PACER: PROGRESSIVE AVAILABILITY-CONDITIONED EVIDENCE ROUTING FOR RADIOLOGY REPORT GENERATION UNDER INCOMPLETE CLINICAL CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation (RRG) increasingly incorporates heterogeneous clinical evidence, such as multi-view radiographs and previous reports, whose availability varies across examinations. However, accommodating different input combinations does not ensure effective evidence use: generated reports may still omit or inaccurately describe clinically relevant findings. To address this problem, we propose PACER, a Progressive Availability-Conditioned Evidence Routing framework for structured incomplete-context RRG that follows a Refine–Calibrate–Commit pipeline. It first refines observed visual representations through endpoint-preserving patchwise routing across frozen encoder depths, incorporating complementary cues while retaining the pretrained terminal representation. It then calibrates the language-model prefix according to the observed evidence and availability state, adapting the shared generator's conditioning as the available source set changes. Finally, it generates polarity-structured clinical commitments before the report in the same autoregressive trajectory, providing structured clinical context for subsequent generation. Experiments demonstrate state-of-the-art clinical efficacy across all four MIMIC-RG4 settings and strong MIMIC-CXR performance, while maintaining competitive language-generation quality.

1 INTRODUCTION

Radiology report generation (RRG) aims to translate radiographic examinations into clinically meaningful reports, potentially reducing repetitive reporting workload and improving interpretation efficiency (Wang et al., 2025b; Liu et al., 2024). In clinical practice, however, radiologists may draw on heterogeneous evidence whose availability varies across examinations. Such source-level variability motivates RRG models that operate under structured incompleteness, where optional evidence sources may be present or absent as whole sources (Liu et al., 2025b;a; 2026a).

Recent RRG research increasingly treats multi-view radiographs and historical context as complementary clinical evidence. Multi-view longitudinal learning, historical constraints, and prior-guided decoding improve the extraction and integration of complementary spatial and longitudinal evidence (Liu et al., 2025a;b; 2026a). To accommodate unavailable context, MLRG introduces tokenized absence encoding, while LLM-RG4 supports four input configurations through adaptive token fusion and token-level weighting (Liu et al., 2025a; Wang et al., 2025b). Related incomplete multimodal methods further improve robustness to missing sources through adaptive fusion, dynamic weighting, retrieval, or prompting (Yao et al., 2024; Li et al., 2025; Lang et al., 2025). These advances substantially improve heterogeneous evidence integration and robustness under varying input conditions. However, as the available source set changes, how a shared generator should adapt its use of observed evidence throughout the generation process remains insufficiently explored. Without such coordinated adaptation, clinically relevant cues may still be underused, resulting in omitted or inaccurately described findings.

To address this gap, we propose PACER, a Progressive Availability-Conditioned Evidence Routing framework organized as a **Refine–Calibrate–Commit** pipeline. At the visual-representation

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

interface, relying only on the terminal encoder representation may leave complementary information distributed across encoder depths underutilized. **Refine** therefore performs endpoint-preserving patchwise depth routing to enrich the observed visual evidence while retaining the pretrained terminal representation. Building on the refined evidence, changes in the observed source set further call for adaptive language-model conditioning. **Calibrate** addresses this by adjusting the language-model prefix according to both the observed evidence and its availability state. Finally, even with adapted representations and conditioning, free-form decoding still lacks an explicit polarity-aware clinical scaffold to guide subsequent report generation. **Commit** addresses this by generating positive, negative, and uncertain clinical commitments before the report within the same autoregressive trajectory, thereby structuring the clinical evidence that conditions the subsequent report. Overall, PACER progressively coordinates available evidence under varying input configurations. We validate the framework through extensive experiments on MIMIC-RG4 and MIMIC-CXR.

Our main contributions are as follows:

- To the best of our knowledge, we are the first to explicitly formulate structured incomplete-context RRG as an availability-conditioned evidence-routing problem, emphasizing how observed evidence should be progressively utilized rather than merely accommodated as source availability changes.
- We develop PACER to implement multi-interface evidence routing through three coordinated mechanisms: endpoint-preserving patchwise depth routing, availability-conditioned low-rank prefix calibration, and polarity-structured clinical commitment.
- Experiments on MIMIC-RG4 and MIMIC-CXR show that PACER achieves state-of-the-art CE F1 across all four MIMIC-RG4 settings while maintaining competitive language-generation performance, together with strong conventional-RRG results. Component and mechanism ablations further validate the proposed routing design.

2 RELATED WORK

We review three lines of work most relevant to our setting: flexible and context-enriched RRG, incomplete multimodal learning, and evidence-grounded structured generation.

Flexible and context-enriched RRG. Multi-view and longitudinal RRG exploit additional projections and historical information to enrich spatial and temporal evidence (Liu et al., 2025b;a; 2026a), while recent methods further model disease evolution through temporal decoupling and progression-aware prompting (Dong et al., 2026; Liu et al., 2026b). LLM-RG4 establishes a flexible four-context setting with adaptive token fusion and token-level weighting for variable inputs (Wang et al., 2025b). These methods improve contextual flexibility and evidence integration under variable input conditions. PACER complements this line by focusing on effective evidence utilization rather than input accommodation alone as source availability changes.

Incomplete multimodal learning. Existing methods address missing inputs through reconstruction, representation decoupling, and adaptive fusion (Liu et al., 2023; Wang et al., 2024; Yao et al., 2024), or through dynamic weighting, retrieval, and prompting (Li et al., 2025; Lang et al., 2025; Pipoli et al., 2025). In RRG, DiA-gnostic VLVAE handles missing clinical context through MoE-based shared-latent inference, allowing the shared posterior to rely on observed modalities when context is unavailable (Shaik et al., 2026). In contrast, PACER focuses on availability-conditioned adaptation of the shared generator, calibrating its language-model prefix according to both observed evidence and source availability.

Evidence grounding and structured generation. RRG methods improve visual grounding through hierarchical representations, anatomical regions, and auxiliary alignment (Huang et al., 2023; Tanida et al., 2023; Gao et al., 2026). Structured generation has explored description planning, diagnosis-derived prompting, observation planning, topic organization, and explicit clinical reasoning (Nishino et al., 2022; Jin et al., 2024; Hou et al., 2023; Cheng & Subramanian, 2026; Zhang et al., 2026). Recent work further explores self-critique and confidence-guided rewriting (Yan et al., 2026; Yu et al., 2026). Compared with prior work, PACER couples endpoint-preserving visual refinement with polarity-structured clinical commitments to guide subsequent report generation.

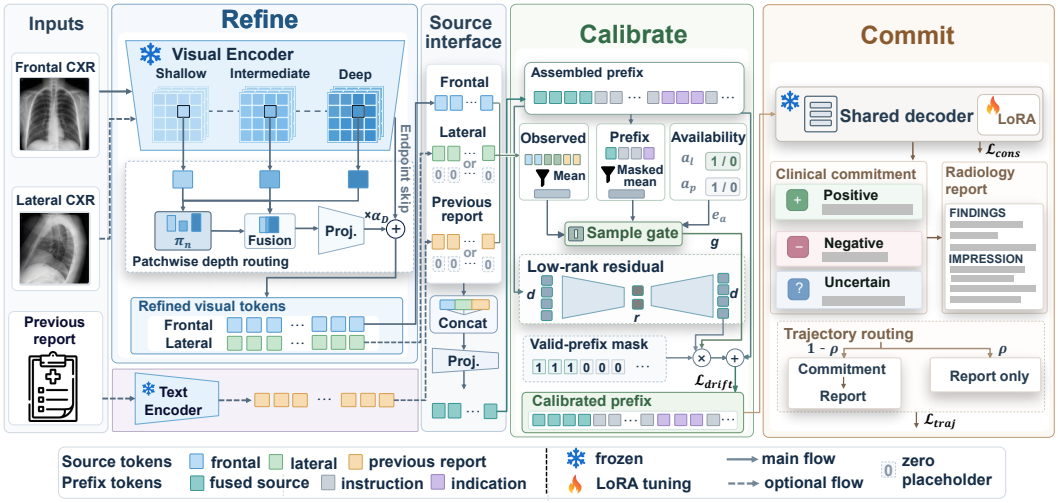

Figure 1: Overview of PACER under structured incomplete clinical context. Refine enhances observed visual evidence through endpoint-preserving patchwise depth routing, Calibrate applies availability-conditioned prefix calibration, and Commit generates polarity-structured clinical commitments before the report in the same autoregressive trajectory.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We study radiology report generation under structured incomplete clinical context, where incompleteness is defined at the source level over three predefined sources: a frontal radiograph $\mathbf{I}_f$, a lateral radiograph $\mathbf{I}_l$, and a previous report $\mathbf{T}_p$. The frontal radiograph is always observed, whereas the lateral radiograph and previous report may be unavailable. Indication/history, when available, is included as prompt context and is not treated as an availability axis.

We represent auxiliary-source availability by $\mathbf{a} = (a_l, a_p) \in \{0,1\}^2$, where a_l and a_p indicate the presence of the lateral view and previous report. SN, SW, MN, and MW correspond to $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, respectively. Here S/M denote frontal-only/frontal–lateral image inputs, and N/W denote the absence/presence of a previous report. Let $\mathbf{Y} = (y_1, \ldots, y_T)$ be the scenario-specific target report. Following LLM-RG4 (Wang et al., 2025b), unavailable sources occupy zero-valued feature slots in a fixed source layout.

