PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

PDD-RRG: Posterior Diagnostic Decision for Study-level Radiology Report Generation

Yang Yu[1] , **Yiming Ji**[1] , **Bin Dai**[2] , **Dong Zhang**[1,3*] , **Zhiyong Zhou**[2] , **Shoushan Li**[1] , **Yakang Dai**[2]
[1]School of Computer Science & Technology, Soochow University
[2]Suzhou Institute of Biomedical Engineering and Technology
[3]Jiangsu Key Lab of Language Computing, Suzhou
dzhang@suda.edu.cn

Abstract

Automatic radiology report generation (RRG) aims to simulate the workflow of radiologists, assisting them in clinical diagnosis. However, existing methods often fall short in utilizing all information relevant to the examination, as is typically done in clinical practice. Although some works attempt to incorporate multi-view images and historical data, these additional inputs may sometimes lead to avoidable diagnostic errors on the contrary. To address these challenges, we introduce a decision-making stage after report generation for the first time and propose a Posterior Diagnostic Decision framework (PDD-RRG) to integrate potentially conflicting diagnoses. Specifically, we create various subsets of input data and utilize an existing RRG model to generate reports from different perspectives. Then the Bayesian posterior probability and the learned thresholds for each clinical observation are calculated to obtain an aggregated diagnostic conclusion, which is subsequently used to refine the generated report. Experiments on MIMIC-CXR demonstrate that our proposed PDD-RRG can effectively enhance the clinical efficacy of existing RRG models without any retraining.

1 Introduction

Radiological examinations, led by chest X-rays, are widely used in clinical for the diagnosis of various diseases. When drafting reports, radiologists must synthesize all information relevant to the current examination. For the example in Figure 1(a), they have to read multi-view images and compare them with prior exams and reports (when available). This process is both time-consuming and error-prone. Consequently, automatic radiology report generation (RRG) which can provide high-quality draft reports, is attracting increasing attention.

RRG is intuitively an end-to-end text generation task. The success of large language models (LLM) across various domains [Liu *et al.*, 2024b; Ju *et al.*, 2025; Tang *et al.*, 2025] has further reinforced this stereotype. As Figure 1(b) shows, early methods [Jing *et al.*, 2018; Liu *et al.*, 2021] typically relied on a single image, which diverges significantly from the real-world diagnostic workflow. These generated drafts still require manual integration to resolve potential conflicts and supplement missing perspective information before serving as a study-level report. Some recent studies [Serra *et al.*, 2023; Wang *et al.*, 2025] shown in Figure 1(c), have incorporated multi-view images and additional historical studies. Nevertheless, due to the model limitation, they may still need to generate multiple reports to cover all information relevant to the examination.

Only a small number of approaches [Nicolson *et al.*, 2024] are capable of ingesting all available cues in one pass to produce a full study-level report (as shown in Figure 1(d)). However, this does not guarantee that their diagnoses are always optimal. In fact, we observed that richer inputs can sometimes yield worse diagnostic results. As the case illustrated in Figure 2 on a Type-2 model MAIRA-2 [Bannur *et al.*, 2024], introducing additional historical view causes a missed diagnosis of cardiomegaly that is correctly identified under simpler input. We counted such cases in the MIMIC-CXR [Johnson *et al.*, 2019] test set and found that they are far from rare. As shown in Figure 3, for each of the 14 clinical observations extracted from reports, over 10% of the samples suffer avoidable diagnostic errors when integrating all available inputs.

Superficially, this reflects a fusion flaw of current approaches, where additional cues may not be fully exploited and can even turn into noise that misleads the model. However, at a deeper level, it reveals a persistent contradiction between the goal and reality of the RRG task. As methods evole, we pack more cues into the input to attempt covering all available views in one pass, avoiding merging drafts generated from fragmented information. But for certain findings, richer input can generate suboptimal results, so we may still need to consult the reports from simpler views to reach a sound diagnosis. In short, to fully leverage a model's diagnostic power, a report-merging step appears inevitable.

Thus, we introduce this integration as an automated decision stage, implemented after report generation. This shifts RRG from a single-step generation task to a pipeline of multi-path generations followed by a decision-level fusion, shown in figure 1(e). On this basis, we propose PDD-RRG, a posterior diagnostic decision framework to aggregate conflicting diagnoses. As Figure 4 shows, we extract 14 observation classes from reports, compute their Bayesian posteriors,

*Corresponding author.

IJCAI-ECAI 2026 PREPRINT

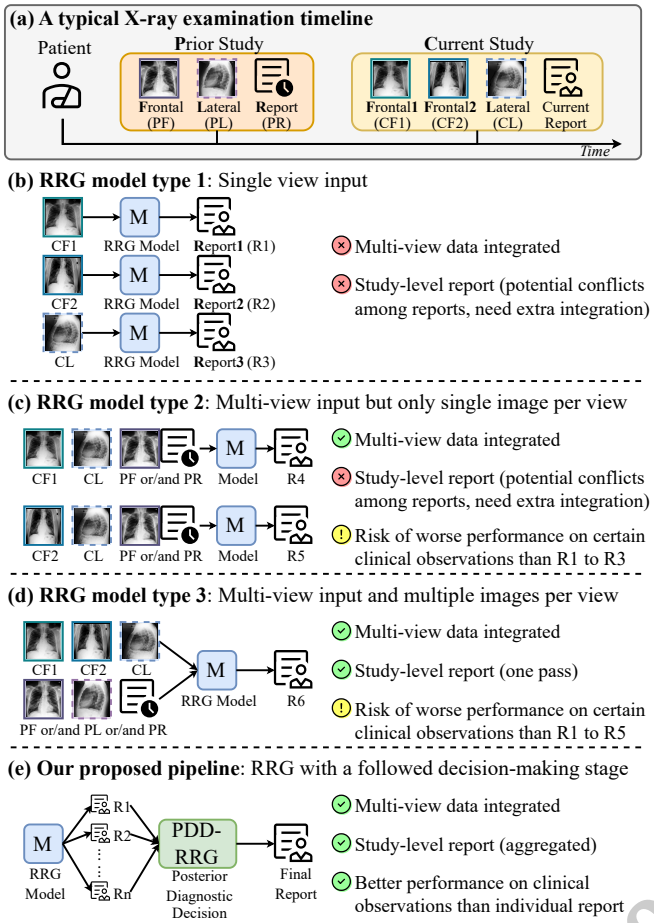

Figure 1: Pipelines of RRG models and our proposed framework.

