

Escaping the Frequent Finding Shortcut: Pathology-Aware Curriculum Learning for Radiology Report Generation

Ziming Song^{1,2*} Mingyang Shen^{1,2*} Chi Man VONG^{1†}

¹University of Macau

²Shenzhen University of Advanced Technology

ziming.song.tech@gmail.com, mingyangshen02@gmail.com, cmvong@um.edu.mo

Abstract

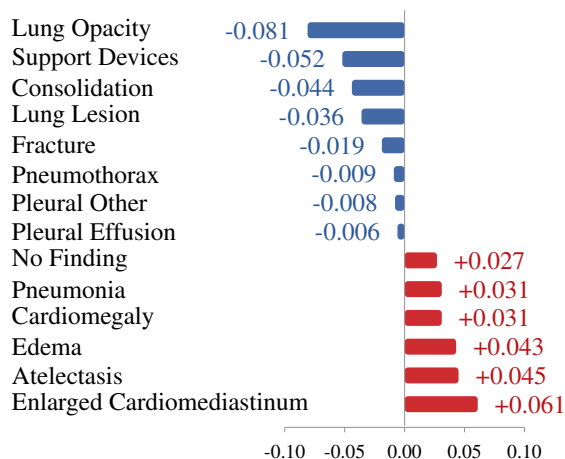
Radiology Report Generation (RRG) has benefited from stronger encoder-decoder pipelines and particularly Large Language Model (LLM) report generators, but most methods still perform training through random mini-batches. We argue that random sampling pushes report generators toward a frequent-finding shortcut, where models quickly learn fluent report narratives and high-frequency disease phrases, then receive limited pressure to recover rare attributes that contribute little to aggregate cross-entropy. To address this problem, we propose PARAC, a pathology-aware replay curriculum for long-tail RRG. PARAC ranks training reports with pathology labels and structured medical knowledge, expands the training process from simple reports to complex long-tail cases, and replays prototype anchors to preserve basic cross-modal grounding. Extensive experiments on IU X-RAY and MIMIC-CXR show that PARAC consistently improves several existing methods, with dedicated diagnostics and ablations attributing the gains to better coverage of low-frequency pathological findings.

1 Introduction

Radiology report generation (RRG) aims to translate chest radiographs into clinically faithful free-text findings. Unlike generic image captioning, a clinical report is normally long in context and usually covers multiple pathological attributes. This domain-specific feature requires automatic RRG systems to generate both lexically coherent and clinically accurate reports. Recent studies improve report fluency by moving from task-specific encoder-decoder models (Jing et al., 2018; Chen et al., 2020, 2021; Liu et al., 2021b) to LLM-enhanced pipelines (Wang et al., 2023b; Liu et al., 2024; Wang et al., 2025), but they still inherit

*Equal contribution.

†Corresponding author.



(a) Label-prevalence gap with the test set of MIMIC-CXR.

Diagnostic slice	Metric	Value
All CheXbert labels	CE F1	58.8
Common labels	Recall	60.4
Rare pathology labels	Recall	14.9
Rare pathology labels	Omission rate	85.1

(b) Analysis of high-frequency and rare findings.

Figure 1: Statistical analysis of (a) CHEXBERT (Smit et al., 2020) labels and (b) metrics of the generated reports by LLM-RG4 (Wang et al., 2025), evaluated between the generated and ground-truth reports on MIMIC-CXR (Johnson et al., 2019).

the highly imbalanced report distribution of available training data in IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). In public chest radiology datasets, e.g., IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019), a large fraction of reports contain patternized descriptions, common support-device statements, and frequent findings, whereas uncommon positive or uncertain pathological attributes appear sparsely. This imbalance creates an optimization problem, where a model can reduce token-level loss by fitting frequent report patterns while still omitting low-frequency attributes.

Fig. 1 presents the detailed statistical analysis

on MIMIC-CXR, where we choose a representative LLM-based method LLM-RG4 (Wang et al., 2025) for evaluation. Results in Fig. 1(a) validate our claim of label imbalance, where LLM-RG4 tends to over-produce frequent attributes such as enlarged cardiomeastinum, atelectasis, and edema, while under-generating long-tail labels including lung opacity, consolidation, and lung lesion. This observation gets more significant when evaluating its reports under settings of all labels or the rare ones. While LLM-RG4 reaches an outstanding recall score of 60.4 on common labels, its performance under rare pathological labels is only 14.9, with the omission rate reaching up to 85.1%.¹ These analyses explain why high lexical scores measured by BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) and competitive clinical metrics do not always lead to reliable reports, as a model might hack these scores by learning to produce frequent phrases while still missing the rare attributes that distinguish hard cases. We regard this as a local optimum induced by random sampling and cross-entropy. An expected direction is curriculum learning (Bengio et al., 2009), which orders examples from easy to hard so that the model first stabilizes basic radiograph-to-report mappings, and then receives denser updates on complex long-tail cases. This is especially suitable to address the issue of long-tail attributes, as it can prevent hard reports from being diluted by high-frequency templates throughout training. Notably, the closest prior method based on this paradigm is CMCL (Liu et al., 2021a), where it shows that curriculum learning indeed helps report generation, yet its difficulty signal does not align well with pathology imbalance, leaving the issue still open to solve.

Therefore in this paper, we propose PARAC, a **Pathology-Aware Replay Curriculum** for RRG. Specifically to address the long-tail attribute issue, our method considers two granularities of in-domain auxiliary to offer balance training, including medical labels and fine-grained structured medical knowledge. In details, these two aspects offer complementary information to build consistent difficulty metric for the curriculum learning process, where the former measures report-level pathology polarity and uncertainty; and the latter builds meticulous entity-relation structure. For op-

timization, we discard the conventional way of conducting randomly sampled training, and instead, we arrange the optimization process as an easy-to-hard sequence, which allows the model to first learn basic cross-modal mappings and then progressively converge to handle long-tail scenarios with more complex pathological complexity. Additionally in this method, we introduce Prototype Anchor Replaying (PAR) to leverage a set of anchor samples, which stabilizes the overall training from catastrophic forgetting when proceeding to harder cases. Extensive experiments on two widely-adopted benchmarks, i.e., IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019), show that our design improves multiple existing methods from different categories. Further ablations offer more insights on how the curriculum learning process resolves this issue.

