CLINICALLY GROUNDED CURRICULUM LEARNING FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current radiology report generation (RRG) systems predominantly rely on random mini-batch sampling, a paradigm that treats all reports equally and ignores the pathological complexity across different patients. Consequently, while producing fluent text, these models struggle to accurately capture fine-grained local pathology attributes, often suffering from severe clinical hallucinations. Curriculum learning presents a promising solution by progressively introducing data from easy to hard; however, defining a valid difficulty metric in the RRG context is nontrivial, as naive proxies like report length cannot effectively reflect true pathological complexity. To address this, we propose a **C**linically-structured **C**urriculum RR**G** framework (C^2-GEN). We instantiate this framework with hierarchical difficulty metrics that assess clinical complexity through both the global semantic density of pathology findings and the structural medical knowledge. Guided by these metrics, our approach replaces conventional uniform sampling with a difficulty-aware training schedule, gradually expanding the training pool towards complex cases while continually replaying simpler anchor samples to prevent catastrophic forgetting. Furthermore, we augment the curriculum from the aspect of network architecture, introducing an adaptive memory mechanism that dynamically adjusts the attention mechanism based on sample complexity to better bridge visual features and complex text generation. Experiments on public datasets demonstrate the effectiveness of C^2-GEN, consistently improving the fidelity of local pathology attributes compared to existing baselines. Comprehensive analyses further illustrate the progressive learning patterns and cross-modal alignment in RRG, where these insights establish curriculum learning as a principled pathway to mitigating hallucinations.

1 INTRODUCTION

RRG stands as a cornerstone task in medical image analysis, aiming to automatically translate visual radiographic findings into clinically accurate free-text reports. Traditionally, clinical reporting relies heavily on human expertise; however, this manual process is normally labor-intensive and error-prone, particularly when patients exhibit highly complex or subtle pathological features. To alleviate this clinical burden, deep learning-based approaches have emerged, leveraging large-scale annotated radiograph-report datasets (Johnson et al., 2019) to automatically capture cross-modal mappings between vision and language. More recently, motivated by the remarkable text-generation capabilities of large language models (Brown et al., 2020; Touvron et al., 2023; Qwen Team, 2025), this research paradigm increasingly shifts toward domain-adapting foundation models for medical text generation and radiology report generation (Li et al., 2023; Liu et al., 2024; Bannur et al., 2024; Chen et al., 2024a; Xiao et al., 2025b; Wang et al., 2023; 2025; Liu et al., 2025; Gao et al., 2026). This paradigm shift has attracted widespread attention, where the generated reports should remain both linguistically coherent and strictly faithful to fine-grained pathologies.

Despite these advancements, the majority of existing training pipelines still rely on standard random mini-batch sampling, a paradigm that implicitly treats all reports as uniformly difficult throughout the optimization process. This uniform treatment is fundamentally flawed given the profound heterogeneity of clinical reports, which vary substantially in pathology complexity, local finding density, anatomy-observation coupling, and the precise articulation of uncertainty or negation. By indiscriminately mixing simple and complex cases in training, the model tends to adopt a globally

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Overview of C^2-GEN, including three main components, namely Hierarchical Clinical Complexity Assessor (HCCA), Difficulty-Aware Anchor Replaying (DAAR), and Complexity-Adaptive Memory (CAM). In this figure, we highlight the corresponding illustrations of the aforementioned components in purple, orange, and green, respectively.

fluent reporting style, while failing to capture fine-grained local pathology attributes—a vulnerability further amplified by the inherent susceptibility of LLMs to hallucinations. Consequently, this mismatch leads to severe clinical hallucinations, thus misaligning the real requirements in clinical practices. Curriculum learning (Bengio et al., 2009) offers a principled solution to this bottleneck by structuring the training data from simple to complex. However, this paradigm hinges entirely on defining a difficulty metric that accurately reflects the clinical intricacies of radiology report generation. Naive proxies, such as report length, are inherently inadequate to reflect the clinical complexity. The only prior curriculum-based report-generation method, CMCL (Liu et al., 2021a), already suggests that training order matters, but it adopts a general competence schedule rather than a clinically grounded pathology-difficulty metric. Moreover, CMCL is tailored to pre-LLM generators and uses a general competence schedule rather than clinically grounded pathology-difficulty metrics. Consequently, developing a specialized curriculum learning framework customized for LLM-based report generation emerges as a highly significant research endeavor.

In this paper, we propose the **C**linically-structured **C**urriculum RR**G** framework (C^2-GEN), tailored for LLM-based RRG. Our framework consists of three core components, namely the Hierarchical Clinical Complexity Assessor (HCCA), the Difficulty-Aware Anchor Replaying (DAAR) strategy, and the Complexity-Adaptive Memory (CAM) module. The key of our approach lies in formulating difficulty metrics that accurately reflect the intricate pathological complexity of clinical reports. Specifically, **HCCA** models clinical difficulty from two primary perspectives, i.e., a multi-label classification paradigm that serves as a semantic proxy by quantifying the breadth of pathological findings, and a structured knowledge graph that captures the structural depth by modeling the topological complexity among medical entities and relations, while using report length only as a control proxy in ablations. Guided by these metrics, **DAAR** executes a progressive sampling schedule. This schedule allows the model to learn the cross-modal mapping smoothly from pathologically simple samples to highly complex ones, while continually replaying simpler anchor samples to prevent catastrophic forgetting and alleviate local optima. Furthermore, we augment the network architecture with the **CAM** module, which dynamically allocates network capacity based on sample complexity to better bridge visual features and complex text generation. Extensive experiments on IU X-Ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) demonstrate that our C^2-GEN significantly outperforms the conventional random mini-batch sampling paradigm in existing methods. Furthermore, comprehensive ablations show that semantic and structural difficulty signals, replay-aware scheduling, and adaptive memory jointly improve local pathology fidelity.[1]

2 METHOD

Given the full training set $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, where $\mathbf{X}_i \in \mathbb{R}^{C\times H\times W}$ denotes a radiology image and $\mathbf{Y}_i$ denotes its corresponding ground-truth report, C^2-GEN framework follows a cohesive cur-

[1]We discuss the relationships with existing methods in Appendix A.

riculum roadmap illustrated below. Fig. 1 shows the overview of C^2-GEN. First, **HCCA** acts as a dataset annotator to process the reference report $\mathbf{Y}_i$ of each training sample to compute a clinically grounded difficulty score d_i. Next, leveraging the global distribution of these difficulty scores, **DAAR** orchestrates the training schedule. At any given global step t, it dynamically curates an active training pool governed by a monotonically expanding curriculum coefficient ρ_t, from which a curriculum-guided mini-batch $\mathcal{B}_t$ is sampled. Finally, **CAM** integrates the visual features with the aforementioned curriculum. Conditioned on the specific difficulty signal of the sample and the current status, CAM modulates the internal complexity-adaptive states before they are consumed by the LLM decoder $\mathcal{M}_\theta$ to produce the final predicted report. Formally, this process is written as:

$$\mathcal{D} \xrightarrow{\text{HCCA}} \{d_i\}_{i=1}^{N} \xrightarrow[\rho_t\uparrow]{\text{DAAR}} \mathcal{B}_t \xrightarrow{\text{CAM}+\mathcal{M}_\theta} \left\{\widehat{\mathbf{Y}}_i\right\}_{(\mathbf{X}_i,\mathbf{Y}_i)\in\mathcal{B}_t}. \tag{1}$$

Here, $\mathcal{B}_t$ denotes the sampled batch at step t and $\widehat{\mathbf{Y}}_i$ denotes the generated report for the i-th sample. Below, we illustrate the details of each component following the pipeline sequence.

