Joint Imbalance Adaptation for Radiology Report Generation

Anonymous ACL submission

Abstract

Radiology report generation, predicting text descriptions for radiological images, may face critical challenges due to data imbalance – medical tokens appear less frequently than regular tokens, and normal labels of images may not equal to abnormal ones. However, existing studies mainly consider label imbalance without mitigating other factors, such as token imbalance. In this study, we jointly consider two imbalance factors, label and token, determining distributions of radiology images and language, two fundamental modalities of the generation task. We propose a **J**oint **Im**balance **A**daptation (*JIMA*) model to promote task robustness by leveraging token and label imbalance. Experiments on two standard evaluation data (IU X-ray (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019)) by automatic and human evaluations demonstrate our significant improvements over current state-of-the-art models. We conduct extensive ablation and case analyses to examine and present dual imbalance effects on the radiology report generation robustness. While data imbalance remains challenging, our approach opens new task directions and shows promising results.

1 Introduction

Radiology report generation is a multimodal and medical image-to-text task that generates text descriptions for radiographs (e.g., X-ray or CT scan), which may reduce the workloads of radiologists (Jing et al., 2018, 2019). The domain-specific task has own unique characteristics than general image-to-text tasks (e.g., image captioning), such as lengthy documents, medical annotations, and clinical terminologies. As demonstrated in Figure 1, *data imbalance* can significantly impact model robustness that prevents model deployment in practice – models can easily overfit on frequent patterns. However, encountering data imbalance to augment the robustness of the radiology report generation task is still in its infancy.

Two major data imbalances exist in the radiology generation task, label and token. *Label imbalance* pertains to a disproportionate ratio of normal and abnormal diagnosis categories, which exist in radiological images and text reports. For instance, in X-ray images, normal regions dominate major areas, and the numbers of normal and abnormal reports may not be equal, leading to failures in disease detection and description. As shown in Table 1, abnormal reports are considerably longer than normal reports, with an average difference of 60.56%. These reports are much harder to generate than shorter reports (Lovelace and Mortazavi, 2020; Tan et al., 2021; Wang et al., 2023).[1] Existing imbalance learning studies of radiology report generation primarily focus on label imbalance (Nishino et al., 2020; Yu and Zhang, 2022). *Token imbalance* is a critical challenge in generation that tokens have varied occurrence frequencies, and the issue is more critical in the medical task. Learning infrequent tokens can be harder than frequent tokens for generation models (Gu et al., 2020; Wu et al., 2023). Medical tokens appear less frequently than regular ones, and the infrequent tokens may contain more medical results, highlighting the domain-specific uniqueness. For example, our empirical analysis in Section 2 has demonstrated that over 80% medical terms are infrequent tokens, while frequent tokens can count over 82% corpus. However, to the best of our knowledge, a joint adaptation of label and token imbalance to enhance the robustness of radiology report generation has not yet been explored.

To jointly model label and token imbalance, we propose a **J**oint **Im**balance **A**daptation (JIMA) model by curriculum learning (Bengio et al., 2009) that assumes infrequent patterns are usually harder to optimize. JIMA dynamically guides the model

[1] Clinical reports are also much longer than general-domain image captions, such as MS-COCO (Lin et al., 2014).

Table 1: Data statistics summary. Variations exist in label (Normal and Abnormal %) and average report length (L).

	Image	Report	Vocab	Normal %	Abnormal %	L	L_{normal}	$L_{abnormal}$
IU X-ray	7,470	3,955	1,517	32.96%	67.04%	35.99	27.76	40.72
MIMIC-CXR	377,110	227,835	13,876	13.97%	86.03%	59.70	34.57	59.36

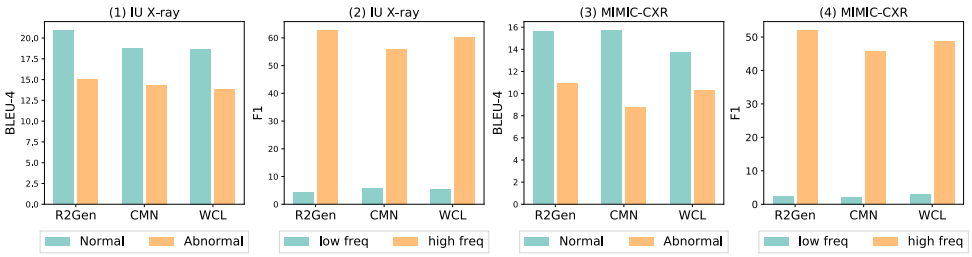


Figure 1: Baselines' BLEU-4 on normal and abnormal samples and F1 scores on low- and high-frequent tokens.

learning process by leveraging optimization difficulties, strengthening learning capability on infrequent samples, and alleviating overfittings on frequent patterns on both label and token. To jointly incorporate token and label imbalance, we propose to measure optimization difficulties by leveraging performance discrepancy across three evaluation aspects, token F1, label F1, and BLEU-4. We incorporate the three metrics as a joint optimization and design a novel Training Scheduler sampling and sorting training instances with the multi- difficulty scores based on performance discrepancies, which dynamically ranks easier samples when the models' performance decreases and vice versa. We conduct experiments on two publicly available datasets, MIMIC-CXR (Johnson et al., 2019) and IU X-ray (Demner-Fushman et al., 2015) with automatic and human evaluations. By comparing with six state-of-the-art baselines on overall and imbalance performance settings, our experiments show the promising results of our proposed approach. Our ablation and qualitative analyses show that JIMA can generate more coherent and precise medical reports, balancing label and token imbalance. Our code and data access will be available at [URL].

2 Data

We collected two publicly accessible datasets for this study, IU X-ray (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019), de-identified chest X-ray datasets to evaluate radiology report generation. IU X-ray (Demner-Fushman et al., 2015), collected from the Indiana Network for Patient Care, includes 7,470 X-ray images and corresponding 3,955 radiology reports. MIMIC-CXR (Johnson et al., 2019), collected from the Beth Israel Deaconess Medical Center, contains 377,110 X-ray images and 227,827 radiology reports for 65,379 patients. Each report is a text document and associates with one or more front and side X-ray images. Table 1 summarizes statistics of data imbalance. We include preprocessing details and imbalance visualizations in Appendix A.