2 Related Work

Radiology report generation. The development of RRG can be broadly divided into the task-specific encoder-decoder and LLM-adapted methods. Early encoder-decoder systems mainly adapt image captioning models to chest radiographs by pairing convolutional visual features with recurrent or attention-based text decoders (Vinyals et al., 2015; Rennie et al., 2017; Lu et al., 2017; Jing et al., 2018). Subsequent works enhance the clinical fidelity of this pipeline, through knowledge graph (Li et al., 2019; Zhang et al., 2020), memory mechanisms (Chen et al., 2020, 2021), contrastive learning (Liu et al., 2021c), posterior-prior knowledge (Liu et al., 2021b; Jin et al., 2024; Bu et al., 2024), enhanced Transformer architectures (Huang et al., 2023; Wang et al., 2023a), region-level grounding (Tanida et al., 2023). Afterwards, RRG systems increasingly replace task-specific decoders with pretrained language generators. This line adapts LLMs through medical instruction tuning and visual connector learning (Li et al., 2023; Wang et al., 2023b; Thawakar et al., 2024; Chen et al., 2024; Tu et al., 2023; Hyland et al., 2023; Bannur et al., 2024), where later studies leverage bootstrapped (Liu et al., 2024) or context-augmented supervision (Wang et al., 2025; Liu et al., 2025b; Gu et al., 2025; Xiao et al., 2025b,a; Liu et al., 2025a; Gao et al., 2026), and efficient generation strategies for long clinical reports (Chen et al., 2026). These studies mainly improve model capacity or alignment objectives, whereas our work examines a funda-

¹See Appendix B for the rare and high-frequency label definitions.

mental failure that neither of them addresses, i.e., the long-tail attribute issue in RRG.

Curriculum learning. This typical technique studies how the order of training examples affects convergence and generalization. Classical curriculum learning (Bengio et al., 2009), self-paced learning (Kumar et al., 2010), and self-paced curriculum learning (Jiang et al., 2015) show that easier examples can provide a smoother optimization path before harder examples dominate training. Despite to it, long-tail learning also rebalances optimization from different perspectives, e.g., Focal loss (Lin et al., 2017), class-balanced loss (Cui et al., 2019), and LDAM (Cao et al., 2019) adjust example contributions or classification margins; Adaptive Gradient Gating (Yu et al., 2022) modifies rare-token gradients; and Rebalanced Contrastive Loss (Alvis et al., 2025) requires a supervised contrastive objective. Specifically for the RRG task, curriculum learning has been used to organize examples according to cross-modal matching difficulty and report semantics (Liu et al., 2021a), and to jointly adapt label and token imbalance through a hard-to-easy multi-task schedule (Li et al., 2025).

3 Method

Fig. 2 presents the overall pipeline of PARAC, consisting of three main components, i.e., Clinical Difficulty Estimation (CDE), Long-Tail Curriculum Scheduling (LTCS), and Prototype Anchor Replay-ing (PAR). Specifically, CDE annotates the difficulty metric for each report in the training set based on its pathology distribution and structured clinical complexity. LTCS serves as the main curriculum learning process, which replaces conventional random sampling with a re-ordered schedule that expands from simple reports to complex long-tail cases. PAR is a vital component to stabilize the curriculum process, where it replays prototype anchors throughout training to preserve basic cross-modal grounding. In the following sections, we first illustrate the preliminaries of the existing RRG pipeline, then elaborate the details of CDE, LTCS, and PAR.

3.1 Preliminaries

In existing methods, most RRG methods are fundamentally constructed as encoder-decoder pipelines. Specifically, they normally contain a visual encoder $\mathcal{E}_{\text{vis}}(\cdot)$ and a report decoder $\mathcal{D}_{\text{report}}(\cdot)$, where the former encode the input radiographs into image features and the latter convert them to explicit re-

port tokens. Here, the decoder parameterized by θ can be either a task-specific Transformer decoder (Vaswani et al., 2017; Chen et al., 2020, 2021), or an LLM-based decoder adapted from pre-trained general LLMs (Wang et al., 2023b, 2025).

Given the training set $\mathcal{D}$, standard training on the aforementioned pipelines randomly samples mini-batches from $\mathcal{D}$, and optimizes the model along the resulting training order. We denote this order by $\mathcal{O}_{\text{rand}} = \{\mathcal{B}_s^{\text{rand}}\}_{s=1}^S$, which is a sequence of S sampled mini-batches. Under this order, the model minimizes the autoregressive cross-entropy:

$$\ell_{\theta}(i) = - \sum_{t=1}^{t_i} \log p_{\theta}(y_{i,t} | y_{i,<t}, \mathcal{Z}_i), \quad (1)$$

Here, $\ell_{\theta}(i)$ denotes the autoregressive token loss of example i ; $\mathcal{Z}_i$ refers to the input context, e.g., the radiology image, the LLM prompt, etc. This objective is statistically simple but clinically biased. If common normality statements and frequent findings dominate $\mathcal{D}$, Eq. (1) gives them more optimization mass than rare positive or uncertain attributes. The model thus reaches a relative low loss state by cutting the shortcut and fitting high-frequency disease phrases, while rare attribute descriptions remain under-optimized, as is analyzed in Fig. 1.

To resolve this issue, our method draw inspiration from curriculum learning (Bengio et al., 2009), in order to offer an easy-to-hard training sequence for the RRG model and allow it reach better convergence to cover the long-tail cases. Before illustrating our details, we elaborate the preliminaries of CL as follows. Specifically, CL replaces the random order with an easy-to-hard one induced by a difficulty metric d . Formally, it assigns the i -th training example a difficulty score d_i and sorts the dataset by d_i , where this process is written as:

$$\begin{aligned} d_i &= d(\mathcal{X}_i, \mathcal{Y}_i), \\ \mathcal{O}_d &= \text{argsort}_{i \in \{1, \dots, n\}}(d_i). \end{aligned} \quad (2)$$

Here, smaller values of d_i indicate that the i -th sample appears earlier in the sequence; and $\mathcal{O}_d$ denotes the easy-to-hard curriculum order that replaces $\mathcal{O}_{\text{rand}}$. The key to address the long-tail issue is how to design d , so that it can serve as an effective indicator of pathological complexity, where we illustrate this in the next section.

3.2 Clinical Difficulty Estimation (CDE)

In this section, we illustrate how we design d_i in Eq. (2) via CDE. To better cover long-tail cases in

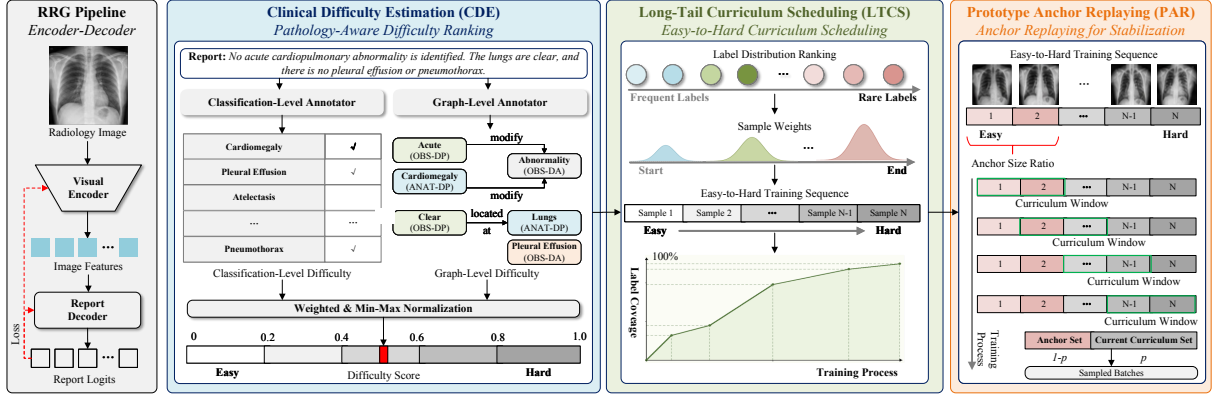


Figure 2: Overview of PARAC for radiology report generation, including the encoder-decoder backbone, Clinical Difficulty Estimation (CDE), Long-Tail Curriculum Scheduling (LTCS), and Prototype Anchor Replaying (PAR).