PACER progressively routes observed evidence through the **Refine–Calibrate–Commit** pipeline (Figure 1). **Refine** processes the observed radiographs, while a separate text encoder handles the previous report; their representations are assembled into the language-model prefix. **Calibrate** adapts this prefix to the observed evidence and availability state, and **Commit** generates polarity-structured commitments to guide report content.

3.2 REFINE: ENDPOINT-PRESERVING PATCHWISE DEPTH ROUTING

Representations from different Transformer depths encode complementary visual information (Ranftl et al., 2021; Lin et al., 2025; Huang et al., 2023). Because clinically relevant radiographic evidence can be spatially localized (Tanida et al., 2023), we allow each patch to select its own depth distribution rather than using a global mixture shared across the image.

For each observed radiograph $\mathbf{I}_s$, $s \in \{f, l\}$, let $\mathbf{V}_s^{(\ell)} = [\mathbf{v}_{s,1}^{(\ell)}, \ldots, \mathbf{v}_{s,N_v}^{(\ell)}]^\top \in \mathbb{R}^{N_v \times d_v}$ denote the non-CLS patch features extracted by frozen RAD-DINO (Pérez-García et al., 2025) at encoder depth $\ell \in \mathcal{K}$. Here, N_v is the number of non-CLS patch tokens, d_v is their feature dimension, and $\mathcal{K}$ is the set of selected encoder depths. The same encoder and depth router are applied independently to the frontal and lateral sources. For readability, we omit the source index s below and write $\mathbf{V}^{(\ell)}$ and

$\mathbf{v}_n^{(\ell)}$. To align depth-dependent feature statistics, we map each patch into a common routing space:

$$\mathbf{u}_n^{(\ell)} = \mathbf{P}_\ell \, \mathrm{LN}_\ell\Big(\mathbf{v}_n^{(\ell)}\Big) + \mathbf{b}_\ell, \tag{1}$$

where $\mathbf{P}_\ell \in \mathbb{R}^{d_r \times d_v}$ is a learnable projection, LN_ℓ is a layer-specific normalization, $\mathbf{b}_\ell \in \mathbb{R}^{d_r}$ is the corresponding bias, and d_r is the routing dimension.

A shared scorer produces a patch-specific distribution over the selected depths:

$$\pi_{n,\ell} = \frac{\exp\Big(\mathbf{w}_\pi^\top \mathbf{u}_n^{(\ell)} + \beta_\ell\Big)}{\sum_{j \in \mathcal{K}} \exp\Big(\mathbf{w}_\pi^\top \mathbf{u}_n^{(j)} + \beta_j\Big)}, \qquad \sum_{\ell \in \mathcal{K}} \pi_{n,\ell} = 1, \tag{2}$$

where $\mathbf{w}_\pi \in \mathbb{R}^{d_r}$ is the shared routing vector, $\beta_\ell \in \mathbb{R}$ is a learnable depth-specific bias, and j indexes candidate depths in $\mathcal{K}$. Thus, $\boldsymbol{\pi}_n = (\pi_{n,\ell})_{\ell \in \mathcal{K}}$ varies across patches, allowing spatially localized evidence to draw differently on representations from different encoder depths rather than sharing a single image-level depth mixture.

To incorporate complementary cross-depth cues without replacing the pretrained terminal representation, we use the routed mixture to predict an additive correction. This preserves a direct path from the pretrained endpoint while allowing each patch to incorporate information from earlier encoder depths:

$$\boldsymbol{\delta}_n = \mathbf{W}_o \, \mathrm{LN}_\delta\left(\sum_{\ell \in \mathcal{K}} \pi_{n,\ell} \mathbf{u}_n^{(\ell)}\right) + \mathbf{b}_o, \tag{3}$$

$$\widehat{\mathbf{v}}_n = \mathbf{v}_n^{\mathrm{end}} + \alpha_D \boldsymbol{\delta}_n, \qquad \alpha_D = \bar{\alpha}_D \, \mathrm{sigmoid}(\eta_D), \tag{4}$$

where $\mathbf{v}_n^{\mathrm{end}}$ denotes the final normalized RAD-DINO patch representation, LN_δ normalizes the routed mixture, $\mathbf{W}_o \in \mathbb{R}^{d_v \times d_r}$ and $\mathbf{b}_o \in \mathbb{R}^{d_v}$ project it back to the encoder feature space, η_D is a learnable scalar, and $\bar{\alpha}_D$ bounds the scalar residual weight α_D.

With $\mathbf{W}_o$ and $\mathbf{b}_o$ initialized to zero, the module starts exactly from the pretrained terminal representation and learns patch-specific corrections during training. Restoring the source index, Refine outputs $\widehat{\mathbf{V}}_f$ and, when available, $\widehat{\mathbf{V}}_l$, which are passed to the source interface used by Calibrate.

3.3 Calibrate: Availability-Calibrated Prefix Routing

Motivated by context-aware prompting for incomplete multimodal inputs (Lang et al., 2025; Pipoli et al., 2025), we adapt the pre-Transformer prefix to observed evidence and source availability. Each observed source—the refined image features $\widehat{\mathbf{V}}_f, \widehat{\mathbf{V}}_l$ or the text encoding of $\mathbf{T}_p$—is compressed and projected into pre-fusion tokens $\mathbf{Z}_s \in \mathbb{R}^{N_q \times d}$, $s \in \{f, l, p\}$, where N_q is the number of compressed tokens per source and d is the language-model embedding dimension. Unavailable sources use zero-valued slots. The source tokens are fused, then assembled with the prompt context into $\mathbf{E}_P$.

During teacher forcing, $\mathbf{E} = [\mathbf{E}_P; \mathbf{E}_T] \in \mathbb{R}^{T' \times d}$ concatenates this prefix with the target-side embeddings $\mathbf{E}_T$ of the routed sequence in Section 3.4, where T' counts input positions. The mask $\mathbf{m} \in \{0,1\}^{T'}$ selects valid prefix positions, excluding target-side positions and padding. We define

$$\begin{aligned} \bar{\mathbf{h}}_P &= \frac{\mathbf{E}^\top \mathbf{m}}{\|\mathbf{m}\|_1}, \qquad \bar{\mathbf{z}}_s = \frac{(\mathbf{Z}_s)^\top \mathbf{1}_{N_q}}{N_q}, \\ \bar{\mathbf{z}}_{\mathbf{a}} &= \frac{\bar{\mathbf{z}}_f + a_l \bar{\mathbf{z}}_l + a_p \bar{\mathbf{z}}_p}{1 + a_l + a_p}, \end{aligned} \tag{5}$$

where $\|\mathbf{m}\|_1$ counts valid prefix positions, $\mathbf{1}_{N_q}$ is an N_q-dimensional all-ones vector, $\bar{\mathbf{h}}_P \in \mathbb{R}^d$ is the prefix summary, and $\bar{\mathbf{z}}_{\mathbf{a}} \in \mathbb{R}^d$ averages only observed sources, avoiding dilution by missing-source zero slots.

Each availability state $\mathbf{a}$ is associated with a learnable embedding $\mathbf{e}_{\mathbf{a}} \in \mathbb{R}^{d_a}$, where d_a is the availability-embedding dimension. The three inputs provide complementary conditioning cues: $\bar{\mathbf{h}}_P$ summarizes the assembled prefix after source fusion and prompting, $\bar{\mathbf{z}}_{\mathbf{a}}$ explicitly summarizes the

observed sources before fusion, and $\mathbf{e_a}$ identifies which optional sources are available. Their combination conditions the adaptation strength on both evidence content and source availability:

$$g = \mathrm{sigmoid}\big(\mathbf{w}_g^\top \,\mathrm{GELU}\big(\mathbf{W}_g[\bar{\mathbf{h}}_P; \bar{\mathbf{z}}_\mathbf{a}; \mathbf{e_a}] + \mathbf{b}_g\big) + b_{\mathrm{out}}\big)\,, \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation, $\mathbf{W}_g \in \mathbb{R}^{d_g \times (2d+d_a)}$ and $\mathbf{b}_g \in \mathbb{R}^{d_g}$ parameterize the hidden layer, $\mathbf{w}_g \in \mathbb{R}^{d_g}$ and $b_{\mathrm{out}} \in \mathbb{R}$ are the output projection and scalar bias, respectively, and d_g is the gate hidden dimension. The resulting scalar $g \in (0,1)$ controls the adaptation strength for each sample.