Table 1 presents imbalance patterns in tokens and labels. Abnormal entries are predominant in both datasets, and MIMIC-CXR displays a more skewed label distribution, as more abnormal samples were collected during diagnosis phases not for screening purposes. MIMIC-CXR has a longer average length than IU X-ray. The lengthier documents may pose a unique multimodal generation challenge in the medical field. To conduct our analysis, we define the low and high frequency using the top 12.5% frequent tokens. Our findings in the Appendix A suggest a joint relation between label and token imbalance and higher ratios of low-frequency tokens in abnormal reports. This observation motivates us to investigate how the imbalance impacts model robustness and reliability.

2.1 Imbalance Effects

We examine the potential impact of label and token imbalance on model performance. To ensure consistency, we keep the top 12.5% to split low- and high-frequent tokens for evaluation purposes. The analysis includes three state-of-the-art models, R2Gen (Chen et al., 2020), WCL (Yan et al., 2021), and CMN (Chen et al., 2021). We either use released source codes and leave implementation details in the Appendix D.2. We use BLEU-4 (Papineni et al., 2002) and F1 scores to measure performance across both token (low vs high frequency) and label (normal vs. abnormal) imbalance. We

visualize performance variations in Figure 1.

The results suggest that the models exhibit significant difficulties in coping under label and token imbalance. Models consistently perform worse on abnormal reports, which are lengthier and have more infrequent tokens than normal reports. For example, the top 12.5% frequent tokens count > 80% tokens in two datasets, and low-frequent tokens have much worse performance than frequent tokens, as infrequent tokens are harder to optimize (Yu et al., 2022). However, infrequent tokens contain higher ratios of medical terms (e.g., silhouettes and pulmonary) describing health states. The significantly varying performance highlights the unique challenges to adapt token and label imbalance. While existing work (Nishino et al., 2020) has considered label imbalance, however, the study did not examine the performance effects of label or token imbalance. The findings inspire us to propose our model **J**oint **Im**balance **A**daptation (*JIMA*) to model token and label imbalance.

3 Joint Imbalance Adaptation

In this section, we present our approach **J**oint **Im**balance **A**daptation (*JIMA*) using *curriculum learning*. JIMA aims to augment model robustness under label and token imbalance. As optimizing data imbalance has been demonstrated difficulty, deploying such a learning strategy will strengthen model robustness and reliability. Our proposed approach deploys curriculum learning (*CL*) (Wang et al., 2022) that automatically adjusts the optimization process by gradually selecting training data entries from learning difficulty — learning from hard to easy samples as our optimization strategy (Zhou et al., 2020). To achieve the goal, we propose two major CL modules, difficulty measurer and training scheduler in Figure 2.

Difficulty measurer is to measure sample difficulties. To diversify learning aspects and jointly incorporate imbalance factors, we deploy three measurement tasks: 1) *Task 1 - Label F1* promotes generating clinically correct reports, 2) *Task 2 - Token F1* adjusts the balance between token infrequency and frequency, and 3) *Task 3 - BLEU-4* is to generate coherently long reports. We start with a pre-train model (e.g., Transformer (Vaswani et al., 2017)), which can perform well on easy samples (e.g., normal samples and frequent tokens). The difficulty measurer will evaluate samples' difficulties by the three metrics, label F1, token F1, and BLEU-4. We feed the difficulty information to the next step, Training Scheduler.

Training Scheduler aims to automatically leverage imbalance effects by selecting training samples via the scores of the measurers. We design our scheduler function, $c(p_t)$ as following:

$$c(p_t) = min(1, [1 - \frac{(p_t - p_{t-1})}{p_{t-1}}] \times c(p_{t-1})), t \geq 1 \quad (1)$$

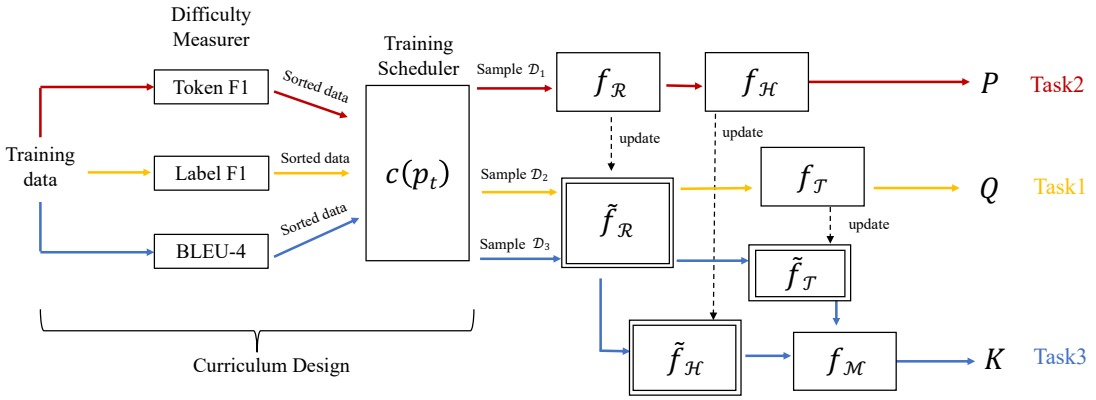
, where p is the average performance of all training samples, measuring the model's learning ability. t is the training step. Our goal is to increase the number of easier samples when the performance decreases and vice versa. Given decreasing performance as an example, $\frac{(p_t - p_{t-1})}{p_{t-1}}$ will be negative. During the process, the ratio $1 - \frac{(p_t - p_{t-1})}{p_{t-1}} > 1$ will allow the model to include more easy training data than the last step $c(p_{t-1})$. Similarly, the scheduler will also feed harder samples when increasing performance. To start our curriculum learning, we record the samples' average performance of the last two regular training epochs as p_0 and p_1, where we empirically initialize $c(p_0)$ as 1.

3.1 CL-Task 1

Our Task 1 is to exploit imbalance patterns of report labels to generate clinically accurate reports. We measure accuracy of the generated reports by comparing the predictions made by the label classifiers in Section 2 with those of the gold truths. If our generated reports are clinically correct, the label classifiers in Section 2 can accurately yield the same prediction on our generated reports as golden truth reports. In order to assess the difficulty level of each sample, we utilize F1 score, which reflects the degree of agreement between the predicted and true labels. The greater the discrepancy between the predicted and true labels indicates harder samples and vice versa. As clinical performance is a critical metric for radiology report generation, we utilize clinical error to sample data for Task 1. We expect this task helps the model leverage label imbalance, as the training scheduler can strengthen model training on the misclassified samples.

