

CONDISTILL: Asymmetric Context Self-Distillation 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

Building fine-grained cross-modal alignment via pure radiograph-report pairs has been a long-standing challenge in the task of Radiology Report Generation (RRG). An intuitive solution is to introduce auxiliary contexts into report generation, yet such clinical contexts are normally unavailable in most real-world deployment scenarios. In this paper, we present an asymmetric context self-distillation framework for RRG, namely CONDISTILL. To facilitate radiograph-only report generation, it refines basic radiograph-report alignment by progressively injecting additional contexts through a multi-stage curriculum. Since these contexts are unavailable at inference, we develop a self-distillation strategy that transfers bridge states from context-aware teacher branches to a student conditioned only on radiographs. For evaluation, we conduct extensive experiments on both radiograph-only and multi-context benchmarks, i.e., MIMIC-CXR and MIMIC-RG4, respectively. The results demonstrate consistent improvements, while further ablations clarify how LLM-based RRG models learn from and distill heterogeneous contexts.

1 Introduction

RRG aims to automatically generate free-text radiology reports from input radiographs. In clinical practice, this task can reduce radiologists' repetitive documentation burden and support image-based diagnosis. Unlike a short image caption, a chest X-ray report covers multiple clinical aspects, including opacities, edema, pleural effusion, cardiomeastinal findings, and support devices. A reliable report generation model therefore needs dense and fine-grained cross-modal alignment that converts pathological evidence in radiographs into explicit textual descriptions.

To perform this task, the development of existing methods can be categorized into two stages. In the pre-LLM stage, task-specific encoder-decoder models are trained from scratch to encode radiographs into implicit visual representations and decode them into report tokens, as exemplified by TIENET (Wang et al., 2018) and R2GEN (Chen et al., 2020). To enhance the cross-modal alignment of the model, subsequent studies introduce memory modules (Chen et al., 2021), knowledge graphs (Zhang et al., 2020; Liu et al., 2021b), reinforcement learning (Nguyen et al., 2021), and regional grounding (Tanida et al., 2023). Nevertheless, these methods still struggle to effectively align localized image evidence with long clinical narratives due to the under-constrained supervision in standard image-report pairs. Normally, subtle abnormalities are only sparsely described, while explicit region-phrase correspondences are rarely available, so report statements often depend on highly contextualized cues rather than only the radiograph input. With Large Language Models (LLMs) demonstrating strong generation ability, recent studies (Wang et al., 2023b; Liu et al., 2024; Bannur et al., 2024; Wang et al., 2025; Li et al., 2024b; Gao et al., 2026) adapt general-domain LLMs to RRG. However, the introduction of LLMs also causes significant language prior bias and hallucination (Favero et al., 2024; Kapadnis et al., 2024; Wang et al., 2025; Artsi et al., 2025), where clinical phrases may be produced without reliable visual-textual grounding. To alleviate this issue, we draw inspiration from real clinical reporting, where radiologists often consult various contextual sources while writing diagnostic reports, e.g., prior diagnoses, multi-view radiographs, historical reports, and categorical clinical labels. Directly adding these contexts can improve report reliability, as shown by longitudinal RRG studies (Wang et al., 2024; Liu et al., 2025b,a; Yun et al., 2025), yet the required inputs are often unavailable, incomplete,

*Equal contribution.

†Corresponding author.

or disallowed at inference. Existing radiograph-only methods fail to internalize such supervision, whereas context-assisted methods typically require dedicated context inputs; this gap motivates our work in this paper.¹

In this paper, we propose CONDISTILL, an asymmetric context self-distillation framework for RRG, with two core components, namely Progressive Context Curriculum (PCC) and Asymmetric Context Self-Distillation (ACSD). Specifically, PCC organizes training into a sequential coarse-to-fine alignment process, where it first learns the basic radiograph-report alignment in a standard radiograph-only setting, then grounds the mapping with instance-level contexts, e.g., multi-view radiographs and historical reports, and eventually refines the cross-modal alignment with structured medical knowledge. ACSD addresses train-test misalignment through an asymmetric architecture: a teacher branch integrates the available training contexts, while their routed representations are distilled into a radiograph-only student. At inference, the teacher context branch is removed and the student generates reports from radiograph-only input. We evaluate radiograph-only inference on MIMIC-CXR (Johnson et al., 2019) and multi-context inference on its contextualized variant MIMIC-RG4 (Wang et al., 2025). Comparisons with existing state-of-the-art methods validate the effectiveness of the asymmetric architecture in our method. Further ablations analyze how the MLLM aligns heterogeneous contexts and transfers their information to the radiograph-only pathway².

2 Method

Fig. 1 presents the overall pipeline of CONDISTILL. In general, CONDISTILL uses PCC to learn coarse-to-fine alignment from progressively injected contexts and ACSD to distill this alignment into a learnable MEMORY BANK, which serves as the context proxy for radiograph-only inference. In the following sections, we sequentially detail PCC and ACSD with the corresponding math expressions.

2.1 Progressive Context Curriculum

To gradually construct a fine-grained cross-modal alignment for report generation, PCC progressively expands the context sets in the teacher branches in

addition to the standard radiograph-only settings. Formally, we define the available context set $\Omega_i^{(j)}$ at the j -th stage as the following expressions:

$$\begin{aligned}\Omega_i^{(1)} &= \{\mathcal{U}_i\}, \\ \Omega_i^{(2)} &= \{\mathcal{U}_i, \mathcal{L}_i, \mathcal{H}_i\}, \\ \Omega_i^{(3)} &= \{\mathcal{U}_i, \mathcal{L}_i, \mathcal{H}_i, \mathcal{G}_i\}.\end{aligned}\tag{1}$$

Here, $\mathcal{U}_i$, $\mathcal{L}_i$, $\mathcal{H}_i$, and $\mathcal{G}_i$ denote the frontal radiograph, lateral radiograph, history report, and structured knowledge graph of the i -th case, respectively. Note that $\mathcal{U}_i$ also serves as the only input for the standard RRG setting, i.e., radiograph-only generation. For each context $\mathcal{X}_i^{(j)} \in \Omega_i^{(j)}$, PCC encodes it according to the prescribed order to progressively refine the alignment from coarse to fine-grained guidance. Particularly, we consider instance-level supervision, e.g., lateral-view radiology images and historical reports, as additional auxiliary information for the radiograph-only setting. On top of this, we further investigate meticulous clinical cues with an off-the-shelf medical labeler (i.e., RadGraph (Jain et al., 2021)) to extract structured annotations from the ground-truth free-text reports. These extracted knowledge graphs are converted into textual narratives from their nodes and relations to highlight critical pathological attributes expressed in the target reports. To process heterogeneous inputs, a context-specific encoder $\mathcal{E}^{(j)}$ converts $\mathcal{X}_i^{(j)}$ into an embedding $\mathcal{O}_i^{(j)}$, which a bridge layer $\mathcal{B}^{(j)}$ then maps into the LLM embedding space:

$$\begin{aligned}\mathcal{O}_i^{(j)} &= \mathcal{E}^{(j)}(\mathcal{X}_i^{(j)}), \\ \mathcal{V}_i^{(j)} &= \mathcal{B}^{(j)}(\mathcal{O}_i^{(j)}),\end{aligned}\tag{2}$$