To provide token-specific corrections with a compact residual parameterization, we use a position-wise bottleneck mapping $\mathcal{B} : \mathbb{R}^{T' \times d} \to \mathbb{R}^{T' \times d}$, scaled by the sample-wise gate g:

$$\begin{aligned} \mathcal{B}(\mathbf{E}) &= \mathrm{Dropout}\big[\mathrm{GELU}\big(\mathrm{LN}(\mathbf{E})\mathbf{W}_\downarrow^\top\big)\big]\,\mathbf{W}_\uparrow^\top, \\ \widetilde{\mathbf{E}} &= \mathbf{E} + g\ \mathrm{Diag}(\mathbf{m})\mathcal{B}(\mathbf{E}), \end{aligned} \tag{7}$$

where LN denotes layer normalization, $\mathbf{W}_\downarrow \in \mathbb{R}^{r \times d}$ and $\mathbf{W}_\uparrow \in \mathbb{R}^{d \times r}$ are the learnable down- and up-projection matrices, $r \ll d$ is the bottleneck dimension, and $\mathrm{Diag}(\mathbf{m}) \in \mathbb{R}^{T' \times T'}$ is the diagonal mask induced by $\mathbf{m}$. Thus, $\mathcal{B}(\mathbf{E})$ is a token-wise correction with the same shape as $\mathbf{E}$, while only valid prefix positions receive this correction. Zero-initializing $\mathbf{W}_\uparrow$ gives $\widetilde{\mathbf{E}} = \mathbf{E}$ at initialization.

The gate is target-blind because it uses only valid-prefix information, observed sources, and availability, not $\mathbf{E}_T$. Together with the position-wise $\mathcal{B}$, this gives $\partial \widetilde{\mathbf{E}}_P / \partial \mathbf{E}_T = \mathbf{0}$. The mask ensures $[\widetilde{\mathbf{E}} - \mathbf{E}]_{t,:} = \mathbf{0}$ whenever $m_t = 0$. The same prefix-only correction is applied before the Transformer during teacher forcing and autoregressive generation.

3.4 COMMIT: POLARITY-STRUCTURED CLINICAL COMMITMENT ROUTING

Building on structured RRG generation (Nishino et al., 2022; Jin et al., 2024; Hou et al., 2023), the shared decoder uses the calibrated prefix $\widetilde{\mathbf{E}}_P$ to generate a polarity-structured clinical commitment before the report, providing structured clinical context for subsequent generation. Supervision is provided by $\mathbf{q}^\star = (\mathbf{q}^{\mathrm{pos}}, \mathbf{q}^{\mathrm{neg}}, \mathbf{q}^{\mathrm{unc}})$, which organizes positive, negative, and uncertain findings, with its construction detailed in Appendix A.

To combine explicit clinical commitment supervision with direct-report supervision, we stochastically route complete supervised trajectories through the same decoder. Let $\xi \sim \mathrm{Bernoulli}(1-\rho)$, where ρ is the probability of selecting the direct-report route, and let $\boldsymbol{\iota}_{\mathrm{C}}$ and $\boldsymbol{\iota}_{\mathrm{D}}$ denote the commitment-first and direct-report instructions, respectively. The corresponding route-specific target sequence $\boldsymbol{\tau}_\xi^\star$ is

$$(\boldsymbol{\iota}_\xi, \boldsymbol{\tau}_\xi^\star) = \begin{cases} (\boldsymbol{\iota}_{\mathrm{C}}, [\langle \textsc{Anchor}\rangle, \mathbf{q}^\star, \langle /\textsc{Anchor}\rangle, \langle \textsc{Report}\rangle, \mathbf{Y}, \langle /\textsc{Report}\rangle])\,, & \xi = 1, \\ (\boldsymbol{\iota}_{\mathrm{D}}, \mathbf{Y}), & \xi = 0. \end{cases} \tag{8}$$

The route is sampled before instruction and target construction, so routing changes the complete supervised sequence rather than isolated tokens. Inference always uses the commitment-first route, with one decoder generating the commitment followed by the report in a single left-to-right pass.

3.5 OVERALL LEARNING OBJECTIVE

Training follows the Refine–Calibrate–Commit design in two optimization stages (phase details in Appendix A). Refine and Commit are first optimized with autoregressive trajectory supervision; the resulting parent is then frozen and only Calibrate is trained with output- and representation-level regularization.

For training example b, let ξ_b denote the route sampled as in Section 3.4, and let $\boldsymbol{\tau}_b^\star = \boldsymbol{\tau}_{\xi_b}^\star$ denote the corresponding routed target sequence. We then let $\Omega_{\mathrm{Q}}^{(b)}$ and $\Omega_{\mathrm{Y}}^{(b)}$ denote its supervised commitment- and report-side positions, respectively, with $\Omega_{\mathrm{Q}}^{(b)} = \varnothing$ for direct-report samples, and define $\mathcal{S}_b = \Omega_{\mathrm{Q}}^{(b)} \cup \Omega_{\mathrm{Y}}^{(b)}$. Given

$$\ell_{bt} = -\log p_\theta\left(\tau_{bt}^\star \mid \boldsymbol{\tau}_{b,<t}^\star, \mathbf{E}_{P,b}\right),$$

where p_θ denotes the complete model; observed-source and availability conditioning is omitted for brevity. To balance commitment and report supervision, we set $\omega_{bt} = \lambda_{\mathrm{Q}}$ on $\Omega_{\mathrm{Q}}^{(b)}$ and $\omega_{bt} = 1$ on $\Omega_{\mathrm{Y}}^{(b)}$. Let $\kappa_b = \lambda_{\mathrm{Q}}$ for commitment-first samples and $\kappa_b = 1$ for direct-report samples denote the ignored-BOS weight retained by the implementation normalizer. The trajectory objective is

$$\mathcal{L}_{\mathrm{traj}} = \frac{\sum_b \sum_{t \in \mathcal{S}_b} \omega_{bt} \ell_{bt}}{\sum_b \sum_{t \in \mathcal{S}_b} \omega_{bt} + \sum_b \kappa_b + \varepsilon} + \lambda_{\mathrm{TLW}} \mathcal{L}_{\mathrm{TLW}}, \tag{9}$$

where ε is a numerical stabilizer and $\mathcal{L}_{\mathrm{TLW}}$ is an independently normalized report-side auxiliary loss using the emphasized-token annotations from LLM-RG4 (Wang et al., 2025b). Exact span boundaries, masking, and token weighting are detailed in Appendix A.

During Calibrate training, the frozen parent provides a reference for limiting output deviation and prefix drift. Let p_0 denote the next-token distribution of the frozen parent model with prefix routing disabled, and p_θ that of the adapted model, both conditioned on the same observed case. Let $\Omega_{\mathrm{cons}}^{(b)}$ denote the valid prediction positions used for output consistency. We regularize output deviation and valid-prefix representation drift by

$$\mathcal{L}_{\mathrm{cons}} = \mathbb{E}_{b, t \in \Omega_{\mathrm{cons}}^{(b)}} D_{\mathrm{KL}}\big(p_0(\cdot \mid \boldsymbol{\tau}^\star_{b,<t}, \mathbf{E}_{P,b}) \| p_\theta(\cdot \mid \boldsymbol{\tau}^\star_{b,<t}, \mathbf{E}_{P,b})\big), \tag{10}$$

$$\mathcal{L}_{\mathrm{drift}} = \mathbb{E}_{b,t:m_{bt}=1} \frac{\left\| \widetilde{\mathbf{E}}_{bt,:} - \mathbf{E}_{bt,:} \right\|_2^2}{d}, \tag{11}$$

where $\mathbf{m}_b$ is the prefix mask defined in Section 3.3. The Calibrate stage minimizes

$$\mathcal{L}_{\mathrm{PACER}} = \mathcal{L}_{\mathrm{traj}} + \lambda_{\mathrm{cons}} \mathcal{L}_{\mathrm{cons}} + \lambda_{\mathrm{drift}} \mathcal{L}_{\mathrm{drift}}, \tag{12}$$

where λ_{cons} and λ_{drift} weight the output-consistency and prefix-drift regularizers, respectively.

4 Experiments

4.1 Experimental Setup

Task and data. We evaluate PACER under both structured incomplete-context and conventional RRG settings. The primary evaluation uses MIMIC-RG4 (Wang et al., 2025b), constructed from MIMIC-CXR (Johnson et al., 2019), under the four availability states defined in Section 3.1. Following the benchmark protocol, the four-context evaluation uses findings-and-impression reports, whereas conventional SN uses findings only. For this fixed-availability setting, we evaluate a dedicated Refine+Commit model, while Calibrate is evaluated in the multi-context setting where structured source availability varies.

Baselines and metrics. For the four-context comparison, we use the published LLM-RG4, CXR-Mate, and RadFM results from Wang et al. (2025b), and include local adaptations of SimMLM (Li et al., 2025) and RAGPT (Lang et al., 2025) under the same four-context MIMIC-RG4 evaluation protocol; provenance and adaptation details are provided in Appendix C. Clinical efficacy (CE) is measured by micro-averaged precision, recall, and F1 from CheXbert-extracted labels (Smit et al., 2020), and language quality by BLEU-1–4 (B@1–4), ROUGE-L (R-L), and METEOR (MTR) (Papineni et al., 2002; Lin, 2004; Banerjee & Lavie, 2005). Ablations are summarized by the arithmetic mean over the four MIMIC-RG4 contexts, computed before rounding.

Implementation. Following LLM-RG4, we use Vicuna-7B v1.5 with LoRA (Hu et al., 2022) as the language decoder and frozen RAD-DINO as the visual backbone. The depth router uses encoder layers $\{4, 8, 12\}$, and the prefix residual branch uses rank $r = 16$. Clinical commitment targets are constructed offline from reference reports using a hybrid LLM-assisted and rule-based extraction pipeline. Training follows the staged optimization in Section 3.5; stage-specific trajectory sampling, loss weights, and other implementation details are provided in Appendix A.