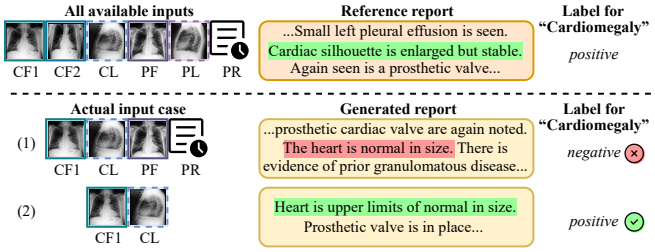

Figure 2: A real case on MAIRA-2 where additional historical information lead to an avoidable false negative for cardiomegaly.

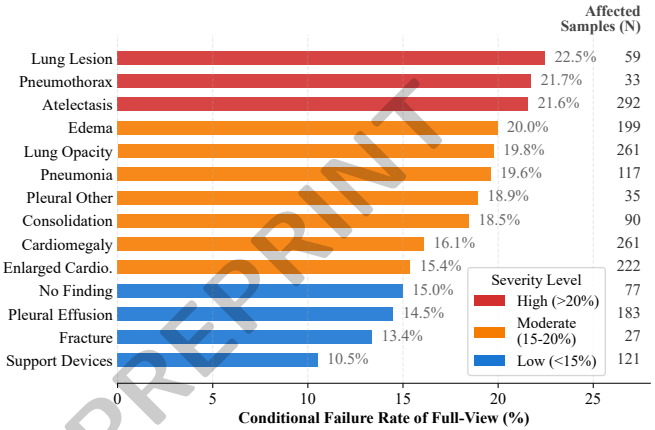

Figure 3: The proportion of diagnostic errors in various clinical observations that can be corrected with fewer inputs on MAIRA-2.

and apply validation-tuned thresholds to finalize each diagnosis, guiding the new study-level report. Experiments on three models demonstrate the effectiveness of our method.

Our contributions are stated as follows: 1) To the best of our knowledge, this is the first work to introduce a diagnostic decision layer into RRG, enabling the reconciliation of heterogeneous inputs and unlocking the latent diagnostic capability of backbone models. 2) We propose PDD-RRG, a likelihood-based aggregation framework that improves the utilization of clinically meaningful signals without requiring any model retraining. 3) We conduct extensive experiments and analyses on MIMIC-CXR, demonstrating the feasibility of posterior decision aggregation for multi-input fusion, bypassing the information fusion bottlenecks in LLM. Code is available at https://github.com/yynj98/PDD-RRG.

2 Related Works

Radiology report generation (RRG) aims at clinically accurate reporting for examinations, yet while the target report is fixed, the input remains loosely defined. Early and many recent studies [Jing *et al.*, 2019; Hou *et al.*, 2023; Bu *et al.*, 2024; Xiao *et al.*, 2025] frame it as single-image captioning, ignoring the extra context reporting demands. To mirror clinical routine, additional views such as lateral images [Miura *et al.*, 2021; Qin and Song, 2022; Li *et al.*, 2023; Nicolson *et al.*, 2025], historical data [Wang *et al.*, 2024; Mei *et al.*, 2024; Liu *et al.*, 2025b; Hou *et al.*, 2025], and auxiliary text [Nguyen *et al.*, 2023; Liu *et al.*, 2024a] are progressively folded into the input. MAIRA-2 [Bannur *et al.*, 2024] is the first to ingest all available views, yet concatenated features limit it to one image per view. MLRG [Liu *et al.*, 2025a] anchors on one image and pools all other views via cross-attention, integrating every accessible cue. Despite their success, no existing work noted that richer input can introduce avoidable errors. To address this deficiency, our PDD-RRG aggregates reports from both lean and rich inputs, delivering a unified diagnosis that outperforms any single path.

3 Methodology

3.1 Problem Formulation

Given a study $\mathcal{S} = (\{x_m\}_{m=1}^{M}, z, \{x_n^{pri}\}_{n=1}^{N}, y^{pri}, y)$, it consists of M current images x_m, an auxiliary text z such as "Indication", N previous images x_n^{pri} with their report y^{pri}, and a reference report y. All images are frontal or lateral views, with M typically ranging from 1 to 4. Each study includes at least one current frontal image (CF), while z, x^{pri}, and y^{pri} could all be absent. The traditional RRG task is to learn a function $f_R(\cdot)$ that maps $(\{x_m\}_{m=1}^{M}, z, \{x_n^{pri}\}_{n=1}^{N}, y^{pri})$ or a portion of it to y in one pass.

As shown in Figure 1(e), now we reframe the single-step RRG to a pipeline involving multi-path generations