The computing process is as the following. Given a radiology image Img and the corresponding report $Z = (z_0, \ldots, z_l)$ with the length l, we extract the features from images with a visual extractor. We use ResNet101 (He et al., 2016) ($f_{\mathcal{R}}$) as our visual extractor and obtain an image's feature

Figure 2: JIMA has three tasks, $\mathbf{P}$ (e.q. 5) as token distribution prediction, $\mathbf{Q}$ (e.q. 3) as label prediction by generated reports, and $\mathbf{K}$ (e.q 8) as regular report generation. We assign one color per task and solid arrows as workflows. The dotted arrow yields new models ($\widetilde{f}$). Frames with double solid lines freeze model parameters. $f_{\mathcal{R}}$, $f_{\mathcal{H}}$, $f_{\mathcal{T}}$, $f_{\mathcal{M}}$ refer to the visual extractor in e.q. 2, token distribution predictor in e.q. 5, transformer in e.q. 8 and memory-driven model in e.q. 7, respectively.



($\mathbf{x}$) from different convolutional channels,

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S\} = f_{\mathcal{R}}(\mathrm{Img}) \quad (2)$$

Then we generate i-th token probability distribution $\mathbf{Q_i}$ from image feature $\mathbf{x}$ and contexts by our text generator ($f_{\mathcal{T}}$),

$$\mathbf{Q_i} = f_{\mathcal{T}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S, z_1, z_2, \ldots, z_{i-1}) \quad (3)$$

To optimize the model, we minimize negative log-likelihood loss (NLL) as follows,

$$\mathcal{L}_{NLL} = -\sum \log(s_i) \quad (4)$$

where s_i is the prediction probability of the i-th token.

3.2 CL-Task 2

The objective of Task 2 is to exploit token imbalance by predicting word occurrences in a given report. We utilize a multi-class binary schema to denote the tokens' occurrence and calculate the token F1 score as the difficult metrics. This approach does not count the tokens' frequency and assigns the same weight to all tokens. As a result, samples with infrequent tokens are identified as difficult and can be used by the training scheduler to enhance the model's performance in handling rare tokens.

To predict token distribution, we feed the average of feature $\mathbf{x}$ into the Token Distribution Predictor ($f_{\mathcal{H}}$) and obtain a token occurrence probability prediction ($\mathbf{P} \in R^{|V|}$),

$$\mathbf{P} = f_{\mathcal{H}}(Avg(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S\})) \quad (5)$$

where V is the vocabulary. We use a feed-forward network as our token distribution predictor since our experimental findings suggest that employing a complex network architecture does not lead to improvements in performance. Samples containing infrequent tokens are prone to obtaining lower F1 scores, and as such, the samples will be prioritized in training data repeatedly. This approach allows the model to devote more attention to learning from samples containing infrequent tokens, particularly when the model struggles to capture the underlying patterns in such tokens. Since infrequent tokens have much higher ratios of medical terms, leveraging token imbalance will be beneficial.

Task 2 is to predict the occurrence probability of a word in a report, which is a multi-classification task. Therefore, we optimize the model by multi-classification loss as follows,

$$\mathcal{L}_{BCE} = [\mathbf{y} \cdot \log \sigma(\mathbf{P}) + (1 - \mathbf{y}) \cdot \log(1 - \sigma(\mathbf{P}))] \quad (6)$$

where $\sigma(\cdot)$ is a sigmoid function. $\mathbf{y} \in R^{|V|}$ is the ground truth and $y_i \in \mathbf{y}$ is represented by,

$$y_i = \begin{cases} 1, & \text{if i-th token in a report} \\ 0, & \text{else} \end{cases}$$

We set the threshold as 0.5 to predict whether a token occurs in a report and choose F1 score as our difficulty evaluator.

3.3 CL-Task 3

Task 3 implements an image-to-text generation pipeline with the objective of enhancing the fluency of generated reports. In text generation training, the model typically predicts i-th tokens based

on 1-th to (i-1)-th tokens from the ground truth. However, these tokens and ground-truth context are not accessible during the test stage — models generate the current position token by previous predictions, which causes the accumulation error for long documents and decreases the generation fluency. To narrow the generation discrepancy between the training and test period, we calculate the BLEU-4 score generation from the beam search to measure the model's performance in the test mode. BLEU-4 score matches four consecutive tokens between prediction generation and reference reports, which can efficiently evaluate the fluency of reports. Thus, we can improve the model's generation fluency by feeding the samples with lower BLEU-4 scores into the model's learning. Also, we propose a Memory-Driven module aiming to self-adjust the current token probability distribution based on the previous predictions instead of the ground truth.

To enable the adjustment, our Memory-Driven takes two contextual inputs, the token occurrence probability prediction $\mathbf{P}$ from Task 1 and a sequence token probability distribution $\mathbf{Q}$ from Task 2. We utilize Gated Recurrent Unit (GRU) (Cho et al., 2014) as our memory-driven encoder to learn a conditional token occurrence probability prediction $\mathbf{h} \in \mathcal{R}^{l \times V}$, where l is the sequence length of a report.[2] The memory-driven model can capture the implicit relationship between a conditional token occurrence probability $\mathbf{h}$ and a sequence token prediction probability $\mathbf{Q_i}$ as follows,

$$h_i = f_{\mathcal{M}}(\mathbf{Q}_i, h_{i-1}), \tag{7}$$

Where $h_i \in \mathcal{R}^{1 \times V}$. We initialize $h_0 = \mathbf{P}$ and obtain $\mathbf{h}$ by stacking all h_i. Then, we obtain our final probability prediction $\mathbf{K}^{l \times V}$ as follows,

$$\mathbf{K} = \mathbf{Q} * sigmoid(\mathbf{h}) \tag{8}$$

This task optimize the model by e.q 4. Finally, we can obtain our generation ($\mathbf{G}$) from $\mathbf{K}$ by beam search,

$$\mathbf{G} = beam\_search(\mathbf{K})$$

To maximize report fluency with the foundation of correct clinical description, we choose BLEU-4 as our difficulty evaluator on $\mathbf{G}$ and ground truth to augment generation ability on lengthier documents.

