ICON: Improving Inter-Report Consistency of Radiology Report Generation via Lesion-aware Mix-up Augmentation

Anonymous ACL submission

Abstract

Previous research on radiology report generation has made significant progress in terms of increasing the clinical accuracy of generated reports. In this paper, we emphasize another crucial quality that it should possess, i.e., *inter-report consistency*, which refers to the capability of generating consistent reports for semantically equivalent radiographs. This quality is even of greater significance than the overall report accuracy in terms of ensuring the system's credibility, as a system prone to providing conflicting results would severely erode users' trust. Regrettably, existing approaches struggle to maintain inter-report consistency, exhibiting biases towards common patterns and susceptibility to lesion variants. To address this issue, we propose ICON, which Improves the inter-report CONsistency of radiology report generation. Aiming at enhancing the system's ability to capture the similarities in semantically equivalent lesions, our approach involves first extracting lesions from input images and examining their characteristics. Then, we introduce a lesion-aware mix-up augmentation technique to ensure that the representations of the semantically equivalent lesions align with the same attributes, by linearly interpolating them during the training phase. Extensive experiments on three publicly available chest X-ray datasets verify the effectiveness of our approach, both in terms of improving the consistency and accuracy of the generated reports.[1]

1 Introduction

Being part of the diagnostic process, radiology report generation (Shin et al., 2016; Zhang et al., 2017; Jing et al., 2018) has garnered significant attention within the research community, due to its large potential to alleviate the heavy strain of radiologists. Recent research (Nishino et al., 2022; Tanida et al., 2023; Hou et al., 2023b) has made noteworthy progress in enhancing the clinical accuracy of the generated reports.

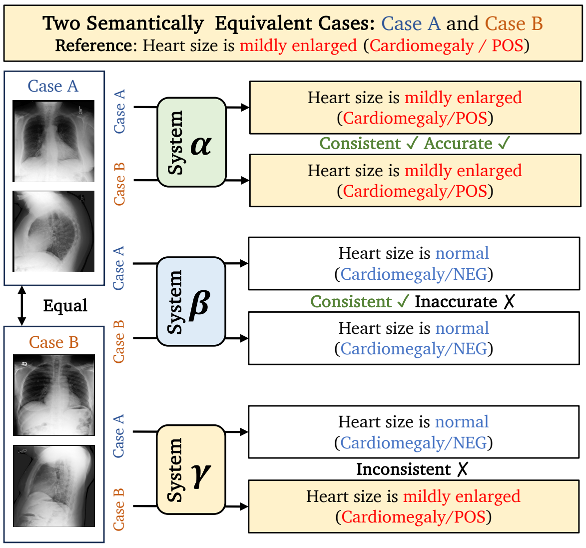


Figure 1: Given two semantically equivalent cases (i.e., Case A and Case B), an example to illustrate the difference between three radiology report generation systems: a consistent and accurate system (i.e., System α) and a consistently inaccurate system (i.e., System β), and an inconsistent system (i.e., System γ).

However, constructing a credible report generation system goes beyond the overall accuracy. There is another crucial quality for report generation systems that has been largely overlooked in the existing literature of medical report generation, which is, *inter-report consistency* (Elazar et al., 2021). To illustrate the disparity between accuracy and inter-report consistency, we exemplify two semantically equivalent cases as shown in Figure 1. Specifically, System α demonstrates the ability to maintain both inter-report consistency and factual accuracy for two similar cases (i.e., *"mildly enlarged"* for positive *Cardiomegaly*), whereas other systems (i.e., β and γ) fail to meet these criteria. These systems might have overfitted to ordinary cases and could be vulnerable to noise or

[1] We will release our codes and model checkpoints after the review process.

attack. In terms of enhancing the system's credibility, inter-report consistency might even hold greater significance than the overall accuracy, since a system prone to providing conflicting results would severely undermine users' trust (Qayyum et al., 2020; Asan et al., 2020). Regrettably, existing report generation systems struggle to maintain this important quality. They tend to exhibit biases towards common patterns, primarily describing normal observations, and are extremely susceptible to lesion variants and context noise (Chen et al., 2020; Qin and Song, 2022; Ma et al., 2021; Kaviani et al., 2022). We argue that this is largely due to their limited capability of capturing shared attributes of similar patterns, which arises from the data scarcity of distributed lesions and their semantically equivalent variants, rendering it challenging for neural models to accurately locate and describe abnormalities.

In this paper, we propose ICON, which aims to Improves inter-report CONsistency of radiology report generation. Our proposed method involves first extracting lesions from given input images, followed by examining the attributes of these lesions. Subsequently, both the radiographs and their associated attributes are utilized as inputs for report generation. To further enhance the inter-report consistency, we introduce a lesion-aware mix-up augmentation technique by learning from linearly interpolated lesions and attributes that belong to the same observation. In summary, the contributions of this paper are as follows:

- We propose ICON which improves the consistency of the radiology report generation system by capturing abnormalities at the lesion level. ICON only requires coarse-grained labels (i.e., image labels) to extract lesions, which can be easily acquired and transferred to other modalities, compared with bounding boxes (Tanida et al., 2023).
- We introduce a lesion-aware mix-up augmentation to capture the shared features between two chest X-ray lesions to further enhance consistency. Besides, two metrics (i.e., CON and R-CON) are devised to quantify the inter-report consistency.
- Extensive experiments are conducted on three publicly available datasets, and experimental results demonstrate the effectiveness of ICON in terms of improving both the consistency and accuracy of the generated reports.