Here, $\mathcal{V}_i^{(j)}$ is the resulting embedding aligned with the LLM embedding. We sequence-concatenate the context, prompt, and ground-truth report embeddings to construct the final LLM input. Eventually, the LLM outputs logits for all sequence positions, and is optimized with the standard Cross-Entropy (CE) loss only on the tokens corresponding to the ground-truth report. Formally, let $\ell_{i,t}^{(j)}$ denote the LLM logit for predicting $y_{i,t}$ at the t -th report token, and let $\mathbf{q}_{i,t}$ denote the one-hot vector of the ground-truth report $\mathcal{Y}_i^*$. The loss function is for-

¹Appendix A offers a comprehensive review of existing related works to CONDISTILL.

²Appendix A discusses these categories and the corresponding literature.

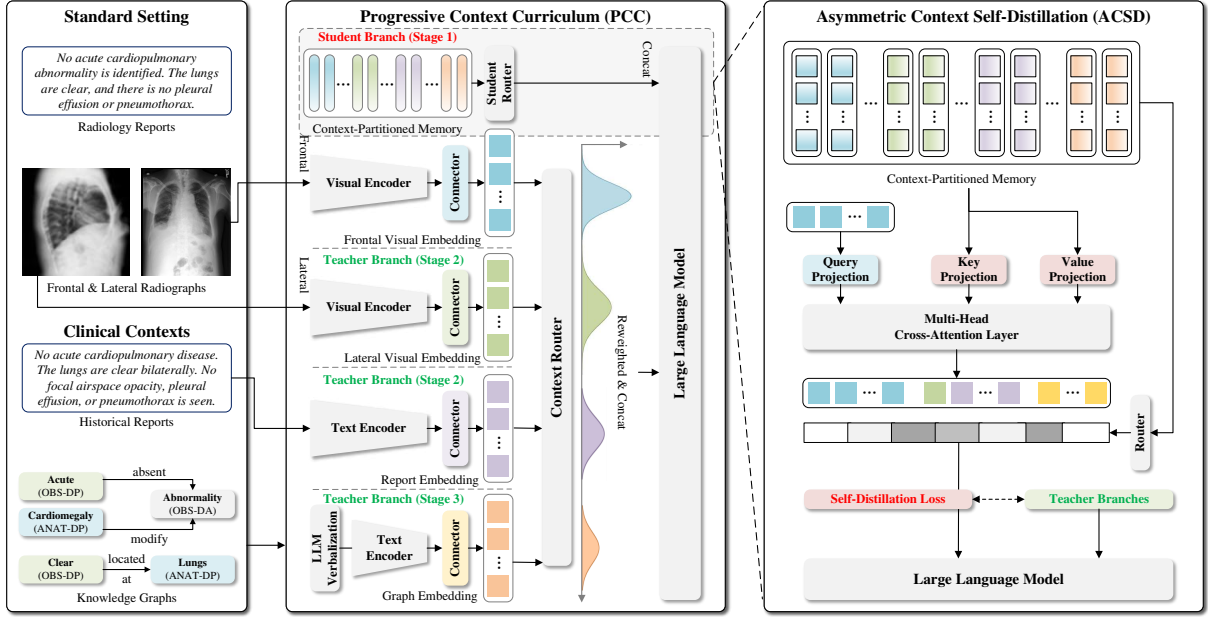


Figure 1: Overview of CONDISTILL. PCC progressively injects lateral radiographs, history reports, and structured knowledge graphs into a training-only teacher branch, where the context router produces routed teacher prefixes. ACSD distills these routed prefixes and routing distributions into a context-partitioned student memory bank, enabling context assistance in radiograph-only report generation, even after the teacher context branch is removed.

ally expressed as:

$$\begin{aligned} \ell_{i,t}^{(j)} &= \mathcal{F}_{\text{LM}} \left(\left\{ \mathcal{V}_i^{(j)} \right\}_{\mathcal{X}_{i,c} \in \Omega_i^{(j)}}, \mathcal{V}^{\mathcal{P}}, \mathcal{V}_i^{*} \right), \\ \mathcal{L}_{\text{CE}}^{(j)} &= -\frac{1}{T_i} \sum_{t=1}^{T_i} \mathbf{q}_{i,t}^{\top} \log \text{Softmax} \left(\ell_{i,t}^{(j)} \right). \end{aligned} \quad (3)$$

In practice, we optimize Eq. (3) through per-stage context injection, i.e., the model is trained with the context set $\Omega_i^{(j)}$ at the j -th stage. Note that the j -th stage inherits the optimized parameters of the previous stage and augments them with the parameters introduced for the newly injected context when moving from stage $j-1$ to stage j .

Context Routing. Jointly treating all injected contexts with the same weight is sub-optimal, as these contexts represent different granularities of clinical knowledge and contribute differently to the radiograph-only branch. After the three per-stage injection steps, the final context set $\Omega_i^{(3)}$ contains all training-time contexts; we then train a router to adjust their contributions adaptively. The router is an L_R -layer Multi-Layer Perceptron (MLP) that takes concatenated pooled context embeddings $\tilde{\mathcal{V}}_i^{(k)}$ as input and outputs a routing distribution α_i . Here, $\alpha_i = \left\{ \alpha_i^{(k)} \right\}_{\mathcal{X}_i^{(k)} \in \Omega_i^{(3)}}$ represents the teacher-side routing distribution over the final

context set $\Omega^{(3)}$. The routed teacher context is then balanced with its corresponding weight, expressed as:

$$\begin{aligned} \tilde{\mathcal{V}}_i^{(k,T)} &= \alpha_i^{(k)} \mathcal{V}_i^{(k)}, \\ \mathcal{Z}_i^T &= \text{Concat} \left(\left\{ \tilde{\mathcal{V}}_i^{(k,T)} \right\}_{\mathcal{X}_i^{(k)} \in \Omega_i^{(3)}} \right). \end{aligned} \quad (4)$$

$\text{Concat}(\cdot)$ refers to the same sequence-wise concatenation operation as that in Eq. (3). Subsequently, the routed sequence $\mathcal{Z}_i^T$ is used in the asymmetric distillation to assist the student branch.