[2]We have experimented more complex models other than GRU such as Transformer, but found GRU is the best option.

3.4 CL-Joint Optimization

We propose a joint optimization approach to integrate three tasks. Algorithm 3.4 summarizes the overall optimization process of our approach. We set the learning rate of task 2 as α and β refers to the learning rate of tasks 1 and 3. In each training step, we sample different data for different tasks and each task focuses on optimizing its own module of the models. For example, we update the visual extractor ($f_{\mathcal{R}}$) and token distribution predictor parameters $f_{\mathcal{H}}$ in task 2. Then we fix the visual extractor parameters ($f_{\mathcal{R}}$) and update transformer parameters ($f_{\mathcal{T}}$) in task 1. Finally, we combine the global token distribution $\mathcal{P}$ from task 2 and the generation $\mathcal{Q}$ from task 1 to optimize the memory-driven model ($f_{\mathcal{M}}$) in task 3.

Optimization Process of JIMA.

Require: learning rate α, β
for each epoch **do**
1. Rank entries by the three difficulty measurers (token F1, label F1 and BLEU-4);
2. Calculate three $c(p_t)$ training schedulers by e.q. 1;
3. Select top $c(p_t)$ samples from the ranked datasets obtained by step 1 as training sets;
4. Sample a batch from $\mathcal{D}_1$ and update Task 1: $\widetilde{f}_{\mathcal{T}} \leftarrow f_{\mathcal{T}} - \beta \nabla_{f_{\mathcal{T}}} \mathcal{L}_{NLL}(\mathbf{P})$;
5. Sample a batch from $\mathcal{D}_2$ and update Task 2: $\widetilde{f}_{\mathcal{R}} \leftarrow f_{\mathcal{R}} - \alpha \nabla_{f_{\mathcal{R}}} \mathcal{L}_{BCE}(\mathbf{Q})$, $\widetilde{f}_{\mathcal{H}} \leftarrow f_{\mathcal{H}} - \alpha \nabla_{f_{\mathcal{H}}} \mathcal{L}_{BCE}(\mathbf{Q})$;
6. Sample a batch in $\mathcal{D}_3$ & update Task 3: $\widetilde{f}_{\mathcal{M}} \leftarrow f_{\mathcal{M}} - \beta \nabla_{f_{\mathcal{M}}} \mathcal{L}_{NLL}(\mathbf{K})$;
end for

Our optimization approach integrates with curriculum learning to tailor joint imbalance learning for each module ($f_{\mathcal{R}}$, $f_{\mathcal{H}}$, $f_{\mathcal{T}}$, $f_{\mathcal{M}}$). Curriculum learning empowers the model to concentrate on optimizing hard samples while mitigating the risk of overfitting to easier samples. The joint optimization scheme facilitates each task to manage different module parameters optimization and learn a transferable knowledge from the simpler to more complex task. As a result, all modules collaborate to enhance error reduction from previous tasks.

Table 2: Overall performance. Δ are averaged percentage improvements over baselines.

Dataset	Model	NLG metrics						CE metrics
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	F1
IU X-ray	R2Gen	48.80	31.93	23.24	17.72	20.21	37.10	63.62
	CMN	45.53	29.50	21.47	16.53	18.99	36.78	64.83
	WCL	44.74	29.30	21.49	16.79	20.45	37.11	49.24
	CMM + RL	49.40	30.08	21.45	16.10	20.10	38.40	40.79
	RRG	49.96	31.44	22.11	17.05	18.81	33.46	49.10
	TIMER	49.34	32.49	23.84	18.61	20.38	38.25	94.52
	JIMA (Ours)	**50.50**	**33.12**	**24.15**	**18.88**	**21.16**	**38.56**	**96.58**
	Δ (%)	5.49	7.74	8.65	10.44	6.86	4.86	72.10
MIMIC-CXR	R2Gen	35.42	21.99	14.50	10.30	13.75	27.24	54.60
	CMN	35.60	21.41	14.07	9.91	14.18	27.14	50.50
	WCL	37.30	23.13	15.49	10.70	14.40	27.39	55.58
	CMM+RL	35.35	21.80	14.82	10.58	14.20	27.37	65.43
	RRG	37.57	19.78	15.87	9.56	14.77	26.81	62.20
	TIMER	38.30	22.49	14.60	10.40	14.70	28.00	75.86
	JIMA (Ours)	**40.07**	**24.83**	**15.66**	**10.99**	**15.25**	**29.05**	**78.25**
	Δ(%)	9.62	14.34	5.33	7.47	6.46	6.33	31.26

4 Experiments

We design our experiments to evaluate performance on both regular and imbalanced settings via automatic and human evaluations. The automatic evaluation includes NLG-oriented and clinical-correctness metrics. NLG-oriented metrics measure the similarity between generated and reference reports. Clinical correctness and human evaluation belong to factually-oriented metrics, and domain-specific evaluation methods. To be consistent with our baselines (Chen et al., 2020; Delbrouck et al., 2022; Wu et al., 2023), we utilize the F1 CheXbert (Smit et al., 2020) for the clinical-correctness metrics. The experiments compare our proposed approach (JIMA) and the state-of-the-art baselines. Two of our five baselines (CMM + RL & RRG) are designed to solve label imbalance by improving the abnormal findings generation. We conduct ablation and case analyses to fully understand the capabilities of our proposed approach. We include more implementation details and hyper-parameter settings in Appendix D.2.

4.1 Baselines

To examine the validity of our method, we include five state-of-the-art baselines under the same experimental settings: R2Gen (Chen et al., 2020), CMN (Chen et al., 2021), WCL (Yan et al., 2021), CMN + RL (Qin and Song, 2022), RRG (Delbrouck et al., 2022), TIMER (Wu et al., 2023) — and obtain from their open-sourced code repositories. Detailed baseline implementations are in the Appendix D.2.

4.2 Imbalance Setting

We evaluate model performance under token and label imbalance settings. For token imbalance, we compare F1-scores of frequent and infrequent tokens separately. We introduce three different scales to define frequency token sets, $1/4$, $1/6$, and $1/8$ respectively. The splits define the top $1/4$, $1/6$, and $1/8$ vocabulary as frequent tokens and the rest vocabulary as infrequent tokens. The setting is to demonstrate the effectiveness of our approach in adapting token imbalance. For label imbalance, we divide our samples into a binary category, normal and abnormal. We reuse labels from the data section and NLG metrics for evaluation.