PRELIMINARY PREPRINT 2026 PREPRINT

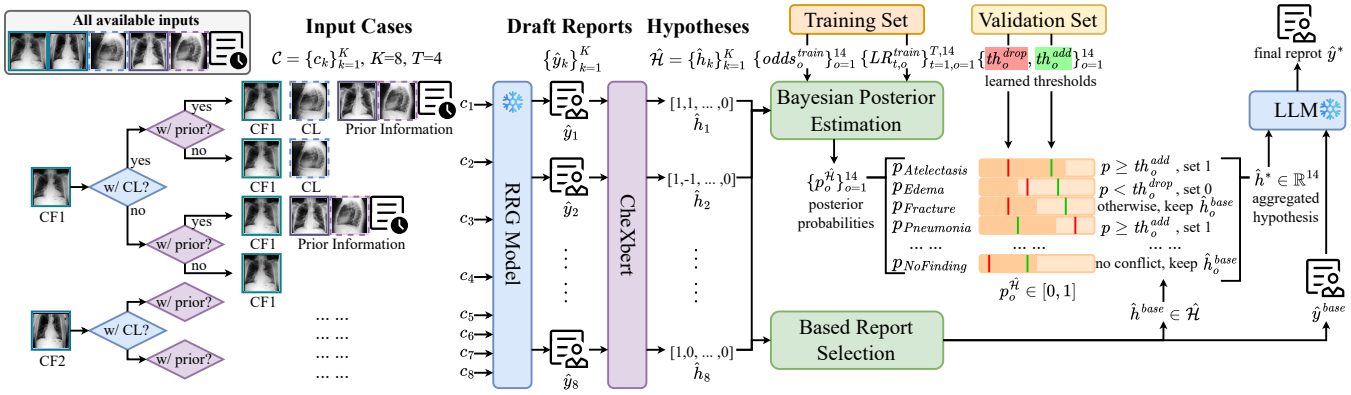

Figure 4: The architecture of our proposed pipeline. Note that $\hat{h}^*_{NoFinding}$ is set to true only when all other revised observations are negative.

and a subsequent decision stage. Given an RRG input $(\{x_m\}_{m=1}^M, z, \{x_n^{pri}\}_{n=1}^N, y^{pri})$, let $\mathcal{C} = \{c_k\}_{k=1}^K$ be a set of clinically plausible input cases derived from it, where each c_k corresponds to a subset of all available inputs (e.g., a specific CF). With a pre-trained RRG model $f_R(\cdot)$, each c_k induces a draft $\hat{y}_k = f_R(c_k)$ which can be regarded as a diagnostic hypothesis. These hypotheses may differ in terms of identified findings, confidence, or completeness. Rather than assuming that any single hypothesis $\hat{y}_k$ is optimal, we aim to infer a final diagnostic decision. Formally,

$$\hat{y}^* = f_D\left(\{\hat{y}_k\}_{k=1}^K\right) \tag{1}$$

where f_D is a posterior decision function that aggregates multiple hypotheses into a unified and robust diagnostic outcome.

3.2 Multi-input Diagnostic Hypothesis Generation

To obtain diverse diagnostic evidence, we need to first generate multiple input cases, where each $c_k \in \mathcal{C}$ reflects a realistic diagnostic scenario. To achieve it, we pair each CF with a lateral image and/or prior information, deriving up to four input configurations as shown in Figure 4 (auxiliary text z is always used if available). All four types exist in real datasets and can be directly received by Type-2 or Type-3 models in Figure 1. We then generate concrete cases according to input types, so that each c_k has its own type too. This is formulated as $type(c_k) \in \{1, 2, \cdots, T\}$, where T denotes the number of input types. Apart from the above four, we can also try different parts of prior reports (e.g., full vs. only "FINDINGS") to obtain eight or even more input types.

For each $c_k \in \mathcal{C}$, the adopted RRG model generates a draft report $\hat{y}_k$. To obtain a structured diagnosis, we utilize CheXbert [Smit *et al.*, 2020] to extract the 14-category labels $\hat{h}_k \in \mathbb{R}^{14}$ from $\hat{y}_k$ and take it as the corresponding hypothesis. Each label $\hat{h}_{k,o} \in \{1, 0, -1, null\}$ in $\hat{h}_k$ encodes the status (among positive, negative, uncertain, and not mentioned) of a specific clinical observation listed in Figure 3 (e.g., Edema). This process results in a set of diagnostic vectors $\hat{\mathcal{H}} = \{\hat{h}_k\}_{k=1}^K$. Notably, all these hypotheses are generated independently based on the same model, without any retraining or architectural modification, allowing PDD-RRG to be applied as a lightweight post-hoc decision module.

This step can be interpreted as eliciting multiple latent reasoning paths of the same model under different informational conditions. Such diversity, instead of being discarded, serves as the foundation for posterior diagnostic decision-making.

3.3 Posterior Diagnostic Decision Framework

In our proposed PDD-RRG framework, we aggregate all hypotheses $\hat{h}_k \in \hat{\mathcal{H}}$ to obtain a revised diagnosis $\hat{h}^*$, and synthesize a revised report $\hat{y}^*$ based on it. Specifically, we first select a based report $\hat{y}^{base}$ as the revision template, and calculate the Bayesian posterior $p_o^{\hat{\mathcal{H}}}$ for each observation o. If there is uncertain or conflicting (1 and 0/$null$) predictions for o, we determine the final diagnosis $\hat{h}_o^*$ (positive or negative) based on its corresponding $p_o^{\hat{\mathcal{H}}}$ and two thresholds (th_o^{drop}, th_o^{add}) learned on the validation set. Finally, we employ a LLM to revise $\hat{y}^{base}$ according to $\hat{h}^*$, resulting in the final report $\hat{y}^*$.

Based Report Selection via Consensus Maximization

For each study, multiple reports are derived from varying input. Instead of relying on the report from the most complete input, we select $\hat{y}^{base}$ as a consensus report that best reflects the model's overall judgment.