2.2 Asymmetric Context Self-Distillation

Now that we have constructed a reliable cross-modal alignment for report generation, the next step is to distill this alignment from the teacher branches to the student one, as these additional contexts are normally unavailable in inference. To address this, we propose ACSD, which stores contextual information in a context-partitioned memory bank and explicitly aligns the encoding behaviors of the teacher and student. We detail these operations below.

Context-Partitioned Memory Bank. The memory bank $\mathcal{M}_{\text{MB}}^{(j)}$ comprises several partitions $\mathcal{M}^{(k)}$ of learnable memory vectors, termed as $\left\{ \mathbf{m}_r^{(k)} \right\}_{r=1}^{R_k}$, where each partition corresponds to

one context type $\mathcal{X}_i^{(k)}$ in $\Omega_i^{(j)}$. Here, R_k denotes the number of memory vectors assigned to this partition. Following the progressive expansion of $\Omega_i^{(j)}$, the memory bank also grows by adding new partitions to model the newly injected contexts:

$$\begin{aligned}\mathcal{M}_{\text{MB}}^{(j)} &= \mathcal{M}_{\text{MB}}^{(j-1)} \cup \Delta\mathcal{M}_{\text{MB}}^{(j)}, \\ \Delta\mathcal{M}_{\text{MB}}^{(j)} &= \left\{ \mathcal{M}^{(k)} \mid \begin{array}{l} \mathcal{X}_i^{(k)} \in \Omega_i^{(j)}, \\ \mathcal{X}_i^{(k)} \notin \Omega_i^{(j-1)} \end{array} \right\}, \quad j > 1.\end{aligned}\quad (5)$$

This design aligns the student memory with the curriculum: the first stage contains the frontal partition, the second adds lateral and history partitions, and the third adds the graph partition.

Memory Routing. To align with the context routing process, we use the same router to perform a similar routing procedure on the memory vectors in different context partitions. Given the current stage $\Omega_i^{(j)}$, the student branch receives only $\mathcal{U}_i$ and its aligned embedding $\mathcal{V}_i^{(1)}$. It first predicts a stage-specific routing distribution over the available memory partitions. Specifically, we feed the concatenated pooled radiograph embedding $\mathcal{V}_i^{(1)}$ and memory-partition representations $\mathcal{M}^{(j)}$ into the MLP router to obtain $\beta_i^{(j)}$. Here, $\beta_i^{(j)} = \left\{ \beta_i^{(j,k)} \right\}_{\mathcal{X}_i^{(k)} \in \Omega_i^{(j)}}$ denotes the student-side routing distribution over the same context order as that in the teacher branch. The student then uses the frontal embedding as the query to read each memory partition via a multi-head attention $\Phi_{\text{Attn}}(\cdot)$ mechanism and builds a context-partitioned memory prefix similar to Eq. (4):

$$\begin{aligned}\mathcal{R}_i^{(j,k)} &= \Phi_{\text{Attn}}\left(\bar{\mathcal{V}}_i^{(1)}, \mathcal{M}^{(k)}, \mathcal{M}^{(k)}\right), \\ \tilde{\mathcal{R}}_i^{(j,k)} &= \beta_i^{(j,k)} \mathcal{R}_i^{(j,k)}, \\ \mathcal{Z}_i^{S,(j)} &= \text{Concat}\left(\left\{ \tilde{\mathcal{R}}_i^{(j,k)} \right\}_{\mathcal{X}_i^{(k)} \in \Omega_i^{(j)}}\right).\end{aligned}\quad (6)$$

Thus, the teacher prefix routes real context embeddings, while the student prefix routes memory vectors that act as context proxies. This symmetry makes contextual knowledge transferable even when the contexts are unavailable at inference.

Asymmetric Distillation Loss. To transfer contextual knowledge from the teacher to the student, we use a distillation objective that makes the student predict the teacher’s routed embedding. In detail, we append a small set of learnable queries to the memory-routed prefix before the embedding predictor, in order to keep the capacity of

the student-side condition close to the teacher one, where this process is formally written as:

$$\begin{aligned}\mathcal{Q}^{(j)} &= \left\{ \mathbf{q}_s^{(j)} \right\}_{s=1}^{S_j}, \\ \hat{\mathcal{Z}}_i^{S,(j)} &= \mathcal{P}_{\text{SD}}^{(j)}\left(\text{Concat}\left(\mathcal{Z}_i^{S,(j)}, \mathcal{Q}^{(j)}\right)\right).\end{aligned}\quad (7)$$

Here, $\mathcal{Q}^{(j)}$ denotes the stage-specific set of S_j learnable queries, and $\mathcal{P}_{\text{SD}}^{(j)}$ maps the query-augmented embeddings. Following the same context order in $\Omega_i^{(j)}$, we obtain the teacher-side routed embedding $\mathcal{Z}_i^{T,(j)}$ by applying Eq. (4) to the current context set. The distillation loss is:

$$\begin{aligned}\mathcal{L}_{\text{ACSD}}^{(j)} &= \frac{1}{N_Z} \left\| \text{sg}\left(\mathcal{Z}_i^{T,(j)}\right) - \hat{\mathcal{Z}}_i^{S,(j)} \right\|_2^2 \\ &\quad + \gamma \text{KL}\left(\text{sg}\left(\alpha_i^{(j)}\right) \parallel \beta_i^{(j)}\right),\end{aligned}\quad (8)$$

where N_Z normalizes the routed-embedding distance; $\text{sg}(\cdot)$ refers to the gradient stopping operation; γ controls the routing-alignment weight. The overall training objective at the j -th stage combines the standard CE loss and the asymmetric distillation loss as $\mathcal{L}^{(j)} = \mathcal{L}_{\text{CE}}^{(j)} + \lambda_{\text{SD}} \mathcal{L}_{\text{ACSD}}^{(j)}$.

Inference Setting. Under standard radiograph-only inference, CONDISTILL uses only the frontal radiograph $\mathcal{U}_i$ and the final memory bank $\mathcal{M}_{\text{MB}}^{(3)}$. The student first computes $\hat{\mathcal{Z}}_i^{S,(3)}$ through Eqs. (6)–(7), and then uses the same LLM input construction as during training. Given the generated token history $\hat{\mathcal{Y}}_{i,<t}$, the entire report $\hat{\mathcal{Y}}_i$ is generated via next-token prediction:

$$\hat{\mathcal{Y}}_i = \text{Decode}\left(\mathcal{F}_{\text{LM}}\left(\mathcal{Z}_i^{S,(j)}, \mathcal{V}^{\mathcal{P}}, \mathcal{V}^{\hat{\mathcal{Y}}_{i,<t}}\right)\right). \quad (9)$$

3 Experiments

In this section, we present the main comparisons and ablation studies.³

3.1 Comparison to Different Baselines

Baseline Setup. Tab. 1 presents the quantitative comparison of CONDISTILL to different baseline variants. Here, we consider 5 main baselines. The first one is BASE, representing the model fine-tuned with only radiograph-report pairs. The second, +PCC, equips BASE with multiple training contexts but trains only the teacher router, without a context-partitioned memory bank. Afterwards, we try two

³Appendix B provides the complete experimental settings.