RRG, the difficulty score d_i should jointly reflect the overall complexity of a free-text report and assign higher difficulty to reports that describe rarer pathological attributes. Specifically, we first use a medical classifier to extract report-level discrete pathology labels. At the classifier level, each report $\mathcal{Y}_i$ is converted into a discrete pathology vector $c_i = [c_{i,1}, \dots, c_{i,m}] \in \{0, 1, u\}^m$ over m clinical attributes, where 1 denotes positive, u denotes uncertain, and 0 denotes negative or unmentioned. Let n_ℓ be the number of training reports that contain a positive or uncertain state for attribute ℓ . For each graph-level label g , let n_g be the number of reports that contain this label after graph canonicalization. We define inverse-frequency weights in a compact form as $a_\ell = \log((n + \epsilon)/(n_\ell + \epsilon))$ and $r_g = \log((n + \epsilon)/(n_g + \epsilon))$, with ϵ as a smoothing constant². The final difficulty metric is composed of both classifier- and graph-level scores, where the classifier-level difficulty of report i is:

$$d_i^{\text{cls}} = \sum_{\ell=1}^m a_\ell [\mathbb{I}(c_{i,\ell} = 1) + \lambda_{\text{unc}} \mathbb{I}(c_{i,\ell} = u)]. \quad (3)$$

This term increases when a report contains more positive or uncertain findings, especially when these findings correspond to rare attributes.

The classifier-level difficulty still lacks some critical task-specific features, e.g., entity spans, anatomical grounding, or modifier relations. Therefore, we also apply a knowledge graph annotator to parse $\mathcal{Y}_i$ into a graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$. Each node in $\mathcal{V}_i$ represents a clinical entity, and each directed edge in $\mathcal{E}_i$ is a relation between two entities. We further canonicalize graph-level labels as $\Gamma(\mathcal{G}_i) = \{g_{i,k}\}_{k=1}^{K_i}$, where each key encodes a fine-

grained clinical pattern. The graph-level difficulty is:

$$d_i^{\text{str}} = w_{\text{obs}} |\mathcal{V}_i^{\text{obs}}| + w_{\text{anat}} |\mathcal{V}_i^{\text{anat}}| + w_{\text{rel}} |\mathcal{E}_i| + w_{\text{rare}} \sum_{g \in \Gamma(\mathcal{G}_i)} r_g. \quad (4)$$

Here, $\mathcal{V}_i^{\text{obs}}$ and $\mathcal{V}_i^{\text{anat}}$ denote observation and anatomy entities; $\Gamma(\mathcal{G}_i)$ denotes the set of canonicalized graph keys; and r_g denotes the inverse-frequency weight of graph key g . Generally, we define the final difficulty metric as:

$$\begin{aligned} \tilde{d}_i &= \beta_{\text{cls}} d_i^{\text{cls}} + \beta_{\text{str}} d_i^{\text{str}}, \\ d_i &= \frac{\tilde{d}_i - \min_{j \in \{1, \dots, n\}} \tilde{d}_j}{\max_{j \in \{1, \dots, n\}} \tilde{d}_j - \min_{j \in \{1, \dots, n\}} \tilde{d}_j + \epsilon}, \\ q_i &= \frac{1}{n} \sum_{j=1}^n \mathbb{I}(d_j \leq d_i). \end{aligned} \quad (5)$$

Here, $\tilde{d}_i$ is the unnormalized clinical difficulty; $d_i \in [0, 1]$ is the min-max normalized difficulty; and $q_i \in (0, 1]$ is the normalized difficulty rank, where a larger value means that the report appears later in the easy-to-hard schedule. After obtaining d_i , CDE substitutes it into Eq. (2) to construct the training order $\mathcal{O}_d$, where reports with common attributes and simple graph structures appear earlier, while reports with rare pathological attributes and complex patterns occur in late training stages.

3.3 Long-Tail Curriculum Scheduling (LTCS)

Once the difficulty metric is defined, the next step is to construct the progressive training sequence to enable full convergence for model training. We achieve this through LTCS, which controls the selections of reports that are eligible for sampling at training step s . Following the standard process of curriculum learning, which uses a coefficient to

²For further illustration, a smaller n_ℓ or n_g yields a larger rarity weight; and the resulting weight makes rare attributes contribute more to the difficulty score.

control the training, we define this coefficient as:

$$\rho_s = \min \left(1, \rho_0 + (1 - \rho_0) \left(\frac{s}{s_{\text{warm}}} \right)^\kappa \right). \quad (6)$$

Here, ρ_0 is the initial fraction of easiest reports; s_{warm} is the warm-up length; and κ controls the pacing curve. The active pool is the subset of reports that is allowed to be sampled at step s . It acts as a curriculum gate, where early steps expose the model only to simple reports, while later steps gradually unlock harder cases according to ρ_s . A uniform distribution over $\mathcal{A}_s = \{i \mid q_i \leq \rho_s\}$ would still under-sample rare labels after they enter the active pool. We therefore define a weight to balance the contribution of rare reports as:

$$v_i = 1 + \lambda_{\text{lt}} \sum_{\ell=1}^m a_\ell \mathbb{I}(c_{i,\ell} \in \{1, u\}), \quad (7)$$

$$p_s^{\text{act}}(i) = \frac{v_i \mathbb{I}(i \in \mathcal{A}_s)}{\sum_{j \in \mathcal{A}_s} v_j}.$$

Here, v_i increases with the total inverse-frequency mass of the positive or uncertain attributes in report i ; λ_{lt} controls how strongly this rarity changes the sampling probability. The indicator $\mathbb{I}(i \in \mathcal{A}_s)$ keeps the easy-to-hard constraint, so a hard report receives zero probability before it enters the active pool. The training distribution by LTCS keeps an easy-to-hard schedule, but gives more exposure to under-represented pathology labels.