2 Priliminaries

2.1 Problem Formulation

Given a set of radiographs $\mathcal{X} = \{X_1, \ldots, X_L\}$ in one study, along with its historical records $\mathcal{X}^p = \{X_1^p, \ldots, X_{|p|}^p\}$ or $\mathcal{X}^p = \emptyset$, and its report $\mathcal{Y} = \{y_1, \ldots, y_T\}$, the task of radiology report generation (RRG) is formulated as $p(\mathcal{Y}|\mathcal{X}, \mathcal{X}^p)$. We elaborate on the justification of using the historical records as context in Appendix A.7. Our proposed method, denoted as ICON, decomposes the RRG task into two stages, i.e., Lesion Extraction (Stage 1) and Report Generation (Stage 2). Specifically, given the input images $\mathcal{X}$, ICON first extracts lesions $\mathcal{Z} = \{Z_1, \ldots, Z_{|O|}\}$ from $\mathcal{X}$, where the probability of a region $R_{i,j}$ from image X_i being identified as a lesion Z_k is estimated as $p(Z_k|X_i)$. Subsequently, in Stage 2, ICON generates a report based on both the input images and the extracted lesions, modeled as $P(\mathcal{Y}|\mathcal{X}, \mathcal{X}^p, \mathcal{Z})$. Finally, our framework aims to maximize the following probability:

$$P(\mathcal{Y}|\mathcal{X}, \mathcal{X}^p) \propto \underbrace{p(\mathcal{Z}|\mathcal{X})}_{\text{Stage 1}} \cdot \underbrace{P(\mathcal{Y}|\mathcal{X}, \mathcal{X}^p, \mathcal{Z})}_{\text{Stage 2}}. \quad 131$$

2.2 Observation Annotation for Lesion Extraction

In order to perform lesion extraction, ICON requires report-level labels for lesion extraction, whereas benchmarking datasets only provide token-level labels. Therefore, it is necessary to annotate the reports to obtain higher-level labels. In this work, we adopt CheXbert (Smit et al., 2020) for this purpose. Specifically, CheXbert first annotates a report with 14 observation categories $O = \{o_1, \ldots, o_{14}\}$ (refer to Appendix A.1 for data statistics). Each observation is then assigned one of four statuses (i.e., *Present*, *Absent*, *Uncertain*, and *Blank*). During training and evaluation, *Present* and *Uncertain* are merged into the *Positive* category, which represents abnormal observation. Note that for the observation category *No Finding*, only two statuses, *Present* or *Absent*, are applicable. Finally, observation information is utilized for lesion extraction as will be illustrated in §3.2.

2.3 Inter-Report Consistency Metrics

To assess the inter-report consistency of a model, we introduce two metrics, i.e., CON and R-CON, similar to Elazar et al. (2021). We start by gathering samples with semantically equivalent images

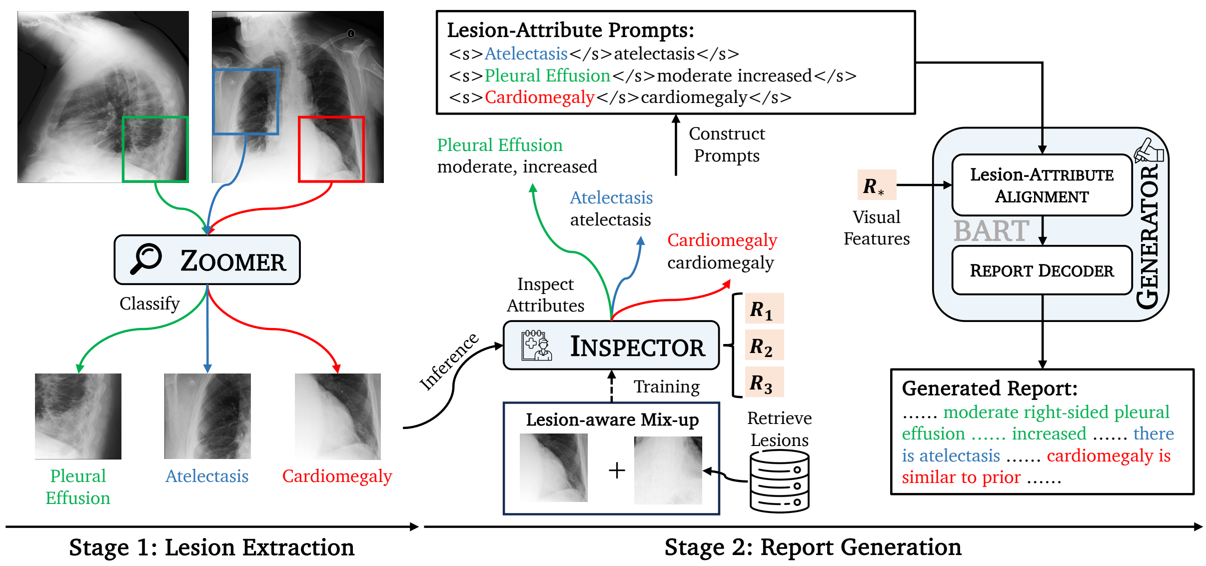


Figure 2: Overview of the ICON framework, which first extracts lesions from radiographs and then generates reports.

based on their reports' similarity. To be specific, given a corpus $\mathcal{D}$ filtered out samples without abnormality, we first build a BM25 retriever (Robertson and Zaragoza, 2009) based on the training data, and then the top-N ($N = 5$) relevant reports are selected as the semantically equivalent samples $\mathcal{K}_i = \{K_i, \dots, K_N\}$ for each report Q_i in $\mathcal{D}$. We also collect the corresponding outputs of $\mathcal{K}_i$ from a model, denoted as $\hat{\mathcal{K}}_i = \{\hat{K}_1, \dots, \hat{K}_N\}$. Afterward, the Jaccard Index (Jaccard, 1901; Wikipedia, 2024) is adopted to measure the similarity between two generated reports $(\hat{Q}_i, \hat{K}_j)$:

$$\texttt{Jaccard}(\hat{Q}_i, \hat{K}_j) = \frac{|\hat{t}_i \cap \hat{t}_j|}{|\hat{t}_i \cup \hat{t}_j|},$$

where $\hat{t}_i$ and $\hat{t}_j$ are tokens[2] in $\hat{Q}_i$ and $\hat{K}_j$, respectively. Then, the consistency is defined as:

$$\textsc{Con} = \frac{1}{|D| \times N} \sum_{i=1}^{|D|} \sum_{j=1}^{N} \texttt{Jaccard}(\hat{Q}_i, \hat{K}_j).$$

As CON only considers the consistency among the model's outputs, ignoring the qualities concerning the references. To consider both consistency and accuracy, for each output $\hat{Q}_i$, we plug a factor $\tau_i = \texttt{Jaccard}(\hat{Q}_i, Q_i)$, which measures the similarity between the hypothesis $\hat{Q}_i$ and reference Q_i. Then, CON is modified as R-CON:

$$\text{R-}\textsc{Con} = \frac{1}{|D| \times N} \sum_{i=1}^{|D|} \sum_{j=1}^{N} \tau_i \cdot \texttt{Jaccard}(\hat{Q}_i, \hat{K}_j).$$

We report the CON and R-CON results in Table 2.