BASELINES	CONTEXT INPUT	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑
BASE	×	0.126	0.168	0.264	0.430	0.512	0.468
+ PCC	×	0.128	0.175	0.291	0.437	0.515	0.473
+ ACSD w/o MEMORY BANK	×	0.122	0.151	0.283	0.336	0.439	0.381
+ ACSD w/o ROUTER	×	0.124	0.153	0.286	0.339	0.443	0.384
+ PCC + ACSD (CONDISTILL)	×	<u>0.154</u>	<u>0.181</u>	<u>0.339</u>	<u>0.584</u>	<u>0.658</u>	<u>0.619</u>
CONTEXT TEACHER UPPER BOUND	✓	0.184	0.207	0.372	0.610	0.696	0.650

Table 1: Quantitative comparison with baselines on MIMIC-CXR (Johnson et al., 2019) with respect to BLEU-4 (B@4) (Papineni et al., 2002), METEOR (MTR) (Banerjee and Lavie, 2005), and the CE metric. Here, we report results under the standard radiograph-only setting, and present the teacher baseline for reference.

CATEGORY	METHOD	PARAMS (B)	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑
PRE-LLM METHODS	ST (Vinyals et al., 2015)	–	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.224
	ATT2IN (Rennie et al., 2017)	–	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.274
	ADAATT (Lu et al., 2017)	–	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.220
	TOPDOWN (Anderson et al., 2018)	–	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.268
	R2GEN (Chen et al., 2020)	–	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.300
	CA (Liu et al., 2021c)	–	0.350	0.219	0.152	0.109	0.151	0.283	–	–	–
	PPKED (Liu et al., 2021b)	–	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
	R2GENCMN (Chen et al., 2021)	–	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.302
	CMCL (Liu et al., 2021a)	–	0.344	0.217	0.140	0.097	0.133	0.281	–	–	–
	R2GENRL (Qin and Song, 2022)	–	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.316
	ITA (Wang et al., 2022)	–	0.395	0.253	0.170	0.121	0.147	0.284	–	–	–
	WARMSTART (Nicolson et al., 2023a)	–	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
	MLRG (Liu et al., 2025a)	–	0.332	0.199	0.132	0.094	0.136	0.248	0.513	0.517	0.515
LLM-BASED METHODS	XRAYGPT (Thawakar et al., 2024)	7	0.128	0.045	0.014	0.004	0.079	0.111	–	–	–
	R2GENGPT (Wang et al., 2023b)	7	0.396	–	–	0.113	–	0.273	0.506	0.414	0.456
	CHEXAGENT (Chen et al., 2024b)	7	0.189	–	–	0.040	–	0.208	0.506	0.306	0.381
	MED-PALM (Tu et al., 2023)	562	0.317	–	–	0.115	–	0.275	–	–	0.378
	MAIRA-1 (Hyland et al., 2023)	7	0.392	–	–	0.142	–	0.289	–	–	0.553
	EKAGEN (Bu et al., 2024)	–	0.419	0.258	0.170	0.119	–	0.287	0.517	0.483	0.499
	R2-LLM (Liu et al., 2024)	14.2	0.402	0.262	0.180	0.128	<u>0.175</u>	0.291	0.465	0.482	0.473
	INVERGE (Deria et al., 2024)	7	<u>0.425</u>	0.240	0.132	0.100	–	0.309	–	–	–
	LLM-RG4 (Wang et al., 2025)	7	0.377	–	–	0.144	–	0.318	0.583	0.593	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>0.613</u>	<u>0.606</u>	<u>0.608</u>
	CONDISTILL	3	0.431	0.271	0.190	0.154	0.181	0.339	0.621	0.617	0.619

Table 2: Radiograph-only evaluation on MIMIC-CXR (Johnson et al., 2019) with respect to the NLG and CE metrics. For LLM-based methods, we report the total parameter count to measure their efficiency.

different variants on the ACSD without either the memory bank or the student router. The baseline +PCC+ACSD refers to our full model. For reference, we also report the results by assuming that all contexts are available in inference, where this baseline denotes the performance upper bound of the teacher model.

Analysis. The radiograph-only baseline obtains low scores across metrics, confirming that paired radiographs and reports alone provide insufficient supervision for fine-grained alignment. Adding multiple contexts, as shown by the baseline +PCC, brings slight gains compared to BASE, indicating that the extra contexts used in training assist the cross-modal alignment, yet the performance is still limited due to the train-test asymmetry. This issue is significantly alleviated by using the memory bank as the context proxy, where either the vectors

stored in the memory bank or the student router offer more fine-grained cues for the radiograph-only report generation process, eventually resulting in higher-quality diagnoses. Because both memory readout and student routing are conditioned on the current radiograph, this gain reflects case-specific proxy selection rather than unconditional decoding from a static dataset prior. Our full model achieves the best performance over all compared baselines, illustrating the synergistic effect of PCC and ACSD to meticulously adapt the learnt context proxies for RRG.⁴

3.2 Comparison to Existing Methods

Radiograph-Only Evaluation. Tab. 2 reports the quantitative comparison between CONDISTILL

⁴Appendix C provides the corresponding sensitivity and training design analyses.

