[2] we map 1 and -1 to the positive label and map 0 and *blank* to the negative label. We then compute the average precision, recall, and F1 score of the 14 classes based on the labels obtained from the predicted reports and the ground truth reports. We do not evaluate clinical accuracy on IU-Xray as CheXpert does not apply to it.

D. Report Generation Performance

We compare our method with a wide range of methods for image captioning and medical report generation. The

[1] github.com/tylin/coco-caption

[2] github.com/zzxslp/WCL/blob/main/chexpert-labeler/calculate_metric.py

TABLE I

REPORT GENERATION PERFORMANCE COMPARISON ON MIMIC-CXR AND IU-XRAY DATASET. BL-N DENOTES BLEU OVER N-GRAM. MTR, RG-L, AND C DENOTE METEOR, ROUGE-L, AND CIDER-D. P, R AND F1 DENOTE AVERAGE PRECISION, RECALL AND F1-SCORE OF THE 14 CLASSES EXTRACTED BY CHEXPERT

Dataset	Method	Extra	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	C	P	R	F1
MIMIC-CXR	ST [10]	-	0.299	0.184	0.121	0.084	0.124	0.263	-	0.249	0.203	0.204
	Att2In [19]	-	0.325	0.203	0.136	0.096	0.134	0.276	-	0.322	0.239	0.249
	AdaAtt [11]	-	0.299	0.185	0.124	0.088	0.118	0.266	-	0.268	0.186	0.181
	TopDown [12]	-	0.317	0.195	0.130	0.092	0.128	0.267	-	0.320	0.231	0.238
	R2Gen [6]	-	0.353	0.218	0.145	0.103	0.142	0.277	-	0.333	0.273	0.276
	R2GenCMN [9]	-	0.353	0.218	0.148	0.106	0.142	0.278	-	0.334	0.275	0.278
	PPKED [40]	CheXpert [13]	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-	-
	WCL [7]	CheXBERT [41]	0.373	-	-	0.107	0.144	0.274	-	0.385	0.274	0.294
	MSAT [25]	Radgraph[42]	**0.413**	0.266	0.186	0.136	**0.170**	0.298	-	-	-	-
	Ours	-	0.406	**0.267**	**0.190**	**0.141**	0.163	**0.309**	**-**	**0.457**	**0.337**	**0.330**
IU-Xray	ST [10]	-	0.216	0.124	0.087	0.066	-	0.306	0.277	-	-	-
	CoAtt [3]	-	0.455	0.288	0.205	0.154	-	0.369	0.277	-	-	-
	MRMA [43]	-	0.457	0.295	0.212	0.157	0.180	0.353	0.244	-	-	-
	HRGR [24]	Templates	0.438	0.298	0.208	0.151	-	0.322	0.343	-	-	-
	CMAS-RL [44]	-	0.464	0.301	0.210	0.154	-	0.362	0.275	-	-	-
	R2Gen [6]	-	0.470	0.304	0.219	0.165	0.187	0.371	-	-	-	-
	PPKED [40]	CheXpert [13]	**0.483**	0.315	0.224	0.168	-	0.376	0.351	-	-	-
	Ours	-	0.472	**0.321**	**0.234**	**0.175**	**0.192**	**0.379**	**0.368**	-	-	-

TABLE II

ABLATION STUDY ON THE MIMIC-CXR DATASET. CL AND RL STAND FOR CONTRASTIVE LEARNING AND REINFORCEMENT LEARNING. TD AND WD DENOTE THE TOPIC DECODER AND WORD DECODER

Model	CL	MRG	RL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	F1
Base				0.345	0.216	0.145	0.103	0.142	0.279	0.277
SCST [19]			✓	0.361	0.228	0.153	0.108	0.144	0.285	0.291
CLIP [32]	CLIP			0.346	0.218	0.148	0.107	0.140	0.281	0.290
SILC	SILC			0.344	0.218	0.150	0.109	0.141	0.289	0.292
MRG	SILC	✓		0.346	0.226	0.159	0.117	0.146	0.290	0.301
MRG w/o L_{topic}	SILC	✓		0.321	0.207	0.145	0.105	0.140	0.287	0.246
RL-TD	SILC	✓	TD	0.392	0.260	0.182	0.133	0.159	0.304	0.314
RL-WD	SILC	✓	WD	0.358	0.234	0.166	0.122	0.148	0.298	0.320
Full	SILC	✓	TD+WD	**0.406**	**0.267**	**0.190**	**0.141**	**0.163**	**0.309**	**0.330**