5 Results and Analysis

In this section, we present overall performance and report results of imbalance evaluations. We conduct an ablation analysis and a case study in Appendix E. Generally, JIMA outperforms the state-of-the-art baselines by a large margin, especially under imbalance settings. Our qualitative studies show our method can achieve more clinically accuracy and generate more precisely clinical terms.

5.1 Overall Performance

Table 2 presents the performance of JIMA by NLG and clinical-correctness metrics. JIMA outperforms baseline models (both imbalance and regular methods) on BLEU scores by a large margin, confirming the validity of selecting training samples by our curriculum learning method. The approach enables the model to learn multiple times from the

Table 3: Label imbalance evaluation with binary types, normal and abnormal.

Dataset	label	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-ray	Normal	R2Gen	50.50	34.91	25.86	**20.93**	23.66	40.56
		CMN	47.42	32.80	25.25	18.72	20.51	38.69
		WCL	49.74	35.44	**28.02**	18.71	**26.88**	**42.09**
		CMM+RL	51.68	**36.65**	21.99	19.47	23.63	40.05
		RRG	50.03	33.76	24.81	19.89	20.43	34.39
		TIMER	51.83	32.43	33.71	20.19	24.43	39.39
		JIMA (ours)	**52.65**	32.14	24.97	18.26	23.73	41.72
	Abnormal	R2Gen	42.67	**27.86**	**18.47**	12.35	15.04	30.10
		CMN	35.09	21.42	14.97	11.32	14.36	29.85
		WCL	32.31	19.93	13.87	10.50	13.81	30.37
		CMM+RL	38.09	25.42	11.17	15.09	13.13	27.64
		RRG	43.38	23.44	10.02	**15.58**	12.43	31.52
		TIMER	44.25	26.73	15.28	10.76	15.43	33.26
		JIMA (ours)	**45.41**	27.25	17.85	12.37	**16.36**	**34.59**
MIMIC-CXR	Normal	R2Gen	40.42	26.76	19.75	15.60	17.58	32.02
		CMN	41.42	27.80	20.25	15.72	17.51	33.69
		WCL	39.74	25.44	18.02	13.71	16.88	32.09
		CMM+RL	17.50	10.11	6.83	14.99	8.05	19.10
		RRG	38.78	21.63	18.04	12.09	18.27	27.56
		TIMER	40.33	27.53	19.88	14.87	17.47	33.08
		JIMA (ours)	**41.79**	**27.87**	**20.49**	**16.00**	**17.93**	**33.87**
	Abnormal	R2Gen	33.97	19.31	12.07	**10.97**	10.98	26.82
		CMN	33.00	19.44	10.02	8.73	10.21	25.16
		WCL	34.56	22.45	14.63	10.26	12.43	26.87
		CMM+RL	27.74	10.87	5.18	3.43	6.11	16.08
		RRG	17.47	9.71	5.78	3.74	8.37	17.59
		TIMER	35.66	21.83	14.25	14.87	9.84	26.77
		JIMA (ours)	**37.81**	**22.46**	**15.26**	10.28	**14.56**	**27.38**

Table 4: Results on high- and low-frequent tokens with three different ratio splits.

		IU X-ray		MIMIC-CXR	
Ratio	Method	infreq	freq	infreq	freq
1/8	R2GEN	4.46	**62.73**	2.52	52.01
	CMN	5.88	55.86	2.23	45.60
	WCL	5.29	60.23	2.91	48.60
	CMN + RL	5.19	49.36	0.21	23.64
	RRG	7.28	41.94	2.50	43.57
	TIMER	13.23	61.89	3.15	52.66
	JIMA (ours)	**14.87**	62.55	**3.58**	**53.06**
1/6	R2GEN	2.80	61.62	2.02	49.86
	CMN	5.75	65.12	0.85	52.02
	WCL	3.72	59.26	2.13	47.88
	CMN + RL	5.19	49.36	0.14	23.36
	RRG	4.55	40.46	2.09	43.56
	TIMER	5.93	67.79	2.02	51.72
	JIMA (ours)	**10.52**	**68.82**	**2.83**	**52.32**
1/4	R2GEN	1.16	59.98	0.00	48.77
	CMN	2.60	63.92	0.33	51.09
	WCL	1.50	56.83	0.30	46.95
	CMN + RL	5.19	49.36	0.07	23.05
	RRG	2.04	38.84	0.39	41.45
	TIMER	8.66	64.00	0.58	51.39
	JIMA (ours)	**9.77**	**66.23**	**0.94**	**51.92**

samples with lower BLEU-4, resulting in a better performance compared to the baseline models. For example, JIMA shows an improvement of 6.84% on average for IU X-ray and 7.10% for MIMIC-CXR. We infer this is as our task 3 improves generated sentence' fluency leading to the improvement of BLEU-(1-4) and ROUGE-L metrics.

Second, Our model achieves the best performance in F1 of the clinical metric. The results clearly indicates the effectiveness of Task 1 (Section 3.1) can enable the model to put more attention on difficult samples with lower F1 scores. Additionally, our method promotes clinical token prediction as performance on infrequent tokens and medical terms have been improved. For example, our generation significantly outperforms the baselines on F1 score by 21.69% on IU X-ray and 17.73% on the MIMIC-CXR average. CMN + RL performs better than other baselines on IU X-ray but not on MIMIC-CXR. In contrast, JIMA maintains a stable performance on both IU X-ray and MIMIC-CXR. We infer this as our joint imbalance adaptation has more improvements than label imbalance adaptation, which has consistent observations with our ablation analysis (Section 5.4).