[2] Stopwords are removed in advanced.

3 Methodology

3.1 Visual Encoding

Given an image X_l, an image processor is first utilized to split X_l into N patches, and then a visual encoder f_θ (e.g., Swin Transformer (Liu et al., 2021d)) is employed to extract visual representations $\boldsymbol{X}_l$ and the pooler output $\boldsymbol{P}_l \in \mathbb{R}^h$:

$$[\boldsymbol{P}_l, \boldsymbol{X}_l] = f_\theta(X_l),$$

where $\boldsymbol{X}_l = \{\boldsymbol{x}_{l,i}, \dots, \boldsymbol{x}_{l,N}\}$ and $\boldsymbol{x}_{l,i} \in \mathbb{R}^h$ is the i-th visual representation.

3.2 Stage 1: Extracting Lesions via Observation Classification (ZOOMER)

Observation Classification. A ZOOMER is a visual encoder parameterized by θ_Z and trained to classify a given input $\mathcal{X}$ into abnormal observations as mentioned in §2.2:

$$p(o_i) = \textsc{Zoomer}(\mathcal{X}).$$

Specifically, ZOOMER first encodes images $\mathcal{X} = \{X_1, \dots, X_L\}$ as outlined in §3.1, and then takes the averaged pooler output for classification, following these steps:

$$[\boldsymbol{P}_l, \boldsymbol{X}_l] = f_{\theta_Z}(X_l),$$
$$\boldsymbol{P} = \frac{1}{L} \sum \boldsymbol{P}_l,$$
$$p(o_i) = \sigma(\boldsymbol{W}_i \boldsymbol{P} + b_i),$$

where $\boldsymbol{W}_i \in \mathbb{R}^h$ is the weight for i-th observation, $b_i \in \mathbb{R}$ is its bias, and σ is the Sigmoid function.

Zooming In for Lesion Extraction. Upon completing training ZOOMER, we can use it to extract lesions without the need for object detectors (Ren et al., 2015). It is worth noting that our method does not require fine-grained labels, such as bounding boxes (Tanida et al., 2023), making it easily adaptable to other modalities, e.g., FFA images (Li et al., 2021).

For an image X_l, a sliding window with a 0.375 ratio of X_l is applied to extract M region candidates $\mathcal{R}_l = \{R_{l,1}, \ldots, R_{l,M}\}$ from X_l, as shown in the left side of Figure 2. These regions are then sequentially fed into ZOOMER for classification. Further details on the extraction of these regions can be found in Appendix A.5. The probability of a region $R_{l,j}$ being classified as an abnormal observation o_i is:

$$p_{l,j}(o_i) = \text{ZOOMER}(R_{l,j}).$$

For each study, all images in $\mathcal{X}$ are iterated, and only the region with the highest $p_{l,j}(o_i)$ is chosen as a lesion Z_i corresponding to the observation o_i. Finally, the set of lesions is denoted as $\mathcal{Z} = \{Z_1, \ldots, Z_{|O|}\}$.

Training ZOOMER. ZOOMER is optimized using the binary cross-entropy (BCE) loss. To handle the class-imbalanced issue (refer to Appendix A.1 for details), a weight factor α_j is plugged for each abnormal observation and the loss function $\mathcal{L}_{\text{S1}}$ is:

$$\begin{aligned}\texttt{BCE}(p(o_*), O) = -\frac{1}{|O|}\sum_j [\alpha_j \cdot o_j \cdot \texttt{log}\, p(o_j) \\ + (1 - o_j)\texttt{log}(1 - p(o_j))],\end{aligned}$$

where $o_j \in \{0, 1\}$ is the label, $\alpha_j = 1 + \texttt{log}(\frac{|\mathcal{D}_{\text{train}}| - w_j}{w_j})$, and $|\mathcal{D}_{\text{train}}|$ and w_j are the number of sample and the number of j-th observation in the training set, respectively.

3.3 Stage 2: Inspecting Lesions (INSPECTOR)

Inspecting Lesions with Attributes. Given that lesions of the same observation can exhibit different characteristics, it is crucial to inspect each lesion and match it with corresponding attributes to differentiate it from other variations. We adopt the attributes released by Hou et al. (2023a) as labels, and for each observation, the top-K attributes are selected as candidates. In specific, an INSPECTOR is a visual encoder parameterized by θ_I, similar to §3.2. INSPECTOR$(\boldsymbol{P}^p, \boldsymbol{P}, Z_j)$ takes prior and current visits chest X-rays as context, along with a lesion region as input:

$$\begin{aligned}[\boldsymbol{P}_{Z_j}, \boldsymbol{Z}_j] &= f_{\theta_I}(Z_j), \\ p_j(a_k) &= \sigma(\texttt{MLP}(\boldsymbol{P}^p, \boldsymbol{P}, \boldsymbol{P}_{Z_j})),\end{aligned}$$

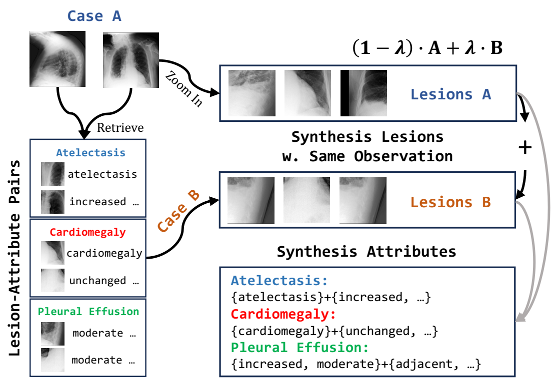


Figure 3: Overview of our proposed lesion-aware mix-up augmentation.

where MLP is a two-layer perception with non-linear activation, and $\boldsymbol{P}^p, \boldsymbol{P}, \boldsymbol{P}_{Z_j} \in \mathbb{R}^h$ are pooler output of prior images, current images, and the lesion, respectively. The lesion features $\boldsymbol{\mathcal{Z}} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{|O|}\}$ are then collected for report generation. For image encoding, we use another visual encoder f_{θ_V} to encoder $\mathcal{X}$ into $\boldsymbol{\mathcal{X}}$ and $\mathcal{X}^p$ into $\boldsymbol{\mathcal{X}}^p$.