image captioning methods include Show-Tell [10], Att2In [19], AdaAtt [11] and TopDown [12], and the their performance on MIMIC-CXR is cited from [6]. The medical report generation methods include MRMA [43], R2Gen [6], R2GenCMN [9], PPKED [40], WCL [7], CoAtt [3], HRGR [24], CMAS-RL [44], and MSAT [25]. Most of these methods were previously evaluated following the same experimental settings (e.g., input image views and generation target) as those used in [6], so we directly cite their performance from the original papers. For [43] which had a different experimental setting (e.g., generating *Findings+Impression* instead of *Findings*), we re-implement their method based on public code[3] and evaluate it using the same setting as in [6]. We stick to all experimental settings including the data pre-processing and testing method[4] used in [6] to make the comparison fair. Note that as some of these works have only evaluated their performances on one dataset and have not published the code, we are unable to compare with them on the other dataset.

The comparison results are shown in Table I. Our method achieves better performance on the two datasets in terms of both NLG metrics and clinical accuracy, especially on BLEU-4 and ROUGE-L. We notice that our method is particularly competitive on high-order BLEU scores compared to low-order ones, which means that it generates semantically meaningful n-grams more effectively than single words or low-order ones. This is important in medical report generation because sometimes the presence of keywords (e.g., *pleural effusion*) may indicate various conditions unless a longer n-gram is known (e.g., *no pleural effusion* or *a left pleural effusion*). The ROUGE-L metric is primarily based on the recall of the generated words, thus the high ROUGE-L score shows our model's ability to generate the keywords better, which sometimes might be ignored by other methods due to their scarcity. In terms of clinical accuracy, our method, without using external knowledge or training data, outperforms previous methods that use extra disease labels or domain-specific prior knowledge during training or pre-training, indicating our method's capability to better learn medical knowledge from the images and reports.

E. Ablation Study

We explore the effectiveness of different components in our approach by performing ablation studies on MIMIC-CXR, as is shown in Table II. Our baseline model ('Base') has exactly the same image encoder architecture as the proposed model and a 6-layer Transformer decoder and is trained with only a generation objective. Because of the imbalance between

[3]https://github.com/tengfeixue-victor/Medical-Report-Generation

[4]github.com/zhjohnchan/R2Gen

normal and abnormal findings and the multi-topic nature of the reports, the base model cannot generate reports accurately enough and gets the lowest scores. We then train models using the same architecture as 'Base', but with additional objectives of CLIP [32] and SILC respectively, trying to help the model learn discriminative features of abnormal findings. Both CLIP and SILC help the model generate better reports, improving high-order BLEU scores and clinical accuracy. As our SILC performs fine-grained contrastive learning, it helps the model to learn better image features and outperforms CLIP which performs contrastive learning between whole reports and images. Then we apply multi-grained generation based on the SILC model ('MRG'), which further improves model performance. We notice that the superiority of our model on high-order BLEU scores is largely brought by the MRG module, as it helps the model generate more accurate topics and contents that are semantically meaningful rather than single keywords. To evaluate the effect of the auxiliary loss L_{topic} in the SILC approach, we removed L_{topic} and re-trained the MRG model. As is shown, the model has a huge performance drop compared to the MRG model, which we find is due to the complicated topic features. While L_{topic} is mainly used to constrain that different sentences in a report have different topic features, it at the same time results in the sparsity of the topic features (i.e., only a few elements of a feature are large while others are very small). This conforms to the characteristic of radiology reports, as a sentence would normally talk about only a few basic topics. Without the auxiliary loss, the topic features are less focused and much more complicated, resulting in bad generation performance. We then apply RL finetuning [19] on the topic decoder or word decoder of the MRG model separately, which can both improve model performance by a large margin. Finetuning the topic decoder brings more performance improvement than finetuning the word decoder, which means that the topics to describe in a report seem to be more important than the wording of the sentences. The final two-stage RL finetuned model achieves the best scores on all metrics. To evaluate the contribution of reinforcement learning in the final model, we finetune the Base model using self-critical sequence training and CIDEr score as the reward ('SCST'), which is the same as how we finetune the two decoders. While the reinforcement fine-tuning improves the performance of the base model, our method substantially outperforms it, demonstrating the effectiveness of our two-stage fine-tuning method.