4.2 Main Results under Structured Incomplete Contexts

Table 1 reports the four-context comparison. PACER achieves the highest CE F1 in all four settings, improving over LLM-RG4 by 0.023–0.028, with gains in both precision and recall. These clinical

Table 1: Four-context results on MIMIC-RG4. Published baselines follow Wang et al. (2025b); $*$ denotes CXRMate retrained therein and † our SimMLM/RAGPT adaptations. Bold/underline indicate the best/second-best non-tied values per context; ties for best are bold.

Setting	Model	CE Metrics			NLG Metrics					
		P	R	F1	B@1	B@2	B@3	B@4	R-L	MTR
SN	CXRMate*	0.572	0.560	0.566	0.421	0.271	0.179	0.122	0.311	0.174
	RadFM	0.413	0.303	0.350	0.188	0.090	0.048	0.028	0.190	0.094
	LLM-RG4	0.588	0.632	0.609	**0.479**	**0.343**	**0.255**	**0.196**	0.384	**0.209**
	SimMLM†	0.568	0.591	0.579	0.425	0.277	0.186	0.128	0.313	0.168
	RAGPT†	0.565	0.564	0.564	0.423	0.278	0.188	0.129	0.314	0.168
	PACER (Ours)	**0.620**	**0.645**	**0.632**	0.468	0.337	0.251	0.192	**0.390**	0.208
SW	CXRMate*	0.573	0.549	0.561	0.361	0.220	0.139	0.093	0.284	0.153
	RadFM	0.508	0.365	0.425	0.211	0.103	0.056	0.033	0.183	0.105
	LLM-RG4	0.599	0.622	0.610	0.455	0.321	0.239	**0.186**	0.382	0.199
	SimMLM†	0.583	0.556	0.569	0.379	0.236	0.152	0.105	0.306	0.152
	RAGPT†	0.568	0.556	0.562	0.387	0.244	0.158	0.106	0.307	0.156
	PACER (Ours)	**0.623**	**0.654**	**0.638**	**0.456**	**0.323**	**0.240**	**0.186**	**0.386**	**0.202**
MN	CXRMate*	0.544	0.522	0.533	0.437	0.289	0.199	0.141	0.332	0.179
	RadFM	0.323	0.187	0.237	0.246	0.113	0.060	0.034	0.194	0.104
	LLM-RG4	0.541	0.578	0.559	**0.491**	**0.359**	**0.274**	**0.216**	0.405	0.214
	SimMLM†	0.512	0.558	0.534	0.429	0.283	0.195	0.138	0.325	0.172
	RAGPT†	0.544	0.531	0.537	0.431	0.288	0.200	0.142	0.327	0.174
	PACER (Ours)	**0.586**	**0.580**	**0.583**	0.483	0.357	0.273	**0.216**	**0.413**	**0.215**
MW	CXRMate*	0.548	0.499	0.523	0.379	0.241	0.158	0.110	0.305	0.159
	RadFM	0.456	0.297	0.360	0.191	0.095	0.054	0.034	0.178	0.095
	LLM-RG4	0.560	0.565	0.563	**0.461**	**0.331**	**0.250**	0.197	0.401	0.204
	SimMLM†	0.529	0.538	0.533	0.387	0.245	0.161	0.113	0.317	0.157
	RAGPT†	0.529	0.523	0.526	0.397	0.255	0.169	0.118	0.324	0.163
	PACER (Ours)	**0.582**	**0.591**	**0.586**	0.458	0.330	0.249	**0.198**	**0.406**	**0.205**

gains largely preserve language-generation quality: BLEU changes marginally, while ROUGE-L improves in all four settings and METEOR in most. Overall, PACER strengthens clinical-label agreement without materially degrading reference-based language quality.

4.3 Conventional SN Evaluation

Table 2 further evaluates PACER under the conventional findings-only SN setting. On the common MIMIC-RG4 test set, PACER achieves the highest CE F1 (0.620) and recall (0.654), exceeding LLM-RG4 by 0.032 and 0.061, respectively, while PromptMRG retains the highest precision. For language generation, PACER obtains the best BLEU-1 and BLEU-4 on clean references and ties LLM-RG4 for the best clean-reference ROUGE-L. Against the corresponding original MIMIC-CXR reports, it achieves the best BLEU-4 and the second-best ROUGE-L while remaining competitive on BLEU-1. Literature-reported results under different original protocols are included in the upper block of Table 2 for broader context rather than direct ranking.

4.4 Component and Mechanism Ablations

Component contributions. Table 3 reveals complementary effects across the three components. Refine primarily improves recall and BLEU overlap, whereas Commit yields a larger gain in precision. Combining them improves both precision and recall over the Base, raising CE F1 from 0.581 to 0.600. Training Calibrate on the frozen Refine+Commit parent further increases F1 to 0.610 while improving the B@1 and B@4 scores of the combined model, supporting complementary roles across the three components.

Table 2: Conventional findings-only SN results. The upper block provides literature context under original protocols, while the lower block evaluates methods on a common findings-only SN test set, with clean and original references reported separately. † denotes local downstream training initialized from released pretrained weights, and ‡ denotes a released checkpoint evaluated by us. Bold/underline indicate best/second-best values within the common-test block.

Model	CE Metrics			Clean NLG			Original NLG		
	P	R	F1	B@1	B@4	R-L	B@1	B@4	R-L
Literature-reported results under original protocols									
KiUT (Huang et al., 2023)	0.371	0.318	0.321	–	–	–	0.393	0.113	0.285
RGRG (Tanida et al., 2023)	0.461	0.475	0.447	–	–	–	0.373	0.126	0.264
EKAGen (Bu et al., 2024)	0.517	0.483	0.499	–	–	–	0.419	0.119	0.287
MAIRA-1 (7B) (Hyland et al., 2023)	–	–	0.553	–	–	–	0.392	0.142	0.289
Med-PaLM M (562B) (Tu et al., 2024)	–	–	0.516	–	–	–	0.317	0.115	0.275
R2-LLM (14.2B) (Liu et al., 2024)	0.465	0.482	0.473	–	–	–	0.402	0.128	0.291
InVERGe (7B) (Deria et al., 2024)	–	–	–	–	–	–	0.425	0.100	0.309
REVTAF (Zhou et al., 2025)	0.628	0.613	0.592	–	–	–	0.465	0.182	0.336
ESC-RL (Zhou et al., 2026)	0.632	0.625	0.608	–	–	–	0.487	0.199	0.352
S2D-Align (Gao et al., 2026)	0.613	0.606	0.608	–	–	–	0.422	0.149	0.332
Evaluation on the common MIMIC-RG4 test set									
R2Gen (Chen et al., 2020)	0.456	0.306	0.366	0.363	0.090	0.269	0.356	0.097	0.267
R2GenCMN (Chen et al., 2021)	0.486	0.400	0.439	0.385	0.102	0.278	0.349	0.094	0.270
CvT2DistilGPT2 (Nicolson et al., 2023)	0.498	0.414	0.452	0.374	0.103	0.272	0.390	0.123	0.282
PromptMRG (Jin et al., 2024)	**0.618**	0.491	0.548	0.326	0.080	0.261	0.381	0.096	0.258
R2GenGPT (7B) (Wang et al., 2023)	0.506	0.414	0.456	0.401	0.118	0.277	**0.396**	0.113	0.273
CheXagent (7B) (Chen et al., 2024)	0.506	0.306	0.381	0.265	0.058	0.239	0.189	0.040	0.208
LLM-RG4 (Wang et al., 2025b)	0.583	0.593	0.588	0.498	0.203	**0.387**	0.377	0.144	**0.318**
MambaXray-VL-Large† (Wang et al., 2025a)	0.555	0.526	0.540	0.449	0.137	0.328	0.376	0.103	0.264
CheXOne‡ (Zhang et al., 2026)	0.613	0.574	0.593	0.428	0.124	0.311	0.199	0.037	0.177
PACER (Ours)	0.590	**0.654**	**0.620**	**0.511**	**0.208**	**0.387**	0.389	**0.146**	0.315

Table 3: Component ablations averaged over the four MIMIC-RG4 contexts. Base denotes the local ablation baseline. Bold/underline indicate the best/second-best values.

Model	P	R	F1	B@1	B@4	R-L
Base	0.576	0.587	0.581	0.461	0.198	0.395
Refine only	0.575	0.598	0.586	**0.470**	**0.200**	0.396
Commit only	0.599	0.587	0.593	0.455	0.194	**0.399**
Refine + Commit	0.598	0.603	0.600	0.452	0.192	0.396
PACER (+ Calibrate)	**0.603**	**0.617**	**0.610**	0.466	0.198	**0.399**

Mechanism analysis. Table 4 further clarifies the role of each design. Global depth weighting improves recall but reduces precision and language overlap, whereas patchwise routing preserves the recall gain while recovering precision, yielding the best CE F1 among the depth variants. For commitment routing, always using the commitment-first trajectory improves precision at the cost of BLEU, while stochastic routing further improves F1 and partially restores language overlap. For prefix routing, Calibrate improves over the no-routing control. Under matched gate architectures and parameter counts, explicitly incorporating availability and observed-evidence cues provides additional gains over prefix-only calibration, with their joint use performing best. This supports conditioning adaptation that accounts for both source availability and observed evidence.