SCENARIO	MODEL	B@1↑	B@2↑	B@3↑	B@4↑	MTR↑	R-L↑	P↑	R↑	F1↑	HALL.↓
SN	CXR-MATE* (Nicolson et al., 2023b)	0.421	0.271	0.179	0.122	0.174	0.311	0.572	0.560	0.566	0.010
	RADFM (Wu et al., 2023)	0.188	0.090	0.048	0.028	0.094	0.190	0.413	0.303	0.350	0.737
	LLM-RG4 (Wang et al., 2025)	<u>0.479</u>	<u>0.343</u>	<u>0.255</u>	<u>0.196</u>	<u>0.209</u>	<u>0.384</u>	<u>0.588</u>	<u>0.632</u>	<u>0.609</u>	0.014
	CONDISTILL	0.506	0.371	0.282	0.221	0.223	0.407	0.636	0.673	0.654	0.010
SW	CXR-MATE* (Nicolson et al., 2023b)	0.361	0.220	0.139	0.093	0.153	0.284	0.573	0.549	0.561	0.009
	RADFM (Wu et al., 2023)	0.211	0.103	0.056	0.033	0.105	0.183	0.508	0.365	0.425	0.092
	LLM-RG4 (Wang et al., 2025)	<u>0.455</u>	<u>0.321</u>	<u>0.239</u>	<u>0.186</u>	<u>0.199</u>	<u>0.382</u>	<u>0.599</u>	<u>0.622</u>	<u>0.610</u>	0.021
	CONDISTILL	0.488	0.351	0.268	0.211	0.219	0.404	0.628	0.654	0.641	<u>0.013</u>
MN	CXR-MATE* (Nicolson et al., 2023b)	0.437	0.289	0.199	0.141	0.179	0.332	<u>0.544</u>	0.522	0.533	0.009
	RADFM (Wu et al., 2023)	0.246	0.113	0.060	0.034	0.104	0.194	0.323	0.187	0.237	0.140
	LLM-RG4 (Wang et al., 2025)	<u>0.491</u>	<u>0.359</u>	<u>0.274</u>	<u>0.216</u>	<u>0.214</u>	<u>0.405</u>	0.541	<u>0.578</u>	<u>0.559</u>	<u>0.008</u>
	CONDISTILL	0.504	0.373	0.284	0.224	0.226	0.410	0.622	0.644	0.633	0.007
MW	CXR-MATE* (Nicolson et al., 2023b)	0.379	0.241	0.158	0.110	0.159	0.305	0.548	0.499	0.523	0.007
	RADFM (Wu et al., 2023)	0.191	0.095	0.054	0.034	0.095	0.178	0.456	0.297	0.360	0.052
	LLM-RG4 (Wang et al., 2025)	<u>0.461</u>	<u>0.331</u>	<u>0.250</u>	<u>0.197</u>	<u>0.204</u>	0.401	<u>0.560</u>	<u>0.565</u>	<u>0.563</u>	0.002
	CONDISTILL	0.471	0.341	0.257	0.200	0.212	<u>0.399</u>	0.616	0.632	0.624	<u>0.004</u>

Table 3: Multi-context evaluation on MIMIC-RG4 (Wang et al., 2025) with respect to the NLG metric, CE metric, and hallucination score. We follow the 4 typical context settings in LLM-RG4 (Wang et al., 2025) for evaluation.

TILL and existing RRG studies on MIMIC-CXR (Johnson et al., 2019) with respect to both NLG and CE metrics. Here, we compare 25 existing studies in total across two categories, i.e., pre-LLM and LLM-based methods. Specifically, pre-LLM studies comprise early RRG systems that train task-specific encoder-decoder networks from scratch and explore basic forms of contextual assistance. Yet, most of these studies obtain limited lexical and clinical accuracy due to their coarse-grained cross-modal mappings. Results are significantly improved in LLM-based methods, as they enhance the report generation process with the strong language prior in pre-trained LLMs. Notably, early-proposed LLM-based studies, e.g., XRAYGPT (Thawakar et al., 2024), R2GENGPT (Wang et al., 2023b), MED-PALM (Tu et al., 2023), MAIRA-1 (Hyland et al., 2023), and R2-LLM (Liu et al., 2024), which mainly focus on adapting general-domain LLMs to the RRG task, show strong lexical coherence with competitive NLG scores. Later studies, which put more emphasis on injecting clinical contexts, particularly known as LLM-RG4 (Wang et al., 2025) and S2D-ALIGN (Gao et al., 2026), further boost the clinical accuracies compared to the early LLM-based ones, supporting the conclusion that dedicated clinical contexts facilitate radiograph-report alignment. Ultimately, our method achieves the strongest clinical accuracies and competitive lexical quality among the compared methods, showing the effectiveness of our proposed asymmetric self-distillation mechanism to boost the radiograph-report alignment with additional contexts.

Multi-Context Evaluation. Tab. 3 presents the multi-context evaluation on MIMIC-RG4 (Wang et al., 2025) using the same metrics. At inference, MIMIC-CXR uses the frontal radiograph and student memory only; MIMIC-RG4 retains memory and additionally routes scenario-provided views/history, without target or future reports.⁵ Unlike MIMIC-CXR, MIMIC-RG4 evaluates context-assisted inference and therefore tests whether distilled memory remains compatible with observed test-time context. Specifically, LLM-RG4 (Wang et al., 2025) already demonstrates significant improvements over RADFM (Wu et al., 2023) and the longitudinal RRG baseline CXR-MATE (Nicolson et al., 2023b). Furthermore, CONDISTILL obtains the best performance of all compared methods and further reaches the teacher performance upper bound in Tab. 1. This result suggests that available contexts and distilled memory proxies provide complementary alignment cues. Also, we follow the standard settings of LLM-RG4 to compute hallucination rate for the generated reports. Since ACS D aligns routed embeddings and routing distributions rather than teacher report tokens, the current radiograph remains the decoding anchor; the competitive hallucination scores in Tab. 3 therefore complement, rather than replace, the targeted factuality tests below.

3.3 Case Study

Tab. 4 presents a selected case study from MIMIC-CXR along with 2 existing studies and 2 baselines for comparison. We compare two recent LLM-

⁵Appendix C presents more results and analyses.

SOURCE	GENERATED REPORTS	
		
RADIOGRAPHS	FRONTAL RADIOGRAPH	LATERAL RADIOGRAPH
GROUND TRUTH	As compared to the previous radiograph, there is no relevant change . Extensive right pleural effusion , potentially combined with pleural thickening , and relatively extensive atelectatic changes in the right lung bases are present. The ventilated right lung parenchyma is small and perihilar. The left lung is normal without parenchymal change or left pleural effusion .	
MLRG (Liu et al., 2025a)	As compared to the previous radiograph, the air collection in the right hemithorax is unchanged . There is no evidence of apical pneumothorax. However, the pleural effusion occupies approximately one-third of the right hemithorax . The left lung is unremarkable . The referring physician was paged for notification .	
LLM-RG4 (Wang et al., 2025)	There is a large right pleural effusion occupying approximately half of the right hemithorax. Extensive atelectasis is present at the right lung bases. The left lung appears normal . The cardiac silhouette is borderline in size with mild tortuosity of the thoracic aorta . Lateral view is recommended to assist in diagnosis .	
BASE	There is a small right pleural effusion with mild left basilar atelectatic opacity . No definite pleural thickening is identified . The right lung is otherwise well aerated, and minimal left pleural fluid cannot be excluded .	
+ PCC	Compared with the prior study, there is slight interval improvement in the right pleural effusion with pleural thickening and right basilar atelectasis . A tiny right apical pneumothorax persists . Mild left basilar opacity may reflect atelectatic change .	
+ PCC + ACSD (CONDISTILL)	The current image again shows no interval change . A large right pleural-fluid collection with probable pleural thickening is present, together with broad right basilar atelectatic change . Only a small perihilar portion of the right lung remains aerated, and the left hemithorax is clear without left pleural fluid .	