Lesion-aware Mix-up Augmentation. To further improve the consistency of the generated outputs, we adopt the Mix-up Augmentation method (Zhang et al., 2018) and devise a Lesion-aware Mix-up during the training phase. To be specific, for a lesion-attribute pair (Z_j, A_j), we retrieve a similar pair (Z_k, A_k) with the same observation from the training data based on report similarity. These lesions are combined using linear interpolation, as illustrated in Figure 3:

$$Z_j^* = \lambda Z_j + (1 - \lambda) Z_k,$$

where λ is set to 0.75. Note that during training, Z_j^* is used for both INSPECTOR and GENERATOR.

Training INSPECTOR. Similar to §3.2, we adopt a linearly interpolated BCE loss to optimize INSPECTOR:

$$\mathcal{L}_{\text{I}} = \lambda\texttt{BCE}_j + (1 - \lambda)\texttt{BCE}_k,$$

where $\texttt{BCE}_j$ and $\texttt{BCE}_k$ take A_j and A_k as their respective labels. Notably, only the attributes that are shared between Z_j and Z_k are fully optimized. Consequently, our lesion-aware mix-up augmentation technique facilitates the improvement of output consistency for two semantically equivalent lesions.

3.4 Generating Consistent Radiology Report (GENERATOR)

Lesion-Attribute Alignment. To bridge the modality gap between lesion representations and text-based attributes, we leverage a BART (Lewis

Dataset	Model	NLG Metrics						CE Metrics (CheXbert)		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -ABN	R2GEN	0.290	0.157	0.093	0.061	0.105	0.208	0.266	0.320	0.272
	R2GENCMN	0.264	0.140	0.085	0.056	0.098	0.212	0.290	0.319	0.280
	ORGAN	0.314	0.180	0.114	0.078	0.120	0.234	0.271	0.342	0.293
	RECAP	0.321	0.182	0.116	0.080	0.120	0.223	0.300	0.363	0.305
	ICON (Ours)	**0.337**	**0.195**	**0.126**	**0.086**	**0.129**	**0.236**	**0.332**	**0.430**	**0.360**
MIMIC -CXR	R2GEN	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
	R2GENCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
	$\mathcal{M}^2$TR	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
	KNOWMAT	0.363	0.228	0.156	0.115	–	0.284	**0.458**	0.348	0.371
	CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
	CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
	KiUT	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	DCL	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
	METrans	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
	RGRG	0.373	0.249	0.175	**0.126**	0.168	0.264	0.380	0.319	0.305
	ORGAN	0.407	0.256	0.172	0.123	0.162	**0.293**	0.416	0.418	0.385
	RECAP	**0.429**	**0.267**	0.177	0.125	0.168	0.288	0.389	0.443	0.393
	ICON (Ours)	**0.429**	0.266	**0.178**	**0.126**	**0.170**	0.287	0.445	**0.505**	**0.464**

Table 1: Experimental results of our model and baselines on the MIMIC-ABN and MIMIC-CXR datasets. The best results are in **boldface**, and the underlined are the second-best results.

Model	MIMIC-ABN		MIMIC-CXR	
	CON	R-CON	CON	R-CON
MAJORITY	100.0	–	100.0	–
R2GEN	16.89	15.60	8.40	9.52
R2GENCMN	16.10	17.12	9.22	11.17
ORGAN	20.15	26.80	18.65	25.09
RECAP	17.02	21.42	16.81	23.80
ICON (Ours)	17.48	**31.15**	16.72	**28.36**
ICON *w/o* ZOOM	11.38	18.93	10.66	15.65
ICON *w/o* INSPECT	12.82	18.83	13.21	19.59
ICON *w/o* MIX-UP	15.72	23.94	15.88	26.93

Table 2: The CON score (in %) and the R-CON score (in ‰). MAJORITY denotes only one generated report is used for evaluation.

Dataset	Model	NLG Metrics		CE Metrics (RadGraph)		
		B-4	R-L	RG_E	RG_ER	$\text{RG}_{\overline{\text{ER}}}$
IU X-RAY	R2GEN	**0.120**	0.298	–	–	–
	$\mathcal{M}^2$TR	0.121	0.288	–	–	–
	$\mathcal{T}_\text{NLL}$	0.114	–	0.230	0.202	0.153
	ICON	0.098	**0.320**	**0.342**	**0.312**	**0.246**
MIMIC -CXR	$\mathcal{T}_\text{NLL}$	0.105	0.253	0.230	0.202	0.153
	ORGAN	0.123	**0.293**	0.303	0.275	0.199
	RECAP	0.125	0.288	0.307	0.276	**0.205**
	ICON	**0.126**	0.287	**0.312**	**0.278**	0.197

Table 3: Radgraph evaluation results on the IU X-RAY and MIMIC-CXR datasets. Results of $\mathcal{T}_\text{NLL}$ are cited from Delbrouck et al. (2022).

et al., 2020) encoder to extract attribute representations. The attributes associated with each lesion are formulated as a prompt: <s>o_j</s>A_j</s>, as depicted in the upper part of Figure 2. Then, a cross-attention module (Vaswani et al., 2017) is inserted after every self-attention module. This module aligns the lesion representations with the attribute representations by querying visual representations using attribute representations, similar to Q-Former (Li et al., 2023a):

$$\boldsymbol{H}_j^a = \texttt{Attention}(\boldsymbol{H}_j^s, \boldsymbol{Z}_j, \boldsymbol{Z}_j),$$

where $\boldsymbol{H}_j^a, \boldsymbol{H}_j^s \in \mathbb{R}^h$ are the aligned attribute representation and the self-attended representation of A_j, respectively. All prompts are encoded and the attribute representations of $\boldsymbol{\mathcal{Z}}$ are denoted as $\boldsymbol{\mathcal{H}}^a$.

Report Generation. Given the input images $\boldsymbol{\mathcal{X}}$, images of prior visit $\boldsymbol{\mathcal{X}}^p$, the lesions $\boldsymbol{\mathcal{Z}}$, and attribute $\boldsymbol{\mathcal{H}}^a$, we utilize a BART decoder in conjunction with the Fusion-in-Decoder (FiD; (Izacard and Grave, 2021)) technique that simply concatenates multiple context sequences for report generation. Then, the probability of t-th step is expressed as:

$$\boldsymbol{h}_t = \texttt{FiD}([\boldsymbol{\mathcal{X}}; \boldsymbol{\mathcal{X}}^p; \boldsymbol{\mathcal{Z}}; \boldsymbol{\mathcal{H}}^a], \boldsymbol{h}_{<t}),$$
$$p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t}) = \texttt{Softmax}(\boldsymbol{W}_g \boldsymbol{h}_t + \boldsymbol{b}_g),$$

where $\boldsymbol{h}_t \in \mathbb{R}^h$ is the t-th hidden representation, $\boldsymbol{W}_g \in \mathbb{R}^{|\mathcal{V}| \times h}$ is the weight matrix, $\boldsymbol{b}_g \in \mathbb{R}^{|\mathcal{V}|}$ is the bias vector, and $\mathcal{V}$ is the vocabulary.