We evaluate consensus through the similarity between the diagnosis vectors $\hat{h}_k \in \hat{\mathcal{H}}$. For each label $\hat{h}_{k,o} \in \{1, 0, -1, null\}$ in $\hat{h}_k$, we define a pairwise label-consistency function $sim(\hat{h}_{i,o}, \hat{h}_{j,o}) \in [0, 1]$ to reward similar or identical findings on the observation o:

$$sim(\hat{h}_{i,o}, \hat{h}_{j,o}) = \begin{cases} 1, & \hat{h}_{i,o} = \hat{h}_{j,o} = 0/1; \\ 0.5, & \hat{h}_{i/j,o} = -1 \wedge \hat{h}_{j/i,o} = 0/1; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Then the $sim(\hat{h}_{i,o}, \hat{h}_{j,o})$ is averaged across all mentioned observations (aka., $\hat{h}_{k,o} \neq null$). Denoting $N_{i,j}^{null}$ as the number of labels unmentioned by either side, $sim(\hat{h}_i, \hat{h}_j)$ is computed as:

$$sim(\hat{h}_i, \hat{h}_j) = \frac{1}{14 - N_{i,j}^{null}} \sum_{o=1}^{14} sim(\hat{h}_{i,o}, \hat{h}_{j,o}) \tag{3}$$

IJCAI-ECAI 2026 PREPRINT

Given a candidate $\hat{h}_k$, we sum up all its similarities between available reports to measure consensus:

$$score(\hat{h}_k) = \sum_{i=1, i\neq k}^{K} sim(\hat{h}_k, \hat{h}_i) \tag{4}$$

and select the report that maximizes the score as the $\hat{y}^{base}$.

$$\hat{h}^{base} = \arg\max_{\hat{h}_k \in \mathcal{H}} score(\hat{h}_k) \tag{5}$$

When multiple candidates tie, the shorter report is chosen to reduce verbosity and potential hallucinations. This yields a report that best aligns with the majority of the model's predictions, providing a stable anchor for posterior refinement.

LR-based Bayesian Posterior Probability Estimation

For each clinical observation in the 14 categories, we estimate the diagnostic reliability of each input type using the training data. We collect hypotheses derived from studies in the training set, constructing a large prediction set $\hat{\mathcal{H}}_{train} = \{\hat{h}_k\}_{k=1}^{K_{train}}$. Given the large number of samples, we build only one input case per current frontal (CF) image that integrates the maximum available views among T input types. After the based report selection, we treat the label "not mentioned" as a negative finding, so all $\hat{h}_{k,o} \in \{1, 0, -1\}$. The ground-truth labels $h_k \in \mathbb{R}^{14}$ for each hypothesis likewise form the $\mathcal{H}_{train} = \{h_k\}_{k=1}^{K_{train}}$, where $h_{k,o} \in \{1, 0\}$.

For each input type $t \in \{1, 2, \ldots, T\}$ and observation $o \in \{1, 2, \ldots, 14\}$, we compute the prior recall R and false positive rate FP over the training set as:

$$R_{t,o}^{train} = P(\hat{h}_{k,o} = 1 \mid h_{k,o} = 1, type(c_k) = t, h_k \in \mathcal{H}_{train})$$

$$FP_{t,o}^{train} = P(\hat{h}_{k,o} = 1 | h_{k,o} = 0, type(c_k) = t, h_k \in \mathcal{H}_{train})$$

These statistics quantify how strongly a positive prediction from a given input type t supports the presence of an observation o. We then separately define the likelihood ratio LR for positive and negative observations.

$$LR_{t,o}^{train+} = \frac{R_{t,o}^{train}}{FP_{t,o}^{train}},\ LR_{t,o}^{train-} = \frac{1 - R_{t,o}^{train}}{1 - FP_{t,o}^{train}} \tag{6}$$

We adopt p_o^{train} to represent the prior positive probability $P(y = 1)$ for each observation o:

$$p_o^{train} = P(h_{k,o} = 1 \mid h_k \in \mathcal{H}_{train}) \tag{7}$$

and obtain the corresponding prior odds as

$$odds_o^{train} = \frac{p_o^{train}}{1 - p_o^{train}} \tag{8}$$

Finally, we estimate the posterior log-odds for the given hypotheses $\hat{\mathcal{H}} = \{\hat{h}_k\}_{k=1}^{K}$ as

$$\log odds_o^{\hat{\mathcal{H}}} = \log odds_o^{train} + \sum_{\hat{h}_k \in \hat{\mathcal{H}}} \alpha \log LR_{t,o}^{train} \tag{9}$$

and achieve the posterior positive probability of each observation based on the definition of odds.

$$p_o^{\hat{\mathcal{H}}} = P(\hat{h}_{k,o} = 1 \mid \hat{h}_k \in \hat{\mathcal{H}}) = \frac{1}{1 + \exp\left(-\log odds_o^{\hat{\mathcal{H}}}\right)} \tag{10}$$

Note that in Equation 9, both the values of $LR_{t,o}^{train}$ and α depend on the current prediction $\hat{h}_{k,o}$. We select $LR_{t,o}^{train}$ via

$$LR_{t,o}^{train} = \begin{cases} LR_{t,o}^{train+}, & \hat{h}_{k,o} \in \{1, -1\} \wedge type(c_k) = t; \\ LR_{t,o}^{train-}, & \hat{h}_{k,o} = 0 \wedge type(c_k) = t. \end{cases} \tag{11}$$

and set α to 1 or 1.2 when $\hat{h}_{k,o}$ equals 1 or 0, where the mild amplification factor is to prevent overconfidence arising from sporadic false positives. For uncertain predictions, we assign the condition-specific α with an empirical probability as:

$$\alpha_{t,o} = \frac{P(\hat{h}_{k,o} = -1 \mid h_{k,o} = 1, type(c_k) = t, h_k \in \mathcal{H}_{train})}{P(\hat{h}_{k,o} = -1, type(c_k) = t, \hat{h}_k \in \hat{\mathcal{H}}_{train})} \tag{12}$$

This effectively softens the contribution of uncertain evidence, allowing it to influence posterior inference while preventing unstable amplification.

Diagnostic Decision and Threshold Calibration

Given posterior probabilities $\{p_o^{\hat{\mathcal{H}}}\}_{o=1}^{14}$, PDD-RRG determines the final status (positive or negative) for each observation based on learned thresholds. Considering that different diseases exhibit highly heterogeneous prevalence, uncertainty, and hallucination patterns, we therefore learn two observation-specific thresholds rather than a global one: 1) a drop threshold th_o^{drop} to suppress unreliable positives, and 2) an add threshold th_o^{add} to introduce missing positives. In this way, PDD-RRG adapts to each disease's posterior uncertainty and enables asymmetric control of false positives (FP) and false negatives (FN), which is critical in clinical practice.