Table 4: A selected case study from MIMIC-CXR, with different pathological attributes highlighted in boldfaces and colors. Note that correct or incorrect diagnoses are marked with backgrounds of green and red, respectively.

based methods, MLRG (Liu et al., 2025a) and LLM-RG4 (Wang et al., 2025), with the BASE and +PCC variants from Tab. 1. Existing methods such as MLRG and LLM-RG4 omit some pathological attributes and additionally produce unsupported statements, e.g., “the air collection in the right hemithorax is unchanged” and “Lateral view is recommended to assist in diagnosis”. This observation supports our motivation that insufficiently fine-grained alignment can produce hallucinated report content. Similar results are seen in BASE, the radiograph-only baseline, with insufficient attributes covered in its resulting report. For example, it only predicts “a small right pleural effusion” while missing the extensive extent of effusion, and further contradicts the ground-truth pleural finding with “No definite pleural thickening is identified”. Introducing additional contexts by +PCC addresses this issue, where the generated pathological attributes are more related to the ground-truth report, e.g., “right pleural effusion”, “pleural thickening”, and “right basilar atelectasis”, yet incorrect diagnoses such as “slight interval improvement” and “A tiny right apical pneumothorax persists” still occur. This might be caused by the train-test mismatch, where the model lacks approximate proxies for the training-time contexts

at inference, leading to reports of inferior quality. All aforementioned issues are alleviated in our full method, i.e., +PCC+ACSD, which preserves all major contextual facts, such as “no interval change”, “large right pleural-fluid collection”, and “the left hemithorax is clear without left pleural fluid”, while avoiding inference-time context dependence, validating the effectiveness of the self-distillation mechanism in constructing more meticulous cross-modal alignment for RRG. This joint recovery supports patient-conditioned memory routing rather than a static report prior.

3.4 Ablation Studies

In this section, we ablate PCC and ACSD to isolate the contribution of each component.

Effect of the PCC. To explore the effect of the PCC, Tab. 5 (a) compares several baseline variants based on different context settings in PCC. The scores improve monotonically as the curriculum exposes the teacher to richer contexts. Specifically, lateral radiographs provide cross-view anatomical evidence, while adding historical reports improves the temporal comparison needed by follow-up studies. Structured knowledge graphs bring a further gain, which is particularly observed in F1, since they force the alignment process to preserve

LAT.	HIST.	KG	ROUT.	B@4↑	F1↑
✓	×	×	×	0.126	0.468
✓	✓	×	×	0.132	0.486
✓	✓	✓	×	<u>0.145</u>	<u>0.598</u>
✓	✓	✓	✓	0.154	0.619

(a) Ablation of PCC.

VARIANT	B@4↑	F1↑
SHARED MEMORY BANK	0.109	0.426
PARTITIONS W/O ROUTER	0.122	0.482
TEACHER ROUTER ONLY	<u>0.149</u>	<u>0.611</u>
STUDENT ROUTER ONLY	0.126	0.497
CONDISTILL	0.154	0.619

(b) Ablation of ACSD.

Table 5: Ablations of (a) PCC and (b) ACSD on MIMIC-CXR w.r.t. BLEU-4 and F1 scores of the CE metric.

abnormality-level entities and relations. Using all contexts with routing performs best, showing that PCC benefits from both richer privileged evidence and adaptive weighting.

Effect of the ACSD. To investigate the impact of different components in ACSD, Tab. 5 (b) tries different memory bank designs as well as router settings. For the memory bank design, we consider two types of variants, i.e., a partitioned-agnostic design without distinguishing different contexts and a memory bank without context routing. A shared memory bank pools learnable vectors without identifying their context type and therefore obtains limited performance. This occurs because a partition-agnostic design mixes newly introduced vectors with earlier ones, allowing irrelevant context proxies to act as noise. Context partitions improve the result, and routing brings further gains, as both the teacher and student use compatible context weights. These results demonstrate the effectiveness of teacher-student routing, which adaptively adjusts the contribution of each context to report generation. The full model performs best because the teacher routes observed contexts while the student routes aligned memory partitions.

4 Conclusion

In this paper, we present CONDISTILL, an asymmetric context self-distillation framework for radiology report generation. The method learns a progressive alignment process from instance-level guidance to fine-grained clinical cues, routes heterogeneous contexts in the teacher and their memory proxies in the student, and distills this cross-modal mapping into a context-partitioned memory bank for inference. Results on MIMIC-CXR and MIMIC-RG4 show improvements in lexical quality and clinical accuracy, while extensive ablations clarify the roles of the progressive curriculum, routing, and self-distillation.

Limitations

CONDISTILL has two main limitations. First, it depends on off-the-shelf annotators such as RAD-GRAPH (Jain et al., 2021), where extraction errors might propagate missing entities or relations and imbalanced label coverage into the structured supervision. Stronger labelers in future studies are therefore important extensions of this work. Second, all experiments use chest radiographs from the MIMIC-CXR family due to the limited open-source data. This scope partly reflects the broader scarcity of large, aligned image-report datasets, and future work should evaluate the framework on additional modalities as data become available.