3.4 Prototype Anchor Replaying (PAR)

In real practice, we observe that pure curriculum learning might cause potential catastrophic forgetting issue, where the model tends to overfit the patterns in harder cases and discards the basic cross-modal alignment that is learnt earlier, eventually leading to obvious performance drop in the final stages (Tab. 1). Thus, we introduce a replay mechanism, i.e., PAR, to prevent a strict easy-to-hard schedule from forgetting prototype reports when the model enters complex schedules. To conduct this, we select the first ρ_{anc} percentage of the whole easy-to-hard schedule as the prototype anchor set, written as $\mathcal{S}_{\text{anc}} = \{i \mid q_i \leq \rho_{\text{anc}}\}$. These anchors correspond to reports from the earliest curriculum stages, where they provide a promising reference distribution for basic cross-modal grounding, preventing the decoder from overwriting simple report patterns when later stages emphasize rare and complex cases. Formally, PAR forms the final sampling

distribution as a mixture of active sampling, prototype anchor replaying, and global exploration:

$$\pi_s(i) = \lambda_{\text{act}} p_s^{\text{act}}(i) + \lambda_{\text{anc}} p^{\text{anc}}(i) + \lambda_{\text{glob}} \frac{1}{n}. \quad (8)$$

Here, p^{anc} is uniform over $\mathcal{S}_{\text{anc}}$; λ_{act} , λ_{anc} , and λ_{glob} are mixture weights that sum to 1; and the three terms define active sampling, prototype anchor replaying, and global exploration. Based on Eq. 1, we train the report decoder with the curriculum-aware objective, formulated by:

$$\mathcal{L}_{\text{PARAC}} = -\mathbb{E}_{i \sim \pi_s} \sum_{t=1}^{t_i} \log p_\theta(y_{i,t} \mid y_{i,<t}, \mathcal{Z}_i). \quad (9)$$

4 Experiments

We keep the main section focused on result analysis and provide the detailed experimental settings separately.³

4.1 Comparison to Baselines

Tab. 1 reports different baseline variants implemented on three representative existing methods, i.e., R2GEN (Chen et al., 2020), R2GENGPT (Wang et al., 2023b), and LLM-RG4 (Wang et al., 2025), evaluated on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). Yet, a counterintuitive comparison is observed in this experiment, where only adding curriculum learning causes performance drop in LLM-based methods. The reason behind this might be that LLMs contain more model parameters than task-specific models and are thus more prone to overfit the medical SFT data, whose scale is relatively smaller than their pre-training corpora (Wang et al., 2023b, 2025). This is also why the replaying mechanism is vital in our method, where it helps to stabilize the model training from catastrophic forgetting, especially for the LLM-based methods. Methods equipped with both curriculum learning and replay show significant improvements. All improvements of the complete method over the corresponding random-sampling backbones satisfy $p < 0.05$. LTCS provides an easy-to-hard sequence, whereas PAR preserves basic cross-modal alignment by replaying prototype anchors throughout training.

In addition to Tab. 1, we analyze the label-prevalence gap between generated reports by different methods and the test set of MIMIC-CXR

³See Appendix A for complete experimental settings.

METHODS & TRAINING SETTINGS	IU X-RAY			MIMIC-CXR				Δ (%) $\uparrow$
	B@1 $\uparrow$	B@4 $\uparrow$	R-L $\uparrow$	B@1 $\uparrow$	B@4 $\uparrow$	R-L $\uparrow$	F1 $\uparrow$	
R2GEN (Random Sampling) (Chen et al., 2020)	0.470	0.165	0.371	0.353	0.103	0.277	0.300	0.0
+CDE+LTCS	0.462	0.158	0.366	0.345	0.099	0.274	0.294	-2.0
+CDE+LTCS+PAR	0.487	0.176	0.383	0.371	0.116	0.290	0.327	+5.4
R2GENGPT (Random Sampling) (Wang et al., 2023b)	0.489	0.180	0.392	0.396	0.113	0.273	0.456	0.0
+CDE+LTCS	0.483	0.177	0.389	0.390	0.111	0.270	0.450	-1.3
+CDE+LTCS+PAR	0.511	0.195	0.409	0.418	0.135	0.313	0.548	+10.0
LLM-RG4 (Random Sampling) (Wang et al., 2025)	-	-	-	0.377	0.144	0.318	0.588	0.0
+CDE+LTCS	-	-	-	0.373	0.142	0.315	0.581	-1.1
+CDE+LTCS+PAR	-	-	-	0.431	0.154	0.339	0.619	+8.1

Table 1: Quantitative comparison of curriculum baselines implemented on three typical existing studies, i.e., R2GEN (Chen et al., 2020), R2GENGPT (Wang et al., 2023b), and LLM-RG4 (Wang et al., 2025), evaluated on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). Results are measured by BLEU- $\{1, 4\}$ (Papineni et al., 2002), ROUGE-L (Lin, 2004), and the F1 score of CHEXBERT labels (Smit et al., 2020). Best and second-best results within each group are highlighted in boldface and underlined. Δ reports the percentage improvements relative to the random-sampling baseline under the same backbone.

(Johnson et al., 2019), where this result is presented in Fig. 3. This figure shows an additional supplement to Fig. 1, with R2GEN and R2GENGPT further investigated, where results show that either task-specific or other LLM-based methods suffer from similar label imbalance issue. Compared to the real statistics of the test set, results reveal that the substantial performance gains of these method come from overfitting the common descriptive patterns and high-frequency pathologies, while leaving the diagnoses of rare medical labels largely neglected. Notably, even only introducing the curriculum learning (+CDE+LTCS) causes slight performance degradation in lexical quality and clinical accuracy, the imbalance issue essentially starts to get alleviated as in shown in Fig. 3. This observation validates our motivation to address the long-tail label issue of RRG with the design of curriculum learning. With the enhancement of the replaying mechanism (+CDE+LTCS+PAR), it is observed that either rare or high-frequency labels are equally generated, indicating that the performance gains in Tab. 1 derive from generating reports with more balanced pathology label distribution.

4.2 Comparison to Existing Methods

Tab. 2 compares the models equipped with PARAC to existing state-of-the-art methods on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). Note that most methods in this table use randomly sampled batches for training, while CMCL (Liu et al., 2021a) is a curriculum-based method. Equipping existing methods with PARAC brings performance gains on most metrics, suggesting similar effectiveness as that in Tab. 1, where introducing our curriculum learning- and replaying-based mechanisms en-

Method	Lung Opac.	Consol.	Lung Les.	Edema	Atelect.	Eml. Card.	Gap MAE
R2GEN	-0.104	-0.066	-0.054	+0.071	+0.074	+0.088	6.8
+CDE+LTCS	-0.055	-0.031	-0.026	+0.035	+0.038	+0.043	3.8
+CDE+LTCS+PAR	-0.026	+0.008	-0.012	+0.014	-0.006	+0.018	1.4
R2GENGPT	-0.096	-0.059	-0.047	+0.058	+0.064	+0.082	5.6
+CDE+LTCS	-0.049	-0.028	-0.022	+0.031	+0.034	+0.039	3.2
+CDE+LTCS+PAR	-0.021	+0.006	-0.009	+0.012	-0.004	+0.016	1.1
LLM-RG4	-0.081	-0.044	-0.036	+0.043	+0.045	+0.061	4.4
+CDE+LTCS	-0.042	-0.023	-0.019	+0.026	+0.028	+0.034	2.6
+CDE+LTCS+PAR	-0.018	+0.005	-0.007	+0.010	-0.003	+0.013	0.9

■ Under-Generated ■ Over-Generated

Figure 3: Label-prevalence gap analysis of R2GEN (Chen et al., 2020), R2GENGPT (Wang et al., 2023b), and LLM-RG4 (Wang et al., 2025), following the same settings as Tab. 1. Here, labels in blue and red indicate rare and high-frequency labels, respectively.

hances both task-specific and LLM-based methods. It is worth noting that these performance gains particularly focus on the CE metrics, where the reason behind this points to a more balanced label distribution in the final generated reports, as is also supported by Fig. 3.