Assuming that we already have the learned thresholds, the decision for each observation o is defined as:

$$\hat{h}_o^* = \begin{cases} 0, & p_o^{\hat{\mathcal{H}}} < th_o^{drop}; \\ 1, & p_o^{\hat{\mathcal{H}}} \geq th_o^{add}; \\ \hat{h}_o^{base}, & \text{otherwise}. \end{cases} \tag{13}$$

If both thresholds are met ($th_o^{add} \leq p_o^{\hat{\mathcal{H}}} < th_o^{drop}$), we set $\hat{h}_o^*$ to 1, since missed diagnoses carry heavier clinical penalties. If neither threshold is met ($th_o^{drop} \leq p_o^{\hat{\mathcal{H}}} < th_o^{add}$), we default to the judgment in $\hat{y}^{base}$. To calibrate them, we apply our PDD-RRG to the validation set, setting each $th_o^{drop/add}$ from 0 to 1 in steps of 0.05, and select the value that maximizes the F_1-score over all hypotheses $\{\hat{h}_o^*\}_{val}$ as the final threshold:

$$th_o^{drop/add} = \arg\max_{th_o^{drop/add} \in [0,1]} F_1\left(\{\hat{h}_o^*\}_{val}, \{h_o\}_{val}\right) \tag{14}$$

Report Refinement Based on Revised Diagnosis

Having the final diagnosis $\hat{h}^*$ for 14 observations, we simply leverage a frozen online LLM (e.g., GPT-4) to revise the based report $\hat{y}^{base}$, achieving the refined report $\hat{y}^*$.

$$\hat{y}^* = \text{LLM}(\hat{h}^*, \hat{y}^{base}) \tag{15}$$

Looking back at our proposed PDD-RRG, we explicitly decouple diagnostic decision-making from report generation, mitigating the instability introduced by input selection and generation bias. As a result, the final diagnosis reflects a more comprehensive utilization of available medical evidence and provides more robust support for clinical judgment.

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Model	Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	Mac-F1$_{14}$	Mic-F1$_{14}$	Mac-F1$_5$	Mic-F1$_5$
MAIRA-2	Raw	**33.89**	**13.19**	**34.30**	**31.25**	39.34	56.09	46.84	56.68
	Selected	28.35	10.05	29.52	27.95	39.34	57.07	47.65	58.24
	PDD	–	–	–	–	**42.94**	**59.85**	**51.18**	**60.03**
	PDD-report	31.20	10.52	31.50	28.03	42.25	59.22	50.82	59.85
LLM-RG4	Raw	**34.92**	**12.42**	**33.85**	**30.24**	41.67	58.70	52.29	59.96
	Selected	31.01	10.29	30.76	28.96	41.72	58.65	51.99	60.49
	PDD	–	–	–	–	**42.56**	59.09	**53.13**	**61.44**
	PDD-report	32.69	10.63	32.16	28.76	42.08	**59.66**	52.95	**61.44**
MLRG	Raw	34.90	**12.14**	33.16	**29.62**	32.65	52.80	44.54	53.95
	Selected	34.17	11.86	32.63	29.55	32.93	53.00	45.30	54.82
	PDD	–	–	–	–	**35.37**	53.69	**46.58**	**55.78**
	PDD-report	**35.13**	11.91	**33.63**	29.33	35.32	**54.50**	46.47	55.72

Table 1: Main performance comparison on the MIMIC-CXR test set. Following common RRG evaluation practice, Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion are used to compute F1$_5$. Best and second best results are highlighted in **red** and blue respectively.

4 Experimentation

4.1 Experimental Setup

Dataset. MIMIC-CXR [Johnson *et al.*, 2019] is a widely used RRG dataset comprising multi-view images and longitudinal patient information. Following previous work, we treat the section "FINDINGS" in reports as our target, conducting all experiments on the filtered official split (resulting in 152,173/1,196/2,347 studies in the train/val/test set).

Base RRG Models. To better evaluate generality, we apply PDD-RRG to three state-of-the-art baselines: MAIRA-2 [Bannur *et al.*, 2024], LLM-RG4 [Wang *et al.*, 2025], and MLRG [Liu *et al.*, 2025a]. MAIRA-2 and LLM-RG4 are Type-2 models shown in Figure 1, we adopt the eight and four input types introduced in Section 3.2. For Type-3 MLRG, we create four input types by pairing each anchor image with optional auxiliary images and/or prior information. Note that PDD-RRG is applied strictly as a post-hoc diagnostic decision module, meaning that all base models remain frozen and no retraining or architectural modification is required.

Evaluation Metrics. Performance is evaluated on both NLG and Clinical Efficacy (CE) metrics. We report BLEU-1, BLEU-4, METEOR, and ROUGE-L, while focusing primarily on the CE scores derived by CheXbert. We map "uncertain" labels to negative and compute macro- and micro-average F1 scores for 14 observations extracted from reports.

4.2 Main Results on Various Base Models

We apply PDD-RRG to three representative RRG models: MAIRA-2, LLM-RG4, and MLRG. For each model, we compare four scenarios under multi-input settings: (i) the original output by richest input (Raw), (ii) the most self-consistent report $\hat{y}^{base}$ selected in our approach (Selected), (iii) the diagnostic labels $\hat{h}^*$ produced by our posterior decision module (PDD), and (iv) the reconstructed report $\hat{y}^*$ generated from $\hat{h}^*$ (PDD-report). Table 1 shows the results of these scenarios across all evaluation metrics. Two key observations emerge.

First, PDD-RRG consistently improves clinical performance across all base models. These uniform gains, achievable regardless of the underlying architecture, demonstrate the strong model-agnostic generalization of our approach.

Second, Selected frequently outperforms Raw on clinical metrics. This confirms that the most information-complete input does not necessarily yield the most reliable diagnosis, validating the necessity of explicit posterior decision-making under heterogeneous inputs.