Our method has a hyperparameter N, representing the number of image contents of an image and the dimension of the topic features, which might directly influence model performance. We hypothesize that N should be similar to the number of different basic topics the reports would describe in order to achieve the best model performance. To check the influence of this hyperparameter, we train our MRG model under different N on the MIMIC-CXR dataset and show the performances of the models before the reinforcement finetuning (Table III). Specifically, the Transformer encoder in the image encoder still takes 49 features as input, but only the first N features in the output 49 features are used as the image contents. As can be seen, the model performs poorly when $N < 10$, because there are too few image content features and the model cannot properly assign the basic topics to them, resulting in bad topic feature representation. When $N > 10$, the increase in N results in a slow improvement but does not influence model performance greatly. We can thus conclude that there are dozens of basic topics in the Chest X-Ray reports and our setting of $N = 49$ is sufficient.

TABLE III
MODEL PERFORMANCE UNDER DIFFERENT HYPERPARAMETER N ON THE MIMIC-CXR DATASET BEFORE REINFORCEMENT FINETUNING

N	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
2	0.138	0.074	0.050	0.034	0.065	0.174
5	0.273	0.169	0.117	0.084	0.112	0.268
10	0.340	0.218	0.149	0.106	0.137	0.277
20	0.336	0.219	0.155	0.114	0.141	0.294
30	0.342	0.222	0.155	0.114	0.141	**0.295**
49	**0.346**	**0.226**	**0.159**	**0.117**	**0.146**	0.290

F. Clinical Accuracy

To give a detailed understanding of the clinical accuracy of our generated reports, we further show the clinical accuracy metrics for each of the 14 specific findings on the MIMIC-CXR test set, which most previous works failed to reveal. As shown in Table IV, we calculate the improvement of our model in precision, recall, and F1 score compared with the baseline (the Base model in Table II). To give a more comprehensive demonstration, we also show the result evaluated by the Matthews correlation coefficient (MCC) [50], which is more robust and informative on imbalanced datasets [51]. The positive rates of the findings are also calculated on the test set. We can see that all findings have positive rates less than 0.5, the imbalance of which our method aims to deal with.

We notice that for two findings, *Pleural Other* and *Fracture*, both models get a zero score. This result is observed even in the training set and also during our replication of previous works [6], [7]. We conjecture that these two findings are too rare and subtle for the models to learn. *Fracture* often corresponds to a tiny area in the image, especially for undisplaced rib fractures. As we use an input image size of 224×224 following previous works to make a fair comparison while the original images are much larger (e.g., 2544×3056), the resized images have lost many local details, making the detection of small and subtle abnormalities such as fractures difficult. *Pleural Other* may correspond to various pleural abnormalities, such as pleural thickening, fibrosis, and pleural scar. This makes it more difficult to learn due to the various features these abnormalities may demonstrate. Increasing the input image size and using more advanced techniques may help tackle this problem.

On the remaining 12 findings other than *Fracture* and *Pleural Other*, our model achieves higher F1 scores on 9 findings and higher MCC scores on 11 findings. Overall on all 14 findings, our model achieves an absolute improvement of 0.087, 0.078, 0.053, and 0.064 in the average precision, recall, F1 score, and MCC score respectively, and a relative improvement of 0.236, 0.303, 0.192, and 0.374 in the four metrics respectively. These results demonstrate our method's