Tab. 3 further supplements the aforementioned comparison with a direct long-tail diagnostic on extracted labels of the generated reports.⁴ Across all three backbones, +CDE+LTCS reduces rare-label omissions and label-prevalence gaps, while +CDE+LTCS+PAR gives the strongest rare-label recall after replaying simple anchors. This observation aligns well with the experimental trends as that in Fig. 3 and offers further supplementary to the results reported in Tab. 2, where PARAC brings substantial performance gain through alleviating the label imbalance issue. Notably, we notice that PARAC functions more effectively on LLM-based methods than task-specific ones, where we attribute

⁴See Appendix B for the label groups used in Tab. 3.

	METHOD	PARAMS (B)	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑
IU X-RAY (Demner-Fushman et al., 2016)	ST (Vinyals et al., 2015)	< 1	0.216	0.124	0.087	0.066	–	0.306	–	–	–
	ADAATT (Lu et al., 2017)	< 1	0.220	0.127	0.089	0.068	–	0.308	–	–	–
	ATT2IN (Rennie et al., 2017)	< 1	0.224	0.129	0.089	0.068	–	0.308	–	–	–
	CoATT (Jing et al., 2018)	< 1	0.455	0.288	0.205	0.154	–	0.369	–	–	–
	R2GEN (Chen et al., 2020)	< 1	0.470	0.304	0.219	0.165	–	0.371	–	–	–
	R2GENCMN (Chen et al., 2021)	< 1	0.475	0.309	0.222	0.170	0.191	0.375	–	–	–
	CMCL (Liu et al., 2021a)	< 1	0.473	0.305	0.217	0.162	0.186	0.378	–	–	–
	PPKED (Liu et al., 2021b)	< 1	0.483	0.315	0.224	0.168	–	0.376	–	–	–
	CA (Liu et al., 2021c)	< 1	0.492	0.314	0.222	0.169	0.193	0.381	–	–	–
	HRGR (Li et al., 2018)	< 1	0.438	0.298	0.208	0.151	–	0.322	–	–	–
	CMAS-RL (Jing et al., 2019)	< 1	0.464	0.301	0.210	0.154	–	0.362	–	–	–
	R2GENRL (Qin and Song, 2022)	< 1	0.494	0.321	0.235	0.181	0.201	0.384	–	–	–
	R2-LLM (Liu et al., 2024)	14.2	0.499	0.323	0.238	0.184	0.208	0.390	–	–	–
	XRAYGPT (Thawakar et al., 2024)	7	0.177	0.104	0.047	0.007	0.105	0.203	–	–	–
	S2D-ALIGN (Gao et al., 2026)	7	0.512	–	–	0.195	–	0.407	–	–	–
	R2GEN+PARAC	< 1	0.487	0.316	0.229	0.176	0.198	0.383	–	–	–
	R2GENGPT+PARAC	7	0.511	0.329	0.240	0.195	0.209	0.409	–	–	–
MIMIC-CXR (Johnson et al., 2019)	MAIRA-1 (Hyland et al., 2023)	7	0.392	–	–	0.142	–	0.289	–	–	0.553
	MED-PALM (Tu et al., 2023)	562	0.317	–	–	0.115	–	0.275	–	–	0.378
	R2GENGPT (Wang et al., 2023b)	7	0.396	–	–	0.113	–	0.273	0.550	0.389	0.456
	CHEXAGENT (Chen et al., 2024)	7	0.189	–	–	0.040	–	0.208	0.552	0.291	0.381
	INVERGE (Deria et al., 2024)	7	<u>0.425</u>	0.240	0.132	0.100	–	0.309	–	–	–
	R2-LLM (Liu et al., 2024)	14.2	0.402	0.262	0.180	0.128	<u>0.175</u>	0.291	0.540	0.421	0.473
	XRAYGPT (Thawakar et al., 2024)	7	0.128	0.045	0.014	0.004	<u>0.079</u>	0.111	–	–	–
	LLM-RG4 (Wang et al., 2025)	7	0.377	–	–	0.144	–	0.318	0.670	0.524	0.588
	S2D-ALIGN (Gao et al., 2026)	7	0.422	<u>0.263</u>	<u>0.183</u>	<u>0.149</u>	–	<u>0.332</u>	<u>0.690</u>	<u>0.543</u>	<u>0.608</u>
	ST (Vinyals et al., 2015)	< 1	0.299	0.184	0.121	0.084	0.124	0.263	0.280	0.187	0.224
	ADAATT (Lu et al., 2017)	< 1	0.299	0.185	0.124	0.088	0.118	0.266	0.290	0.177	0.220
	ATT2IN (Rennie et al., 2017)	< 1	0.325	0.203	0.136	0.096	0.134	0.276	0.345	0.227	0.274
	R2GEN (Chen et al., 2020)	< 1	0.353	0.218	0.145	0.103	0.142	0.277	0.380	0.248	0.300
	R2GENCMN (Chen et al., 2021)	< 1	0.353	0.218	0.148	0.106	0.142	0.278	0.382	0.250	0.302
	CA (Liu et al., 2021c)	< 1	0.350	0.219	0.152	0.109	0.151	0.283	–	–	–
	CMCL (Liu et al., 2021a)	< 1	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
	PPKED (Liu et al., 2021b)	< 1	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
	KIUT (Huang et al., 2023)	< 1	0.393	0.243	0.159	0.113	–	0.285	0.410	0.293	0.342
	RGRG (Tanida et al., 2023)	< 1	0.373	0.249	0.175	0.126	0.168	0.264	0.530	0.419	0.468
	EKAGEN (Bu et al., 2024)	< 1	0.419	0.258	0.170	0.119	–	0.287	0.570	0.444	0.499
	PROMPTMRG (Jin et al., 2024)	< 1	0.398	–	–	0.112	–	0.258	0.590	0.441	0.505
	ITA (Wang et al., 2022)	< 1	0.395	0.253	0.170	0.121	0.147	0.284	–	–	–
	R2GENRL (Qin and Song, 2022)	< 1	0.381	0.232	0.155	0.109	0.151	0.287	0.385	0.268	0.316
	TOPDOWN (Anderson et al., 2018)	< 1	0.317	0.195	0.130	0.092	0.128	0.267	0.335	0.223	0.268
	WARMSTART (Nicolson et al., 2023)	< 1	0.392	0.245	0.169	0.124	0.153	0.285	0.470	0.325	0.384
	R2GEN+PARAC	< 1	0.371	0.229	0.156	0.116	0.154	0.290	0.405	0.274	0.327
	R2GENGPT+PARAC	7	0.418	0.259	0.179	0.135	0.170	0.313	0.650	0.474	0.548
	LLM-RG4+PARAC	7	0.431	0.271	0.190	0.154	0.181	0.339	0.710	0.549	0.619

Table 2: Quantitative comparisons with existing methods on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) w.r.t. the NLG and CE metrics.