Acknowledgments

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

References

- 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.
- Yaara Artsi, Eyal Klang, Jeremy D. Collins, Benjamin S. Glicksberg, Girish N. Nadkarni, Panagiotis Korfiatis, and Vera Sorin. 2025. Large Language Models in Radiology Reporting: A Systematic Review of Performance, Limitations, and Clinical Implications. *Intell.-Based Med.*, 12:100287.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *IEEE-valuation@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*, abs/2406.04449.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie L. Hyland, Maria Wetscherek, Tristan Naumann, Aditya V. Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing. In *ECCV*, pages 1–21.
- 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.
- 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*, abs/2604.09450.
- Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. 2024a. Act Like a Radiologist: Radiology Report Generation across Anatomical Regions. In *ACCV*, pages 1–17.
- 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. 2024b. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *arXiv*, abs/2401.12208.
- Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandiyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Pontes Reis, Mohamed S. Muneer, Maya Varma, and Curtis P. Langlotz. 2025. Automated Structured Radiology Report Generation. In *ACL*, pages 26813–26829.
- 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.
- Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-Modal Hallucination Control by Visual Information Grounding. In *CVPR*, pages 14303–14312.
- 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*, pages 484–494.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.
- Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. Fin-eRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. In *MLHC*, volume 252 of *PMLR*, pages 1–29.
- 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*, abs/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*.
- Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In *AAAI*, pages 2607–2615.
- 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.
- Manav Nitin Kapadnis, Sohan Patnaik, Abhilash Nandy, Sourjyadip Ray, Pawan Goyal, and Debdoot Sheet. 2024. SERPENT-VLM: Self-Refining Radiology Report Generation Using Vision Language Models. In *ClinicalNLP@NAACL*, pages 283–291.
- 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.
- Chunyu 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.
- Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb El Saddik, and Xiaojun Chang. 2024a. Contrastive Learning with Counterfactual Explanations for Radiology Report Generation. In *ECCV*, pages 162–180.
- Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024b. KARGEN: Knowledge-Enhanced Automated Radiology Report Generation Using Large Language Models. In *MIC-CAI*, pages 382–392.
- 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.
- 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 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.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.
- 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.
- Microsoft. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv*, abs/2404.14219.
- Hoang T. N. Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated Generation of Accurate & Fluent Medical X-ray Reports. In *EMNLP*, pages 3552–3569.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023a. Improving Chest X-Ray Report Generation by Leveraging Warm Starting. *Artif. Intell. Med.*, 144:102633.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023b. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. *arXiv*, arXiv:2307.09758.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael E. Moseley, Curtis P. Langlotz, Akshay Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. In *Findings of EMNLP*, pages 374–390.

- 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.
- 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 Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. 2025. Exploring Scalable Medical Image Encoders Beyond Text Supervision. *Nat. Mach. Intell.*, 7(1):119–130.
- Han Qin and Yan Song. 2022. Reinforced Cross-Modal Alignment for Radiology Report Generation. In *Findings ACL*, pages 448–458.
- Vignav Ramesh, Nathan Chi, and Pranav Rajpurkar. 2022. CXR-PRO: MIMIC-CXR with Prior References Omitted. *PhysioNet*. Version 1.0.0.
- 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*, abs/2307.14334.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*, pages 3156–3164.
- Fuying Wang, Shenghui Du, and Lequan Yu. 2024. HERGen: Elevating Radiology Report Generation with Longitudinal Data. In *ECCV*, pages 183–200.
- 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.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In *CVPR*, pages 9049–9058.
- 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*, abs/2309.09812.
- Zhuohao 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.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards Generalist Foundation Model for Radiology. *arXiv*, arXiv:2308.02463.
- Joy T. Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest ImaGenome Dataset for Clinical Reasoning. In *NeurIPS Datasets and Benchmarks*.
- Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Zhe Wang, and Chenjia Bai. 2025. Online Iterative Self-Alignment for Radiology Report Generation. In *ACL*, pages 27799–27814.
- Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-II Suk. 2025. Diff-RRG: Longitudinal Disease-Wise Patch Difference as Guidance for LLM-Based Radiology Report Generation. In *MICCAI*, pages 152–161.
- Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. 2025. Libra: Leveraging Temporal Images for Biomedical Radiology Analysis. In *Findings ACL*, pages 17275–17303.
- 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.
- Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A Metric for Radiology Report Generation. In *EMNLP*, pages 15004–15019.

A Related Work

Pre-LLM methods. Early RRG research mainly follows an encoder-decoder paradigm that maps chest radiographs into visual features and decodes free-text reports from these representations. This line of work improves cross-modal alignment through joint disease prediction and report generation (Wang et al., 2018), template retrieval (Li et al., 2018), abnormality-aware graph reasoning (Li et al., 2019; Zhang et al., 2020), cross-modal memory (Chen et al., 2020, 2021), contrastive and posterior knowledge alignment (Liu et al., 2021c,b), curriculum training (Liu et al., 2021a), reinforcement learning (Nguyen et al., 2021; Qin and Song, 2022), architectural improvement (Huang et al., 2023), regional grounding (Tanida et al., 2023; Chen et al., 2024a), prompting techniques (Wang et al., 2023a; Jin et al., 2024; Li et al., 2024a). These designs gradually enrich the visual-textual interface, but they still train task-specific decoders from paired radiograph-report data and therefore provide limited help when the target report depends on subtle, multi-source clinical context.

LLM-based RRG methods. Recent RRG systems adapt large language models to radiology by mapping radiograph features into an instruction tuning paradigm. These systems use frozen or lightly tuned language backbones (Wang et al., 2023b; Thawakar et al., 2024; Li et al., 2023; Chen et al., 2024b; Liu et al., 2024; Tu et al., 2023), grounded radiology encoders (Hyland et al., 2023; Bannur et al., 2024), clinical knowledge prompts (Li et al., 2024b), flexible input contexts (Wang et al., 2025), longitudinal representations (Liu et al., 2025b; Yun et al., 2025), preference alignment (Xiao et al., 2025), retrieval-guided diagnostic concepts (Gu et al., 2025), and efficient generative decoding (Chen et al., 2026) to improve fluency and factuality. Although these approaches substantially strengthen report language quality, many of them still require explicit context, retrieval evidence, concept prompts, or specialized decoding during inference. CONDISTILL instead uses context-rich teacher branches only during training and deploys a radiograph-only student.

Context assistance in RRG. Context assistance studies how signals beyond the current frontal radiograph can improve clinical faithfulness. Prior work explores structured abnormality knowledge (Li et al., 2019; Zhang et al., 2020; Jain et al., 2021; Delbrouck et al., 2025), anatomy-centered scene

DATASET	SPLIT USED
MIMIC-CXR (Johnson et al., 2019)	222,758 train, 1,808 validation, 3,269 test studies after section normalization and view filtering
MIMIC-RG4 (Wang et al., 2025)	172.6K train, 1.4K validation, 2.4K test reports

Table 6: Dataset statistics for MIMIC-CXR (Johnson et al., 2019) and MIMIC-RG4 (Wang et al., 2025)

graphs (Wu et al., 2021), medical terms (Li et al., 2018; Liu et al., 2021b; Huang et al., 2023; Jin et al., 2024), regional visual evidence (Tanida et al., 2023; Chen et al., 2024a; Bannur et al., 2024; Gao et al., 2026), longitudinal image-report pairs and history-aware representations (Wang et al., 2024, 2025; Liu et al., 2025b,a; Zhang et al., 2025; Yun et al., 2025), reinforcement learning rewards (Nguyen et al., 2021; Qin and Song, 2022; Xiao et al., 2025). These studies confirm that richer medical context improves RRG, but the context often remains an inference-time input.