Regarding language quality, PDD-report shows slightly lower NLG scores than Raw but consistently outperforms Selected, which serves as the backbone for report reconstruction. This indicates that label-aligned report reconstruction preserves, and in some cases improves, textual quality. Overall, *PDD-RRG provides robust, architecture-independent improvements in clinical performance while maintaining competitive language performance*, validating its suitability as a general plug-and-play diagnostic refinement layer for RRG.

Method	Mac-F1$_{14}$	Mic-F1$_{14}$	Mac-F1$_5$	Mic-F1$_5$
Baseline (Selected)	39.34	57.07	47.65	58.24
PDD	42.94	**59.85**	**51.18**	**60.03**
PDD w/o LR	42.59	59.17	50.90	59.32
PDD w/o OSDT	**43.48**	59.15	51.16	59.55

Table 2: Ablation study of different components on MAIRA-2.

4.3 Ablation Study

We conduct ablation studies to evaluate two key components of PDD-RRG: Likelihood-Ratio (LR)-based posterior modeling and Observation-Specific Decision Thresholds (OSDT). Results on MAIRA-2 are reported in Table 2.

Effect of Likelihood-Ratio Modeling. We replace the LR-based formulation with a frequency-based aggregation scheme ("w/o LR"), computing disease confidence as the frequency of non-negative predictions in $\hat{\mathcal{H}}$ (uncertain as 0.5). Although frequency aggregation improves over the selected report, it fails to match LR modeling. These results confirm that *PDD-RRG's gains stem from explicitly modeling heterogeneous diagnostic evidence rather than naive aggregation*.

Effect of Observation-Specific Decision Thresholds. We further compare PDD with a variant that applies two global thresholds $th^{drop/add}$ to all observations ("w/o OSDT"). While the unified-threshold variant yields a slightly higher macro F1$_{14}$, *the gain is unstable and likely coincidental, failing to generalize to other clinically relevant metrics*. This demonstrates that *a global decision boundary is insufficient to capture the substantial inter-disease variability*.

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Model	Method	Mac-F1$_{14}$	Mic-F1$_{14}$	Mac-F1$_5$	Mic-F1$_5$
MAIRA-2	Vote	38.98	57.89	47.13	58.24
	PDD	**42.94**	**59.85**	**51.18**	**60.03**
LLM-RG4	Vote	41.30	58.81	51.18	60.41
	PDD	**42.56**	**59.09**	**53.13**	**61.44**
MLRG	Vote	32.64	53.14	45.23	54.60
	PDD	**35.37**	**53.69**	**46.58**	**55.78**

Table 3: Comparison between Voting and PDD-RRG across three backbones. Best results for each backbone are highlighted in **red**.

Error Type	Total	Recoverable	Rate (%)
False Negative (FN)	2315	877	37.88
False Positive (FP)	2019	1361	67.41

Repair Type	Vote	PDD	Vote-only	PDD-only
FN recovery	155	**538**	**4**	**387**
FP recovery	**542**	203	384	45

Table 4: Theoretical recoverability of test-set errors and comparison of FN and FP recovery under Voting and PDD-RRG. "Vote-only" and "PDD-only" denote cases uniquely repaired by each method.

4.4 Comparison with Majority Voting

Conventional RRG is typically formulated as an end-to-end text generation problem without an explicit diagnostic decision layer, and thus prior work has not addressed how to reconcile potentially conflicting outputs arising from multiple inputs. For a principled comparison, we adopt majority voting as a representative and widely used aggregation baseline.

As shown in Table 3, PDD consistently outperforms Voting across all base models. This demonstrates that *equal-weight voting fails to effectively aggregate heterogeneous diagnostic evidence*, especially in clinical scenarios where different findings are preferentially revealed by different input configurations. In such cases, Voting suffers from a *vote dilution* effect, where highly informative but infrequent signals are overwhelmed by numerous weak or uninformative votes.

Overall, these results confirm that majority voting is suboptimal in medical decision aggregation, whereas *PDD-RRG provides a more reliable and clinically aligned alternative for multi-input radiology report generation*.

4.5 Decision Behavior Under Conflicting Inputs

While Section 4.2 demonstrates the effectiveness of PDD-RRG, this section investigates the underlying decision mechanisms. We analyze how different aggregation strategies resolve conflicting hypotheses, revealing a fundamental asymmetry in distinct error types handing. We compare the selected base output with two post-hoc strategies on MAIRA-2:

Silence-biased Perception (Base). The selected report $\hat{y}^{base}$ inherits the model's training bias. It defaults to silence under uncertainty, leading to high precision but low recall.

Consensus-driven Conservatism (Vote). Majority voting aggregates views based on agreement. While effective at suppressing noise, it indiscriminately penalizes minority but potentially correct signals, reinforcing collective silence.

Evidence-weighted Decision (PDD). In contrast, PDD-RRG re-weights alternative views via likelihood ratios, allowing reliable minority evidence to override defaults, shifting the decision boundary from silence toward detection.

Asymmetric Error Recoverability in Multi-view RRG

Errors in $\hat{h}^{base}$ are categorized as false negatives (FN) , where GT is positive but predicted negative, and false positives (FP), where GT is negative but predicted positive. An error is deemed recoverable if at least one $\hat{h}_k \in \hat{\mathcal{H}}$ predicts correctly, defining the theoretical upper bound achievable by post-hoc decision rules in our multi-view setting.

We first establish the theoretical upper bound for post-hoc correction. As shown in Table 4, all based reports contain 2,315 FNs and 2,019 FPs, of which 37.88% and 67.41% are recoverable, respectively. This indicates that hallucinated findings are more easily contradicted across views, while missed diagnoses require explicit positive evidence. Consequently, FNs are intrinsically harder to recover than FPs.

Within this recoverable pool, Table 4 reveals a fundamental asymmetry in how Voting and PDD repair errors:

PDD Strictly Dominates in FN Recovery. Among recoverable FNs, PDD recovers 61.35% of cases, compared to only 17.67% achieved by Voting. More importantly, *PDD uniquely recovers 387 cases, whereas Voting contributes virtually no unique value (4 cases)*.