Training GENERATOR. The generation process is optimized using the negative log-likelihood loss, given each token's probability $p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t})$:

$$\mathcal{L}_\text{G} = -\sum_{t=1}^{T} \texttt{log}\, p(y_t|\mathcal{X}, \mathcal{X}^p, \mathcal{Z}, \mathcal{Y}_{<t}).$$

Dataset	Model	NLG Metrics						CE Metrics		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
MIMIC -ABN	ICON	0.337	0.195	0.126	0.086	0.129	0.236	0.332	0.430	0.360
	ICON *w/o* ZOOM	0.310	0.181	0.119	0.084	0.120	0.243	0.306	0.353	0.306
	ICON *w/o* INSPECT	0.315	0.182	0.117	0.081	0.121	0.236	0.338	0.401	0.352
	ICON *w/o* MIX-UP	0.335	0.192	0.124	0.085	0.129	0.239	0.332	0.413	0.356
MIMIC -CXR	ICON	0.429	0.266	0.178	0.126	0.170	0.287	0.445	0.505	0.464
	ICON *w/o* ZOOM	0.377	0.237	0.162	0.119	0.149	0.288	0.363	0.280	0.278
	ICON *w/o* INSPECT	0.399	0.248	0.168	0.122	0.157	0.287	0.444	0.447	0.423
	ICON *w/o* MIX-UP	0.427	0.264	0.176	0.124	0.169	0.285	0.444	0.502	0.462

Table 4: Ablation results of our model and its variants on the MIMIC-ABN and MIMIC-CXR datasets. ICON *w/o* ZOOM (§3.2) is the standard encoder-decoder model, *w/o* INSPECT stands for without INSPECTOR (§3.4), and *w/o* MIX-UP stands for without lesion-aware Mix-up augmentation (§3.4).

Model	B-4	R-L	CE-F_1	TEM
CXR-RePaiR-2	0.021	0.143	0.281	0.125
BioViL-NN	0.037	0.200	0.283	0.111
BioViL-T-NN	0.045	0.205	0.290	0.130
BioViL-AR	0.075	0.279	0.293	0.138
BioViL-T-AR	0.092	**0.296**	0.317	0.175
RECAP	0.118	0.279	0.400	0.304
ICON (Ours)	**0.120**	0.279	**0.468**	**0.335**

Table 5: Progression modeling results on the MIMIC-CXR dataset. The results of BioViL-* models are cited from Bannur et al. (2023).

The loss function of Stage 2 is: $\mathcal{L}_{S2} = \mathcal{L}_I + \mathcal{L}_G$.

4 Experiments

4.1 Datasets

Three public datasets are used to evaluate our models, i.e., IU X-RAY[3] (Demner-Fushman et al., 2016), MIMIC-CXR[4] (Johnson et al., 2019), and MIMIC-ABN[5] (Ni et al., 2020). We follow previous research (Chen et al., 2020) to preprocess these datasets, and provide other details in Appendix A.6.

- IU X-RAY consists of 3,955 reports. We follow previous research (Chen et al., 2020) and split the dataset into train/validation/test sets with a ratio of 7:1:2.
- MIMIC-CXR consists of 377,110 chest X-ray images and 227,827 reports.
- MIMIC-ABN is modified from the MIMIC-CXR dataset and its reports only contain abnormal part. We adopt the data-split as used in Hou et al. (2023a), and the data-split is 71,786/546/806 for train/validation/test sets.

Unlike previous research (Chen et al., 2020; Hou et al., 2023a) which only used one view for report generation on MIMIC-CXR and MIMIC-ABN, we collect all views for each visit in experiments. The justification is provided in Appendix A.7.

4.2 Evaluation Metrics and Baselines

NLG Metrics. To assess the quality of generated reports, we adopt several natural language generation (NLG) metrics for evaluation. BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) are selected as NLG Metrics, and we use the MS-COCO caption evaluation tool[6] to compute the results.

CE Metrics. Following previous research (Chen et al., 2020, 2021), we adopt macro-weighted clinical efficacy (CE) metrics to evaluate the observation-level factual accuracy, and CheXbert (Smit et al., 2020) is used in this paper. To measure the entity-level factual accuracy, we leverage the RadGraph (Jain et al., 2021; Delbrouck et al., 2022) and temporal entity matching (TEM) scores for evaluation.

Consistency Metrics. CON and R-CON (mentioned in §2.3) are utilized to measure the inter-report consistency.

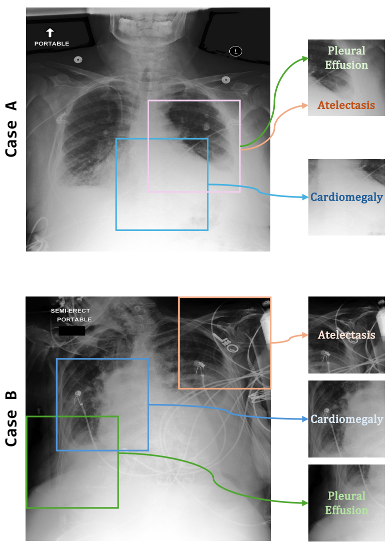
Baselines. We compare our models with the following state-of-the-art baselines: R2GEN (Chen et al., 2020), R2GENCMN (Chen et al., 2021), KNOW-MAT (Yang et al., 2021), $\mathcal{M}^2$TR (Nooralahzadeh et al., 2021), CMM-RL (Qin and Song, 2022), CMCA (Song et al., 2022), CXR-RePaiR-Sel/2 (Endo et al., 2021), BioViL-T (Bannur et al., 2023), DCL (Li et al., 2023b), METrans (Wang et al., 2023c), KiUT (Huang et al., 2023), RGRG (Tanida et al., 2023), ORGAN (Hou et al., 2023b), and RECAP (Hou et al., 2023a).