Table 5: Ablation analysis.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CE - F1
IU X-ray	-task1	**50.65**	31.64	23.55	17.89	20.01	36.92	92.28
	-task2	48.82	31.28	23.04	18.05	**21.56**	37.86	96.82
	-task3	50.44	32.56	23.12	16.59	18.26	32.44	**97.73**
	full	50.50	**33.12**	**24.15**	**18.88**	21.16	**38.56**	96.58
MIMIC-CXR	-task1	37.66	23.30	15.34	10.58	15.08	27.29	69.28
	-task2	34.58	22.58	15.64	10.79	14.89	27.43	68.79
	-task3	37.77	22.86	15.32	9.44	14.78	26.89	69.56
	full	**40.07**	**24.83**	**15.66**	**10.99**	**15.25**	**29.05**	**78.25**

5.2 Token Imbalance

Our method consistently outperforms baselines in the low-frequent tokens across frequency splits ($\frac{1}{4}$, $\frac{1}{6}$, and $\frac{1}{8}$) on IU X-ray and MIMIC-CXR. While RRG and CMN + RL approaches have adapted label imbalance, the approaches may not be able to adapt the token imbalance. Our approach achieves better performance on the token imbalance.

Generating rare tokens with accuracy remains a difficult task despite the high performance achieved on frequent tokens. Common tokens are prone to overfitting while rare tokens are predicted with less precision. For example, the 0.00 score by R2GEN on $3/4$ split of the MIMIC-CXR vocabulary. Performance imbalance can deteriorate the clinical correctness of generated reports as medical terminologies are usually infrequent. Nonetheless, our joint imbalance adaptation approach has shown considerable improvements in this area, indicating a promising direction to enhance the robustness of radiology report generation, a critical clinical task.

5.3 Label Imbalance

We report NLG evaluations on label imbalance (normal vs. abnormal) in Table 3. JIMA significantly outperforms baseline models both on normal and abnormal splits, which demonstrates its effectiveness under label imbalance. JIMA also performs better than the label imbalance methods, RRG and CMM+RL, indicating that the joint imbalance adaptation is a promising direction to improve model robustness. It is worth noting that models generally perform better on normal samples than on abnormal ones. We infer this for two reasons: 1) abnormal reports contain more infrequent medical tokens, and 2) abnormal reports are longer, as discussed in Section 2. JIMA shows more improvements on abnormal samples over baselines while maintains a similar performance on samples with normal labels. The observations suggest that our approach can successfully learn from lengthier documents with more medical tokens.

5.4 Ablation Analysis

To measure each task's contribution, we report ablation analysis in Table 5. Overall, our full model performs best in most evaluations. However, when we remove Task 1, there is a significant decrease in clinical metrics (F1). This task is crucial as it allows JIMA to focus on learning from samples with low clinical correctness. Without Task 1, the model treats all samples equally, resulting in failure to capture useful features from complex samples. Furthermore, removing Task 2 leads to a decrease in BLEU-1, as this task augments samples with low token F1 scores. We infer that Task 2 is highly relevant to BLEU-1 since they both measure single token accuracy. This task optimizes the model by promoting infrequent tokens through e.q. 1, making it important in improving the BLEU-1 score. Similarly, removing Task 3 results in a decline in BLEU-4 score, which indicates that JIMA can reduce the generation discrepancy between the training and test period. These results demonstrate strong evidence that our proposed joint adaptation approach can effectively learn from multiple imbalance settings.

6 Conclusion

In this study, we have illustrated the critical challenges of label and token imbalance. We proposed a curriculum learning-based model to jointly adapt label and token imbalance. Our diverse analysis can demonstrate the effectiveness of our approach (JIMA) on radiology report generation. Extensive experiments and ablation analysis show that JIMA leads to significant improvements in handling token and label imbalance. Appendix presents implementations, data analysis, and results to allow for full replication.

7 Limitations

Limitations should be fully acknowledged before fully interpreting this study, as no research can be fully perfect. Our study conducts experiments on English data without *multilingual* coverage. We expect to extend our study to other languages in the future when we have publically available datasets. However, releasing and accessing new clinical data can face privacy and ethical challenges as we also discuss in our Appendix. The second challenge is the *large-scaled human evaluation*. Our study invited an expert from a medical institution. Having annotations from one expert may face subjective effects. However, limited fund prevents us to scale our human evaluations. For example, the last author requested evaluations from multiple clinicians, while most of them said they were "very busy". We expect to extend our human evaluations in our future work. Finally, we are also aware of *other evaluation metrics*, such as RadGraph (Jain et al., 2021) and CheXpert (Irvin et al., 2019). However, additional metrics may only be applicable to the MIMIC-CXR or have overlapped with our existing method, such as CheXpert and CheXbert (Smit et al., 2020). We have included diverse metrics, including NLG, clinical correctness, and human evaluations. To keep consistency with our state-of-the-art baselines, we utilize a similar evaluation schema. Having consistent observations between our human and automatic evaluations may also prove our evaluation validity.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

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 (Volume 1: Long Papers)*, pages 5904–5914, Online. 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 (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Pritam Deka, Anna Jurek-Loughrey, et al. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

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.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the RadSum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482, Toronto, Canada. 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. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

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.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, Online. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. 2023. Multimodal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy. Association for Computational Linguistics.

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

Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Jadhav. 2023. Replace and report: NLP assisted radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10731–10742, Toronto, Canada. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Fenglin Liu, Shen Ge, and Xian Wu. 2021. 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 (Volume 1: Long Papers)*, pages 3001–3012, Online. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. 2020. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2223–2236, Online. 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, volume 32, pages 8024–8035. Curran Associates.

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

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

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.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. 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.

Yuexin Wu, I-Chan Huang, and Xiaolei Huang. 2023. Token imbalance adaptation for radiology report generation. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 72–85. PMLR.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-nan Hsu. 2021. Weakly supervised contrastive learning for chest x-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015.

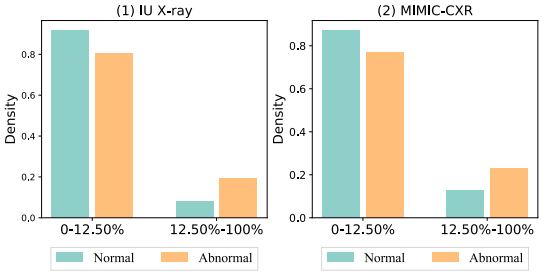
H. Yu and Q. Zhang. 2022. Clinically coherent radiology report generation with imbalanced chest x-rays. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1781–1786, Los Alamitos, CA, USA. IEEE Computer Society.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. 2020. Curriculum learning by dynamic instance hardness. In *Advances in Neural Information Processing Systems*, volume 33, pages 8602–8613. Curran Associates, Inc.

A Data

Figure 3: Frequent and infrequent token distributions conditioning on report label.