This strict dominance highlights a limitation of consensus-based decision rules in high-uncertainty medical settings. FN errors typically arise from weak or view-dependent evidence, where only a minority of input configurations reveal the abnormality. Voting indiscriminately suppresses minority signals, while PDD-RRG's likelihood-ratio weighting exploits weak but reliable evidence to rescue missed diagnoses.

Voting only Looks Better in FP Suppression. Although Voting suppresses a larger fraction of FPs than PDD (39.82% vs. 14.92%), but *this apparent advantage is largely an artifact of protocol-driven silence mapping and passive consensus bias, rather than explicit diagnostic refutation*.

First, radiology data are highly imbalanced toward negative labels and silence. Trained on such data, models tend to default to silence when visual evidence is weak or ambiguous. Voting amplifies this bias by interpreting cross-view omission as negative evidence. Since the test set follows the same negative-dominant distribution, Voting gains a natural albeit passive advantage by simply aligning with the majority class.

Second, many "suppressed" FPs are not truly contradicted by other views. In most alternative $\hat{h}_k$, the model omits the finding rather than negating it. Standard evaluation protocols conventionally map unmentioned categories to negative labels. Consequently, when the majority of views are silent, they form a "consensus of omission". If the ground truth is also unmentioned, this is counted as a correct suppression. However, this represents a *coincidental agreement* driven by default mapping rules, not a confirmed diagnostic exclusion.

Third, hallucinated positives are typically *stochastic and unstable across views*, making them particularly vulnerable to consensus-based dilution.

In contrast, PDD-RRG enables targeted recovery of clinically critical abnormalities that are weakly expressed and thus silenced by the majority. Notably, approximately 20% of the FPs repaired by PDD are unique cases that Voting fails to re-

IJCAI-ECAI 2026 PREPRINT

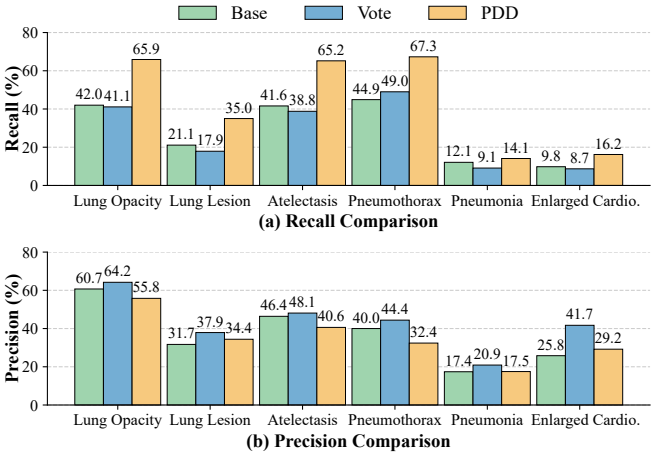

Figure 5: Recall (a) and Precision (b) comparisons. The numbers above bars indicate the corresponding post-decision values.

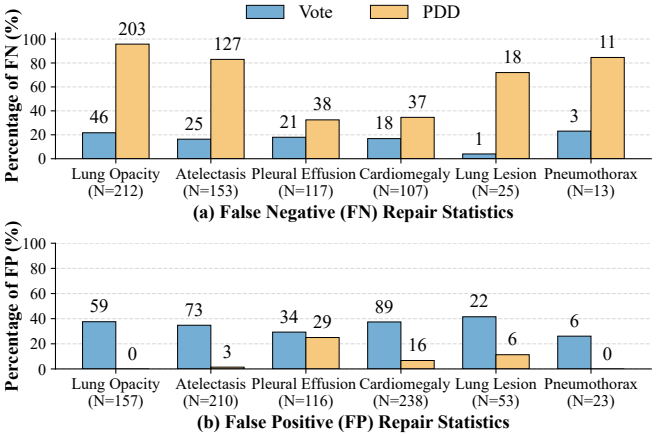

Figure 6: Error repair statistics for FN (a) and FP (b). The numbers above bars indicate the exact count of repaired samples.

cover, confirming that PDD-RRG provides essential corrective power where conservative consensus fails.

Clinical Trade-Offs Under Asymmetric Error Correction

Figure 5 presents per-disease Recall and Precision for the selected $\hat{h}^{base}$, Voting, and PDD under multi-view inputs. Due to space constraints, we show six representative diseases here (full results are provided in our code repository).

As visualized in Figure 5, both Base and Voting strategy exhibit a "false-good" phenomenon: high Precision but very low Recall (e.g., $<$25% for Pneumonia and Lung Lesion). In stark contrast, *PDD achieves a systematic recall surge across all diseases*. For instance, Recall for Lung Opacity and Pneumothorax jumps by over 20 percentage points (e.g., Pneumothorax: 44.9% $\rightarrow$ 67.3%). This demonstrates that PDD-RRG effectively activates valid positive signals that were suppressed by the consensus mechanism. These gains are most pronounced for *subtle or focal abnormalities* such as Lung Opacity, Atelectasis and Lung Lesion.

Figure 5(b) shows a moderate Precision drop for PDD, expected due to the denominator effect: Base and Voting predict few positives (small denominator), resulting in inflated precision. By activating potential findings, PDD-RRG expands the denominator. *From a clinical perspective, this trade-off is highly desirable: missing a life-threatening condition is far costlier than raising a false alarm.*

Overall, PDD-RRG transforms the model *from a passive silence-biased predictor into a clinically viable screening system*, trading a marginal precision drop for a critical breakthrough in sensitivity.

Lesion-level Repair Analysis

To investigate specific repair behaviors across diverse pathologies, we analyze error repair on the lesion-level. Due to space limits, Figure 6 reports six representative diseases, with full results in our code repository.

Figure 6(b) shows that Voting is highly effective at FP suppression for both global findings such as Cardiomegaly and regional findings like Lung Opacity and Atelectasis. For global abnormalities, Voting leverages their expected cross-view consistency to eliminate projection artifacts; for regional findings, it acts as a denoiser, filtering out spurious single-view signals that lack corroboration.