[3] https://openi.nlm.nih.gov/

[4] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

[5] https://github.com/zzxslp/WCL

[6] https://github.com/tylin/coco-caption



Reference A: in comparison with the study of there is little overall change. continued enlargement of the cardiac silhouette with pulmonary vascular congestion and bilateral pleural effusions with compressive atelectasis. central catheter remains in place.

ICon *w/o* Zoom: as compared to the previous radiograph the patient has received a right-sided picc line. the course of the line is unremarkable the tip of the line projects over the mid svc. there is no evidence of complications notably no pneumothorax. otherwise the radiograph is unchanged.

ICon: in comparison with the study of there is little overall change. continued enlargement of the cardiac silhouette with pulmonary vascular congestion and mild to moderate cardiomegaly. small bilateral pleural effusions with areas of compressive atelectasis. the right picc line has been removed. nasogastric tube remains in place.

Reference B: in comparison with the study of there is again enlargement of the cardiac silhouette with extensive bilateral pleural effusions and compressive atelectasis combined with pulmonary vascular congestion.

ICon *w/o* Zoom: as compared to the previous radiograph there is no relevant change. moderate cardiomegaly with bilateral pleural effusions and subsequent areas of atelectasis. moderate pulmonary edema. no newly appeared focal parenchymal opacity suggesting pneumonia.

ICon: as compared to the previous radiograph there is unchanged evidence of moderate cardiomegaly with pulmonary vascular congestion and moderate pulmonary edema. unchanged moderate bilateral pleural effusions with areas of compressive atelectasis at the lung bases. no new parenchymal opacities. no pneumothorax.

Figure 4: A case study of ICon on two semantically equivalent cases (i.e., Case A and Case B), given their radiographs and lesions. Spans with the same color (*Cardiomegaly*, *Pleural Effusion*, *Atelectasis*, and *Others*) represent the same positive observation. Consistent outputs are highlighted with underline.

4.3 Implementation Details

The small and tiny version of Swin Transformer V2 (Liu et al., 2022) are used as the visual backbone for Zoomer and Inspector, respectively. The Generator is initialized with the base version of BART pretrained on biomedical corpus (Yuan et al., 2022). Other parameters are randomly initialized. For Stage 2 training, the learning rate is $5e-5$ with linear decay, the batch size is 32, and the models are trained for 20 and 5 epochs on MIMIC-ABN and MIMIC-CXR with early stopping, respectively. Since the number of samples in IU X-Ray is too small to train a multimodal model, we only provide results produced by models trained on MIMIC-CXR as a reference, similar to (Delbrouck et al., 2022). For other training details (e.g., training Zoomer), and the resources used in this paper, we list them in Appendix A.2.

5 Results

5.1 Quantitative Analysis

NLG Results. The NLG results are presented in the left section of Table 1. Our proposed method, ICon, achieves SOTA performance, demonstrating significant improvements on the MIMIC-ABN dataset. Regarding the MIMIC-CXR dataset, our model achieves BLEU-3 and METEOR scores of 0.178 and 0.170 respectively, surpassing previous top-performing models, while maintaining strong performance on other metrics. Additionally, we provide experimental results on the IU X-Ray dataset as a reference in Table 3.

Inter-Report Consistency Analysis. Table 2 provides Con and R-Con scores, where our models achieve the highest R-Con for both datasets. In terms of the Con score, ICon demonstrates competitive performance when compared with ORGan. However, it is important to note that the Majority baseline achieves perfect (100%) consistency when generating only one report. Hence, the Con score primarily serves as a reference.

CE (CheXbert and RadGraph) Results. In the right section of Table 1, we observe that our proposed model, ICon, achieves SOTA clinical accuracy, increasing CE F_1 from 0.393 to 0.464 on the MIMIC-CXR dataset and rising it by 5.5% on the MIMIC-ABN dataset. These results indicate that our model is capable of generating accurate and consistent radiology reports. Furthermore, Table 3 presents the RadGraph F_1 on both the IU X-Ray and MIMIC-CXR dataset. Our model achieves competitive performance compared with the non-RL-optimized baselines.

Temporal Modeling Results. Since ICon takes longitudinal information as context, we present the results of our temporal modeling analysis in Table 5. Notably, ICon exhibits significant improvements over other baselines in terms of BLEU score, clinical accuracy, and TEM score while maintaining competitive performance on the ROUGE score, indicating its enhanced capacity to effectively utilize historical records.

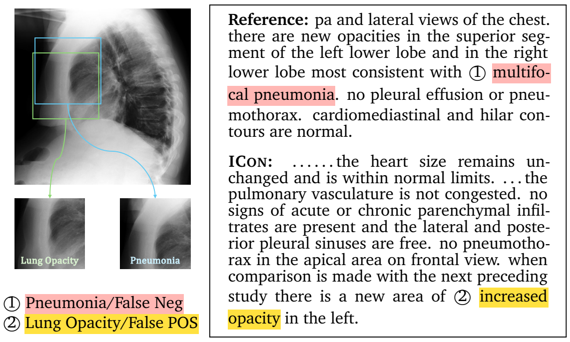


Figure 5: An error case, where AP and Lateral views are available, produced by ICON. The span and the span denote false negative observation and false positive observation, respectively.

Ablation Results. The ablation results for MIMIC-ABN and MIMIC-CXR are listed in Table 2 and Table 4. The performance of the ablated model *w/o* ZOOM drops significantly for both datasets, while the variant *w/o* INSPECT achieves competitive results on clinical accuracy. This suggests that the ZOOMER effectively extracts lesions and provides relevant abnormal information for report generation. In addition, the variant *w/o* MIX-UP further improves the performance, which demonstrates the effectiveness of INSPECTOR in transforming concise lesion information into precise free-text reports. Moreover, introducing lesion-aware mix-up augmentation strengthens the consistency of generated outputs, indicating the effectiveness of ICON.

5.2 Qualitative Analysis

Case Study. Figure 4 showcases two semantically equivalent cases, i.e., Case A and Case B, extracted from the test set of MIMIC-CXR. In both instances, ICON successfully identifies abnormal observations (e.g., *Cardiomegaly*, *Pleural Effusion*, and *Atelectasis*) and generates consistent phrases including "*pulmonary vascular congestion*", "*bilateral pleural effusions*", and "*compressive atelectasis.*" Conversely, variant *w/o* ZOOM fails to produce these descriptions given Case A. This demonstrates that ZOOMER plays a crucial role in identifying lesions and also highlighting the ability of the mix-up augmentation to ensure the alignment of lesions with their corresponding attributes.