METHOD	CE F1↑	HEAD R↑	RARE R↑	RARE OMIT↓	GAP MAE↓
R2GEN	30.0	44.9	7.8	92.2	6.8
+CDE+LTCS	29.4	42.7	16.5	83.5	3.8
+CDE+LTCS+PAR	32.7	47.8	28.4	28.6	1.4
R2GENGPT	45.6	51.8	10.7	89.3	5.6
+CDE+LTCS	46.8	51.2	24.6	75.4	3.2
+CDE+LTCS+PAR	54.8	56.5	38.7	24.7	1.1
LLM-RG4	58.8	60.4	14.9	85.1	4.4
+CDE+LTCS	60.4	59.6	32.7	67.3	2.6
+CDE+LTCS+PAR	61.9	61.7	43.1	21.8	0.9

Table 3: Analysis of high-frequency and rare findings on MIMIC-CXR (Johnson et al., 2019).

this observation as the overfitting-prone nature of LLMs.

Tab. 4 visualizes a selected case study from MIMIC-CXR with rare labels. It is observed that although reports generated by baselines preserve high-frequency findings, but they particularly under-specify the rare ones, e.g., “*edema*” and “*bilateral pleural effusions*”. This confirms the observation in Fig. 1 again from the aspect of qualitative

results, where training with random sampling leads to high-frequency label hacking in existing methods. The case also explains why aggregate lexical overlap and fitting common clinical labels might reach high overall performance in Tab. 2, where a generated report might repeat frequent observations, and still miss the rare findings. With the assistance of our curriculum learning, these rare findings are produced more accurately, showing that it changes the optimization paradigm to allow rare attributes receive equal weights in training.

4.3 Ablation Studies

Effect of the Difficulty Metrics. Difficulty metric is a vital factor to reflect the curriculum schedule and impact whether it could address the label imbalance issue. To explore its effect, we try several variants in Tab. 5. It is observed that using

RADIOGRAPH AND CASE	METHOD	REPORT COMPARISON
 Rare Labels: multifocal pneumonia, asymmetric edema, bilateral effusions.	Ground-truth Report	Right-base pneumonia is seen. Right-upper reticulonodular opacity suggesting asymmetric edema or multifocal pneumonia . There is small bilateral effusions and no pneumothorax.
	R2GEN	Reports bilateral airspace opacities compatible with pneumonia , but does not describe asymmetric edema or bilateral pleural effusions .
	R2GEN+PARAC	Recovers right-base pneumonia , right-upper opacity suggesting asymmetric edema , small bilateral effusions , and no pneumothorax.
	R2GENGPT	Bilateral opacities compatible with pneumonia . Heart size is mildly enlarged. No pleural effusion is seen.
	R2GENGPT+PARAC	There is multifocal pneumonia with right-upper reticulonodular opacity. Asymmetric edema and small bilateral effusions are present.
	LLM-RG4	Lungs are clear. Heart size is normal. There is no pneumonia and pleural effusion. Vascular congestion and misses asymmetric edema is present.
	LLM-RG4+PARAC	Right-base pneumonia is present. There is asymmetric edema and small bilateral effusions . Absence of pneumothorax is seen.

Table 4: A selected case study from MIMIC-CXR (Johnson et al., 2019) with rare labels, compared to the same baselines of Tab. 1. Here, we highlight rare findings with bold colored texts and shade the PARAC rows in light gray.

DIFFICULTY METRIC	B@4↑	CE F1↑	RARE REC.↑
random sampling	0.144	0.588	0.342
report length	0.143	0.586	0.339
medical labels	0.150	0.606	0.386
structured knowledge	0.151	0.608	0.397
full model	0.154	0.619	0.431

Table 5: Quantitative results under the curriculum learning with different difficulty metric designs.

a naive proxy such as report length obtains similar results to the random-order baseline, indicating that longer reports do not always cover more rare pathologies, as they might contain patternized descriptive texts that commonly occur. Medical labels and structured knowledge graphs offer information for the curriculum from different aspects. On one hand, they have several overlapped pathological attributes, and thus both contribute positively compared to the baseline. On the other hand, structured knowledge not only indicates pathological entities via its nodes (which overlaps with the medical labels), but also illustrates the relations between them, while medical labels contain predefined aspects such as device state that structured knowledge lacks. Our full method reaches the best performance, where using both representations give the curriculum a synergistic effect.

Effect of the anchor set size. The prototype anchor set serves as a crucial role in stabilizing the training schedule. Fig. 4 presents the performance curve with respect to different anchor set size. There appears to be an optimal interval where anchor set functions the best for each method. Increasing the anchor ratio first improves the score by replaying stable prototype reports, but performance drops rapidly after the optimal region, since too many anchors dilute the training of harder cases. The optimal ratio is larger for R2GEN, whereas R2GENGPT and LLM-RG4 show a narrower high-performing region.

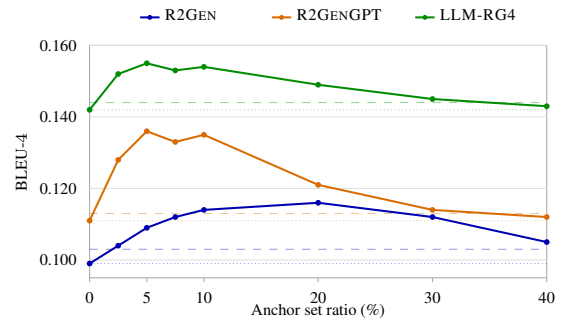


Figure 4: BLEU-4 score curves with respect to different anchor set ratio. Here, we report the results of the original baselines (highlighted in dash lines) for reference.

5 Conclusion

In this paper, we present PARAC, a pathology-aware replay curriculum to address the label imbalance issue in RRG. We investigate the long-tail failures as a distribution-related problem and use clinically ordered exposure and prototype anchor replaying to give rare attributes more effective optimization. Experimental results on IU X-RAY and MIMIC-CXR support the performance gains and attribute coverage.