2.1 PRELIMINARIES

Standard VLM for Radiology Report Generation. To establish the foundation, we formulate the RRG task under the standard Vision-Language Model (VLM) paradigm. Given a radiology image $\mathbf{X}_i \in \mathbb{R}^{C\times H\times W}$ as input, a frozen visual encoder $\mathcal{E}_{\text{vis}}$ extracts the dense image features $\mathbf{V}_i = \mathcal{E}_{\text{vis}}(\mathbf{X}_i)$. Subsequently, a trainable cross-modal connector $\mathcal{B}_\theta$ maps $\mathbf{V}_i$ into the embedding space of the LLM, yielding the visual conditioning states $\mathbf{Z}_i = \mathcal{B}_\theta(\mathbf{V}_i)$. During training, these visual tokens are fused with the ground-truth report sequence $\mathbf{Y}_i = (y_{i,1}, \ldots, y_{i,T_i})$ via sequence-wise token concatenation. The LLM decoder $\mathcal{M}_\theta$ then autoregressively predicts the text tokens:

$$p_\theta(\mathbf{Y}_i \mid \mathbf{X}_i) = \prod_{t=1}^{T_i} p_\theta(y_{i,t} \mid y_{i,<t}, \mathbf{Z}_i). \tag{2}$$

Under the conventional training recipe, the model is optimized by minimizing the standard cross-entropy loss over uniformly sampled data from $\mathcal{D}$, formally expressed as:

$$\mathcal{L} = -\mathbb{E}_{(\mathbf{X}_i,\mathbf{Y}_i)\sim\mathcal{U}(\mathcal{D})} \sum_{t=1}^{T_i} \log p_\theta(y_{i,t} \mid y_{i,<t}, \mathbf{Z}_i). \tag{3}$$

Crucially, this objective assumes a uniform distribution $\mathcal{U}(\mathcal{D})$, treating all reports equally from the very first optimization step. This uniform difficulty motivates our curriculum design in this work.

Curriculum Learning and Dataset Scheduling. Classical curriculum learning (Bengio et al., 2009) fundamentally replaces uniform data exposure with a progressive, easy-to-hard training schedule. Rather than presenting the model with all samples from $\mathcal{D}$ uniformly, it formalizes a scalar difficulty score d_i for each training instance, where lower values refer to simpler samples and higher values indicate more complex ones. Sorting the entire dataset according to d_i yields a normalized ranking sequence $q_i \in (0, 1]$ for each sample. To regulate the data exposure, a standard pacing function (or curriculum coefficient) $\rho_t \in (0, 1]$ can determine the proportion of the ranked dataset at any given training step t, where we formally express this process as:

$$\begin{aligned} \rho_t &= \rho_0 + (1-\rho_0)\min\left(\frac{t}{T_{\text{warmup}}}, 1\right), \\ \alpha_i^{(t)} &= \mathbb{I}(q_i \le \rho_t). \end{aligned} \tag{4}$$

Here, ρ_0 denotes the initial fraction of the easiest examples; T_{warmup} denotes the duration over which the curriculum unfolds; $\alpha_i^{(t)}$ serves as a binary indicator to show the visibility of the sample i at step t. As ρ_t monotonically increases, more challenging samples progressively are used for optimization. Consequently, the active training subset (i.e., the active pool) at step t is strictly defined as:

$$\mathcal{A}_t = \left\{(\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{D} \mid \alpha_i^{(t)} = 1\right\}. \tag{5}$$

In the subsequent sections, we detail this generic paradigm for the specific context of C^2-GEN. Specifically, we first elaborate how the intrinsic difficulty d_i is quantified via clinical structures, then illustrate how dynamic scheduling is seamlessly integrated with the VLM-based RRG pipelines.

2.2 HIERARCHICAL CLINICAL COMPLEXITY ASSESSOR

The fundamental prerequisite for an effective curriculum is quantifying the intrinsic difficulty of each sample. HCCA instantiates the generic difficulty variable for the curriculum schedule of C^2-GEN. Given a ground-truth report $\mathbf{Y}_i$, HCCA computes a scalar clinical complexity score $d_i = \phi_{\mathrm{HCCA}}(\mathbf{Y}_i)$. The key here is the formulation of ϕ_{HCCA}—which is expected to capture the authentic clinical information from free-text reports and subsequently construct the schedule. Accordingly, we decompose HCCA into a hierarchy of metrics, including a baseline metric via report length, a multi-label classification-based metric to represent the pathology density, and a structural metric to indicate structured medical knowledge.

Length-based Control Metric. The most naive proxy for difficulty is report length. Given a free-text report $\mathbf{Y}_i$, this baseline outputs a normalized token-length score d_i^{len}:

$$d_i^{\mathrm{len}} = \mathrm{Norm}\left(|\mathrm{Tokenizer}\left(\mathbf{Y}_i\right)|\right). \tag{6}$$

While sorting free-text reports provides a direct easy-to-hard ordering based on sequence length, it is inherently flawed, since radiology reports normally contain large proportions of descriptive texts. Therefore, longer reports do not strongly correlate to higher pathological complexity.

Multi-Label Classification-based Metric. To ground the difficulty with clinical knowledge, the task of multi-label classification appears to be a natural representation to indicate the pathological attributes through different classified labels. This type of metric models the clinical complexity through the semantic density of pathology attributes. Using an off-the-shelf pathology classifier (e.g., CheXbert (Smit et al., 2020)), this metric extracts a multi-label clinical summary $\mathbf{c}_i \in \{0,1\}^K$ from $\mathbf{Y}_i$, indicating the presence of various abnormalities. The semantic complexity score d_i^{sem} is then computed based on the following equation:

$$d_i^{\mathrm{sem}} = \mathbf{w}_{\mathrm{sem}}^{\top}\mathbf{c}_i. \tag{7}$$

Here, $\mathbf{w}_{\mathrm{sem}}$ denotes the weight vector controlling the relative importance of different pathology labels. This branch significantly upgrades the difficulty definition by measuring exactly *how many* clinically meaningful findings the model is expected to simultaneously align.