Error Analysis. Figure 5 presents an error case produced by ICON. Although ZOOMER successfully identifies *Pneumonia* in the given radiographs, GENERATOR fails to realize it into descriptions "*multifocal pneumonia*" (i.e., false negative observation). We notice that the lesion of this observation is inaccurately identified. Additionally, ZOOMER outputs a false positive observation *Lung Opacity*, leading to an inaccurate phrase "*increased opacity*". In light of this, a better ZOOMER trained with larger datasets could be a way to mitigate it.

6 Related Works

Radiology report generation (Jing et al., 2018; Li et al., 2018) has gained much attention recently. Prior research either devised various memory mechanisms to record key information (Chen et al., 2020, 2021; Qin and Song, 2022; Wang et al., 2023c; Zhao et al., 2023) or proposed different learning methods to enhance the performance (Liu et al., 2021c,a,b). In addition, Yang et al. (2021); Li et al. (2023b); Huang et al. (2023); Yan et al. (2023) proposed to utilize knowledge graphs for report generation. Liu et al. (2019); Lovelace and Mortazavi (2020); Miura et al. (2021); Nishino et al. (2022); Delbrouck et al. (2022) designed various rewards for reinforcement learning to improve clinical accuracy. Tanida et al. (2023) proposed an explainable framework for report generation. Hou et al. (2023b) proposed to introduce observations to improve factual accuracy. Additionally, Ramesh et al. (2022); Bannur et al. (2023); Hou et al. (2023a); Dalla Serra et al. (2023) focus on exploring the temporal structure. Wang et al. (2023b,a) utilize CLIP (Radford et al., 2021) to bridge the modality gap. Mix-up augmentation is also closely related to this research Zhang et al. (2018), and this method has been adopted in various NLP research (Sun et al., 2020; Yoon et al., 2021; Yang et al., 2022).

7 Conclusion and Future Works

In this paper, we propose ICON comprising three components to improve the inter-report consistency between semantically equivalent lesions. ICON first extracts lesions and then matches fine-grained attributes for report generation. A lesion-aware mix-up augmentation method is devised for attribute alignment. Experimental results on three datasets demonstrate the effectiveness of ICON. In the future, we plan to explore incorporating large language models (LLMs) into our framework, given their advanced capabilities in planning and generation, to further enhance the performance of the task of radiology report generation. Leveraging the strengths of LLMs could provide more refined signals to enhance the performance of ZOOMER.

Limitations

Although ICON can improve the consistency of radiology report generation, it still exhibits some limitations. Since our lesion extraction method is based on coarse-grained labels (i.e., image labels), training such a model requires annotations for images. However, annotations could be difficult to obtain in some medical settings. Recent advances in foundation vision models (Kirillov et al., 2023) and open-set learning (Zara et al., 2023) could be a potential direction to handle this issue. In addition, since our framework consists of two stages, prediction errors will propagate through the pipeline, and as a result, the final performance of our framework largely depends on Stage 1. Reinforcement learning (Nishino et al., 2022) that takes factual improvement as a reward could be a solution to optimize the framework in an end-to-end manner.

Ethics Statement

The IU X-RAY (Demner-Fushman et al., 2016), MIMIC-ABN (Ni et al., 2020), and MIMIC-CXR (Johnson et al., 2019) are publicly available datasets and have been automatically de-identified to protect patient privacy. Our goal is to enhance the inter-report consistency of radiology report generation systems. Despite the substantial improvement of our framework over SOTA baselines, the performance still lags behind the requirements of real-world deployment and could lead to unexpected failures in an untested environment. Thus, we urge the readers of this paper and the potential users of this system to cautiously check the generated outputs and should inquire about suggestions of experts when using it.

References

Onur Asan, Alparslan Emrah Bayrak, and Avishek Choudhury. 2020. Artificial intelligence and human trust in healthcare: Focus on clinicians. *J Med Internet Res*, 22(6):e15154.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to exploit temporal structure for biomedical vision-language processing.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5904–5914. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. 2023. Controllable chest X-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4891–4904, Singapore. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. Recap: Towards precise radiology report generation via dynamic disease progression reasoning.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. ORGAN: Observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8108–8122, Toronto, Canada. Association for Computational Linguistics.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19809–19818.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *CoRR*, abs/2106.14463.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2577–2586. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Sara Kaviani, Ki Jin Han, and Insoo Sohn. 2022. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. *Expert Systems with Applications*, 198:116815.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, yingfeng zheng, Yizhi Liu, Flora D. Salim, Karin Verspoor, Xiaodan Liang, and Xiaojun Chang. 2021. FFA-IR: Towards an explainable and reliable medical report generation benchmark . In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3001–3012. Association for Computational Linguistics.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13753–13762. Computer Vision Foundation / IEEE.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021c. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 269–280. Association for Computational Linguistics.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *CoRR*, abs/1904.02633.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021d. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.

Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. 2020. Learning visual-semantic embeddings for reporting abnormal findings on chest X-rays. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1954–1960, Online. Association for Computational Linguistics.

Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. 2022. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2020. Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, 14:156–180.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 448–458. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. 2022. Improving radiology report generation systems by removing hallucinated references to non-existent priors.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA. MIT Press.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. 2022. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2388–2397, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Siyuan Wang, Zheng Liu, and Bo Peng. 2023a. A self-training framework for automated medical report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16443–16449, Singapore. Association for Computational Linguistics.

Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. 2023b. Fine-grained medical vision-language representation learning for radiology report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15949–15956, Singapore. Association for Computational Linguistics.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023c. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11558–11567.

Wikipedia. 2024. Jaccard index — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Jaccard%20index&oldid=1196092673`. [Online; accessed 12-February-2024].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. 2023. Style-aware radiology report generation with RadGraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688, Singapore. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Shuxin Yang, Xian Wu, Shen Ge, Shaohua Kevin Zhou, and Li Xiao. 2021. Knowledge matters: Radiology report generation with general and specific knowledge. *CoRR*, abs/2112.15009.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. SSMix: Saliency-based span mixup for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3225–3234, Online. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Giacomo Zara, Subhankar Roy, Paolo Rota, and Elisa Ricci. 2023. Autolabel: Clip-based framework for open-set video domain adaptation.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3549–3557.