Limitations

In this section, we would like to illustrate the limitations of this work. First, statistical rarity is an optimization criterion in this work. ParaC gives low-frequency attributes more effective supervision, but our results do not establish that every rare finding is more consequential than a frequent one. Second, we observe that the proposed CDE depends on automatic report annotators, i.e., CheXbert and RadGraph, which might introduce annotation noise and affect the precision. Nevertheless, experiments with perturbation and alternative labelers show that the training benefit persists with imperfect signals. Third, the experiments cover chest radiography only due to the limitation of available open-source data, where we will extend this work

to more modalities when data is available.

Acknowledgments

This work was supported by the Macau Science and Development Fund (File Nos. 0216/2024/AGJ and 0118/2024/RIA2).

References

- Charika De Alvis, Dishanika Denipitiyage, and Suranga Seneviratne. 2025. Long-Tail Learning with Rebalanced Contrastive Loss. *Neurocomputing*, 657:131601.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, pages 6077–6086.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, pages 65–72.
- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. MAIRA-2: Grounded Radiology Report Generation. *arXiv*, arXiv:2406.04449.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *ICML*, pages 41–48.
- Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 2024. Instance-Level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation. In *CVPR*, pages 14194–14204.
- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*, pages 1567–1578.
- Lifeng Chen, Tianqi You, Hao Liu, Zhimin Bao, Jile Jiao, Xiao Han, Zhicai Ou, Tao Sun, Xiaofeng Mou, Xiaojie Jin, and Yi Xu. 2026. ECHO: Efficient Chest X-Ray Report Generation with One-Step Block Diffusion. *arXiv*, arXiv:2604.09450.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-Modal Memory Networks for Radiology Report Generation. In *ACL-IJCNLP*, pages 5904–5914.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-Driven Transformer. In *EMNLP*, pages 1439–1449.
- Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Bao Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. 2024. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv*, arXiv:2401.12208.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*, pages 9268–9277.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Med. Inform. Assoc.*, 23(2):304–310.
- Ankan Deria, Komal Kumar, Snehashis Chakraborty, Dwarikanath Mahapatra, and Sudipta Roy. 2024. IN-VERGE: Intelligent Visual Encoder for Bridging Modalities in Report Generation. In *CVPRW*, pages 2028–2038.
- Jiechao Gao, Chang Liu, and Yuangang Li. 2026. S2D-Align: Shallow-to-Deep Auxiliary Learning for Anatomically-Grounded Radiology Report Generation. In *AAAI*, pages 30780–30788.
- Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, and Dimitris N. Metaxas. 2025. RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment. In *MICCAI*.
- Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. KiUT: Knowledge-Injected U-Transformer for Radiology Report Generation. In *CVPR*, pages 19809–19818.
- Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. 2023. MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation. *arXiv*, arXiv:2311.13668.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre J. 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. In *NeurIPS Datasets and Benchmarks*.

- Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. 2015. Self-Paced Curriculum Learning. In *AAAI*, pages 2694–2700.
- Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In *AAAI*, pages 2607–2615.
- Baoyu Jing, Zeya Wang, and Eric P. Xing. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *ACL*, pages 6570–6580.
- Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the Automatic Generation of Medical Imaging Reports. In *ACL*, pages 2577–2586.
- 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. *Sci. Data*, 6:317.
- M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-Paced Learning for Latent Variable Models. In *NeurIPS*, pages 1189–1197.
- Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *AAAI*, pages 6666–6673.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *NeurIPS*, pages 28541–28564.
- Wang Li, Guangzeng Han, Yuexin Wu, I-Chan Huang, and Xiaolei Huang. 2025. Joint Imbalance Adaptation for Radiology Report Generation. *J. Healthc. Inform. Res.*, 9(4):720–742.
- Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *NeurIPS*, 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.
- Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*, pages 2999–3007.
- Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *AAAI*, pages 18635–18643.
- Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-Based Multimodal Curriculum Learning for Medical Report Generation. In *ACL-IJCNLP*, pages 3001–3012.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *CVPR*, pages 13753–13762.
- 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 ACL-IJCNLP*, pages 269–280.
- Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025a. Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation. In *CVPR*, pages 10348–10359.
- Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. 2025b. HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation. In *AAAI*, pages 5595–5603.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *CVPR*, pages 3242–3250.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving Chest X-ray Report Generation by Leveraging Warm Starting. *Artif. Intell. Med.*, 144:102633.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318.
- Han Qin and Yan Song. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of ACL*, pages 448–458.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*, pages 7008–7024.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *EMNLP*, pages 1500–1519.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*, pages 7433–7442.
- Omkar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, Jorma Laaksonen, and Fahad Khan. 2024. XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models. In *BioNLP@ACL*, pages 440–448.

- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, Philip Andrew Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, and 13 others. 2023. Towards Generalist Biomedical AI. *arXiv*, arXiv:2307.14334.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NeurIPS*, pages 5998–6008.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pages 3156–3164.
- Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. 2022. An Inclusive Task-Aware Framework for Radiology Report Generation. In *MICCAI*, pages 568–577.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In *CVPR*, pages 11558–11567.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. R2GenGPT: Radiology Report Generation with Frozen LLMs. *arXiv*, arXiv:2309.09812.
- Zhu hao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. 2025. LLM-RG4: Flexible and Factual Radiology Report Generation Across Diverse Input Contexts. In *AAAI*, pages 8250–8258.
- Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. 2025a. Radiology Report Generation via Multi-Objective Preference Optimization. In *AAAI*, pages 8664–8672.
- Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Zhe Wang, and Chenjia Bai. 2025b. Online Iterative Self-Alignment for Radiology Report Generation. In *ACL*, pages 27799–27814.
- 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 *ACL*, pages 29–45.
- Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. 2020. When Radiology Report Generation Meets Knowledge Graph. In *AAAI*, pages 12910–12917.

DATASET	TRAIN	VAL.	TEST
IU X-RAY (Demner-Fushman et al., 2016; Jing et al., 2018; Chen et al., 2020)	2,069	296	590
MIMIC-CXR (Johnson et al., 2019)	270,790	2,130	3,858

Table 6: Dataset statistics of IU X-RAY (Demner-Fushman et al., 2016; Jing et al., 2018; Chen et al., 2020) and MIMIC-CXR (Johnson et al., 2019).

GROUP	LABEL ATTRIBUTES
Rare labels	<i>Lung Opacity; Consolidation; Lung Lesion</i>
High-frequency labels	<i>Edema; Atelectasis; Enlarged Cardiomediastinum</i>

Table 7: Label groups used by the long-tail diagnostic in Tab. 3 on MIMIC-CXR (Johnson et al., 2019). No ranking highlight is used because the table defines diagnostic label groups only.

A Experimental Settings

Datasets. We select two widely adopted benchmarks, IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019), to conduct all experiments. Specifically, IU X-RAY contains 7,470 images and 3,955 reports, where we follow the splits of conventional studies, resulting in 2,069, 296, and 590 cases in the training, validation, and testing sets, respectively (Jing et al., 2018; Chen et al., 2020). MIMIC-CXR comprises 377,110 images from 227,835 studies, where we follow its official splits, consisting of 270,790, 2,130, and 3,858 cases in the training, validation, and testing sets, respectively (Johnson et al., 2019). Tab. 6 summarizes these dataset statistics.