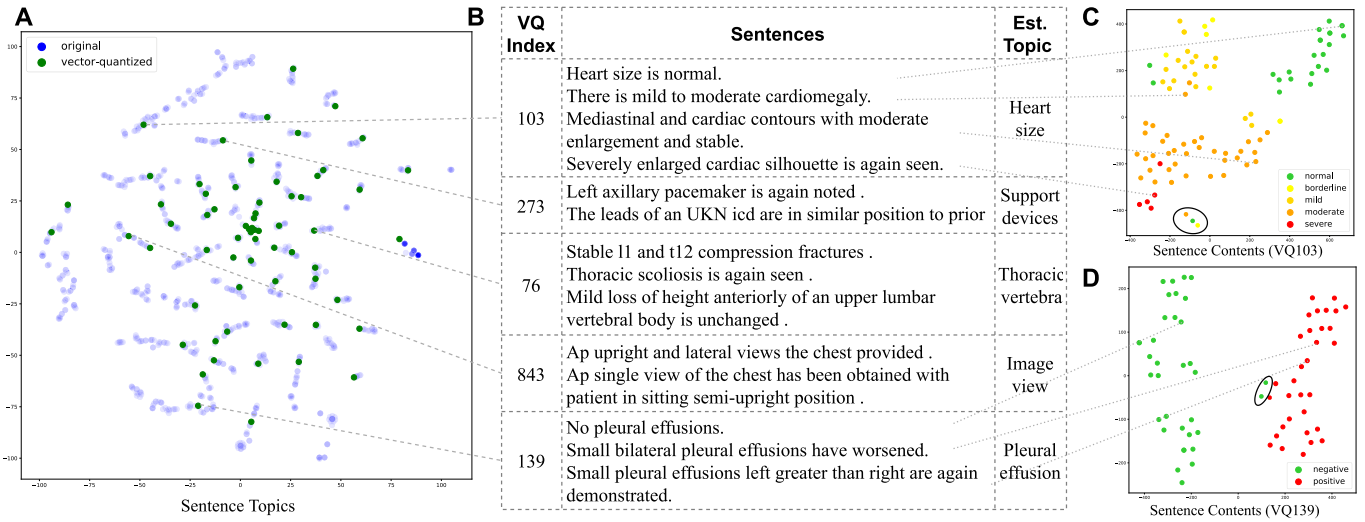

Fig. 4. A visualization of the sentence topic and content features learned by the SILC approach. **A**, the t-SNE visualization of sentence topic features randomly sampled in the MIMIC-CXR test set. Blue points represent the original topics and green points represent the vector-quantized topics. **B**, examples of sentences whose topic features are mapped to some specific VQ embeddings and their manually estimated topics. **C-D**, t-SNE visualization of sentence content features for two topics.

TABLE IV
CLINICAL ACCURACY OF OUR MODEL IN PREDICTING EACH FINDING ON THE MIMIC-CXR TEST SET COMPARED WITH THE BASELINE MODEL

Finding	Positive	Baseline				Ours			
		P	R	F1	MCC	P	R	F1	MCC
Enlarged cardiomediastinum	0.301	0.355	0.239	0.286	0.059	0.415 (↑0.060)	0.276 (↑0.037)	0.332 (↑0.046)	0.124 (↑0.065)
Cardiomegaly	0.436	0.621	0.616	0.618	0.325	0.645 (↑0.024)	0.733 (↑0.117)	0.686 (↑0.068)	0.417 (↑0.092)
Lung lesion	0.071	0.292	0.026	0.047	0.068	1.000 (↑0.708)	0.004 (↓0.022)	0.007 (↓0.040)	0.058 (↓0.010)
Lung opacity	0.477	0.611	0.147	0.237	0.097	0.697 (↑0.086)	0.124 (↓0.023)	0.211 (↓0.026)	0.134 (↑0.037)
Edema	0.244	0.513	0.350	0.416	0.280	0.446 (↓0.067)	0.701 (↑0.351)	0.545 (↑0.129)	0.372 (↑0.092)
Consolidation	0.108	0.104	0.069	0.083	-0.004	0.250 (↑0.146)	0.002 (↓0.067)	0.005 (↓0.078)	0.015 (↑0.019)
Pneumonia	0.169	0.315	0.117	0.170	0.101	0.353 (↑0.038)	0.156 (↑0.039)	0.217 (↑0.047)	0.140 (↑0.039)
Atelectasis	0.311	0.465	0.251	0.326	0.150	0.489 (↑0.024)	0.598 (↑0.347)	0.538 (↑0.212)	0.302 (↑0.152)
Pneumothorax	0.052	0.198	0.119	0.149	0.118	0.173 (↓0.025)	0.214 (↑0.095)	0.192 (↑0.043)	0.143 (↑0.025)
Pleural effusion	0.365	0.698	0.511	0.590	0.418	0.733 (↑0.035)	0.710 (↑0.199)	0.722 (↑0.132)	0.566 (↑0.148)
Pleural other	0.042	0.000	0.000	0.000	0.000	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)
Fracture	0.067	0.000	0.000	0.000	0.000	0.000 (-)	0.000 (-)	0.000 (-)	0.000 (-)
Support devices	0.424	0.858	0.624	0.722	0.586	0.884 (↑0.026)	0.693 (↑0.069)	0.777 (↑0.055)	0.657 (↑0.071)
No finding	0.068	0.148	0.552	0.233	0.185	0.316 (↑0.168)	0.506 (↓0.046)	0.389 (↑0.156)	0.344 (↑0.159)
Average	-	0.370	0.259	0.277	0.170	0.457 (↑0.087)	0.337 (↑0.078)	0.330 (↑0.053)	0.234 (↑0.064)

capability to help tackle the imbalance of the findings and predict more clinically accurate reports.