However, a joint inspection of Figure 6(a) and (b) reveals a critical asymmetry: the categories where Voting excels at FP suppression are precisely those where PDD achieves massive FN recovery. This suggests that *Voting's "success" largely stems from discarding weak but genuine positive evidence*. Consequently, Voting fails to repair FNs effectively (e.g., Lung Opacity: 46 vs. 203 recovered by PDD).

This consensus bias *is particularly harmful for focal and subtle lesions, where cross-view agreement is rare*. For Lung Lesion, Voting's FN recovery collapses to near zero (1 case fixed), whereas PDD rescues 18 cases, *an order of magnitude improvement*. Most critically, for Pneumothorax, a life-threatening emergency, PDD rescues 11 missed cases compared to only 3 by Voting, demonstrating its superior sensitivity to clinically critical abnormalities.

Overall, these lesion-level analyses reveal a fundamental trade-off: *Voting favors conservatism by suppressing positives, while PDD-RRG prioritizes patient safety by recovering clinically meaningful findings*, especially when diagnostic evidence is sparse, localized, or view-dependent.

5 Conclusion

Our work provides the first systematic investigation of the post-processing decision layer for RRG. To address limitations in multi-view report generation, we propose the PDD-RRG, a likelihood-based aggregation framework that reconciles divergent inference paths from multiple input configurations of the same study. PDD-RRG enhances the use of clinically meaningful signals and unlocks the latent diagnostic capabilities of backbone models. Crucially, PDD-RRG requires no model retraining, effectively mitigates the information fusion bottlenecks in existing multi-view RRG systems. Nevertheless, PDD-RRG is inherently limited to resolving conflicts within the model's own outputs, but cannot fully correct severe hallucinations or intrinsic diagnostic blind spots. Future work will explore incorporating reliable external knowledge or multi-agent collaboration to overcome these limitations.

IJCAI-ECAI 2026 PREPRINT

Acknowledgments

This work was supported by Jiangsu Key Technology Research Development Program (BF2025036), and Hong Kong RGC grant GRF #15611021. They are also with Jiangsu Key Lab of Language Computing, Suzhou.

Contribution Statement

Yang Yu and Yiming Ji contributed equally to this work.

References

[Bannur *et al.*, 2024] 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. *CoRR*, abs/2406.04449, 2024.

[Bu *et al.*, 2024] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14194–14204. IEEE, 2024.

[Hou *et al.*, 2023] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. ORGAN: observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8108–8122. Association for Computational Linguistics, 2023.

[Hou *et al.*, 2025] Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. RADAR: enhancing radiology report generation with supplementary knowledge injection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 26366–26381. Association for Computational Linguistics, 2025.

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

[Jing *et al.*, 2019] Baoyu Jing, Zeya Wang, and Eric P. Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6570–6580. Association for Computational Linguistics, 2019.

[Johnson *et al.*, 2019] 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. *CoRR*, abs/1901.07042, 2019.

[Ju *et al.*, 2025] Xincheng Ju, Dong Zhang, Junhui Li, Shoushan Li, and Guodong Zhou. Enhanced generative framework with llms for multimodal emotion-cause pair extraction in conversations. *IEEE Trans. Multim.*, 27:4924–4935, 2025.

[Li *et al.*, 2023] Qingqiu Li, Jilan Xu, Runtian Yuan, Mohan Chen, Yuejie Zhang, Rui Feng, Xiaobo Zhang, and Shang Gao. Enhanced knowledge injection for radiology report generation. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Istanbul, Turkiye, December 5-8, 2023*, pages 2053–2058. IEEE, 2023.

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

[Liu *et al.*, 2024a] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison X. Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part III*, volume 15003 of *Lecture Notes in Computer Science*, pages 433–443. Springer, 2024.

[Liu *et al.*, 2024b] Liang Liu, Dong Zhang, Shoushan Li, Guodong Zhou, and Erik Cambria. Two heads are better than one: Zero-shot cognitive reasoning via multi-llm knowledge fusion. In Edoardo Serra and Francesca Spezzano, editors, *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1462–1472. ACM, 2024.

[Liu *et al.*, 2025a] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10348–10359. Computer Vision Foundation / IEEE, 2025.

[Liu *et al.*, 2025b] 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-25,*

IJCAI-ECAI 2026 PREPRINT

Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, pages 5595–5603. AAAI Press, 2025.

[Mei *et al.*, 2024] Xin Mei, Rui Mao, Xiaoyan Cai, Libin Yang, and Erik Cambria. Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 4699–4708. ACM, 2024.

[Miura *et al.*, 2021] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5288–5304. Association for Computational Linguistics, 2021.

[Nguyen *et al.*, 2023] Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. Pragmatic radiology report generation. In *Machine Learning for Health, ML4H@NeurIPS 2023, 10 December 2023, New Orleans, Louisiana, USA*, volume 225 of *Proceedings of Machine Learning Research*, pages 385–402. PMLR, 2023.

[Nicolson *et al.*, 2024] Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585, 2024.

[Nicolson *et al.*, 2025] Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. The impact of auxiliary patient data on automated chest x-ray report generation and how to incorporate it. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 177–203. Association for Computational Linguistics, 2025.

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

[Serra *et al.*, 2023] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest x-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4891–4904. Association for Computational Linguistics, 2023.

[Smit *et al.*, 2020] 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1500–1519. Association for Computational Linguistics, 2020.

[Tang *et al.*, 2025] Quanwei Tang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. A comprehensive graph framework for question answering with mode-seeking preference alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 21504–21523. Association for Computational Linguistics, 2025.

[Wang *et al.*, 2024] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LV*, volume 15113 of *Lecture Notes in Computer Science*, pages 183–200. Springer, 2024.

[Wang *et al.*, 2025] 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 Toby Walsh, Julie Shah, and Zico Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 8250–8258. AAAI Press, 2025.

[Xiao *et al.*, 2025] Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 8664–8672. AAAI Press, 2025.

IJCAI-ECAI 2026 PREPRINT