Structural Metric from Medical Knowledge Graphs. However, merely counting findings only indicates whether pathological attributes appear, and fails to characterize the structural relations among different clinical entities. To provide more clinically structured information for the training schedule, we use medical knowledge graphs such as RadGraph (Jain et al., 2021) to model difficulty through structural depth. Specifically, this branch parses the free-text report $\mathbf{Y}_i$ into a structured clinical graph $\mathcal{G}_i$ with RadGraph (Jain et al., 2021), summarizes its topological properties into a feature vector $\mathbf{s}_i^{\mathrm{str}}$, and derives the structural complexity score as:

$$d_i^{\mathrm{str}} = \mathbf{w}_{\mathrm{str}}^{\top}\mathbf{s}_i^{\mathrm{str}}. \tag{8}$$

This vector encodes intricate graph statistics, such as the number of anatomical entities, clinical observations, and the frequency of anatomy–observation–uncertainty relational edges. The final HCCA difficulty score combines the semantic and structural branches, while the length-based score is retained only as a control metric:

$$d_i = \mathrm{Norm}\left(\beta_{\mathrm{sem}} d_i^{\mathrm{sem}} + \beta_{\mathrm{str}} d_i^{\mathrm{str}}\right). \tag{9}$$

Here, β_{sem} and β_{str} control the relative contribution of semantic pathology density and structured medical knowledge, respectively. In general, these metrics explicitly translate qualitative pathological features into mathematical formulations of clinical complexity, providing the difficulty signals required to drive the curriculum sampler and adaptive architecture in the subsequent processes.

2.3 DIFFICULTY-AWARE ANCHOR REPLAYING

Once HCCA defines the difficulty scores $\{d_i\}_{i=1}^N$, DAAR translates these static indicators into a dynamic optimization trajectory. Its role is to construct the progressive sampling distribution π_t, effectively guiding the model to reach improved convergence in an easy-to-hard sequence.

Initially, DAAR converts the raw score d_i into a normalized rank percentile $q_i \in (0, 1]$ across the dataset, where this process is defined as: $q_i = \mathrm{RankNorm}(d_i)$. As defined in Eq. (4) of Sec. 2.1, the binary visibility indicator $\alpha_i^{(t)} = \mathbb{I}(q_i \leq \rho_t)$ explicitly constructs an active pool based on the current curriculum coefficient ρ_t. Here, a small ρ_t at the beginning exposes only the pathologically simplest reports; while as ρ_t monotonically increases, more complex cases are progressively offered.

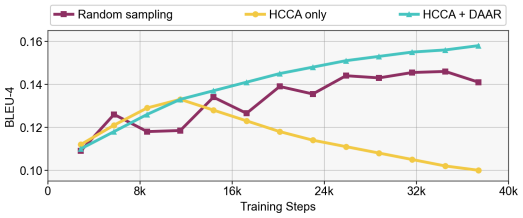


Figure 2: BLEU-4 curves of baseline variants under different sampling strategies, i.e., random sampling (purple), HCCA-only (yellow), and HCCA + DAAR (cyan) to illustrate the motivation of DAAR in Sec. 2.3.

However, adhering strictly to a monotonically expanding active pool introduces a critical vulnerability: catastrophic forgetting. As the training progresses and the model is forced to focus exclusively on highly complex cases, we empirically observe a noticeable catastrophic forgetting issue in the trained model, as shown in Fig. 2. The reason behind might be that, while pathologically simple reports may lack fine-grained clinical intricacies, they serve as indispensable anchors for maintaining foundational vision-language grounding. To mitigate this severe forgetting, DAAR incorporates a strategic anchor replaying mechanism. Let $\mathcal{E} \subset \mathcal{D}$ denote a fixed anchor subset of the easiest samples, with a replay distribution η_i over $\mathcal{E}$. DAAR formulates the sampling distribution $\pi_t(i)$ at step t as:

$$\pi_t(i) = \lambda_{\mathrm{act}} \frac{\alpha_i^{(t)}}{\sum_{j=1}^{N} \alpha_j^{(t)}} + \lambda_{\mathrm{easy}} \eta_i + \lambda_{\mathrm{rand}} \frac{1}{N}. \tag{10}$$

Here, the first term drives the progressive curriculum expansion; the second term explicitly replays easy anchors to consolidate the fundamental cross-modal alignment; and the third term injects a marginal global exploration noise to preserve data diversity.

By drawing mini-batches $\mathcal{B}_t \sim \pi_t$ rather than from random batches, the LLM is optimized along a smooth learning trajectory. The standard objective is thus upgraded to a curriculum-aware equation:

$$\mathcal{L}_{\mathrm{daar}} = -\mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \pi_t} \sum_{\tau=1}^{T_i} \log p_\theta \left(y_{i,\tau} \mid y_{i,<\tau}, \mathbf{Z}_i \right). \tag{11}$$

2.4 Complexity-Adaptive Memory

While DAAR modulates the *external* data distribution, we further introduce CAM to ensure the *internal* network architecture, in order to dynamically adapt to the curriculum difficulty. Given an input radiology image with its visual embedding as $\mathbf{Z}_i$, CAM outputs a complexity-adaptive conditioning state $\widetilde{\mathbf{Z}}_i$. This allows the model to allocate minimal compute for easy cases while unleashing extra representation capacity for complex ones.

To prevent the adaptive mechanism from only relying on arbitrary visual feature magnitudes, CAM introduces an auxiliary structured-summary head. Driven by $\mathbf{Z}_i$, this head predicts a graph-level summary $\widehat{\mathbf{s}}_i^{\mathrm{str}}$ to approximate the ground-truth structural features $\mathbf{s}_i^{\mathrm{str}}$, i.e., the ones extracted by HCCA. This is supervised via an auxiliary loss:

$$\mathcal{L}_{\mathrm{str}} = \ell_{\mathrm{str}} \left(\widehat{\mathbf{s}}_i^{\mathrm{str}}, \mathbf{s}_i^{\mathrm{str}} \right). \tag{12}$$

To couple the network directly with the curriculum scheduler, CAM incorporates the difficulty indicator and pacing parameters of the current curriculum difficulty into a unified state vector. Since reference reports are only available during training, CAM uses the HCCA difficulty from the reference report during training and uses the structured-summary head to estimate difficulty from the visual condition in inference:

$$\begin{aligned} \bar{d}_i &= \begin{cases} d_i, & \text{training}, \\ \widehat{d}_i, & \text{inference}, \end{cases} \\ \widehat{d}_i &= \mathrm{Norm} \left(\mathbf{w}_{\mathrm{str}}^{\top} \widehat{\mathbf{s}}_i^{\mathrm{str}} \right), \\ \bar{q}_i &= \mathrm{RankNorm}_{\mathcal{D}} \left(\bar{d}_i \right). \end{aligned} \tag{13}$$

Table 1: Quantitative comparisons with variant baselines on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) with respect to NLG and CE metrics. Here, we highlight the **best** and second-best results are in **boldfaces** and underlines. Here, † indicates that gains of C^2-GEN over the best baseline are statistically significant ($p < 0.05$).