Guosheng Zhao, Yan Yan, and Zijian Zhao. 2023. Normal-abnormal decoupling memory for medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1962–1977, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Abnormal Observation Statistics

The abnormal observation statistics of MIMIC-ABN, MIMIC-CXR, and IU X-RAY are listed in Table 6.

#Observation	MIMIC-ABN	MIMIC-CXR	IU X-RAY
No Finding	5002/32/22	64,677/514/229	744/108/318
Cardiomegaly	16,312/118/244	70,561/514/1,602	244/38/61
Pleural Effusion	10,502/80/186	56,972/477/1,379	60/13/15
Pneumothorax	1,452/24/4	8,707/62/106	9/2/5
Enlarged Card.	5,202/40/90	49,806/413/1,140	159/29/28
Consolidation	4,104/36/96	14,449/119/384	17/1/3
Lung Opacity	22,598/166/356	67,714/497/1,448	295/35/57
Fracture	4,458/32/76	11,070/59/232	84/6/15
Lung Lesion	5,612/54/112	11,717/123/300	85/14/17
Edema	8,704/76/168	33,034/257/899	28/2/7
Atelectasis	19,132/134/220	68,273/515/1,210	143/15/37
Support Devices	9,886/58/196	60,455/450/1,358	89/20/16
Pneumonia	17,826/138/260	23,945/184/503	20/2/1
Pleural Other	2,850/30/62	7,296/70/184	32/4/7

Table 6: Observation distribution in train/valid/test split of three datasets. *Enlarged Card.* refers to *Enlarged Cardiomediastinum.*

A.2 Implementation Details and Related Pretrained Models

For Stage 1, all three datasets use the same hyperparameters for training ZOOMER, with a learning rate of $1e-4$, batch size of 128, and dropout rate of 0.1, and the number of training epochs is adjusted accordingly. We train ZOOMER for 5, 10, and 15 epochs on MIMIC-ABN, MIMIC-CXR, and IU X-RAY, respectively. During training, several data augmentation methods are applied. The input resolution of Swin Transformer is 256×256, and we first resize an image to 288×288, and then randomly crop it to 256×256 with random horizontal flip. All experiments are conducted using one NVIDIA-3090 GTX GPU. For Stage 2, no data augmentation is applied, and we conduct experiments on MIMIC-ABN and IU X-RAY using two NVIDIA-3090 GTX GPUs, and on MIMIC-CXR using four NVIDIA-V100 GPUs, both with half precision. Our model has 328.38M trainable parameters, and the implementations are based on the HuggingFace's Transformers (Wolf et al., 2020). Here are the pretrained models we used:

- Small version of Swin Transformer V2: https://huggingface.co/microsoft/swinv2-small-patch4-window8-256
- Tiny version of Swin Transformer V2: https://huggingface.co/microsoft/swinv2-tiny-patch4-window8-256
- Base Version of Biomedical BART: https://huggingface.co/GanjinZero/biobart-v2-base

Observation	P	R	F_1
Enlarged Card.	0.442	0.525	0.428
Cardiomegaly	0.630	0.822	0.714
Lung Opacity	0.542	0.563	0.552
Lung Lesion	0.321	0.177	0.228
Edema	0.464	0.784	0.583
Consolidation	0.275	0.162	0.204
Pneumonia	0.341	0.350	0.345
Atelectasis	0.539	0.620	0.577
Pneumothorax	0.400	0.444	0.421
Pleural Effusion	0.721	0.827	0.770
Pleural Other	0.295	0.315	0.304
Fracture	0.225	0.164	0.190
Support Devices	0.785	0.784	0.785
No Finding	0.263	0.535	0.352
Macro Average	0.445	0.505	0.464

Table 7: Experimental results of each observation on the MIMIC-CXR dataset.

A.3 Detailed CE Results of ICON on the MIMIC-CXR Dataset

A.4 Experimental Results of Stage 1

The experimental results are provided in Table 8. Results on the IU X-RAY dataset are only provided for reference.

Dataset	P	R	F_1
IU X-RAY	0.223	0.243	0.225
MIMIC-ABN	0.379	0.472	0.411
MIMIC-CXR	0.454	0.550	0.491

Table 8: Abnormal observation prediction results of ZOOMER at Stage 1.

A.5 Lesion Extraction

There are two steps in extraction lesions: candidate generation and candidate classification. Given an image with a resolution of 1024×1024, padding if needed, we apply a sliding window of 384×384, with a step size of 128 to extract candidates for classification. This operation results in 36 regions. Then, each region is fed into the ZOOMER for classification, and only the top-1 lesion is selected for each observation. Note that before extracting lesions, each input case is first assigned with their observations by ZOOMER, and as a result, the number of lesions corresponds to the number of observations.

The *No Finding* observation is excluded for lesion extraction, as it estimates the overall conditions of a patient, which makes it difficult to locate at specific regions.

A.6 Other Preprocessing Details

We adopt the same preprocessing setup used in Chen et al. (2020), and the minimum count of each token is set to 3/3/10 for IU X-RAY/MIMIC-ABN/MIMIC-CXR, respectively. Other tokens are replaced with a special token <unk>.

A.7 Justifications for Additional Data Processing

Justification for Using Historical Records. As stated in Hou et al. (2023a), without historical information, it is unreasonable to generate reports with comparisons between two consecutive visits and will lead to hallucinations (Ramesh et al., 2022). As a result, we include historical records as context information for report generation.

Justification for Using All Views. Prior research (Chen et al., 2020, 2021; Hou et al., 2023b,a) treated different views of radiographs in one visit as different samples. However, this is unreasonable to generate a report with only one view position, since different diseases could be observed from different view positions. For example, most of the devices can not be observed from a Lateral view. Given a lateral view radiograph, writing a sentence of "*A right chest tube is in unchanged position.*" is not acceptable.

In addition, some reports describe how many views are provided at the beginning, e.g., "*PA and lateral views are provided.*" Above all, we have justified reasons to use all the views in one visit of a patient to generate the target report. Note that previous work treated each image as a sample and their settings have more samples than ours. For a fair comparison, each generated output of a study with L images is duplicated L times so that the number of samples in evaluation is consistent with previous research.