G. Qualitative Analysis

The sentence topic features play an important role in our overall approach. It serves as a soft representation of sentence topics in SILC and an intermediate feature to bridge the two decoders in the multi-grained report generation. However, we do not inject any prior knowledge into the model but rely solely on SILC to learn it from the dataset automatically. To check the quality of the learned topic features, we randomly sample 200 reports in the MIMIC-CXR test set and use the t-SNE method [52] to visualize the learned topic features of the sentences in Fig.4A. We can see that the topic features of different sentences do not distribute uniformly but form into many small clusters. This is consistent with the patterned feature of medical reports, as radiologists would write the reports based on rules or using templates, trying to make the topic of each sentence focused. We then randomly sample sentences whose topic features are mapped by the VQ module to the same embedding (Fig.4B). We find that sentences describing similar topics are mapped by our SILC method to similar topic features and correctly grouped together by the VQ module, even if they vary a lot in text representation or actual contents. Although we use a $1,024$ embedding space for the VQ module, we find that empirically only about 100 vectors are used. This roughly represents the number of different topics that the model learned from the dataset. To check the learned sentence content features, we choose two topics (VQ 103 and VQ 139) and manually label the sentences mapped to these topics into several classes based on their contents. The t-SNE visualization of their content features and their manually labeled classes are shown in Fig.4C-D. We can see that sentences with the same content have similar content features, suggesting that the learned content features can effectively represent the actual sentence contents. There

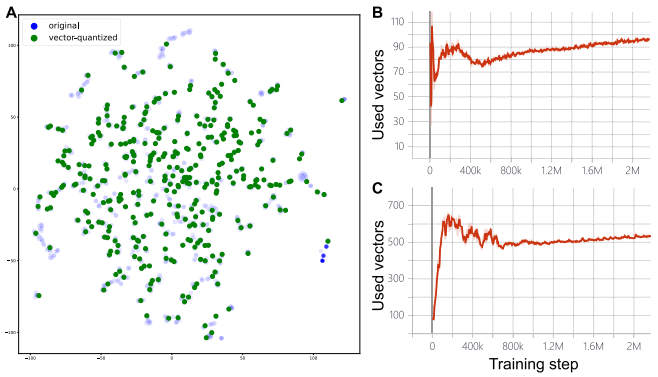

Fig. 5. Influence of the topic loss in our approach. **A**, the t-SNE visualization of sentence topic features in the MIMIC-CXR test set learned without L_{topic}. **B-C**, the number of vectors used in the VQ module during training with and without L_{topic} respectively.

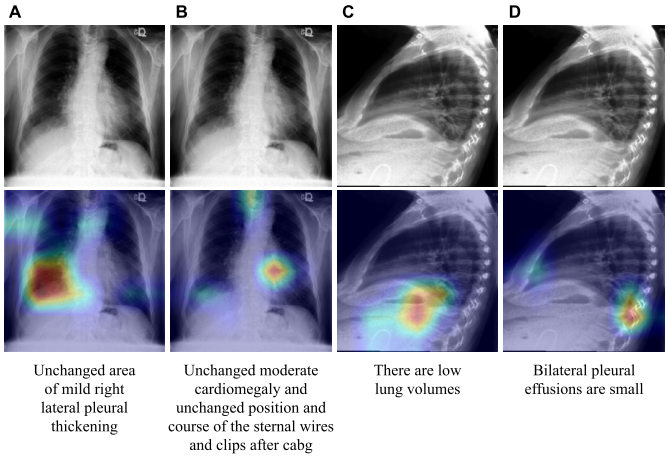

Fig. 6. Grad-CAM visualization of the image encoder's attention in SILC.

are a few outliers due to fair reasons. For example, the three points in the ellipse in Fig.4C all correspond to sentences with mixed topics of heart size and lung volume (e.g., "*Cardiac silhouette appears moderately enlarged likely accentuated due to low lung volumes and ap technique*"), while the two points in the ellipse in Fig.4D are due to uncertain descriptions (e.g., "*There is no definite left pleural effusion*"). Note that the estimated topics and content labels are just for illustration and are not used in model training.

To demonstrate the influence of the auxiliary loss L_{topic} on SILC, we show the topics learned without L_{topic} in Fig.5 A. Compared with Fig.4A, the topic features learned without L_{topic} are much more scattered and complicated, and do not effectively form into small clusters based on the topics. This also results in much more vector-quantized features being used in the VQ module (from around 100 to around 500 as in Fig.5B and C). As a result, the burden of the topic decoder is greatly increased, and the performance of MRG drops substantially, as shown in the ablation study.

To further demonstrate that SILC enables fine-grained encoder learning, we use Grad-CAM [53] to visualize the attention of the image encoder in the SILC learning task. Specifically, for a sentence and an image, we use their similarity score as the logit to generate a 7×7 attention map at the last layer of the CNN in the image encoder, as shown in Fig.6. We can see that even for the same image, the image encoder learns to focus on different locations when it is contrasted with different sentences. For example, the model focuses on the right pleural thickening in Fig.6A, and focuses on the heart and the sternal wires in Fig.6B. In Fig.6C and D, the image encoder focuses on the area beneath the lung when it needs to assess the lung volume, and focuses on the posterior costophrenic angle for the pleural effusions.