METHOD	IU X-RAY (Demner-Fushman et al., 2016)						MIMIC-CXR (Johnson et al., 2019)								
	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑
UNIFORM RANDOM SAMPLING	0.472	0.305	0.219	0.166	0.188	0.375	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.468
+ HCCA	0.458	0.294	0.208	0.156	0.182	0.367	0.359	0.236	0.162	0.114	0.160	0.255	0.442	0.456	0.449
+ HCCA + DAAR	0.499	0.323	0.238	0.184	0.208	0.390	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473
+ HCCA + DAAR + CAM (C^2-GEN)†	**0.526**	**0.338**	**0.247**	**0.204**	**0.214**	**0.418**	**0.431**	**0.271**	**0.190**	**0.154**	**0.181**	**0.339**	**0.621**	**0.617**	**0.619**

Here, $\mathrm{RankNorm}_{\mathcal{D}}$ maps the estimated difficulty to the rank statistics of the training set. During training, we set $\bar{\rho} = \rho_t$ and $\bar{\alpha}_i = \alpha_i^{(t)}$; during inference, we set $\bar{\rho} = 1$ and $\bar{\alpha}_i = 1$, so that CAM remains difficulty-aware without requiring a reference report. The unified curriculum state becomes:

$$\mathbf{u}_i^{(t)} = \left[\bar{d}_i,\, \bar{q}_i,\, \bar{\rho},\, \bar{\alpha}_i\right]. \tag{14}$$

Leveraging the visual feature, the curriculum state, and the structural summary, CAM predicts a dynamic memory gate $\mathbf{g}_i \in [0, 1]$ via:

$$\mathbf{g}_i = \sigma\left(\mathrm{MLP}\left(\left[\mathrm{pool}\left(\mathbf{Z}_i\right), \mathbf{u}_i^{(t)}, \widehat{\mathbf{s}}_i^{\mathrm{str}}\right]\right)\right), \tag{15}$$

where this gate actively modulates the integration of learnable key-value memory banks $(\mathbf{K}_{\mathrm{mem}}, \mathbf{V}_{\mathrm{mem}})$ into the standard multi-head attention layers of the LLM:

$$\widetilde{\mathbf{Z}}_i = \mathbf{Z}_i + \mathbf{g}_i \odot \mathrm{Attn}\left(\mathbf{Z}_i, \mathbf{K}_{\mathrm{mem}}, \mathbf{V}_{\mathrm{mem}}\right). \tag{16}$$

Through this architectural design, we enable the LLM to leverage $\widetilde{\mathbf{Z}}_i$ and become explicitly curriculum-aware. When generating a highly complex report (high d_i), the gate activates heavily, routing rich auxiliary memory to support the attention layers of the LLM. Conversely, for simple anchors, the gate suppresses the memory and lets the LLM output through its basic capability. Ultimately, by integrating the DAAR sampler and the CAM architectural enhancements, we construct the final optimization objective of the C^2-Gen framework based on Eq. (3), formulated by:

$$\mathcal{L}_{\mathrm{full}} = \mathcal{L}_{\mathrm{daar}} + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{mem}}\mathcal{L}_{\mathrm{mem}}. \tag{17}$$

Here, $\mathcal{L}_{\mathrm{mem}} = \mathbb{E}_{i\sim\mathcal{B}_t} \|\mathbf{g}_i\|_1$ is a gate sparsity regularizer that stabilizes the memory-augmented attention mechanism.

3 EXPERIMENTS

In this section, we first evaluate different curriculum-driven baselines to isolate the effect of each component, then compare C^2-GEN with state-of-the-art radiology report generation methods, and finally provide a qualitative case study and ablation studies to intuitively analyze clinical correctness and component-level contributions. Notably, we provide detailed experimental settings in Sec. B of our supplementary materials, including the used datasets, evaluation metrics, and implementation details of C^2-GEN.

3.1 COMPARISON WITH BASELINES

Tab. 1 reports the quantitative comparison with several baselines. Specifically, we consider 4 baseline variants of our full model, where the first (UNIFORM RANDOM SAMPLING) employs the standard uniform random sampling as most existing methods do; the second (+HCCA) is equipped with our HCCA, meaning the model learns strictly from a re-ordered easy-to-hard sequence, rather than a random one; the third (+HCCA+DAAR) further integrates the DAAR into HCCA to stabilize the model training; the fourth (+HCCA+DAAR+CAM) represents our full model, with all components incorporated. From the results, we observe that solely integrating HCCA conversely leads

Table 2: Quantitative comparisons with state-of-the-art methods on IU X-RAY (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) with respect to NLG and CE metrics. Note that we illustrate the number of overall model parameters of each LLM-based method in parentheses.