We extract labels of each data entry and follow baseline studies (Chen et al., 2020, 2021; Qin and Song, 2022) to preprocess the report documents to ensure comparisons under same settings. In order to ensure data format consistency, we include and infer two primary labels of radiology reports, normality and abnormality. To obtain labels for IU X-ray, we build a supervised classifier using

BioBert-PubMed200kRCT (Deka et al., 2022) to extract the binary labels on the Medical Subject Heading (MESH)[3] and RadLex[4] labels (normal and abnormal). To obtain labels for MIMI-CXR, we utilize CheXbert (Smit et al., 2020) to extract the binary categories, disease types and "no finding". We define "no finding" as normality and disease types as abnormality. In this study, we conducted text preprocessing by utilizing the Natural Language Toolkit (NLTK) (Loper and Bird, 2002) to lowercase and tokenize documents. Furthermore, we removed redundant spaces, empty lines, serial numbers, and punctuation marks from the documents. We visualize the distributions of frequent (ranked in the top 12.5% of the vocabulary) and infrequent tokens in Figure 3.

B Ethic, Privacy, and IRB

We follow data agreement and training to access the two radiology report datasets. To protect user privacy, we ensure proper data usage and experiment with de-identified data. Our experiments do not store any data and only use available multimodal entries for research demonstrations. Due to privacy and ethical considerations, we will not release any clinical data associated with patient identities. Instead, we will release our code and provide detailed instructions to replicate our study. This study only uses publicly available and de-identified data. Our study focuses on computational approaches and does not collect data from human subjects. Our institutional IRB determines that IRB approval is not required for this study.

C Related Work

Radiology report generation is a domain-specific image-to-text task that has two major directions, retrieval- (Endo et al., 2021; Jeong et al., 2023) and generation-based (Chen et al., 2020; Qin and Song, 2022; Kale et al., 2023). The retrieval-based approach compares similarities between an input radiology image and a set of report candidates, ranks the candidates, and returns the most similar one (Liu et al., 2021; Endo et al., 2021; Jeong et al., 2023; Wang et al., 2023; Delbrouck et al., 2023). In contrast, our study focuses on the generation-based task, which automatically generates a precise report from an input image. The task has domain-specific characteristics in the clinical field. The clinical data contains many infrequent medical terminologies and longer documents than image captioning from general domains (Lin et al., 2014). As radiology report generation can reduce the workloads of radiologists, generating highly qualified and precise can be a critical challenge, especially under the imbalance settings. Differing from previous work, we aim to promote model robustness and reliability under imbalance settings, which have been rarely studied in the radiology report generation.

Imbalance learning aims to model skewed data distributions. The primary focus of imbalance learning is on class or label imbalance, such as positive or negative reviews in sentiment analysis (Li et al., 2022). While previous studies proposed new objective functions (e.g., focal-loss (Lin et al., 2020)) or oversampling (Chawla et al., 2002), those methods may not be applicable to our primary generation unit, token, which has large vocabulary sizes and extreme sparsity. In terms of radiology report generation, reports may have disease-related labels. Recent studies have augmented model robustness by balancing performance between disease and normal by reinforcement learning (Nishino et al., 2020; Yu and Zhang, 2022). However, those methods ignore a fundamental challenge of generation task, token imbalance – a long-tail distribution. The token imbalance can be even more critical for the clinical domain, as medical tokens appear less frequently than regular tokens in radiology reports. Our study makes *a unique contribution* to the radiology report generation that jointly incorporates token and label imbalance via curriculum learning.

D Experiment

D.1 Baselines

R2Gen (Chen et al., 2020) is a transformer-based model with ResNet101 (He et al., 2016) as the visual extractor. To capture some patterns in medical reports, R2Gen proposes a relational memory to enhance the transformer so that the model can learn from the patterns' characteristics. Furthermore, R2Gen deploys a memory-driven conditional layer normalization to the transformer decoder facilitating incorporating the previous step generation into the current step.

CMN (Chen et al., 2021) is a novel extension to the transformer architecture that facilitates the alignment of textual and visual modalities. The

[3] https://www.nlm.nih.gov/mesh/meshhome.html
[4] https://radlex.org/

cross-modal memory network record the shared information of visual and textual features. The alignment process is carried out via memory querying and responding. The model maps the visual and textual features into the same representation space in memory querying and learns a weighted representation of these features in memory responding.

WCL (Yan et al., 2021) utilizes the R2Gen framework and incorporates a weakly supervised contrastive loss. Specifically, WCL leverages the contrastive loss to enhance the similarity between a given source image and its corresponding target sequence. Furthermore, the model enhances its ability to learn from difficult samples by assigning more weights to instances sharing common labels.

CMM + RL (Qin and Song, 2022) is a cross-modal memory-based model with reinforcement learning for optimization. CMM + RL designs a cross-modal memory model to align the visual and textual features and deploy reinforcement learning to capture the label imbalance between abnormality and normality. The author uses BLEU-4 as a reward to guide the model to generate the next word from the image and previous words.

RRG (Delbrouck et al., 2022, 2023) aims to generate clinically correct reports by weakly-supervised learning of the entities and relations from reports. RRG is a BERT-based model with Densenet-121 (Huang et al., 2017) as a visual extractor. RRG leverages RadGraph (Jain et al., 2021) to extract the entities and relation labels in a report. RRG utilizes reinforcement learning to optimize the model. The reward assesses the consistency and completeness of entities and the relation set between generated reports and reference radiology reports. RRG addresses label imbalance issues by maximizing the reward of predicting more complicated entities and relations in abnormal samples.

TIMER (Wu et al., 2023) aims to decrease the over-fitting of frequent tokens by introducing unlikelihood loss to punish the error on these tokens. The tokens set of unlikelihood loss is dynamically adjusted by maximizing the average F1 score on different frequency tokens.

D.2 Implementation Details

In our model architecture, we set the transformer structure with 3 layers and 8 attention heads, 512 dimensions for hidden states. The memory-driven model is a single-layer GRU network with a hidden size equal to vocabulary size. We set the α learning rate as $4e-4$ and β learning rate as $1e-5$ and decay them by a 0.8 rate per epoch for all datasets. The pre-training epoch is 30 in IU X-ray and 10 in MIMIC-CXR. Then we adopt curriculum learning to optimize our pre-trained model. The maximum training epoch is 70 for the IU X-ray and 50 for the MIMIC-CXR datasets. We keep the learning rate the same as in the pre-trained stage.