To give a more intuitive understanding of our multi-grained report generation method and its performance, we show some examples of the generated reports in Fig.7. Besides our approach and the baseline generation approach ('Base'), we also show the reports generated using the ground truth sentence topics, bypassing the topic decoder ('Ours - GT Topic'). We can find that: 1) Compared with the baseline generation approach, our method generates more accurate reports describing the correct abnormal findings. 2) Our word decoder could generate accurate sentences in correspondence with both the topics and the image contents, as is shown in 'Ours - GT Topic' (see also topic [103] in 'Ours' of Fig.7A and B). However, when the original sentence is too long with a complicated topic (e.g., topic [375] in Fig.7A), it is more difficult to generate the correct content. 3) There is a high similarity between the generated topic indexes and the ground truth topic indexes, demonstrating the ability of the topic decoder to predict accurate topics to describe for given images. As there are some similar topic features that are mapped by the VQ module to different indexes, the topic decoder sometimes generates alternative topic tokens that also result in accurate sentences (e.g., topic [677] in Fig.7A). 4) The model may not predict findings with correct severity or uncertainty. As shown in Fig.7A, the severity of the pleural effusion and the uncertainty of the atelectasis are mistaken. The severity of findings is more frequently described in the prediction than the uncertainty and is also correct in many cases (e.g., the mild vascular congestion and mild cardiomegaly in Fig.7A). This suggests that predicting the uncertainty of findings is more difficult than just mentioning the findings or describing their severity, as it relies on more subtle features in the images and probably also the relations between different abnormalities. 5) The model may predict partly overlapped or redundant content in different sentences conditioned on similar or related topics in a few cases. For example, in Fig.7C, "*no pleural effusion*" is predicted twice conditioned on topic [139] (mainly pleural effusion) and [62] (mainly pneumothorax but sometimes also describes pleural effusion). This is because of the independent decoding of the sentences.

H. Influence of Input Views

To make a fair comparison with previous works, the above experiments are all performed under the same setting where single images are used as model input on MIMIC-CXR and two image views and the average-fuse operation are used on IU-Xray. Although the way to manage the input views does not play a key role in our method or contribution, it is beneficial to evaluate the influence of the input views on the report

Input Image

Ground Truth	Ours	Ours - GT Topic	Base
[843] frontal and lateral views of the chest are obtained. [776] the patient is status post median sternotomy and CABG. [273] dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and likely right ventricle. [139] there is blunting of the left costophrenic angle most consistent with a small left pleural effusion. [375] left base opacity may be due to combination of *pleural effusion* and *atelectasis*, although *consolidation* is not excluded. [972] there is mild central pulmonary vascular congestion. [103] the cardiac silhouette is mildly enlarged.	[843] pa and lateral views of the chest provided. [776] the patient is status post median sternotomy and cabg. [273] left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle. [677] there is a moderate to large left pleural effusion with overlying atelectasis . [972] there is mild pulmonary vascular congestion. [103] the heart is mildly enlarged. [62] there is no pneumothorax.	[843] pa and lateral views of the chest provided. [776] the patient is status post median sternotomy and cabg. [273] left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle.[139] there are small bilateral pleural effusions. [375] there is a new opacity in the left lower lobe which is concerning for *pneumonia*. [972] there is mild pulmonary vascular congestion. [103] the heart is mildly enlarged.	frontal and lateral views of the chest were obtained. dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle. there are low lung volumes. there is moderate pulmonary edema. there is a small left pleural effusion. the cardiac silhouette is top normal to mildly enlarged. mediastinal contours are stable. there is no pneumothorax.

B

Input Image

Ground Truth	Ours	Ours - GT Topic	Base
[43] there are low lung volumes. [49] bibasilar atelectasis have minimally improved. [149] mild vascular congestion has minimally improved. [62] there are no new lung abnormalities or pneumothorax. [139] bilateral pleural effusions are small. [493] right PICC tip is at the cavoatrial junction.	[843] pa and lateral views of the chest were obtained. [294] right internal jugular central venous catheter terminates in the mid svc. [43] lung volumes are low. [49] there is mild bibasilar atelectasis. [139] there are small bilateral pleural effusions. [62] there is no pneumothorax. [103] the heart size is normal. [16] the mediastinal and hilar contours are normal.	[43] lung volumes are low. [49] there is mild bibasilar atelectasis. [149] there is no pulmonary edema. [62] there is no pneumothorax. [139] there are small bilateral pleural effusions. [493] right picc line terminates in the mid svc.	pa and lateral views of the chest provided. a right upper extremity picc line terminates in the low svc. there is a new retrocardiac opacity which may represent *pneumonia*. there is no pleural effusion or pneumothorax. the cardiomediastinal silhouette is unchanged. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