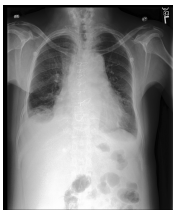
METHOD	IU X-RAY (Demner-Fushman et al., 2016)						MIMIC-CXR (Johnson et al., 2019)								
	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑
EARLY IMAGE CAPTIONING METHODS															
ST (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	–	0.306	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.224
ATT2IN (Rennie et al., 2017)	0.224	0.129	0.089	0.068	–	0.308	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.274
ADAATT (Lu et al., 2017)	0.220	0.127	0.089	0.068	–	0.308	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.220
COATT (Jing et al., 2018)	0.455	0.288	0.205	0.154	–	0.369	–	–	–	–	–	–	–	–	–
HRGR (Liu et al., 2024)	0.438	0.298	0.208	0.151	–	0.322	–	–	–	–	–	–	–	–	–
CMAS-RL (Liu et al., 2024)	0.464	0.301	0.210	0.154	–	0.362	–	–	–	–	–	–	–	–	–
TOPDOWN (Liu et al., 2024)	–	–	–	–	–	–	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.268
FROM-SCRATCH RRG METHODS															
R2GEN (Chen et al., 2020)	0.470	0.304	0.219	0.165	–	0.371	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.300
CA (Liu et al., 2021c)	0.492	0.314	0.222	0.169	0.193	0.381	0.350	0.219	0.152	0.109	0.151	0.283	–	–	–
PPKED (Liu et al., 2021b)	0.483	0.315	0.224	0.168	–	0.376	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
R2GENCMN (Chen et al., 2021)	0.475	0.309	0.222	0.170	0.191	0.375	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.302
R2GENRL (Liu et al., 2024)	0.494	0.321	0.235	0.181	0.201	0.384	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.316
ITA (Liu et al., 2024)	–	–	–	–	–	–	0.395	0.253	0.170	0.121	0.147	0.284	–	–	–
WARMSTART (Liu et al., 2024)	–	–	–	–	–	–	0.392	0.245	0.169	0.124	0.153	0.285	0.359	0.412	0.384
KIUT (Huang et al., 2023)	–	–	–	–	–	–	0.393	0.243	0.159	0.113	–	0.285	0.371	0.318	0.342
RGRG (Tanida et al., 2023)	–	–	–	–	–	–	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.468
PROMPTMRG (Jin et al., 2024)	–	–	–	–	–	–	0.398	–	–	0.112	–	0.258	0.501	0.509	0.505
LARGE LANGUAGE MODEL-BASED METHODS															
XRAYGPT (7B) (Thawakar et al., 2024)	0.177	0.104	0.047	0.007	0.105	0.203	0.128	0.045	0.014	0.004	0.079	0.111	–	–	–
R2GENGPT (7B) (Wang et al., 2023)	–	–	–	–	–	–	0.396	–	–	0.113	–	0.273	0.506	0.414	0.456
CHEXAGENT (7B) (Chen et al., 2024b)	–	–	–	–	–	–	0.189	–	–	0.040	–	0.208	0.506	0.306	0.381
MED-PALM (562B) (Tu et al., 2023)	–	–	–	–	–	–	0.317	–	–	0.115	–	0.275	–	–	0.378
MAIRA-1 (7B) (Hyland et al., 2023)	–	–	–	–	–	–	0.392	–	–	0.142	–	0.289	–	–	0.553
EKAGEN (Bu et al., 2024a)	**0.526**	**0.361**	**0.267**	0.203	**0.214**	0.404	0.419	0.258	0.170	0.119	0.157	0.287	0.517	0.483	0.499
R2-LLM (14.2B) (Liu et al., 2024)	0.499	0.323	0.238	0.184	0.208	0.390	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473
INVERGE (7B) (Deria et al., 2024)	–	–	–	–	–	–	0.425	0.240	0.132	0.100	–	0.309	–	–	–
LLM-RG4 (7B) (Wang et al., 2025)	–	–	–	–	–	–	0.377	–	–	0.144	–	0.318	0.583	0.593	0.588
S2D-ALIGN (7B) (Gao et al., 2026)	0.512	–	–	0.195	–	0.407	0.422	0.263	0.183	0.149	–	0.332	0.613	0.606	0.608
CURRICULUM LEARNING-BASED METHODS															
CMCL (Liu et al., 2021a)	0.473	0.305	0.217	0.162	0.186	0.378	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
C^2-GEN† (3B)	**0.526**	0.338	0.247	**0.204**	**0.214**	**0.418**	**0.431**	**0.271**	**0.190**	**0.154**	**0.181**	**0.339**	**0.621**	**0.617**	**0.619**

to performance degradation in all metrics. This demonstrates a behavior consistent with Fig. 2, indicating that a naive difficulty-aware ordering is insufficient; it inadvertently amplifies catastrophic forgetting, causing the model to discard foundational cross-modal alignments as it attempts to handle harder inputs. Furthermore, the baseline with both HCCA and DAAR equipped shows substantial improvements. It is clearly demonstrated that the replay mechanism is the key principle that transforms difficulty-aware scheduling from an unstable sequence into a robust training paradigm. Finally, our full model yields the best performance, showing the performance gains are realized not just via data sampling order, but by ensuring the LLM adapts its internal capacity to the clinical complexity of each specific case.

3.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Tab. 2 compares our proposed C^2-GEN with 3 types of main existing methods, i.e., early image captioning methods, from-scratch RRG methods, and LLM-based methods. Additionally, we explicitly isolate CMCL into a distinct category—curriculum learning-based methods—given its direct relevance to our work. There are several observations from the results. First, early image captioning models and from-scratch RRG methods inherently struggle to achieve competitive performance. This limitation primarily stems from their reliance on learning constrained cross-modal mappings strictly within closed-set datasets. LLM-based methods largely mitigate this bottleneck, illustrating substantial overall improvements. However, while most LLM-based pipelines achieve comparable NLG scores, their performance on the CE metric remains relatively sub-optimal, indicating a vulnerability to clinical hallucinations as is discussed in Sec. 1.

Table 3: A selected case study from the official testing split of MIMIC-CXR (Johnson et al., 2019). Descriptive phrases corresponding to different pathological attributes are highlighted with different colors. Underlined green phrases indicate correctly captured key attributes, while red phrases indicate incorrect diagnoses.

Radiograph	Reports
	**Ground-truth report:** Interval increase/development of **bilateral**, right greater than left, pleural effusions with overlying atelectasis. Right base opacity may be due to a combination of pleural effusion and atelectasis; underlying consolidation cannot be excluded. Cardiomegaly and minimal pulmonary vascular congestion are present.
	UNIFORM RANDOM SAMPLING: The heart size is normal. Mild bibasilar atelectasis is present. No pleural effusion or pneumothorax is seen. No focal consolidation or edema is identified.
	+ HCCA: There is bibasilar atelectasis with hazy right basilar opacity. A trace right pleural effusion is suspected. The cardiac silhouette is mildly enlarged, while no vascular congestion is described.
	+ HCCA + DAAR: **Bilateral** pleural effusions with adjacent basilar atelectasis are observed. A right basilar opacity is present, likely related to pleural fluid and atelectatic change. The cardiac silhouette is enlarged. Superimposed consolidation and vascular congestion are not clearly reported.
	FULL MODEL (C^2-GEN): **Bilateral** pleural effusions are present. There is adjacent basilar atelectatic change. A hazy right basilar opacity is present, and superimposed consolidation remains possible. The cardiac silhouette is enlarged, with minimal pulmonary vascular congestion.