4.5 Qualitative Analysis

Figure 2 illustrates an example of longitudinal evidence utilization. Without the previous report, both PACER and LLM-RG4 largely agree on the current PICC position and the absence of complications. The key difference emerges when the previous report becomes available: PACER introduces the reference-supported new PICC insertion, whereas the fixed LLM-RG4 checkpoint describes the line as having been pulled back. Meanwhile, the current-image findings remain largely consistent across

Table 4: Mechanism ablations averaged over the four MIMIC-RG4 contexts. For prefix routing, h, z, and e denote the prefix summary, observed-evidence summary, and availability embedding. Bold/underline indicate best/second-best values within each group.

Mechanism	Variant	P	R	F1	B@1	B@4	R-L
Refine	No Refine (endpoint only)	**0.599**	0.587	0.593	**0.455**	**0.194**	**0.399**
	Global depth weighting	0.588	**0.603**	0.595	0.444	0.185	0.390
	Patchwise depth routing	0.598	**0.603**	**0.600**	0.452	0.192	0.396
Commit	No Commit (direct report)	0.575	0.598	0.586	**0.470**	**0.200**	**0.396**
	Always commitment-first	0.594	0.596	0.595	0.448	0.190	0.394
	Stochastic trajectory routing	**0.598**	**0.603**	**0.600**	0.452	0.192	**0.396**
Calibrate	No Calibrate (no prefix routing)	0.598	0.603	0.600	0.452	0.192	0.396
	$g(h)$	0.601	0.611	0.606	0.464	0.196	0.398
	$g(h, e)$	0.602	0.612	0.607	0.465	0.197	0.398
	$g(h, z)$	0.601	0.613	0.607	0.463	0.196	0.398
	$g(h, z, e)$ (PACER)	**0.603**	**0.617**	**0.610**	**0.466**	**0.198**	**0.399**

Controls. Refine variants share the Commit training protocol, and Commit variants share the Refine architecture and the checkpoint obtained after the SN Refine warm-up, with Calibrate disabled in both; Calibrate variants share the same frozen Refine+Commit parent.



Current frontal CXR

Previous report (SW only)

Dual-chamber pacemaker leads are in appropriate position in the right atrium and right ventricle. No pneumothorax, mediastinal widening, or hemothorax.

Context	Reference	LLM-RG4	PACER
SN: no previous report	The right PICC line projects over the mid SVC. The course is unremarkable. There is no evidence of complication, notably no pneumothorax.	The PICC tip projects over the mid-to-low SVC. There is no evidence of complications, notably no pneumothorax. The pacemaker leads are in position.	The patient has a right-sided PICC line. The course is unremarkable and the tip projects over the mid SVC. No evidence of complications, notably no pneumothorax.
SW: with previous report	As compared with the previous radiograph, **a new right PICC line has been inserted**. The tip projects over the mid SVC. The course is unremarkable and there is no complication.	As compared with the previous radiograph, the PICC line has been **pulled back**. The tip now projects over the mid SVC. There is no evidence of complications, notably no pneumothorax.	As compared with the previous radiograph, **the patient has received a right-sided PICC line**. The course is unremarkable and the tip projects over the mid SVC. There is no evidence of complications, notably no pneumothorax.

Color coding: reference-supported shared content; **reference longitudinal event**; **PACER-captured longitudinal event**; **reference-unverified description**.

Longitudinal change: the reference indicates a new PICC insertion; PACER captures the same event, whereas LLM-RG4 describes the line as having been pulled back.

Figure 2: Qualitative comparison of longitudinal evidence utilization, showing how the generated report changes when the previous report becomes available.

the two contexts. This example suggests that PACER can adapt the longitudinal interpretation of the report to newly available previous-report evidence while preserving the description of current findings. Further discussion of the findings and study limitations is provided in Appendix D.

5 CONCLUSION

In this work, we presented PACER, a Progressive Availability-Conditioned Evidence Routing framework for radiology report generation under structured incomplete clinical context. Its Refine–Calibrate–Commit pipeline coordinates available evidence across visual representation, language-model conditioning, and clinical report generation within a unified model. PACER achieves state-of-the-art CE F1 across all four MIMIC-RG4 settings while maintaining competitive language-generation performance, with strong results in the conventional MIMIC-CXR setting. These findings highlight effective evidence utilization, beyond accommodating variable inputs alone, as an important design dimension for flexible-context RRG.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, research ideation and methodological or experimental feedback, manuscript organization, language editing, LaTeX preparation, and consistency checking. All AI-assisted suggestions, reported experimental results, scientific claims, citations, and implementation descriptions were reviewed and verified by the authors. Separately, a hybrid LLM-assisted and rule-based extraction pipeline was used to construct offline clinical commitment supervision, as described in Appendix A; this external pipeline is not required during inference. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the proposed routing mechanisms, information-flow constraints, staged optimization, and learning objectives. Appendix A provides implementation details including module initialization, prefix masking, commitment supervision, trajectory routing, report extraction, and loss configuration. Tables 1 and 2 report the complete main comparisons, while Appendices B and C document evaluation protocols, result provenance, local baseline adaptations, and ablation settings. Together, these sections specify the methodological and evaluation procedures used for the reported experiments.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14194–14204, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.459.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation, 2024.

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ankan Deria, Komal Kumar, Snehashis Chakraborty, Dwarikanath Mahapatra, and Sudipta Roy. InVERGe: Intelligent visual encoder for bridging modalities in report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2028–2038, 2024. doi: 10.1109/CVPRW63382.2024.00208.

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. TIM: Temporal decoupling with iterative mutual-refinement model for longitudinal radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6951–6961, 2026.

Jiechao Gao, Chang Liu, and Yuangang Li. S2D-Align: Shallow-to-deep auxiliary learning for anatomically-grounded radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30780–30788, 2026. doi: 10.1609/aaai.v40i36.40335.

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)*, pp. 8108–8122. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.451.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. KiUT: Knowledge-injected U-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19809–19818, 2023.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. MAIRA-1: A specialised large multimodal model for radiology report generation, 2023.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 2607–2615, 2024. doi: 10.1609/aaai.v38i3.28038.

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 de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

Ashwin Kumar, Robbie Holland, Corey Barrett, Jangwon Kim, Maya Varma, Zhihong Chen, Yunhe Gao, Greg Zaharchuk, Tara Taghavi, Krishnaram Kenthapadi, et al. CheXmix: Unified generative pretraining for vision language models in medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9466–9476, 2026. arXiv preprint arXiv:2604.22989.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18035–18043, 2025. doi: 10.1609/aaai.v39i17.33984.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal LLMs: Methods, analysis, and best practices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4156–4166, 2025.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643, 2024. doi: 10.1609/aaai.v38i17.29826.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1657–1665, 2023. doi: 10.1609/aaai.v37i2.25253.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10359, 2025a. doi: 10.1109/CVPR52734.2025.00968.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. PriorRG: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7206–7214, 2026a. doi: 10.1609/aaai.v40i9.37657.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5595–5603, 2025b. doi: 10.1609/aaai.v39i6.32596.

Tengfei Liu, Yijian Fan, Boyue Wang, Yongli Hu, Mingjie Li, Jinghua Li, Junbin Gao, Xiaojun Chang, Zhihui Li, and Baocai Yin. BiOTPrompt: Bidirectional optimal transport guided prompting for disease evolution-aware radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13755–13765, 2026b.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. doi: 10.1016/j.artmed.2023.102633.

Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7123–7138, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.480.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040/`.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119–130, 2025. doi: 10.1038/s42256-024-00965-w.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. MissRAG: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3215–3224, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Nagur Shareef Shaik, Teja Krishna Cherukuri, Adnan Masood, and Dong Hye Ye. DiA-gnostic VLVAE: Disentangled alignment-constrained vision language variational autoencoder for robust radiology reporting with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):8814–8823, 2026. doi: 10.1609/aaai.v40i11.37835.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew 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)*, pp. 1500–1519, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15483–15491, 2024. doi: 10.1609/aaai.v38i14.29474.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. CXPMRG-Bench: Pre-training and benchmarking for x-ray medical report generation on CheXpert Plus dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5123–5133, 2025a.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. doi: 10.1016/j.metrad.2023.100033.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8250–8258, 2025b. doi: 10.1609/aaai.v39i8.32890.

Sixing Yan, Ziao Wang, Kejing Yin, William Kwok-wai Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest x-ray report generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16416–16424, 2024. doi: 10.1609/aaai.v38i15.29578.

Shiying Yu, Jielei Wang, and Guoming Lu. AnchorDiff: Topology-aware masked diffusion with confidence-based rewriting for radiology report generation, 2026.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, et al. A reasoning-enabled vision-language foundation model for chest x-ray interpretation. arXiv preprint arXiv:2604.00493, 2026.

Qin Zhou, Guoyan Liang, Xindi Li, Jingyuan Chen, Zhe Wang, Chang Yao, and Sai Wu. Learnable retrieval enhanced visual-text alignment and fusion for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22529–22538, 2025.