Baselines. To evaluate the proposed curriculum strategy, we implement PARAC on three representative methods, R2GEN (Chen et al., 2020), R2GENGPT (Wang et al., 2023b), and LLM-RG4 (Wang et al., 2025), covering both task-specific and LLM-based decoders. For the baseline variants, we consider 3 settings in Tab. 1: the standard training strategy with randomly sampled batches, adding CDE and LTCS on top of each method, and further equipping PAR. For existing RRG methods, we compare with 28 studies in Tab. 2.

Evaluation metrics. Following previous studies, we measure the report quality from two aspects,

lexical coherence and clinical accuracy, with Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. For NLG metrics, we include BLEU- $\{1, 2, 3, 4\}$ (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). For CE metrics, we use CheXbert (Smit et al., 2020) to extract medical labels from both generated and ground-truth reports, and compute precision, recall, and F1 scores as the final CE measurements.

Implementation details. To obtain the medical labels and structured medical knowledge in PARAC, we use CHEXBERT (Smit et al., 2020) and RADGRAPH (Jain et al., 2021) in our experiments. For the hyper-parameter settings in the curriculum learning, we try various experiments on the validation set and select the best-performing one as the final setup. Specifically, we set the initial accessible ratio ρ_0 to 0.2, the warm-up epoch to 1, and use $(\lambda_{\text{act}}, \lambda_{\text{anc}}, \lambda_{\text{glob}}) = (0.7, 0.2, 0.1)$ for PAR. For optimization, we follow the implementation settings reported in the official papers, and re-train the methods equipped with PARAC. All experiments are conducted on one NVIDIA A40 GPU.

Protocol of Statistical Testing. For all reported experiments in this paper, we apply paired significance tests to measure their contributions from matched test studies. Consequently, we observe that all reported improvements satisfy $p < 0.05$.

Computational Cost of CDE. In our method, CDE is constructed once before training, so that it introduces extra computational cost in an offline manner. Tab. 11 presents the corresponding results. One can see that it takes 1,245 seconds for the medical label extraction procedure, and costs 3,862 seconds for structured knowledge extraction, and eventually, consumes 18 seconds for sorting, resulting in approximately 0.019 seconds per training report in total.

B Long-Tail Diagnostic Label Groups

Tab. 7 lists the rare and high-frequency labels used in the corresponding experiments.

C More Results and Analyses

We offer more detailed analyses to validate our design. Specifically, to explore the effect of resampling strategies, Tab. 8 compares ParaC with direct alternatives to address the label imbalance issue. Tab. 9 investigates the effect of noisy labels under

METHOD	B@4↑	CE F1↑	RARE R↑	RARE OMIT↓	GAP MAE↓
REPORT-LEVEL					
LLM-RG4	0.144	58.8	14.9	85.1	4.4
+RARE-LABEL REPLICATION	0.146	59.4	24.6	59.8	2.9
+INVERSE-FREQUENCY	<u>0.147</u>	<u>59.8</u>	<u>27.3</u>	<u>55.6</u>	<u>2.5</u>
RE-SAMPLING					
+LOSS RE-WEIGHTING	0.145	59.7	24.8	60.7	3.0
+PARAC	0.154	61.9	43.1	21.8	0.9
LOSS-LEVEL					
LLM-RG4	0.144	58.8	14.9	85.1	4.4
+RARE-LABEL	0.143	58.5	16.1	83.6	4.2
OVERSAMPLING					
+INVERSE-FREQUENCY	<u>0.145</u>	<u>59.0</u>	<u>17.4</u>	<u>82.8</u>	<u>4.1</u>
BATCH SAMPLER					
+TOKEN-LEVEL CE	0.144	58.7	14.2	86.0	4.6
RE-WEIGHTING					
+FOCAL LOSS	0.142	58.4	15.7	84.0	4.3
+PARAC	0.154	61.9	43.1	21.8	0.9

Table 8: Comparison of LLM-RG4 (Wang et al., 2025) with report- and loss-level alternatives on MIMIC-CXR (Johnson et al., 2019), measured by BLEU-4, CE F1, rare-label recall and omission, and Gap MAE. Best and second-best results within each block are highlighted in boldface and underlined.

CDE SIGNAL / LABEL SOURCE	CE F1↑	RARE R↑	RARE OMIT↓	GAP MAE↓
NOISE PERTURBATION				
clean CDE (0%)	61.9	43.1	21.8	0.9
perturbed labels/graphs (10%)	61.3	40.8	25.2	1.1
perturbed labels/graphs (20%)	60.6	37.2	30.6	1.5
perturbed labels/graphs (30%)	59.8	33.6	36.9	2.0
random sampling	58.8	14.9	85.1	4.4
ANNOTATOR TYPE				
random sampling	58.8	14.9	85.1	4.4
rule-based labels (word frequency)	60.7	38.8	31.9	1.7
CheXbert (noise ratio 20%)	60.6	37.2	30.6	1.5
CheXbert (noise ratio 10%)	<u>61.3</u>	<u>40.8</u>	<u>25.2</u>	<u>1.1</u>
CheXbert (noise ratio 0%)	61.9	43.1	21.8	0.9

Table 9: Robustness of PARAC to CDE-signal perturbations and annotator type on MIMIC-CXR (Johnson et al., 2019), measured by CE F1, rare-label recall and omission, and Gap MAE. Best and second-best results are highlighted in boldface and underlined.

$(\lambda_{\text{act}}, \lambda_{\text{anc}}, \lambda_{\text{glob}})$	B@4↑	CE F1↑	RARE R↑	GAP MAE↓
(1.0, 0.0, 0.0)	0.142	60.4	32.7	2.6
(0.8, 0.1, 0.1)	0.150	61.1	39.2	1.4
(0.7, 0.2, 0.1)	0.154	61.9	43.1	0.9
(0.6, 0.3, 0.1)	<u>0.151</u>	<u>61.2</u>	<u>41.0</u>	<u>1.2</u>
(0.7, 0.1, 0.2)	0.149	60.8	38.4	1.5

Table 10: Sensitivity analysis of PAR to the active, anchor, and global mixture weights on MIMIC-CXR (Johnson et al., 2019), measured by BLEU-4 and long-tail clinical metrics. Best and second-best results are highlighted in boldface and underlined.

STAGE	TIME COST (S)	TIME/REPORT (S)
medical label extraction	1,245	0.0046
structured knowledge extraction	3,862	0.0143
difficulty sorting	18	0.0001

Table 11: Computational cost for constructing the curriculum signals on MIMIC-CXR (Johnson et al., 2019).

different magnitudes of perturbations and the annotator type. Tab. 10 provides a sensitivity analysis of three hyper-parameters in PAR. Tab. 11 reports the computational cost of our label extraction.