Particularly, S2D-ALIGN (Gao et al., 2026) emerges as the most competitive LLM-based method. Yet, it relies on the conventional random sampling paradigm, it fails to fully resolve the fine-grained pathology alignment issue. In contrast, C^2-GEN consistently outperforms this method across both NLG and CE metrics, validating that our clinically structured curriculum learning schedule effectively guides the model to capture accurate pathological attributes. Most notably, we notice that existing LLM-based methods often adopt computationally heavy LLMs, whose model scales range from 7B to 562B (e.g., Med-PaLM (Tu et al., 2023)), whereas we only use a lightweight LLM backbone with 3B. This direct efficiency comparison thoroughly illustrates the proficiency of C^2-GEN in data utilization and dynamic network capacity allocation, proving the great potentials of a difficulty-aware curriculum to maximize the clinical reasoning capability of LLMs. Finally, a direct comparison with existing curriculum-based RRG method, i.e., CMCL, confirms the necessity of clinically grounded difficulty metrics. While CMCL serves as the closest prior work to ours, it exhibits a substantial performance gap compared to our method across both datasets. This gap is particularly significant on MIMIC-CXR, which is characterized by highly complex and entangled pathological attributes. The fundamental reason is that CMCL relies on a general schedule, which is not sufficiently fine-grained to accurately reflect the clinical complexity of free-text reports. In contrast, the HCCA provides precise difficulty signals to unlock the full potential of the curriculum learning.

3.3 CASE STUDY

Tab. 3 presents a selected example from the testing split of MIMIC-CXR (Johnson et al., 2019). The qualitative trend is consistent with Tab. 1, where uniform random sampling produces a fluent but clinically incomplete report. +HCCA increases the number of captured pathological attributes, but the strictly ordered curriculum still causes inaccurate severity and missing vascular descriptions. After DAAR introduces anchor replaying, the model recovers most major findings, yet it still omits the uncertain consolidation and vascular congestion. In contrast, the full C^2-GEN report covers all highlighted pathological attributes, where all aforementioned issues are alleviated.

3.4 ABLATION STUDIES

Tab. 4 reports three ablations on MIMIC-CXR, validating the core components of C^2-GEN through various variants, where we analyze their corresponding behaviors as follows.

Table 4: Quantitative results for the ablation studies of different components, i.e., (a) the difficulty metric setup in HCCA, (b) the sampler design in DAAR, and (c) the memory design in CAM, with all experimental results conducted on MIMIC-CXR with respect to the BLEU-4 score and the F1 score of the CE metric.

METRICS	B@4↑	F1↑
RANDOM SAMPLING	0.126	0.468
REPORT LENGTH	0.121	0.434
CHEXBERT TERMS	0.136	0.468
RADGRAPH	0.139	0.479
HCCA	**0.142**	**0.486**

(a) Ablation of the difficulty metrics.

SAMPLER	B@4↑	F1↑
RANDOM SAMPLING	0.126	0.468
ORDERED SCAN ONLY	0.114	0.418
ACTIVE POOL ONLY	0.121	0.432
RANDOM REPLAY	0.138	0.472
DAAR	**0.142**	**0.486**

(b) Ablation of the DAAR sampler.

VARIANTS	B@4↑	F1↑
RADGRAPH	0.134	0.463
+ SUMMARY CONSISTENCY	0.137	0.472
$+\mathbf{g}_i$ IN EQ. (15)	0.140	0.480
$+\widetilde{\mathbf{Z}}_i$ IN EQ. (16)	**0.142**	**0.486**

(c) Ablation of the CAM design.

Effect of the Difficulty Metric in HCCA. In Tab. 4a, we replace HCCA with difficulty metrics of different granularity. Among them, report length serves as the most natural proxy, yet it performs even worse than random sampling. This result is consistent with the highly templated nature of radiology reports, indicating that a longer report does not necessarily indicate richer pathological attributes, because the additional contents might contain descriptive texts rather than from effective clinical findings. In contrast, using CHEXBERT (Smit et al., 2020) terms and RADGRAPH (Jain et al., 2021) both outperform random sampling, and the full HCCA achieves the best results by jointly modeling the semantic density and structured knowledge.

Effect of the Sampler in DAAR. Tab. 4b explores different samplers for DAAR. Both ORDERED SCAN ONLY and ACTIVE POOL ONLY remain below random sampling, which is consistent with the performance trend shown in Fig. 2. This indicates that a monotonic easy-to-hard schedule might cause catastrophic forgetting of the easy samples that stabilize the basic cross-modal alignment. By comparing RANDOM REPLAY to these baselines, the performance improves substantially, which supports the necessity of replay in the curriculum schedule. However, DAAR still yields further gains because its replay is more fine-grained. It explicitly preserves easy anchors for stable image-report grounding, whereas random replaying might sample middle or hard cases and provides a weaker refinement to forgetting.

Effect of the Memory Design in CAM. Finally, Tab. 4c analyzes the design of CAM. Starting from the RadGraph-based curriculum objective, adding summary consistency already improves both lexical coherence (BLEU-4) and clinical accuracy (F1), which shows that the optimization target benefits from stronger report-level structural supervision. Building on this objective, introducing the gating variable $\mathbf{g}_i$ in Eq. (15) brings further gain, indicating that the network benefits from explicitly filtering the conditioning pathway with curriculum-aware signals. Replacing this partial gating with the full adaptive state $\widetilde{\mathbf{Z}}_i$ in Eq. (16) leads to the best performance, confirming that the role of CAM is to make the LLM itself difficulty-aware.

4 CONCLUSION

In this paper, we propose a curriculum-driven framework for LLM-based RRG, namely C^2-GEN. Specifically in our proposed method, HCCA translates the pathological complexities of doctor-written radiology reports through difficulty metrics; DAAR then serves as a replay-aware progressive sampler to re-order the overall data; and CAM further injects the reformulated data sequence into the LLM to make it difficulty-aware. Extensive experiments on two benchmark datasets validate the effectiveness of our method, along with comprehensive ablations analyzing the internal impacts of different components. The results reveal that curriculum learning benefits RRG only when clinically grounded difficulty signals are paired with replay-aware scheduling and difficulty-aware architecture adaptation. However, due to the limitation of available data modality, the current framework is only validated on chest X-ray radiographs. Future work should extend clinically structured curriculum learning to more diverse scenarios, where the findings of this work can serve as a reference for follow-up studies.

AI DISCLOSURE

In this work, we used generative AI tools (OpenAI Codex, powered by GPT-5.5) for implementing experiment code. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims or proofs, proposing or refining hypotheses, designing the research methodology or experiments, assisting with translation, cleaning or reformatting datasets, supporting qualitative data analysis, or interpreting results, and the remaining tasks requiring disclosure under the ICLR 2027 AI Policy for Authors are not applicable to this work. Additionally, we used generative AI tools for editing software code and editing the manuscript to improve readability. All AI-assisted work was reviewed by the authors: AI-generated code was executed and its outputs validated, and AI-edited text was checked for correctness. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65–72, 2005.

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. MAIRA-2: Grounded Radiology Report Generation. *arXiv*, abs/2406.04449, 2024.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *ICML*, pp. 41–48, 2009.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models Are Few-Shot Learners. In *NeurIPS*, 2020.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation. In *CVPR*, pp. 14194–14204, 2024a.