Qin Zhou, Guoyan Liang, Qianyi Yang, Jingyuan Chen, Sai Wu, Chang Yao, and Zhe Wang. Enhancing reinforcement learning for radiology report generation with evidence-aware rewards and self-correcting preference learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37044–37056, 2026. doi: 10.18653/v1/2026.acl-long.1718.

A Implementation Details

Interface and visual routing. Following LLM-RG4, source encodings are query-compressed and arranged in a fixed frontal–lateral–previous-report layout, with unavailable lateral and previous-report branches represented by zero-valued slots before fusion (Wang et al., 2025b). Images are processed with the RAD-DINO image processor at 518×518 resolution without data augmentation in the reported runs. The visual router uses frozen RAD-DINO features from depths $\mathcal{K} = \{4, 8, 12\}$, removes the CLS token, and preserves patch correspondence across depths. The same RAD-DINO encoder and depth router are applied to frontal and lateral views. Refine produces 1,369 non-CLS patch tokens of dimension 768 per image. The frontal branch is compressed by 128 learned visual queries through cross-attention; the resulting 128 frontal query features are reused as queries for lateral-image and previous-report compression. Previous reports are first encoded by frozen CXR-BERT (Boecking et al., 2022) with a maximum input length of 100 tokens. Each compressed source is subsequently projected from 768 to the 4096-dimensional language-model space, yielding $\mathbf{Z}_s \in \mathbb{R}^{128\times4096}$ for $s \in \{f, l, p\}$.

In the reported implementation, $d_r = d_v = 768$, so each alignment projection $\mathbf{P}_\ell \in \mathbb{R}^{768\times768}$ is initialized as a square identity matrix with $\mathbf{b}_\ell = \mathbf{0}$. The affine scale and bias of each layer-specific normalization module are initialized to one and zero, respectively, while the shared routing scorer and depth biases are initialized to zero. The routed residual projection is also zero-initialized, so the initial correction is exactly zero. The residual scale is initialized to $\alpha_D = 0.1$ with upper bound $\bar{\alpha}_D = 0.5$. The routing candidates are taken from RAD-DINO hidden states at depths 4, 8, and 12, whereas the residual endpoint is the encoder's final hidden state after final normalization. The depth-12 routing candidate and the normalized endpoint preserve the same patch indexing but need not be numerically identical.

Prefix routing. The Calibrate module is inserted once at the pre-Transformer input-embedding level and is used at the same site during teacher forcing and autoregressive generation. The low-rank residual branch uses rank $r = 16$ and dropout 0.05. The availability embedding and gate hidden dimension are both 4096 in the reported model. With $d = d_a = d_g = 4096$, the gate projection has $\mathbf{W}_g \in \mathbb{R}^{4096\times12288}$. The gate output vector $\mathbf{w}_g$ is initialized to zero with $b_{\text{out}} = -2.2$, while $\mathbf{W}_\uparrow$ is zero-initialized; the module therefore starts from an identity mapping with an initial sample-wise gate of approximately 0.1.

The valid-prefix mask follows the prompt attention mask and excludes padding and all teacher-forcing target-side positions. The routing gate is computed from the mean valid-prefix representation, the mean of the actually observed source representations, and the learned availability embedding. Missing lateral or previous-report sources are omitted from this observed-evidence average. A single scalar gate is produced per sample and broadcast across valid prefix positions, whereas the low-rank residual branch produces token-specific corrections.

Commitment supervision. Clinical commitments are supervised by a hybrid offline pipeline. We construct a cache of 10,000 training records, with 2,500 records from each of SN, SW, MN, and MW. The selected records are drawn only from the training split using a fixed stratified sampling procedure that emphasizes clinically relevant patterns such as uncertainty, negation, devices, common findings, and temporal comparisons. The preserved cache contains no validation- or test-split records.

For cached records, an external model with API identifier `gpt-4.1-mini` is queried with temperature 0 and a maximum of 600 output tokens. The request contains the current reference report together with a rule-based candidate anchor and sampling metadata. The output schema contains positive, negative, and uncertain lists over the fixed 13-finding vocabulary reproduced in Appendix A.2, together with a short rationale used only for validation. Outputs are normalized, restricted to the predefined vocabulary, and deduplicated across polarity fields before entering the cache.

For a training sample not covered by the cache, a deterministic rule-based extractor constructs the commitment over the same 13 findings. Reports are lower-cased, whitespace-normalized, and segmented at sentence or semicolon boundaries. For each finding, uncertainty cues such as `possible`, `probable`, `may represent`, `could represent`, `questionable`, and `suspect` are checked before local negation cues. Negation is detected within local windows around the finding using expres-

sions such as no, not, without, absent, negative for, and no evidence of, with an exception for phrases such as no change. The formal fallback does not apply a separate temporal-state resolver: expressions such as stable, resolved, or no new are handled only through these surface matching rules. Negative matches remain candidates while subsequent mentions are inspected, whereas the first positive or uncertain match terminates the search for that finding. Support devices are matched through explicit tube, catheter, PICC, central-line, device, and pacemaker patterns. Each polarity field is deduplicated and serialized with at most eight findings; completely empty anchors use none for all three fields. The external model is never called online during PACER training or inference.

The commitment-first target has the form

```
<ANCHOR> positive: ...; negative: ...; uncertain: ... </ANCHOR>
<REPORT>
report text
</REPORT></s>
```

whereas the direct-report target contains only the cleaned report followed by </s>. Both routes use the same task-specific base instruction inherited from LLM-RG4. The commitment-first instruction ι_{C} appends the exact suffix First output <ANCHOR> positive, negative and uncertain findings, then output the final report in <REPORT>., whereas the direct-report instruction ι_{D} appends no additional suffix. The same commitment-first instruction is used during inference. At inference, the Vicuna decoder generates the commitment and report in a single left-to-right call; neither the external model nor a cached ground-truth commitment is used as inference guidance.

A.1 Optimization Protocol

The two optimization stages described in the main text map to three physical training phases. Paper Stage 1 consists of an SN Refine warm-up followed by four-context Refine+Commit parent construction. Paper Stage 2 then freezes that parent and optimizes only Calibrate. Table 5 summarizes the run-level settings that are directly supported by the saved checkpoints and logs.

Table 5: Optimization protocol used for PACER. "Batch" denotes per-device batch size and "Accum." gradient accumulation.

Phase	Context	LR	Batch	Accum.	Updates
Refine warm-up	SN	3×10^{-4}	24	2	14,384
Refine+Commit parent	SN/SW/MN/MW	3×10^{-4}	16	2	43,152
Calibrate	SN/SW/MN/MW	1×10^{-4}	16	2	21,576

All three training phases were executed as single-GPU runs on NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPUs. Accordingly, the effective global batch sizes, computed as per-device batch size times gradient accumulation, are 48 for Refine warm-up and 32 for both Refine+Commit and Calibrate. During Refine warm-up, the depth router and frontal visual interface are optimized while RAD-DINO, CXR-BERT, and the base Vicuna parameters remain frozen. Refine+Commit parent construction continues to train Refine and the source interfaces, introduces the commitment trajectory, and applies LoRA to Vicuna with rank 32, scaling 64, and dropout 0.1 on the effective q_proj/v_proj target modules. During Calibrate training, the full parent—including Refine, source interfaces, fusion modules, and Vicuna LoRA—is frozen, and only the Calibrate parameters, including the availability embeddings, gate network, and low-rank residual branch, are updated.

Trajectory-routing and commitment-loss weights are stage-specific. During Refine+Commit parent construction, the direct-report probability increases linearly from 0.1 to 0.5 over the first 16,000 optimizer updates and remains at 0.5 thereafter, with commitment-token weight $\lambda_{\mathrm{Q}} = 0.35$. During Calibrate training, the direct-report probability is fixed at $\rho = 0.5$ and $\lambda_{\mathrm{Q}} = 0.20$. The report-side TLW coefficient is $\lambda_{\mathrm{TLW}} = 0.75$ in both stages that use the report-generation objective. Under trajectory weighting, the commitment-side span extends through the opening <REPORT> delimiter and its following newline, while the report-side span begins with the report content and includes the

closing delimiter and EOS. The tokenizer-added BOS label is ignored by cross-entropy but retains its route-dependent weight in the normalization denominator.

Inherited token-level weighting. We reuse the report-side emphasized-token mask released with LLM-RG4 (Wang et al., 2025b) and incorporate it through the independently normalized auxiliary term $\mathcal{L}_{\mathrm{TLW}}$ defined below. In the original construction, CheXbert first identifies positive or uncertain observations, Integrated Gradients provides token attribution scores, and Gaussian smoothing is applied within the report. If an attribution score in a sentence exceeds the threshold of 0.4, the tokens of that sentence receive the elevated coefficient 1.75 rather than the default coefficient 1. In the released annotations used here, these emphasized report positions are stored as binary `new_scores`. Let $s_{bt} \in \{0, 1\}$ denote this stored mask after alignment to the report span. The auxiliary term used by PACER is

$$\mathcal{L}_{\mathrm{TLW}} = \frac{\sum_{b,t \in \Omega_{\mathrm{Y}}^{(b)}} s_{bt}\, \ell_{bt}}{\sum_{b,t \in \Omega_{\mathrm{Y}}^{(b)}} s_{bt} + \varepsilon}.$$

PACER preserves the inherited report-side scores but shifts them to the report positions of the routed trajectory; commitment positions have $s_{bt} = 0$ and are supervised separately through the trajectory loss. We use $\lambda_{\mathrm{TLW}} = 0.75$ and $\varepsilon = 10^{-3}$.