C

Input Image

Ground Truth	Ours	Ours - GT Topic	Base
[273] a left pectoral pacemaker device with leads through the left transvenous approach end into the right atrium and right ventricle respectively. [776] the patient is status post median sternotomy with intact sternal sutures. [16] heart size mediastinal and hilar contours are normal. [49] left lung is remarkable for mild left basal atelectasis. [43] right lung is clear. [710] no pneumonia or pulmonary edema. [139] there is no pleural abnormality.	[843] pa and lateral chest views were obtained with patient in upright position. [273] the pacemaker device is noted with leads terminating in the right atrium and right ventricle. [776] the patient is status post median sternotomy and cabg. [43] lung volumes are low. [49] there is atelectasis at the left lung base. [139] no pleural effusions. [62] there is no pleural effusion or pneumothorax. [103] the heart size is normal. [16] the mediastinal and hilar contours are normal.	[273] the pacemaker device is noted with leads terminating in the right atrium and right ventricle. [776] the patient is status post median sternotomy and cabg. [16] the mediastinal and hilar contours are normal. [49] there is atelectasis at the left lung base. [43] lung volumes are low. [710] no pulmonary edema. [139] no pleural effusions.	there is a new left pectoral pacemaker with leads terminating in the right atrium and right ventricle. there is no pneumothorax. there is no pleural effusion. the heart size is normal . the mediastinal and hilar contours are normal.

Fig. 7. Examples of generated reports in the MIMIC-CXR test set using different methods. Abnormal findings are highlighted. Findings that are both certain and correct are highlighted. Findings that are suspected with uncertainty are italicized. For our results and the ground truth reports, we show the VQ topic index of each sentence ahead of the sentence.

generation performance, as different image views may contain different information. To this end, we further train several models using different input views and fusion methods and evaluate their performance.

First, we train multi-view models on MIMIC-CXR as opposed to the single-view model. For each study, the multi-view model takes 3 images as input, as more than 98% of the studies have no more than 3 images. For those with more than 3 images, we randomly sample 3 images, and for those with less than 3, blank images are padded. We use the maximum and average operation to fuse the 3 feature maps respectively in two models. The max-fuse can preserve the salient features better while the average-fuse results in more balanced features. The performance of the multi-view models is shown in Table V. The multi-view model using max-fuse performs better than the single-view model, suggesting that using more image views does improve the report generation performance, as it complements the model with more information that might be missing in single-view input. The multi-view model using average-fuse performs slightly better than the single-view model but inferior to the max-fuse model. This is probably due to the variance of the input image numbers, as there might be padded images whose features would affect the resulting feature greatly through the average-fuse operation.

TABLE V
MODEL PERFORMANCE UNDER USING DIFFERENT INPUT VIEWS AND FUSING METHODS. MIMIC STANDS FOR MIMIC-CXR AND IU STANDS FOR IU-XRAY

Dataset	View	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
MIMIC	1	0.406	0.267	0.190	0.141	0.163	0.309
	3-max	**0.417**	**0.276**	**0.196**	**0.145**	**0.166**	**0.315**
	3-ave	0.416	0.273	0.191	0.141	0.165	0.306
IU	Frontal	0.475	0.316	0.224	0.164	**0.196**	0.365
	Lateral	0.459	0.300	0.213	0.158	0.187	0.362
	2-max	**0.476**	0.318	0.229	0.171	**0.196**	0.369
	2-ave	0.472	**0.321**	**0.234**	**0.175**	0.192	**0.379**

For the IU-Xray dataset, we train two single-view models using frontal and lateral images as input respectively (Table V). Both models perform inferior to the two-view model, which is consistent with the result on MIMIC-CXR. The frontal-view model achieves slightly better performance than the lateral-view one, suggesting that the frontal view might be more relevant for report generation. We then train a multi-view model using max-fuse operation. Compared with the default average-fuse model (already trained and shown in Table I), the max-fuse model outperforms it on BLEU-1 and METEOR but loses on other metrics. We conclude that on IU-Xray where the number of input images remains constant,

the max-fuse does not have evident superiority over the average-fuse.