Shenshen Bu, Yujie Song, Taiji Li, and Zhiming Dai. Dynamic Knowledge Prompt for Chest X-ray Report Generation. In *LREC/COLING*, pp. 5425–5436, 2024b.

Yiming Cao, Lizhen Cui, Fuqiang Yu, Lei Zhang, Zhen Li, Ning Liu, and Yonghui Xu. KdTNet: Medical Image Report Generation via Knowledge-Driven Transformer. In *DASFAA*, pp. 117–132, 2022.

Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. Act Like a Radiologist: Radiology Report Generation Across Anatomical Regions. In *ACCV*, pp. 1–17, 2024a.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating Radiology Reports via Memory-Driven Transformer. In *EMNLP*, pp. 1439–1449, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-Modal Memory Networks for Radiology Report Generation. In *ACL-IJCNLP*, pp. 5904–5914, 2021.

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. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv*, abs/2401.12208, 2024b.

Jean-Benoit Delbrouck, Pierre J. Chambon, Christian Bluethgen, Emily Bao Tsai, Omar Almusa, and Curtis P. Langlotz. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In *EMNLP (Findings)*, pp. 4348–4360, 2022.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Pontes Reis, Mohamed S. Muneer, Maya Varma, and Curtis P. Langlotz. Automated Structured Radiology Report Generation. In *ACL*, pp. 26813–26829, 2025.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Medical Informatics Assoc.*, 23 (2):304–310, 2016.

Ankan Deria, Komal Kumar, Snehashis Chakraborty, Dwarikanath Mahapatra, and Sudipta Roy. InVERGe: Intelligent Visual Encoder for Bridging Modalities in Report Generation. In *CVPR Workshops*, pp. 2028–2038, 2024.

Jiechao Gao, Chang Liu, and Yuangang Li. S2D-Align: Shallow-to-Deep Auxiliary Learning for Anatomically-Grounded Radiology Report Generation. In *AAAI*, pp. 30780–30788, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *CVPR*, pp. 19809–19818, 2023.

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. MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation. *arXiv*, abs/2311.13668, 2023.

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. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *NeurIPS Datasets and Benchmarks*, 2021.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-Paced Curriculum Learning. In *AAAI*, pp. 2694–2700, 2015.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In *AAAI*, pp. 2607–2615, 2024.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the Automatic Generation of Medical Imaging Reports. In *ACL*, pp. 2577–2586, 2018.

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. MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv*, abs/1901.07042, 2019.

Navdeep Kaur and Ajay Mittal. CADxReport: Chest X-Ray Report Generation Using Co-Attention Mechanism and Reinforcement Learning. *Comput. Biol. Medicine*, 145:105498, 2022.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-Paced Learning for Latent Variable Models. In *NeurIPS*, pp. 1189–1197, 2010.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *NeurIPS*, 2023.

Xiao Liang, Yanlei Zhang, Di Wang, Haodi Zhong, Ronghan Li, and Quan Wang. Divide and Conquer: Isolating Normal-Abnormal Attributes in Knowledge Graph-Enhanced Radiology Report Generation. In *ACM MM*, pp. 4967–4975, 2024.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping Large Language Models for Radiology Report Generation. In *AAAI*, pp. 18635–18643, 2024.

Fenglin Liu, Shen Ge, and Xian Wu. Competence-Based Multimodal Curriculum Learning for Medical Report Generation. In *ACL-IJCNLP*, pp. 3001–3012, 2021a.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *CVPR*, pp. 13753–13762, 2021b.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive Attention for Automatic Chest X-ray Report Generation. In *ACL-IJCNLP (Findings)*, pp. 269–280, 2021c.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation. In *AAAI*, pp. 5595–5603, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *CVPR*, pp. 3242–3250, 2017.

Microsoft. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv*, abs/2404.14219, 2024.

Johannes Moll, Louisa Fay, Asfandyar Azhar, Sophie Ostmeier, Sergios Gatidis, Tim C. Lueth, Curtis Langlotz, and Jean-Benoit Delbrouck. Structuring Radiology Reports: Challenging LLMs with Lightweight Models. In *EMNLP*, pp. 7707–7724, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*, pp. 311–318, 2002.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision. *arXiv*, abs/2401.10815, 2024.

Qwen Team. Qwen3 Technical Report. *arXiv*, abs/2505.09388, 2025.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, pp. 7008–7024, 2017.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *EMNLP*, pp. 1500–1519, 2020.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*, pp. 7433–7442, 2023.

Omkar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, Jorma Laaksonen, and Fahad Khan. XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models. In *BioNLP@ACL*, pp. 440–448, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv*, abs/2302.13971, 2023.

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, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards Generalist Biomedical AI. *arXiv*, abs/2307.14334, 2023.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Manuel Zambrano Chaves, Curtis P. Langlotz, Akshay Chaudhari, and John M. Pauly. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. In *BioNLP*, pp. 449–460, 2023.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pp. 3156–3164, 2015.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology Report Generation with Frozen LLMs. *arXiv*, abs/2309.09812, 2023.

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. LLM-RG4: Flexible and Factual Radiology Report Generation Across Diverse Input Contexts. In *AAAI*, pp. 8250–8258, 2025.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology Report Generation via Multi-objective Preference Optimization. In *AAAI*, pp. 8664–8672, 2025a.

Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Zhe Wang, and Chenjia Bai. Online Iterative Self-Alignment for Radiology Report Generation. In *ACL*, pp. 27799–27814, 2025b.

Dexuan Xu, Yanyuan Chen, Jiayu Zhang, Yiwei Lou, Hanpin Wang, Jing He, and Yu Huang. Radiology Report Generation via Structured Knowledge-Enhanced Multi-Modal Attention and Contrastive Learning. In *BIBM*, pp. 2320–2325, 2023.

Sixing Yan. Memory-aligned Knowledge Graph for Clinically Accurate Radiology Image Report Generation. In *BioNLP@ACL*, pp. 116–122, 2022.

Longzhen Yang, Zhangkai Ni, Ying Wen, Yihang Liu, Lianghua He, and Heng Tao Shen. Self-Supervised Anatomical Consistency Learning for Vision-Grounded Medical Report Generation. In *ACM MM*, pp. 2958–2967, 2025.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Knowledge Matters: Chest Radiology Report Generation with General and Specific Knowledge. *Med. Image Anal.*, 80:102510, 2022.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology Report Generation with a Learned Knowledge Base and Multi-Modal Alignment. *Med. Image Anal.*, 86: 102798, 2023.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. When Radiology Report Generation Meets Knowledge Graph. In *AAAI*, pp. 12910–12917, 2020.