For all baselines, we set the maximum training epoch as 100 and 60 for IU X-ray and the MIMIC-CXR datasets, respectively. Also, we use the same preprocessing, optimizer, batch size, maximum length of training data, sampling method, and machine learning framework in all experiments. Specifically, we optimize models by ADAM (Kingma and Ba, 2015) with 16 batch sizes. The maximum length of training data is 60. In the test stage, we generate tokens by beam search (Sutskever et al., 2014) with 3 beam sizes for all experiments. All implementations are on PyTorch (Paszke et al., 2019). In implementing baselines, we keep all the model architecture and optimization parameters the same as in their papers. In R2Gen, CMN, and RRG, we generate reports by using the code and the pre-trained models published by the authors. For the other baselines (WCL & CMM+RL & TIMER), we use the released code to train and generate reports.

We personalize the following setting in baselines. In WCL, we use the basic contrastive learning loss without assigning a hardness weight to different samples in IU X-ray dataset. Because the file measuring the similarity among different samples is inaccessible. We set the contrastive embedding size as 256 and the weight of contrastive loss is 0.2. In CMM + RL, the reinforcement learning reward is based on evaluation metrics and we select BLEU-4 in this case.

D.3 Evaluation Metrics

Automatic Evaluation includes seven evaluation methods from two major categories, *NLG* and *Clinical metrics*. We first evaluate our model and the baseline models on *natural language generation (NLG) metrics*, including BLEU (-1, -2, -3, and -4) (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and ROUGE-L (Lin, 2004). BLEU score measures the precision of prediction with a penalty for the reference-to-prediction length ratio. METEOR computes the harmonic mean of unigram precision and recall. Unlike BLEU, which considers only single words, METEOR incorporates a penalty to account for the importance of

word order. ROUGE-L takes into account sentence-level structure similarity naturally and identifies the longest co-occurring in sequence n-grams automatically. *Clinical metrics* is a domain-specific evaluation method to measure the factual completeness and consistency of generated reports. We use CheXbert (Smit et al., 2020) to extract the labels of ground truth and prediction and evaluate clinical efficacy (CE) metrics by F1. We do not present clinical F1 score in the label imbalance experiment since we can not access recall in separate normal and abnormal sample sets.

Human Evaluation To verify the factual correctness, we invite a radiological professional from a medical institution to perform evaluation using Google Forms. First, we randomly select 50 test instances per dataset from IU X-ray and MIMIC-CXR respectively. We choose CMM+RL as our targeting comparison, as the model achieves comparatively better performance than other baselines by automatic metrics. In evaluation, we show the X-ray images, corresponding ground truth reports, and two generated reports (one from our model and the other from CMM+RL) to the expert without disclosing their sources. The expert selects a better description from two candidate reports or chooses the "Same" option if both reports are of similar quality.

E Result Analysis

E.1 Human Evaluation

We present our human evaluation results in Table 6, which shows a consistent result with automatic evaluation results. Generally, JIMA outperforms the baseline with 11 reports in total. Notably, our approach exhibits significant improvements in abnormal samples. Even though JIMA has only one more vote than the baseline in normal samples, our model secures ten more votes in abnormal samples. This is because abnormal samples have lengthier reports on average and encompass more medical entities, indicating that our approach generates more clinically precise reports. Furthermore, our human evaluation is consistent with the automated evaluation results shown in Table 2.

E.2 Case Study

To verify our model's effectiveness in generating clinically correct descriptions, we perform a case study in this section and present the result in Fig 4.

Table 6: Human evaluation. "Same" means two generated reports have the same quality by the clinician.

Dataset	Label	CMM+RL	Same	JIMA (Ours)
IU X-ray	Normal	6	12	6
	Abnormal	4	10	**12**
MIMIC-CXR	Normal	6	15	**7**
	Abnormal	5	10	**7**
Overall	Normal	12	27	**13**
	Abnormal	9	20	**19**
	All	21	47	**32**

We select four samples from IU X-ray and MIMIC-CXR datasets and compare the normal and abnormal samples' performance separately. The correct pathological and anatomical entity predictions are remarked in blue color. Generally, our predictions cover more than 90% entities in reference reports. Compared to normal samples, abnormal samples have longer descriptions and contain more complex entities. These entities usually are rare in corpus and suffer under-fitting from models. Therefore, models underperform in abnormal samples. However, JIMA can capture most of the entities in all kinds of samples and achieve similar performance in both normal and abnormal samples, which proves our model's effectiveness in improving the factual completeness and correctness of generated radiology reports.

Figure 4: Qualitative comparison between JIMA and CMM+RL. We highlight correct predictions of pathological and anatomical entities in blue color.

Label	Baseline	JIMA	Ground Truth
Normal	the lungs are clear . the cardiomediastinal silhouette is within normal limits . no acute osseous abnormalities .	in comparison with the study of from an outside institution there is little change . cardiac silhouette is within normal limits and there is no evidence of acute pneumonia vascular congestion or pleural effusion .	in comparison with the study of there is little change and no evidence of acute cardiopulmonary disease . no pneumonia vascular congestion or pleural effusion .
	the cardiomediastinal silhouette and pulmonary vascularity are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen .	the lungs and pleural spaces show no acute abnormality . heart size and pulmonary vascularity within normal limits .
Abnormal	as compared to the previous radiograph there is no relevant change . extensive emphysematous lung parenchymal destruction in both upper lobes right more than left . subsequent distortion of vascular and airway structures at the lung bases . no pulmonary edema . no pneumonia . borderline size of the cardiac silhouette .	as compared to the previous radiograph there is no relevant change . low lung volumes with areas of atelectasis at both lung bases . no new parenchymal opacities . no larger pleural effusions . no pneumothorax .	as compared to the previous radiograph there is no relevant change . low lung volumes with minimal atelectasis at the lung bases . no evidence of pneumonia . no pulmonary edema . no pleural effusions . normal size of the cardiac silhouette .
	dual lead left-sided pacemaker is stable in position with leads extending to the expected positions of the right atrium and right ventricle . the patient is status post median sternotomy . there is minimal left base atelectasis . no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are stable . no displaced fracture is seen .	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 . the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are stable .	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 . no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are unremarkable .