Other methods might also be adopted to manage the input images, such as the attention-based fusion [33]. However, as the way to manage input images is not a main part of our method, we leave it to future study.

I. Limitations and Discussions

Despite the high performance quantitatively and qualitatively demonstrated, there are also some limitations of our approach. First, as our word decoder independently generates the words of different sentences, detailed relations between words in the sentences are not utilized. This can result in the model predicting partly overlapped or redundant descriptions in a few cases. This problem could probably be solved by introducing cross-sentence attention. Second, the generation of very long sentences with complicated topics is challenging. These limitations are exclusive to our method as we introduce the MRG approach. However, as the NLG and clinical metrics have demonstrated, our gain by generating multi-sentence reports with more accurate topics and contents outweighs the disadvantages. There are also some limitations observed that may exist in other works as well. First, our model may not predict findings with correct severity, and the uncertainty of findings is often ignored by the model. Correct prediction of these extent-related contents is important in clinical scenarios, as the severity can affect the physician's diagnosis and treatment plans and the uncertainty may encourage further examinations. Second, the model is unable to correctly describe findings that are extremely rare or subtle. These limitations suggest that, apart from technical innovation, future works for report generation may explore and introduce changes in experimental settings, such as using larger input image sizes and evaluating the severity and uncertainty of the predicted findings.

Our approach could be potentially used on other tasks such as paragraph generation, and our SILC could be used alone for model pretraining. Further refinement and utilization of the method are left for future studies.

V. Conclusion

In this paper, we present a multi-grained report generation method, which is based on a sentence-level image-language contrastive learning approach, in order to cope with the imbalance and multi-topic nature of radiology reports. We first propose SILC, a fine-grained contrastive learning approach that automatically learns the topic of each sentence and forces the model to learn refined image features for each specific topic. We then present a multi-grained report generation approach based on SILC, which is mainly composed of two decoders and uses the learned sentence topics as intermediate supervision. Experiments on two large-scale datasets MIMIC-CXR and IU-Xray demonstrate that our approach outperforms previous image captioning and report generation methods in terms of both language generation and clinical accuracy metrics. Ablation studies and qualitative analysis also show the effectiveness of our SILC learning approach and multi-grained generation architecture.

References

[1] X. Chen et al., "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[2] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2641–2649.

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

[4] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019.

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

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

[7] A. Yan et al., "Weakly supervised contrastive learning for chest X-ray report generation," 2021, *arXiv:2109.12242*.

[8] Z. Wang, L. Zhou, L. Wang, and X. Li, "A self-boosting framework for automated radiographic report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2433–2442.

[9] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," 2022, *arXiv:2204.13258*.

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

[11] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 375–383.

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

[13] 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.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, San Francisco, CA, USA, 2018.

[16] C. Yin et al., "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 728–737.

[17] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[18] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[19] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7008–7024.

[20] A. E. W. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[21] 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.

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

[23] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional GAN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2970–2979.

[24] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[25] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, Singapore. Cham, Switzerland: Springer, 2022, pp. 655–664.

[26] H. T. N. Nguyen et al., "Automated generation of accurate & fluent medical X-ray reports," 2021, *arXiv:2108.12126*.

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

[28] J. Lovelace and B. Mortazavi, "Learning to generate clinically coherent chest X-ray reports," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1235–1243.
[29] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder–decoder for medical report generation," *World Wide Web*, vol. 26, no. 1, pp. 253–270, Jan. 2023.
[30] B. Yan and M. Pei, "Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2982–2990.
[31] N. Kaur and A. Mittal, "RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings," *J. Biomed. Informat.*, vol. 135, Nov. 2022, Art. no. 104220.
[32] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
[33] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Mach. Intell.*, vol. 4, no. 1, pp. 32–40, Jan. 2022.
[34] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," 2020, *arXiv:2010.00747*.
[35] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3922–3931.
[36] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," 2022, *arXiv:2210.10163*.
[37] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 317–325.
[38] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[39] 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.
[40] 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.
[41] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXBERT: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," 2020, *arXiv:2004.09167*.
[42] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," 2021, *arXiv:2106.14463*.
[43] Y. Xue et al., "Multimodal recurrent model with attention for automated radiology report generation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, Granada, Spain. Cham, Switzerland: Springer, 2018, pp. 457–466.
[44] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," 2020, *arXiv:2004.12274*.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.
[46] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, Philadelphia, PA, USA, Jul. 2002, pp. 311–318.
[48] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *Proc. 6th Workshop Stat. Mach. Transl.*, 2011, pp. 85–91.
[49] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.
[50] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[51] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.
[52] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[53] 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.