Calibrate regularization. During the final Calibrate stage, the residual-off frozen parent provides the teacher next-token distribution and the adapted model provides the student distribution on the same observed case and routed target. The consistency term uses $\mathrm{KL}(p_0 \| p_\theta)$ with temperature 1 and coefficient $\lambda_{\mathrm{cons}} = 0.05$. The consistency set $\Omega_{\mathrm{cons}}^{(b)}$ excludes the complete `<ANCHOR>...</ANCHOR>` prediction span and includes the subsequent report structural tokens, report content, and EOS when present. The prefix-drift term is the mean squared magnitude of the actual gated residual over valid prefix positions, with coefficient $\lambda_{\mathrm{drift}} = 0.01$. We use $\varepsilon = 10^{-3}$ in the weighted trajectory-loss normalization.

Trajectory routing and report extraction. Route selection precedes both instruction and target construction, so a direct-report sample contains neither commitment nor report delimiters, whereas a commitment-first sample contains the complete commitment–report trajectory. At evaluation, configurations with Commit use commitment-first decoding, whereas configurations without Commit use direct-report decoding. For the four-context evaluation, PACER uses deterministic beam search with three beams, `do_sample=False`, 80–260 generated tokens, repetition penalty 2.0, and length penalty 2.0. The conventional findings-only SN evaluation uses five beams, `do_sample=False`, 50–200 generated tokens, repetition penalty 2.0, and length penalty 2.15. Because commitment and report are produced by a single autoregressive generation call, the repetition penalty is applied over the complete generated history rather than being reset at the beginning of the report.

The decoded report is extracted from the text following the last `<REPORT>` marker and ends at the corresponding `</REPORT>` marker when present. If a valid opening marker is absent or the extracted report is empty, the raw decoded generation is retained as a fallback. We audited all reported PACER outputs under both the four-context MIMIC-RG4 evaluation and the conventional SN evaluation. Every output contained a valid opening marker and yielded a nonempty extracted report, so the raw-generation fallback was never invoked for the reported CE or NLG results.

A.2 Preserved Commitment-Labeling Prompt

For transparency, we reproduce below the preserved system prompt associated with the offline commitment-labeling pipeline. The instruction content and ordering are preserved; the internal implementation header is normalized for presentation, and line wrapping is adjusted only for typesetting.

```
# Clinical Commitment Labeling Prompt

You are labeling chest X-ray report facts for a radiology report generation
experiment. Your task is to produce a compact clinical anchor from the
provided report. The anchor is used for training only. Do not add facts that
are not supported by the report.
```

```
Allowed finding vocabulary:

- atelectasis
- cardiomegaly
- consolidation
- edema
- enlarged mediastinum
- fracture
- lung lesion
- lung opacity
- pleural effusion
- pleural abnormality
- pneumonia
- pneumothorax
- support devices

Output exactly one JSON object with this schema:

{
  "sample_uid": "<copy from input>",
  "positive": ["finding_name"],
  "negative": ["finding_name"],
  "uncertain": ["finding_name"],
  "rationale": {
    "finding_name": "short evidence phrase copied or paraphrased from report"
  }
}

Rules:

1. Use only the allowed finding names. Do not invent new labels.
2. A finding must appear in at most one of ‘positive‘, ‘negative‘, or
   ‘uncertain‘.
3. Put a finding in ‘positive‘ only when the report states it is present.
4. Put a finding in ‘negative‘ only when the report explicitly denies that
   specific finding or a very direct synonym of that finding.
   - Good: "no pleural effusion" -> pleural effusion negative.
   - Good: "no focal consolidation" -> consolidation negative.
   - Good: "heart size is normal" -> cardiomegaly negative.
   - Bad: "lungs are clear" -> do not automatically list every lung finding
     as negative.
   - Bad: a finding is not mentioned -> do not label it negative.
   - Bad: "no support devices mentioned" -> do not label support devices
     negative.
   - Bad: "normal chest" -> do not enumerate all findings as negative.
   - Bad: "mediastinal and hilar contours are normal" -> do not list
     unrelated lung findings as negative.
5. Put a finding in ‘uncertain‘ for possible/probable/suspected/
   difficult-to-exclude findings, including wording such as:
   - possible
   - possibly
   - may represent
   - could represent / could reflect / could be seen with
   - cannot be excluded / difficult to exclude / not excluded
   - suspicious for / concerning for / worrisome for
6. ‘no new‘, ‘unchanged‘, ‘stable‘, and ‘no interval change‘ do not mean
   negative. If the finding is still described as present, label it positive.
   If the report only says no new finding but does not state a current finding
   is present, do not label that finding.
7. ‘resolved‘, ‘cleared‘, ‘resolution of‘, and ‘no longer seen‘ usually mean
   the finding is currently absent. Put it in ‘negative‘ only if the wording
   clearly says it has resolved or is no longer present.
8. Support devices include tubes, catheters, PICC lines, central lines, ports,
   pacemakers, ICDs, leads, drains, chest tubes, and feeding/enteric/NG/ET
```

```
   tubes.
9. Do not mark ‘support devices‘ positive just because the report says support
   devices are unchanged if no device is named.
10. Do not mark ‘fracture‘ positive when the report says no fracture or no
    displaced fracture.
11. Hydropneumothorax should imply ‘pneumothorax‘ positive and usually
    ‘pleural effusion‘ positive unless the report clearly separates them.
12. Pleural thickening/scarring/plaques should be ‘pleural abnormality‘, not
    ‘pleural effusion‘, unless fluid/effusion is also stated.
13. Use ‘lung lesion‘ for nodules, masses, lesions, or metastatic pulmonary
    nodules.
14. Use ‘lung opacity‘ for opacities, infiltrates, airspace disease, or
    opacification. If the report says opacity may represent pneumonia, put
    ‘lung opacity‘ positive and ‘pneumonia‘ uncertain.
15. Keep each list concise and clinically important. Empty lists are allowed.
16. The negative list should usually be short. Do not enumerate all absent or
    unmentioned findings. For this training anchor, unknown/unmentioned means
    omitted, not negative.
    Prefer negative labels only for directly denied high-value findings such
    as pneumothorax, pleural effusion, consolidation, edema, pneumonia,
    fracture, cardiomegaly, or enlarged mediastinum.
    Do not put ‘lung opacity‘, ‘lung lesion‘, ‘atelectasis‘,
    ‘support devices‘, or ‘pleural abnormality‘ in negative unless those exact
    concepts are clearly denied in the report.
    If the negative list would exceed 6 findings, keep only the most explicit
    direct negations and omit broad inferred negatives.
17. If a report says vascular congestion, elevated pulmonary venous pressure,
    pulmonary vascular engorgement, or fluid overload, label ‘edema‘ positive
    or uncertain depending on certainty. Do not label ‘edema‘ negative.
    - If the phrase is hedged, such as "could reflect elevated pulmonary
      venous pressure", label ‘edema‘ uncertain.
    - If the report explicitly says "pulmonary edema" or "vascular
      congestion", label ‘edema‘ positive.
18. If a report says "no focal consolidation concerning for pneumonia",
    label ‘consolidation‘ negative and ‘pneumonia‘ negative, not uncertain.
19. If the report says the mediastinum is widened or the cardiomediastinal
    silhouette is enlarged, label ‘enlarged mediastinum‘ positive.
20. If the report says heart size is normal, you may label ‘cardiomegaly‘
    negative. If the report says mediastinal contours are normal, you may
    label ‘enlarged mediastinum‘ negative. Do not use these normal statements
    to infer other negative findings.
21. Never copy a finding into multiple state lists. If evidence is mixed,
    choose the most clinically accurate single state in this priority:
    positive if definitely present; uncertain if possible/suspected;
    negative if explicitly denied and not also described as present.

Input format:

{
  "sample_uid": "...",
  "scenario": "sn|sw|mn|mw",
  "report": "...",
  "regex_anchor": {
    "positive": [],
    "negative": [],
    "uncertain": []
  },
  "risk_buckets": []
}

Return only the JSON object. No markdown and no explanation outside JSON.
```

B Evaluation Protocol and Result Provenance

B.1 Dataset Splits and Availability Settings

We use the existing MIMIC-RG4 train/validation/test splits rather than creating a new random split. Table 6 reports the scenario-record counts used by the four availability settings. Records can correspond to the same underlying examination across multiple availability states, so the sum across scenarios should not be interpreted as the number of independent examinations. An audit of the current annotations found no direct train–validation, train–test, or validation–test overlap at the study or subject level.

Table 6: Scenario-record counts in the four MIMIC-RG4 availability states.

Context	Train	Validation	Test
SN	172,608	1,391	2,357
SW	112,776	937	2,026
MN	91,341	701	1,004
MW	47,686	371	828

The structured availability axes are only the lateral radiograph and previous report. Indication/history may be included in the prompt when available but is not treated as an availability axis. The previous-report source in SW/MW is encoded with frozen CXR-BERT; a previous image is not used as an additional model input in the PACER configuration reported here.