A Related Work

Radiology Report Generation. Early radiology report generation methods primarily build task-specific encoder-decoder pipelines to convert chest X-ray images into free-text reports. These methods focus on establishing effective radiograph-report alignment through cross-modal memory (Chen et al., 2020; 2021), medical knowledge injection (Cao et al., 2022), attention mechanisms (Kaur & Mittal, 2022), and reinforcement learning. However, these pre-LLM methods inherently learn constrained cross-modal mappings strictly within closed-set data, severely limiting their generalizability to real-world clinical scenarios. To overcome this bottleneck, recent studies attempt to integrate the strong text generation capability of LLMs into the RRG task, where they mostly focus on how to effectively perform a general-to-medical domain adaptation. For instance, LLaVA-Med (Li et al., 2023), Bootstrapping LLMs (Liu et al., 2024), and RadAdapt (Van Veen et al., 2023) demonstrate general-domain vision-language models can be adapted to the medical domain. Subsequent works, e.g., MAIRA-2 (Bannur et al., 2024) and X-RGen (Chen et al., 2024a), continue to push the boundaries of grounded, multimodal clinical reporting. Furthermore, recent studies have explored iterative self-alignment (Xiao et al., 2025b), multi-objective preference optimization (Xiao et al., 2025a), automated structured report generation (Delbrouck et al., 2025), shallow-to-deep auxiliary supervised fine-tuning (Gao et al., 2026), and lightweight model deployment (Moll et al., 2025) for improving factual and structured radiology reports. While LLM-based methods significantly boost the quality of generated reports, the standard training recipe of all aforementioned methods still relies on standard random mini-batch sampling. This practice indiscriminately mixes simple and complex cases, failing to address the inherent mismatch in pathology difficulty and often leading to severe clinical hallucinations, as shown by the experiments in the main paper.

Structured Clinical Supervision and Curriculum Learning. Early methods such as MKG (Zhang et al., 2020) and MaKG (Yan, 2022) demonstrate that external knowledge graphs help models explicitly organize abnormal findings and clinical relations. Subsequent works, including GSKEMHA (Yang et al., 2022), M2KT (Yang et al., 2023), and SKMACL (Xu et al., 2023), further verify that structured knowledge significantly improves radiograph-report alignment. More recently, methods like DKP (Bu et al., 2024b), EKAGen (Bu et al., 2024a), and DCG (Liang et al., 2024) continue this trajectory by explicitly organizing knowledge around report prompts, expert knowledge, or normal-abnormal attributes. Parallel to knowledge graphs, works leveraging structured pathology labels and anatomical signals—such as Semantic Rewards (Delbrouck et al., 2022), CheXbert (Smit et al., 2020), RadGraph (Jain et al., 2021), and SS-ACL (Yang et al., 2025)—confirm that clinical entities reveal critical fine-grained information that plain text overlap cannot capture. Despite their progress in improving report factuality, most of these methods predominantly utilize structured signals as auxiliary supervision or direct feature injection into the generator. Crucially, they do not explicitly use these clinical structures to quantify the varying degrees of clinical difficulty, and therefore fail to systematically guide the optimization trajectory across these difficulty levels. To explicitly address the optimization order, Curriculum Learning (CL) (Bengio et al., 2009) offers a principled paradigm by exposing the learner to training examples in a progressive, easy-to-hard sequence. This concept has been further refined by Self-Paced Learning (SPL) (Kumar et al., 2010) and Self-Paced Curriculum Learning (SPCL) (Jiang et al., 2015), which dynamically adjust the learning pace based on the evolving competence of the trained model. Within this context, the closest prior work to ours is CMCL (Liu et al., 2021a), which also analyzes that training order matters. Yet, it utilizes a general competence schedule rather than explicit clinical structures. In contrast, we formulate a clinically-grounded curriculum, specifically designed to unlock the full potential of LLM-based report generators.

B Experimental Settings

Datasets. We conduct experiments on IU X-Ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) following conventional studies (Wang et al., 2025; Liu et al., 2025; Gao et al., 2026). Specifically, IU X-Ray is a chest radiograph benchmark collected from the Indiana University hospital system, and it contains 3,955 reports paired with 7,470 frontal and lateral images. MIMIC-CXR is a large-scale chest X-ray benchmark released from Beth Israel Deaconess Medical Center through PhysioNet, and it contains 377,110 images from 227,827 imaging studies of 65,379 patients. Following the standard radiology report generation protocol, we use the standard split

protocol adopted by prior studies for IU X-Ray and the official train, validation, and test split for MIMIC-CXR. Importantly, we only construct curriculum- and difficulty-related annotations on the training split and we keep the validation and test splits untouched. After applying the standard RRG preprocessing protocol, the IU X-Ray split contains 2,069 training, 296 validation, and 590 test report–study pairs. The corresponding MIMIC-CXR split contains 270,790 training, 2,130 validation, and 3,858 test report–study pairs.

Evaluation Metrics. Following existing studies (Chen et al., 2021; 2020; Cao et al., 2022; Huang et al., 2023; Liu et al., 2024; Wang et al., 2025; Gao et al., 2026), we use two types of metrics, i.e., Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics. Specifically, the NLG metric mainly measures lexical coherence via comparing the n-gram overlapping between the generated and ground-truth reports, where we use BLEU-n with $n = \{1, 2, 3, 4\}$ (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Lin, 2004). The CE metric evaluates the accuracies of capturing medical classification labels in the generated reports, where we use CheXbert (Smit et al., 2020) as the label extractor and compute precision, recall, and F1 for the final scores.

Implementation Details. For our full model, we conduct experiments with various hyperparameters on the validation set, and choose the best-performing setting for our final model. In detail, we initialize the LLM with `Phi-3-mini-4k-instruct` (Microsoft, 2024) and use an off-the-shelf visual encoder RAD-DINO (Pérez-García et al., 2024). We freeze the visual encoder and optimize the trainable connector together with LoRA adapters (Hu et al., 2022) on the LLM. To achieve the best performance, we set the rank, alpha, and dropout rate of the LoRA weights to 32, 64, and 0.1, respectively. For DAAR, we set $\rho_0 = 0.2$, T_{warmup} to 1 epoch, and $(\lambda_{\text{act}}, \lambda_{\text{easy}}, \lambda_{\text{rand}}) = (0.7, 0.2, 0.1)$. For CAM, we set $\lambda_{\text{str}} = 0.2$ and $\lambda_{\text{mem}} = 0.1$. In training, we use AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 3×10^{-4}, a batch size of 6, and a cosine learning rate schedule. All experiments are executed on a single NVIDIA A40 GPU. For significance testing, we perform paired bootstrap resampling over test studies with 1,000 resamples and regard $p < 0.05$ as statistically significant.