B Experimental Settings

Datasets. We conduct all experiments on two publicly available benchmarks, i.e., MIMIC-CXR (Johnson et al., 2019) and MIMIC-RG4 (Wang et al., 2025). The former represents the classical radiograph-only RRG setting and the latter is a context-enhanced variant based on MIMIC-CXR. Specifically, MIMIC-CXR contains 377,110 chest radiographs from 227,835 studies of 65,379 patients. We follow the official patient-level split in all our experiments. MIMIC-RG4 re-organizes the reports and contexts in MIMIC-CXR into 4 different scenarios, including single-view no-history (SN), multi-view no-history (MN), single-view with-history (SW), and multi-view with-history (MW). In addition to the radiographs and the ground-truth reports, each case contains the aligned lateral-view radiographs and historical reports⁶. Tab. 6 summarizes the dataset statistics.

Evaluation metrics. Following prior studies (Chen et al., 2020, 2021; Wang et al., 2023b; Liu et al., 2024; Gao et al., 2026), we consider Natural Language Generation (NLG) and Clinical Efficacy

⁶We do not use IU X-RAY (Demner-Fushman et al., 2016) because it does not contain context annotations. Also note that it is unsuitable for existing MIMIC-CXR-oriented annotators, i.e., RADGRAPH (Jain et al., 2021) and CHEXBERT (Smit et al., 2020), used to construct graphs and clinical labels.

COMPLEXITY	SCENARIO	Avg. Report Length	Avg. Abnormal Labels	F1↑
LOW	SN	38.4	1.7	0.674
	MW	41.9	1.8	0.681
MEDIUM	SN	52.6	2.8	0.646
	MW	57.1	2.9	0.652
HIGH	SN	68.3	4.1	0.604
	MW	74.5	4.3	0.616

Table 7: Complexity-stratified comparison of SN and MW cases on MIMIC-RG4.

(CE) metrics to measure the lexical and clinical accuracies, respectively. For the lexical quality, we report BLEU- n , with n ranging from 1 to 4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). For clinical correctness, we leverage an existing medical annotator CHEXBERT (Smit et al., 2020) to extract labels from both the generated and ground-truth reports. Then, we compute precision, recall, and F1 between these labels as the final scores. We additionally report FINERADSCORE (Huang et al., 2024), GREEN (Ostmeier et al., 2024), and RATESCORE (Zhao et al., 2024) to assess line-level error severity, clinically significant generation errors, and radiology-specific semantic agreement, respectively.

Implementation details. For the pre-trained components in CONDISTILL, we use RAD-DINO (Pérez-García et al., 2025) to encode radiographs, CXR-BERT (Boecking et al., 2022) to encode reports, and PHI-3 (Microsoft, 2024) to initialize the LLM report decoder. In training, the visual encoder is frozen, while the LLM is updated through LORA adapters (Hu et al., 2022). We set the bridge layer as a cross-attention layer with 32 learnable queries following LLM-RG4 (Wang et al., 2025). In PCC, the training epochs are all set to 5 for best performance. We optimize all trainable parameters with the ADAMW optimizer (Loshchilov and Hutter, 2019), using stage-wise learning rates of 2×10^{-4} , 1×10^{-4} , and 5×10^{-5} , a batch size of 64, and 4 gradient-accumulation steps. For the LORA adapters, we use a rank of 16, a scaling factor of 32, and a dropout rate of 0.05. We run all experiments on a single NVIDIA A40 GPU. The full model uses 32 memory vectors per partition and a routing-alignment weight of $\gamma = 0.20$.

C More Results and Analyses

We provide more detailed results and analyses to validate our design. Specifically, Tab. 7 controls for case complexity when comparing SN and MW under matched report difficulty. Tab. 8 complements the primary evaluation with FINERADSCORE, GREEN, RATESCORE, CXR-PRO (Ramesh et al., 2022), and prior-reference statistics. Tab. 9 analyzes memory-bank capacity, the routing coefficient, curriculum order, and robustness to corrupted training contexts.

METHOD	CLINICAL METRICS			CXR-PRO		PRIOR-REFERENCE RATE (%)					
	FINERADSCORE↓	GREEN↑	RATESCORE↑	B@4↑	F1↑	prior	previous	interval	unchanged	improved	worsened
GROUND TRUTH	–	–	–	–	–	16.2	11.8	6.9	25.8	4.2	0.8
MLRG (Liu et al., 2025a)	0.645	0.342	0.292	–	–	–	–	–	–	–	–
LLM-RG4 (Wang et al., 2025)	0.612	0.371	0.315	0.068	0.286	23.5	17.1	9.8	37.2	6.1	1.2
BASE	0.690	0.315	0.270	0.052	0.232	31.8	23.4	13.7	50.6	8.5	1.7
+ PCC	<u>0.579</u>	<u>0.398</u>	<u>0.338</u>	<u>0.079</u>	<u>0.319</u>	6.5	4.7	2.9	10.4	1.6	0.3
+ PCC + ACSD (CONDISTILL)	0.552	0.420	0.362	0.091	0.352	6.2	4.9	2.6	10.0	1.8	0.4

Table 8: Additional evaluations on MIMIC-CXR, with respect to FINERADSCORE (Huang et al., 2024), GREEN (Ostmeier et al., 2024), RATESCORE (Zhao et al., 2024), the prior-filtered CXR-PRO benchmark (Ramesh et al., 2022), and report-level occurrence rates of prior-reference terms. Bold and underlined values denote the best and second-best results.

ANALYSIS	SETTING OR VARIANT	B@4↑	F1↑
REFERENCE	FULL MODEL (32 vectors, $\gamma = 0.20$, proposed order, clean contexts)	0.154	0.619
MEMORY VECTORS PER PARTITION	8	0.121	0.501
	16	0.145	0.585
	64	0.149	0.602
	128	0.115	0.462
ROUTING WEIGHT γ	0.05	0.129	0.476
	0.10	0.145	0.590
	0.50	0.148	0.603
	1.00	0.127	0.469
CURRICULUM ORDER	SIMULTANEOUS INJECTION	0.129	0.476
	GRAPH $\rightarrow$ INSTANCE-LEVEL	0.112	0.452
	INSTANCE-LEVEL $\rightarrow$ GRAPH	0.115	0.459
CONTEXT CORRUPTION (CONDISTILL)	20% MISSING CONTEXTS	0.151	0.611
	20% INCORRECT CONTEXTS	0.147	0.600
	40% NOISY CONTEXTS	0.141	0.576
CONTEXT CORRUPTION W/O MEMORY BANK	CLEAN CONTEXTS	0.122	0.381
	20% MISSING CONTEXTS	0.115	0.342
	20% INCORRECT CONTEXTS	0.108	0.304
	40% NOISY CONTEXTS	0.097	0.261

Table 9: More quantitative results on MIMIC-CXR, measured by BLEU-4 and CE F1.