B.2 Clinical and Language Metrics

Clinical efficacy is computed with the repository CheXbert evaluator over 14 report labels. Blank/negative outputs are treated as absent, while uncertain and positive outputs are treated as present for the reported CE scores. Precision, recall, and F1 are micro-averaged over cases and labels. Four-context ablation summaries are arithmetic means of the four per-context metrics, giving equal weight to SN, SW, MN, and MW rather than re-micro-averaging all scenario records jointly.

Language metrics use the repository evaluation implementation for BLEU-1–4, ROUGE-L, and METEOR 1.5. The reported four-context results are scored against the complete cleaned MIMIC-RG4 references used by the formal evaluation pipeline.

B.3 Four-Context Result Provenance

Table 1 reports the complete CE and NLG comparison under the four MIMIC-RG4 input contexts. The published CXRMate, RadFM, and LLM-RG4 values follow the corresponding four-context comparison in Wang et al. (2025b). SimMLM and RAGPT are our local adaptations described in Appendix C, and the PACER row is obtained from our formal four-context evaluation.

Per-context values are reported to three decimal places. Ablation averages are computed from unrounded per-context outputs and may therefore differ by 0.001 from values recomputed from the rounded entries in Table 1. The *Base* row in Table 3 is the controlled local ablation parent, whereas Table 1 reports the published LLM-RG4 result.

B.4 Conventional SN Evaluation

For this fixed-availability findings-only setting, PACER is trained separately with the Refine+Commit configuration, without Calibrate. Table 2 separates literature-reported results under their original protocols from results evaluated on the common findings-only SN test set. Clean NLG metrics use the cleaned findings references, whereas Original NLG metrics use the corresponding original MIMIC-CXR findings reports. Because the upper block preserves the original evaluation protocols, it is included for broader literature context rather than direct ranking; direct comparisons in Section 4.3 refer to the common-test block. CE values in the literature-only block retain the metric definitions and aggregation protocols reported by the original studies; our unified micro-averaged CheXbert evaluation applies to the common-test evaluation.

Table 7: Supplementary checkpoint-transfer and common-test SN evaluations. RGRG decoding settings and EKAGen missing-entry handling are protocol variants rather than independent models.

Method	P	R	F1	B@1	B@4	R-L
MLRG checkpoint	0.536	0.402	0.460	0.365	0.097	0.269
RGRG (official beam4)	0.515	0.523	0.519	0.366	0.105	0.264
RGRG (greedy)	0.489	0.603	0.540	0.362	0.109	0.271
RGRG (dedup.)	0.517	0.515	0.516	0.294	0.093	0.254
EKAGen (missing→empty)	0.528	0.442	0.481	0.346	0.089	0.275
EKAGen (nearest-neighbor)	0.529	0.444	0.483	0.346	0.089	0.276
CheXmix	0.180	0.077	0.108	0.045	0.014	0.142

The common-test baselines through LLM-RG4 follow Table 2 of Wang et al. (2025b). The literature rows through InVERGe use the same comparison set, with KiUT BLEU-1 and EKAGen BLEU-4 reported according to their original sources. REVTAF, ESC-RL, and S2D-Align are additional recent literature references (Zhou et al., 2025; 2026; Gao et al., 2026). The additional MambaXray-VL-Large and CheXOne results are produced locally under the protocols detailed in Appendix C.2.

C Local Adaptations and Ablation Evaluation

C.1 Incomplete-Multimodal Baseline Adaptations

The four-context SimMLM and RAGPT rows in Table 1 are local mechanism-transfer adaptations of Li et al. (2025) and Lang et al. (2025), respectively, to the MIMIC-RG4 structured incomplete-context setting. They are not values reported in the original papers and should not be interpreted as full official end-to-end reproductions.

For SimMLM, we retain a frozen common RRG parent and introduce source-specific residual experts for frontal, lateral, and previous-report representations. A learned availability-aware gate mixes the observed experts, and training uses paired more-evidence/fewer-evidence examples with a ranking term while retaining the parent report-generation objective. For RAGPT, we transfer the published missing-modality generation and contextual adaptive prompting modules to the frontal/lateral/previous-report source layout; the adapted model uses a train-only retrieval bank and trains only the added RAGPT modules while the common parent remains frozen.

Both adaptations use the same canonical four-context test populations and the same CE/NLG evaluator used for the PACER comparison. Their final adaptation budgets are fixed at 20,000 optimizer updates with seed 42.

Additional checkpoint-transfer and common-test evaluations. We additionally evaluated several released checkpoints and implementations as supplementary transfer references using the same frozen CE/NLG evaluator. The released MLRG checkpoint (Liu et al., 2025a) is evaluated without local parameter updates. Because it was trained for MLRG's original multi-view longitudinal setting rather than the four-context MIMIC-RG4 training protocol, we do not include it in the direct ranking of Table 1. In addition to its SN result reported in Table 7, its arithmetic mean over the four MIMIC-RG4 contexts is 0.535/0.394/0.454 for P/R/F1 and 0.297/0.080/0.250 for B@1/B@4/R-L.

We further evaluate RGRG (Tanida et al., 2023), EKAGen (Bu et al., 2024), and CheXmix (Kumar et al., 2026) on the common findings-only SN test population. For RGRG, the official beam-search configuration is treated as the primary transfer setting, while greedy decoding and complete-deduplication are included only as inference-sensitivity variants. For EKAGen, 11 test cases have unavailable knowledge-base entries; we therefore report both mapping these cases to empty outputs and a nearest-neighbor handling variant. CheXmix produces empty native Findings for 1,217 cases, contributing to its substantially lower transfer performance. The complete supplementary SN results are summarized in Table 7.

C.2 Additional Conventional-SN Baselines

MambaXray-VL-Large. The reported MambaXray-VL-Large result (Wang et al., 2025a) is obtained by locally training its Stage-3 downstream model from the released `MambaXrayCLIP-L.pth` Stage-2 visual weights with Llama-2-7B-chat. The language model is frozen while the visual encoder, projection, and normalization modules are optimized using AdamW with learning rate 10^{-4}, weight decay 0.01, batch size 6, no gradient accumulation, and seed 42. Training runs for five epochs (111,545 optimizer updates) with cosine annealing to a minimum learning rate of 10^{-6}. Checkpoint selection maximizes $0.8\,\mathrm{BLEU4} + 0.2\,\mathrm{CIDEr}$ on validation data. Evaluation uses deterministic three-beam decoding with 80–120 generated tokens and repetition/length penalties of 2.0/2.0.

CheXOne. The CheXOne result uses the released `StanfordAIMI/CheXOne` checkpoint (Zhang et al., 2026) with no local parameter updates. The inference prompt is `Write an example findings section for the CXR. Please reason step by step, and put your final answer within \boxed {{}}`. Generation is deterministic with one beam and at most 1,024 new tokens. The final findings are extracted from the last `\boxed {...}` block using brace-depth matching rather than a regular expression; an absent or unclosed box yields an empty extraction. Both baselines are scored with the same frozen SN evaluator used for the common-test block of Table 2.

C.3 Ablation Evaluation and Shared Configurations

Component and mechanism ablations are summarized by the arithmetic mean over the four MIMIC-RG4 contexts. Table 3 starts from a local ablation parent and evaluates Refine alone, Commit alone, their combination, and the final Calibrate stage. Table 4 varies one mechanism group at a time. Refine variants use the same Commit training protocol with Calibrate disabled; Commit variants use the same Refine architecture and the checkpoint obtained after the SN Refine warm-up with Calibrate disabled; and Calibrate variants are trained from the same frozen Refine+Commit parent.

Shared control configurations are intentionally repeated across mechanism groups for within-group comparison. Consequently, identical values for *Patchwise depth routing*, *Stochastic trajectory routing*, and *No Calibrate (no prefix routing)* correspond to the same evaluated Refine+Commit control configuration rather than independent configurations that happened to produce the same rounded scores. Global depth weighting learns a single set of softmax-normalized depth weights shared across all samples and spatial positions. For the Calibrate input ablations, all gate variants use the same architecture and parameter count; disabled cues are replaced by zero vectors while the full 12,288-dimensional gate input is retained.

D Discussion and Limitations

The results reinforce a key distinction in flexible-context RRG: supporting heterogeneous input combinations does not by itself ensure effective evidence utilization. The consistent gains across availability states suggest that adaptation should extend beyond input fusion to visual representation, language-model conditioning, and report generation.

Several limitations remain. Our evaluation is centered on MIMIC-RG4 and the conventional MIMIC-CXR setting, leaving generalization to other institutions and less structured forms of incomplete or corrupted context for future study. Clinical efficacy follows the benchmark CheXbert binarization and therefore does not directly measure the full three-way polarity distinction among positive, negative, and uncertain findings, nor finer-grained properties such as anatomical grounding or temporal relations. The commitment ablations instead evaluate the downstream effect of polarity-structured supervision on report generation, while Figure 2 provides qualitative evidence of temporal evidence utilization. Finally, commitment supervision is constructed with a hybrid offline pipeline that combines GPT-assisted labels for a stratified 10k subset with deterministic rule-based fallback targets for the remaining training records. Future work should evaluate expert-verified commitment supervision, polarity-sensitive and relation-aware clinical metrics, and external clinical distributions.